MLC - PHYSICAL SCIENCE



Heath Dobberpuhl
Martin Luther College

LibreTexts™

Martin Luther College

MLC - Physical Science

Heath Dobberpuhl

This text is disseminated via the Open Education Resource (OER) LibreTexts Project (https://LibreTexts.org) and like the thousands of other texts available within this powerful platform, it is freely available for reading, printing, and "consuming."

The LibreTexts mission is to bring together students, faculty, and scholars in a collaborative effort to provide an accessible, and comprehensive platform that empowers our community to develop, curate, adapt, and adopt openly licensed resources and technologies; through these efforts we can reduce the financial burden born from traditional educational resource costs, ensuring education is more accessible for students and communities worldwide.

Most, but not all, pages in the library have licenses that may allow individuals to make changes, save, and print this book. Carefully consult the applicable license(s) before pursuing such effects. Instructors can adopt existing LibreTexts texts or Remix them to quickly build course-specific resources to meet the needs of their students. Unlike traditional textbooks, LibreTexts' web based origins allow powerful integration of advanced features and new technologies to support learning.



LibreTexts is the adaptable, user-friendly non-profit open education resource platform that educators trust for creating, customizing, and sharing accessible, interactive textbooks, adaptive homework, and ancillary materials. We collaborate with individuals and organizations to champion open education initiatives, support institutional publishing programs, drive curriculum development projects, and more.

The LibreTexts libraries are Powered by NICE CXone Expert and was supported by the Department of Education Open Textbook Pilot Project, the California Education Learning Lab, the UC Davis Office of the Provost, the UC Davis Library, the California State University Affordable Learning Solutions Program, and Merlot. This material is based upon work supported by the National Science Foundation under Grant No. 1246120, 1525057, and 1413739.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation nor the US Department of Education.

Have questions or comments? For information about adoptions or adaptions contact info@LibreTexts.org or visit our main website at https://LibreTexts.org.

This text was compiled on 06/11/2026

TABLE OF CONTENTS

Licensing

1: Matter and Measurement

- 1.1: Measurement and Significant Figures
  - 1.1.1: Accuracy, Precision, and Significant Figures
- 1.2: Matter, Radioactive Decay, and Experimentation
  - 1.2.1: What is Matter?
  - 1.2.2: Classifying Matter According to Its Composition
  - 1.2.3: The Properties of Protons, Neutrons, and Electrons
  - 1.2.4: Elements- Defined by Their Number of Protons
  - 1.2.5: Ions - Losing and Gaining Electrons
  - 1.2.6: Isotopes - When the Number of Neutrons Varies
  - 1.2.7: The Discovery of Radioactivity
  - 1.2.8: Types of Radioactivity- Alpha, Beta, and Gamma Decay
  - 1.2.9: Detecting Radioactivity
  - 1.2.10: Natural Radioactivity and Half-Life
- 1.3: Fluids
  - 1.3.1: What Is a Fluid?
  - 1.3.2: Pressure
  - 1.3.3: Pressure Due to the Weight of Fluid
  - 1.3.4: Archimedes' Principle
  - 1.3.5: Bernoulli's Equation
  - 1.3.6: Pascal's Principle

2: Motion

- 2.1: Introduction to One-Dimensional Kinematics
  - 2.1.1: Displacement
  - 2.1.2: Time, Velocity, and Speed
  - 2.1.3: Acceleration
- 2.2: Motion Equations for Constant Acceleration in One Dimension
- 2.3: Falling Objects
  - 2.3.1: Projectile Motion
  - 2.3.E: Kinematics (Exercise)
- 2.4: Introduction to Dynamics- Newton's Laws of Motion
  - 2.4.1: Development of Force Concept
  - 2.4.2: Newton's First Law of Motion- Inertia
  - 2.4.3: Newton's Second Law of Motion- Force and Acceleration
  - 2.4.4: Newton's Third Law of Motion- Symmetry in Forces
  - 2.4.5: Friction
  - 2.4.E: Dynamics (Exercise)

3: Electricity and Magnetism

- 3.1: Static Electricity and Charge- Conservation of Charge
  - 3.1.1: Ohm's Law- Resistance and Simple Circuits

- 3.2: Resistors in Series and Parallel
- 3.3: Capacitors and Dielectrics
- 3.4: Magnets
- 3.5: Ferromagnets and Electromagnets

4: Waves

- 4.1: Oscillations and Waves
  - 4.1.1: Introduction to Oscillatory Motion and Waves
  - 4.1.2: Period and Frequency in Oscillations
  - 4.1.3: Simple Harmonic Motion- A Special Periodic Motion
  - 4.1.4: Forced Oscillations and Resonance
  - 4.1.5: Waves
  - 4.1.6: Wave Interference- Standing Waves and Beats
  - 4.1.7: Sound
  - 4.1.8: Speed of Sound, Frequency, and Wavelength
  - 4.1.9: Doppler Effect and Sonic Booms
  - 4.1.E: Oscillations and Waves (Exercise)
- 4.2: Light
  - 4.2.1: Introduction to Light
  - 4.2.2: Maxwell's Equations- Electromagnetic Waves Predicted and Observed
  - 4.2.3: Production and Properties of Electromagnetic Waves
  - 4.2.4: The Electromagnetic Spectrum- an Overview
  - 4.2.5: The Electromagnetic Spectrum- Application Notes
  - 4.2.6: Reflection
  - 4.2.7: Refraction
  - 4.2.8: Dispersion- The Rainbow and Prisms
  - 4.2.9: Image Formation by Mirrors
  - 4.2.10: Polarization
- 4.3: Image Formation by Lenses

5: Energy

- 5.1: Prelude to Thermodynamics
  - 5.1.1: The First Law of Thermodynamics
  - 5.1.2: The First Law of Thermodynamics and Some Simple Processes
  - 5.1.3: Gravitational Potential Energy
- 5.2: Introduction to the Second Law of Thermodynamics - Heat Engines and their Efficiency
  - 5.2.1: Carnot's Perfect Heat Engine- The Second Law of Thermodynamics Restated
  - 5.2.2: Applications of Thermodynamics- Heat Pumps and Refrigerators
  - 5.2.3: Entropy and the Second Law of Thermodynamics- Disorder and the Unavailability of Energy
- 5.3: Simple Machines

Index

Glossary

Detailed Licensing

Detailed Licensing

Licensing

*A detailed breakdown of this resource's licensing can be found in **Back Matter/Detailed Licensing**.*

1: Matter and Measurement

1: Matter and Measurement is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

SECTION OVERVIEW

1.1: Measurement and Significant Figures

1.1.1: Accuracy, Precision, and Significant Figures

1.1: Measurement and Significant Figures is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

1.1.1: Accuracy, Precision, and Significant Figures

Learning Objectives

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

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

Accuracy and Precision of a Measurement

Science is based on observation and experiment—that is, on measurements. **Accuracy** is how close a measurement is to the correct value for that measurement. For example, let us say that you are measuring the length of standard computer paper. The packaging in which you purchased the paper states that it is 11.0 inches long. You measure the length of the paper three times and obtain the following measurements: 11.1 in., 11.2 in., and 10.9 in. These measurements are quite accurate because they are very close to the correct value of 11.0 inches. In contrast, if you had obtained a measurement of 12 inches, your measurement would not be very accurate.



Figure 1.1.1.1: A double-pan mechanical balance is used to compare different masses. Usually an object with unknown mass is placed in one pan and objects of known mass are placed in the other pan. When the bar that connects the two pans is horizontal, then the masses in both pans are equal. The “known masses” are typically metal cylinders of standard mass such as 1 gram, 10 grams, and 100 grams. (credit: Serge Melki)

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

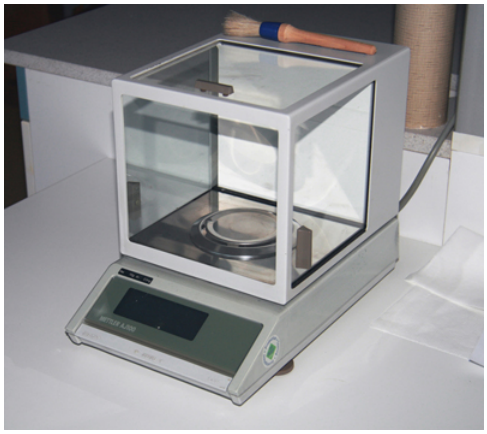


Figure 1.1.1.2: Many mechanical balances, such as double-pan balances, have been replaced by digital scales, which can typically measure the mass of an object more precisely. Whereas a mechanical balance may only read the mass of an object to the nearest tenth of a gram, many digital scales can measure the mass of an object up to the nearest thousandth of a gram. (credit: Karel Jakubec)

The measurements in the paper example are both accurate and precise, but in some cases, measurements are accurate but not precise, or they are precise but not accurate. Let us consider an example of a GPS system that is attempting to locate the position of a restaurant in a city. Think of the restaurant location as existing at the center of a bull's-eye target, and think of each GPS attempt to locate the restaurant as a black dot. In Figure 1.1.1.3, you can see that the GPS measurements are spread out far apart from each other, but they are all relatively close to the actual location of the restaurant at the center of the target. This indicates a low precision, high accuracy measuring system. However, in Figure 4, the GPS measurements are concentrated quite closely to one another, but they are far away from the target location. This indicates a high precision, low accuracy measuring system.



Figure 1.1.1.3: A GPS system attempts to locate a restaurant at the center of the bull's-eye. The black dots represent each attempt to pinpoint the location of the restaurant. The dots are spread out quite far apart from one another, indicating low precision, but they are each rather close to the actual location of the restaurant, indicating high accuracy. (credit: Dark Evil)



Figure 1.1.1.4: In this figure, the dots are concentrated rather closely to one another, indicating high precision, but they are rather far away from the actual location of the restaurant, indicating low accuracy. (credit: Dark Evil)

Accuracy, Precision, and Uncertainty

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

The factors contributing to uncertainty in a measurement include:

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

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

MAKING CONNECTIONS: REAL-WORLD CONNECTIONS – FEVERS OR CHILLS?

Uncertainty is a critical piece of information, both in physics and in many other real-world applications. Imagine you are caring for a sick child. You suspect the child has a fever, so you check his or her temperature with a thermometer. What if the uncertainty of the thermometer were 3.0ºC? If the child's temperature reading was 37.0ºC (which is normal body temperature), the "true" temperature could be anywhere from a hypothermic 34.0ºC to a dangerously high 40.0ºC. A thermometer with an uncertainty of 3.0ºC would be useless.

Percent Uncertainty

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

$$\%\,\text{unc} = \frac{\delta A}{A} \times 100\% \tag{1.1.1.1}$$

Example 1.1.1.1: Calculating Percent Uncertainty: A Bag of Apples

A grocery store sells 5-lb bags of apples. You purchase four bags over the course of a month and weigh the apples each time. You obtain the following measurements:

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

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

Strategy

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

$$= \tfrac{\delta A}{A} \times 100.$$

Solution

Plug the known values into the equation:

$$= \tfrac{0.4lb}{5lb} \times 100.$$

Discussion

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

Uncertainties in Calculations

There is an uncertainty in anything calculated from measured quantities. For example, the area of a floor calculated from measurements of its length and width has an uncertainty because the length and width have uncertainties. How big is the uncertainty in something you calculate by multiplication or division? If the measurements going into the calculation have small uncertainties (a few percent or less), then the **method of adding percents** can be used for multiplication or division. This method says that *the percent uncertainty in a quantity calculated by multiplication or division is the sum of the percent uncertainties in the items used to make the calculation*. For example, if a floor has a length of 4.00m and a width of 3.00m, with uncertainties of 2% and 1%, respectively, then the area of the floor is 12.0m2 and has an uncertainty of 3%. (Expressed as an area this is 0.36m2, which we round to $0.4\,m^2$ since the area of the floor is given to a tenth of a square meter.)

Exercise 1.1.1.1

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

Answer

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

Precision of Measuring Tools and Significant Figures

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

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

Zeros

Special consideration is given to zeros when counting significant figures. The zeros in 0.053 are not significant, because they are only placekeepers that locate the decimal point. There are two significant figures in 0.053. The zeros in 10.053 are not placekeepers but are significant—this number has five significant figures. The zeros in 1300 may or may not be significant depending on the style of writing numbers. They could mean the number is known to the last digit, or they could be placekeepers. So 1300 could have two, three, or four significant figures. (To avoid this ambiguity, write 1300 in scientific notation.) *Zeros are significant except when they serve only as placekeepers*.

Exercise 1.1.1.2

Determine the number of significant figures in the following measurements:

a. 0.0009
b. 15,450.0
c. 6×10^3
d. 87.990
e. 30.42

Solutions

a. 1; the zeros in this number are placekeepers that indicate the decimal point
b. 6; here, the zeros indicate that a measurement was made to the 0.1 decimal point, so the zeros are significant
c. 1; the value 3 signifies the decimal place, not the number of measured values
d. 5; the final zero indicates that a measurement was made to the 0.001 decimal point, so it is significant
e. 4; any zeros located in between significant figures in a number are also significant

Significant Figures in Calculations

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

1. For multiplication and division: *The result should have the same number of significant figures as the quantity having the least significant figures entering into the calculation.* For example, the area of a circle can be calculated from its radius using A=πr2. Let us see how many significant figures the area has if the radius has only two—say, r=1.2m. Then,

$$A = \pi r2 = (3.1415927...) \times (1.2m)^2 = 4.5238934\,m^2 \tag{1.1.1.2}$$

is what you would get using a calculator that has an eight-digit output. But because the radius has only two significant figures, it limits the calculated quantity to two significant figures or

$$A = 4.5\,m^2 \tag{1.1.1.3}$$

even though it is good to at least eight digits.

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

$$7.56\,\text{kg} - 6.052\,\text{kg} + 13.7\,\text{kg} = 15.208\,\text{kg} = 15.2\,\text{kg} \tag{1.1.1.4}$$

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

Significant Figures in this Text

In this text, most numbers are assumed to have three significant figures. Furthermore, consistent numbers of significant figures are used in all worked examples. You will note that an answer given to three digits is based on input good to at least three digits, for example. If the input has fewer significant figures, the answer will also have fewer significant figures. Care is also taken that the number of significant figures is reasonable for the situation posed. In some topics, particularly in optics, more accurate numbers are needed and more than three significant figures will be used. Finally, if a number is *exact*, such as the two in the formula for the circumference of a circle, \(=2πr,\) it does not affect the number of significant figures in a calculation.

Exercise 1.1.1.3

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

a. A woman has two bags weighing 13.5 pounds and one bag with a weight of 10.2 pounds. What is the total weight of the bags?
b. The force F on an object is equal to its mass m multiplied by its acceleration a. If a wagon with mass 55 kg accelerates at a rate of $0.0255 m/s^2$, what is the force on the wagon? (The unit of force is called the newton, and it is expressed with the symbol N.)

Answer

(a) 37.2 pounds; Because the number of bags is an exact value, it is not considered in the significant figures.

(b) 1.4 N; Because the value 55 kg has only two significant figures, the final value must also contain two significant figures.

PHET EXPLORATIONS: ESTIMATION

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

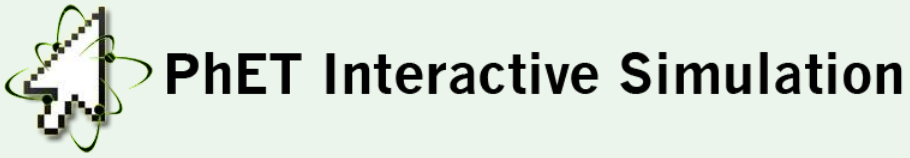


Figure 1.1.1.5: Estimation

Summary

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

Glossary

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

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

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

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

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

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

Contributors and Attributions

Paul Peter Urone (Professor Emeritus at California State University, Sacramento) and Roger Hinrichs (State University of New York, College at Oswego) with Contributing Authors: Kim Dirks (University of Auckland) and Manjula Sharma (University of Sydney). This work is licensed by OpenStax University Physics under a Creative Commons Attribution License (by 4.0).

This page titled 1.1.1: Accuracy, Precision, and Significant Figures is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **1.3: Accuracy, Precision, and Significant Figures** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/college-physics.

SECTION OVERVIEW

1.2: Matter, Radioactive Decay, and Experimentation

1.2.1: What is Matter?

1.2.2: Classifying Matter According to Its Composition

1.2.3: The Properties of Protons, Neutrons, and Electrons

1.2.4: Elements- Defined by Their Number of Protons

1.2.5: Ions - Losing and Gaining Electrons

1.2.6: Isotopes - When the Number of Neutrons Varies

1.2.7: The Discovery of Radioactivity

1.2.8: Types of Radioactivity- Alpha, Beta, and Gamma Decay

1.2.9: Detecting Radioactivity

1.2.10: Natural Radioactivity and Half-Life

1.2: Matter, Radioactive Decay, and Experimentation is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

1.2.1: What is Matter?

Learning Objectives

- Define matter and explain how it is composed of building blocks known as "atoms".

We are all familiar with matter. The definition of **Matter** is anything that has mass and volume (takes up space). For most common objects that we deal with every day, it is fairly simple to demonstrate that they have mass and take up space. You might be able to imagine, however, the difficulty for people several hundred years ago to demonstrate that air had mass and volume. Air (and all other gases) are invisible to the eye, have very small masses compared to equal amounts of solids and liquids, and are quite easy to compress (change volume). Without sensitive equipment, it would have been difficult to convince people that gases are matter. Today, we can measure the mass of a small balloon when it is deflated and then blow it up, tie it off, and measure its mass again to detect the additional mass due to the air inside. The mass of air, under room conditions, that occupies a one quart jar is approximately 0.0002 pounds. This small amount of mass would have been difficult to measure in times before balances were designed to accurately measure very small masses. Later, scientists were able to compress gases into such a small volume that the gases turned into liquids, which made it clear that gases are matter.



Figure 1.2.1.1: Everything from an ant, to a truck, to the earth, and even the entire galaxy is composed of matter. Images used with permission from Wikipedia (CC_SA-BY-3.0; credit High Contrast).

Even though the universe consists of "things" as wildly different as ants and galaxies, the matter that makes up all of these "things" is composed of a very limited number of building blocks. These building blocks are known as **atoms**, and so far, scientists have discovered or created a grand total of 118 different types of atoms. Scientists have given a name to each different type of atom. A substance that is composed of only one type of atom is called an element. At this point, what should amaze you is that all forms of matter in our universe are made with only 118 different building blocks. In some ways, it's sort of like cooking a gourmet, five-course meal using only three ingredients! How is it possible? To answer that question, you have to understand the ways in which different elements are put together to form matter.

The most important method that nature uses to organize atoms into matter is the formation of **molecules**. Molecules are groups of two or more atoms that have been bonded together. There are millions of different ways to bond atoms together, which means that there are millions of different possible molecules. Each of these molecules has its own set of chemical properties, and it's these properties with which chemists are most concerned. You will learn a lot more about atoms and molecules, including how they were discovered, in a later part of the textbook.

Summary

All matter has mass and occupies space. All physical objects are made of matter. Matter itself is composed of tiny building blocks known as "atoms". There are only 118 different types of atoms known to man. Frequently, atoms are bonded together to form "molecules".

Contributions & Attributions

- Wikipedia

1.2.1: What is Matter? is shared under a CK-12 license and was authored, remixed, and/or curated by Henry Agnew & Marisa Alviar-Agnew.

- **3.2: What is Matter?** by Henry Agnew, Marisa Alviar-Agnew is licensed CK-12. Original source: https://www.ck12.org/c/chemistry/.

1.2.2: Classifying Matter According to Its Composition

Learning Objectives

- Explain the difference between a pure substance and a mixture.
- Explain the difference between an element and a compound.
- Explain the difference between a homogeneous mixture and a heterogeneous mixture.

One useful way of organizing our understanding of matter is to think of a hierarchy that extends down from the most general and complex to the simplest and most fundamental (Figure 1.2.2.1). Matter can be classified into two broad categories: pure substances and mixtures. A **pure substance** is a form of matter that has a constant composition (meaning that it is the same everywhere) and properties that are constant throughout the sample (meaning that there is only one set of properties such as melting point, color, boiling point, etc. throughout the matter). A material composed of two or more substances is a **mixture**. Elements and compounds are both examples of pure substances. A substance that cannot be broken down into chemically simpler components is an **element**. Aluminum, which is used in soda cans, is an element. A substance that can be broken down into chemically simpler components (because it has more than one element) is a **compound**. For example, water is a compound composed of the elements hydrogen and oxygen. Today, there are about 118 elements in the known universe. In contrast, scientists have identified tens of millions of different compounds to date.

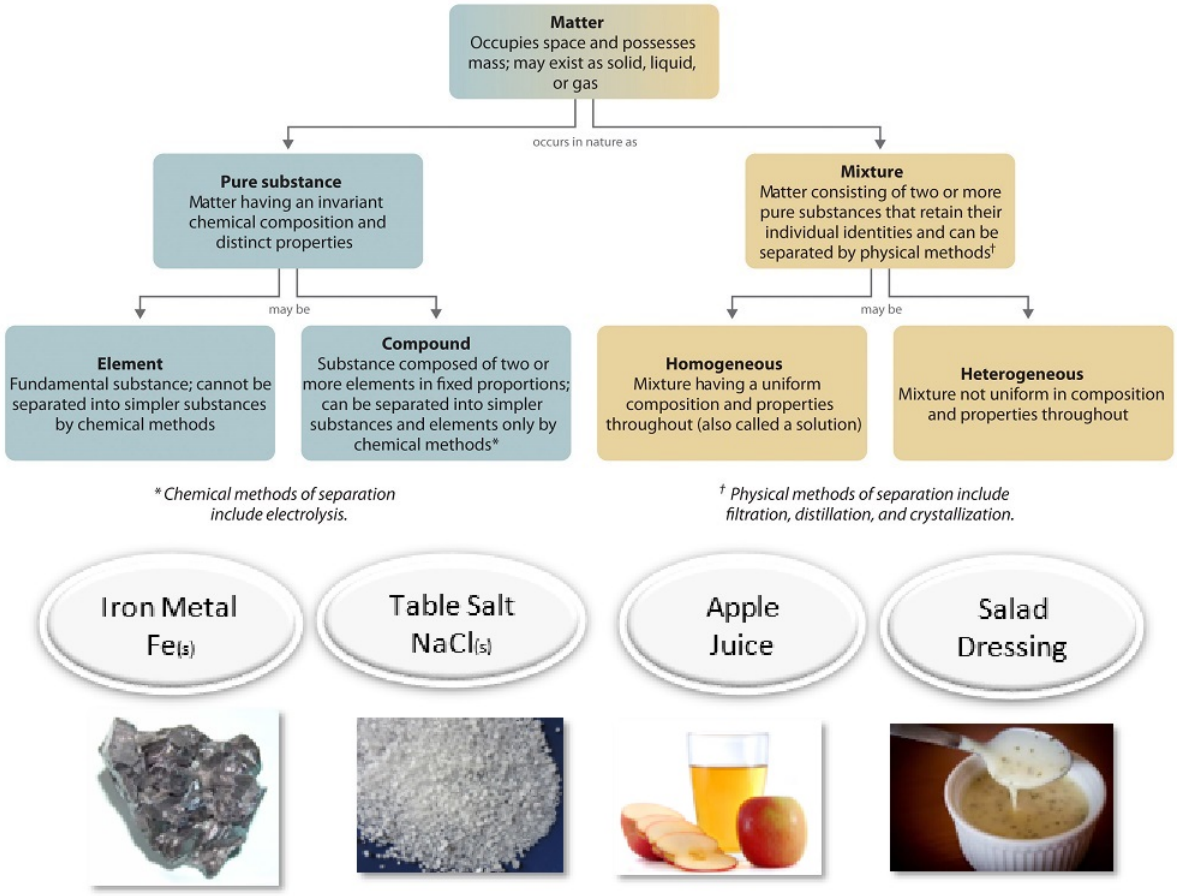


Figure 1.2.2.1: Relationships between the Types of Matter and the Methods Used to Separate Mixtures

Ordinary table salt is called sodium chloride. It is considered a **substance** because it has a uniform and definite composition. All samples of sodium chloride are chemically identical. Water is also a pure substance. Salt easily dissolves in water, but salt water cannot be classified as a substance because its composition can vary. You may dissolve a small amount of salt or a large amount into a given amount of water. A mixture is a physical blend of two or more components, each of which retains its own identity and

properties in the **mixture**. Only the form of the salt is changed when it is dissolved into water. It retains its composition and properties.

A **homogeneous** mixture is a mixture in which the composition is uniform throughout the mixture. The salt water described above is homogeneous because the dissolved salt is evenly distributed throughout the entire salt water sample. Often it is easy to confuse a homogeneous mixture with a pure substance because they are both uniform. The difference is that the composition of the substance is always the same. The amount of salt in the salt water can vary from one sample to another. All solutions are considered homogeneous because the dissolved material is present in the same amount throughout the solution.

A **heterogeneous mixture** is a mixture in which the composition is not uniform throughout the mixture. Vegetable soup is a heterogeneous mixture. Any given spoonful of soup will contain varying amounts of the different vegetables and other components of the soup.

Phase

A phase is any part of a sample that has a uniform composition and properties. By definition, a pure substance or a homogeneous mixture consists of a single phase. A heterogeneous mixture consists of two or more phases. When oil and water are combined, they do not mix evenly, but instead form two separate layers. Each of the layers is called a phase.

Example 1.2.2.1

Identify each substance as a compound, an element, a heterogeneous mixture, or a homogeneous mixture (solution).

a. filtered tea
b. freshly squeezed orange juice
c. a compact disc
d. aluminum oxide, a white powder that contains a 2:3 ratio of aluminum and oxygen atoms
e. selenium

Given: a chemical substance

Asked for: its classification

Strategy:

A. Decide whether a substance is chemically pure. If it is pure, the substance is either an element or a compound. If a substance can be separated into its elements, it is a compound.
B. If a substance is not chemically pure, it is either a heterogeneous mixture or a homogeneous mixture. If its composition is uniform throughout, it is a homogeneous mixture.

Solution

a. **A)** Tea is a solution of compounds in water, so it is not chemically pure. It is usually separated from tea leaves by filtration.
B) Because the composition of the solution is uniform throughout, it is ***a homogeneous mixture***.
b. **A)** Orange juice contains particles of solid (pulp) as well as liquid; it is not chemically pure.
B) Because its composition is not uniform throughout, orange juice is ***a heterogeneous mixture***.
c. **A)** A compact disc is a solid material that contains more than one element, with regions of different compositions visible along its edge. Hence, a compact disc is not chemically pure.
B) The regions of different composition indicate that a compact disc is ***a heterogeneous mixture.***
d. **A)** Aluminum oxide is a single, chemically ***pure compound***.
e. **A)** Selenium is one of the known ***elements***.

Exercise 1.2.2.1

Identify each substance as a compound, an element, a heterogeneous mixture, or a homogeneous mixture (solution).

a. white wine
b. mercury
c. ranch-style salad dressing

d. table sugar (sucrose)

Answer a:
homogeneous mixture (solution)

Answer b:
element

Answer c:
heterogeneous mixture

Answer d:
compound

Example 1.2.2.2

How would a chemist categorize each example of matter?

a. saltwater
b. soil
c. water
d. oxygen

Solution

a. Saltwater acts as if it were a single substance even though it contains two substances—salt and water. Saltwater is a homogeneous mixture, or a solution.
b. Soil is composed of small pieces of a variety of materials, so it is a heterogeneous mixture.
c. Water is a substance. More specifically, because water is composed of hydrogen and oxygen, it is a compound.
d. Oxygen, a substance, is an element.

Exercise 1.2.2.2

How would a chemist categorize each example of matter?

a. coffee
b. hydrogen
c. an egg

Answer a:
a homogeneous mixture (solution), assuming it is filtered coffee

Answer b:
element

Answer c:
heterogeneous mixture

Summary

Matter can be classified into two broad categories: pure substances and mixtures. A pure substance is a form of matter that has a constant composition and properties that are constant throughout the sample. Mixtures are physical combinations of two or more elements and/or compounds. Mixtures can be classified as homogeneous or heterogeneous. Elements and compounds are both examples of pure substances. Compounds are substances that are made up of more than one type of atom. Elements are the simplest substances made up of only one type of atom.

Vocabulary

- Element: a substance that is made up of only one type of atom.
- Compound:a substance that is made up of more than one type of atom bonded together.

- Mixture: a combination of two or more elements or compounds which have not reacted to bond together; each part in the mixture retains its own properties.

Contributions & Attributions

- Stephen Lower, Professor Emeritus (Simon Fraser U.) Chem1 Virtual Textbook

---

1.2.2: Classifying Matter According to Its Composition is shared under a CK-12 license and was authored, remixed, and/or curated by Henry Agnew, Marisa Alviar-Agnew, Stephen Lower, & Stephen Lower.

- **3.4: Classifying Matter According to Its Composition** by Henry Agnew, Marisa Alviar-Agnew, Stephen Lower is licensed CK-12. Original source: https://www.ck12.org/c/chemistry/.

1.2.3: The Properties of Protons, Neutrons, and Electrons

Learning Objectives

- Describe the locations, charges, and masses of the three main subatomic particles.
- Determine the number of protons and electrons in an atom.
- Define atomic mass unit (amu).

Dalton's Atomic Theory explained a lot about matter, chemicals, and chemical reactions. Nevertheless, it was not entirely accurate, because contrary to what Dalton believed, atoms can, in fact, be broken apart into smaller subunits or subatomic particles. We have been talking about the electron in great detail, but there are two other particles of interest to us: protons and neutrons. We already learned that J. J. Thomson discovered a negatively charged particle, called the **electron**. Rutherford proposed that these electrons orbit a positive nucleus. In subsequent experiments, he found that there is a smaller positively charged particle in the nucleus, called a **proton**. There is also a third subatomic particle, known as a neutron.

Electrons

Electrons are one of three main types of particles that make up atoms. Unlike protons and neutrons, which consist of smaller, simpler particles, electrons are fundamental particles that do not consist of smaller particles. They are a type of fundamental particle called leptons. All leptons have an electric charge of -1 or 0. Electrons are extremely small. The mass of an electron is only about 1/2000 the mass of a proton or neutron, so electrons contribute virtually nothing to the total mass of an atom. Electrons have an electric charge of -1, which is equal but opposite to the charge of a proton, which is $+1$. All atoms have the same number of electrons as protons, so the positive and negative charges "cancel out", making atoms electrically neutral.

Unlike protons and neutrons, which are located inside the nucleus at the center of the atom, electrons are found outside the nucleus. Because opposite electric charges attract one another, negative electrons are attracted to the positive nucleus. This force of attraction keeps electrons constantly moving through the otherwise empty space around the nucleus. The figure below is a common way to represent the structure of an atom. It shows the electron as a particle orbiting the nucleus, similar to the way that planets orbit the sun. However, this is an incorrect perspective, as quantum mechanics demonstrates that electrons are more complicated.



Figure 1.2.3.1: Electrons are much smaller than protons or neutrons. If an electron was the mass of a penny, a proton or a neutron would have the mass of a large bowling ball!

Protons

A proton is one of three main particles that make up the atom. Protons are found in the nucleus of the atom. This is a tiny, dense region at the center of the atom. Protons have a positive electrical charge of one $(+1)$ and a mass of 1 atomic mass unit (amu), which is about 1.67×10^{-27} kilograms. Together with neutrons, they make up virtually all of the mass of an atom.

Neutrons

Atoms of all elements—except for most atoms of hydrogen—have neutrons in their nucleus. Unlike protons and electrons, which are electrically charged, neutrons have no charge—they are electrically neutral. That's why the neutrons in the diagram above are labeled n^0. The zero stands for "zero charge". The mass of a neutron is slightly greater than the mass of a proton, which is 1 atomic mass unit (amu). (An atomic mass unit equals about 1.67×10^{-27} kilograms.) A neutron also has about the same diameter as a proton, or 1.7×10^{-15} meters.

As you might have already guessed from its name, the neutron is neutral. In other words, it has no charge whatsoever and is therefore neither attracted to nor repelled from other objects. Neutrons are in every atom (with one exception), and they are bound together with other neutrons and protons in the atomic nucleus.

Before we move on, we must discuss how the different types of subatomic particles interact with each other. When it comes to neutrons, the answer is obvious. Since neutrons are neither attracted to nor repelled from objects, they don't really interact with protons or electrons (beyond being bound into the nucleus with the protons).

Even though electrons, protons, and neutrons are all types of subatomic particles, they are not all the same size. When you compare the masses of electrons, protons, and neutrons, what you find is that electrons have an extremely small mass, compared to either protons or neutrons. On the other hand, the masses of protons and neutrons are fairly similar, although technically, the mass of a neutron is slightly larger than the mass of a proton. Because protons and neutrons are so much more massive than electrons, almost all of the mass of any atom comes from the nucleus, which contains all of the neutrons and protons.

Table 1.2.3.1: Properties of Subatomic Particles

Particle	Symbol	Mass (amu)	Relative Mass (proton = 1)	Relative Charge	Location
proton	p^+	1	1	+1	inside the nucleus
electron	e^-	5.45×10^{-4}	0.00055	−1	outside the nucleus
neutron	n^0	1	1	0	inside the nucleus

Table 1.2.3.1 gives the properties and locations of electrons, protons, and neutrons. The third column shows the masses of the three subatomic particles in "atomic mass units." An **atomic mass unit (**amu**)** is defined as one-twelfth of the mass of a carbon-12 atom. Atomic mass units (amu) are useful, because, as you can see, the mass of a proton and the mass of a neutron are almost exactly 1 in this unit system.

Negative and positive charges of equal magnitude cancel each other out. This means that the negative charge on an electron perfectly balances the positive charge on the proton. In other words, a neutral atom must have exactly one electron for every proton. If a neutral atom has 1 proton, it must have 1 electron. If a neutral atom has 2 protons, it must have 2 electrons. If a neutral atom has 10 protons, it must have 10 electrons. You get the idea. In order to be neutral, an atom must have the same number of electrons and protons.

Summary

- Electrons are a type of subatomic particle with a negative charge.
- Protons are a type of subatomic particle with a positive charge. Protons are bound together in an atom's nucleus as a result of the strong nuclear force.
- Neutrons are a type of subatomic particle with no charge (they are neutral). Like protons, neutrons are bound into the atom's nucleus as a result of the strong nuclear force.
- Protons and neutrons have approximately the same mass, but they are both much more massive than electrons (approximately 2,000 times as massive as an electron).
- The positive charge on a proton is equal in magnitude to the negative charge on an electron. As a result, a neutral atom must have an equal number of protons and electrons.
- The atomic mass unit (amu) is a unit of mass equal to one-twelfth the mass of a carbon-12 atom

1.2.3: The Properties of Protons, Neutrons, and Electrons is shared under a CK-12 license and was authored, remixed, and/or curated by Henry Agnew & Marisa Alviar-Agnew.

- **4.4: The Properties of Protons, Neutrons, and Electrons** by Henry Agnew, Marisa Alviar-Agnew is licensed CK-12. Original source: https://www.ck12.org/c/chemistry/.

1.2.4: Elements- Defined by Their Number of Protons

Learning Objectives

- Define atomic number.
- Define mass number.
- Determine the number of protons, neutrons, and electrons in an atom.

It's important to be able to distinguish atoms of one element from atoms of another element. Elements are pure substances that make up all other matter, so each one is given a unique name. The names of elements are also represented by unique one- or two-letter symbols, such as H for hydrogen, C for carbon, or He for helium. However, it would more powerful if these names could be used to identify the numbers of protons and neutrons in the atoms. That's where atomic number and mass number are useful.



Figure 1.2.4.1: It is difficult to find qualities that differ between each element, and to distinguish one element from another. Each element, however, does have a unique number of protons. Sulfur has 16 protons, silicon has 14 protons, and gold has 79 protons. Images used with permission (public domain for sulfur and silicon, gold is licensed by CC-BY-SA-NC-ND; Alchemist-hp).

Atomic Number

Scientists distinguish between different elements by counting the number of protons in the nucleus (Table 1.2.4.1). If an atom has only one proton, we know that it's a hydrogen atom. An atom with two protons is always a helium atom. If scientists count four protons in an atom, they know it's a beryllium atom. An atom with three protons is a lithium atom, an atom with five protons is a boron atom, an atom with six protons is a carbon atom . . . the list goes on.

Since an atom of one element can be distinguished from an atom of another element by the number of protons in its nucleus, scientists are always interested in this number, and how this number differs between different elements. The number of protons in an atom is called its **atomic number** (Z). This number is very important because it is unique for atoms of a given element. All atoms of an element have the same number of protons, and every element has a different number of protons in its atoms. For example, all helium atoms have two protons, and no other elements have atoms with two protons.

Table 1.2.4.1: Atoms of the First Six Elements

Name	Protons	Neutrons	Electrons	Atomic Number (Z)	Mass Number (A)
Hydrogen	1	0	1	1	1
Helium	2	2	2	2	4
Lithium	3	4	3	3	7
Beryllium	4	5	4	4	9
Boron	5	6	5	5	11
Carbon	6	6	6	6	12

Of course, since neutral atoms have to have one electron for every proton, an element's atomic number also tells you how many electrons are in a neutral atom of that element. For example, hydrogen has an atomic number of 1. This means that an atom of hydrogen has one proton, and, if it's neutral, one electron as well. Gold, on the other hand, has an atomic number of 79, which means that an atom of gold has 79 protons, and, if it's neutral, 79 electrons as well.

Neutral Atoms

Atoms are neutral in electrical charge because they have the same number of negative electrons as positive protons (Table 1.2.4.1). Therefore, the atomic number of an atom also tells you how many electrons the atom has. This, in turn, determines many of the atom's chemical properties.

Mass Number

The **mass number** (A) of an atom is the total number of protons and neutrons in its nucleus. The mass of the atom is a unit called the atomic mass unit (amu). One atomic mass unit is the mass of a proton, or about 1.67×10^{-27} kilograms, which is an extremely small mass. A neutron has just a tiny bit more mass than a proton, but its mass is often assumed to be one atomic mass unit as well. Because electrons have virtually no mass, just about all the mass of an atom is in its protons and neutrons. Therefore, the total number of protons and neutrons in an atom determines its mass in atomic mass units (Table 1.2.4.1).

Consider helium again. Most helium atoms have two neutrons in addition to two protons. Therefore the mass of most helium atoms is 4 atomic mass units (2 amu for the protons + 2 amu for the neutrons). However, some helium atoms have more or less than two neutrons. Atoms with the same number of protons but different numbers of neutrons are called isotopes. Because the number of neutrons can vary for a given element, the mass numbers of different atoms of an element may also vary. For example, some helium atoms have three neutrons instead of two (these are called isotopes and are discussed in detail later on).

Why do you think that the "mass number" includes protons and neutrons, but not electrons? You know that most of the mass of an atom is concentrated in its nucleus. The mass of an atom depends on the number of protons and neutrons. You have already learned that the mass of an electron is very, very small compared to the mass of either a proton or a neutron (like the mass of a penny compared to the mass of a bowling ball). Counting the number of protons and neutrons tells scientists about the total mass of an atom.

$$\text{mass number } A = (\text{number of protons}) + (\text{number of neutrons})$$

An atom's mass number is very easy to calculate, provided that you know the number of protons and neutrons in an atom.

Example 4.5.1

What is the mass number of an atom of helium that contains 2 neutrons?

Solution

$(\text{number of protons}) = 2$ (Remember that an atom of helium always has 2 protons.)

$(\text{number of neutrons}) = 2$

$\text{mass number} = (\text{number of protons}) + (\text{number of neutrons})$

$$\text{mass number} = 2 + 2 = 4$$

A **chemical symbol** is a one- or two-letter designation of an element. Some examples of chemical symbols are O for oxygen, Zn for zinc, and Fe for iron. The first letter of a symbol is always capitalized. If the symbol contains two letters, the second letter is lower case. The majority of elements have symbols that are based on their English names. However, some of the elements that have been known since ancient times have maintained symbols that are based on their Latin names, as shown in Table 1.2.4.2.

Table 1.2.4.2: *Symbols and Latin Names for Elements*

Chemical Symbol	Name	Latin Name
Na	Sodium	Natrium
K	Potassium	Kalium
Fe	Iron	Ferrum
Cu	Copper	Cuprum
Ag	Silver	Argentum

Chemical Symbol	Name	Latin Name
Sn	Tin	Stannum
Sb	Antimony	Stibium
Au	Gold	Aurum
Pb	Lead	Plumbum

Summary

- Elements are pure substances that make up all matter, so each one is given a unique name.
- The names of elements are also represented by unique one- or two-letter symbols.
- Each element has a unique number of protons. An element's atomic number is equal to the number of protons in the nuclei of any of its atoms.
- The mass number of an atom is the sum of the protons and neutrons in the atom.
- Isotopes are atoms of the same element (same number of protons) that have different numbers of neutrons in their atomic nuclei.

1.2.4: Elements- Defined by Their Number of Protons is shared under a CK-12 license and was authored, remixed, and/or curated by Henry Agnew & Marisa Alviar-Agnew.

- **4.5: Elements- Defined by Their Number of Protons** by Henry Agnew, Marisa Alviar-Agnew is licensed CK-12. Original source: https://www.ck12.org/c/chemistry/.

1.2.5: Ions - Losing and Gaining Electrons

Learning Objectives

- Define the two types of ions.

Atoms tend to lose, gain, or share some valance electrons, making bonds to acquire the electron configuration of the nearest noble gas, i.e., usually eight valence electrons for the main group elements, called an octet. Most atoms do not have eight electrons in their valence electron shell. Some atoms have only a few electrons in their outer shell, while others lack only one or two electrons to have an octet. In cases where an atom has three or fewer valence electrons, the atom may lose those valence electrons quite easily until what remains is a lower shell that contains an octet. Atoms that lose electrons acquire a positive charge because they are left with fewer negatively charged electrons to balance the positive charges of the protons in the nucleus. Positively charged ions are called **cations**. Most metals become cations when they make ionic compounds.

Cations

A neutral sodium atom is likely to achieve an octet in its outermost shell by losing its one valence electron.

$$\mathrm{Na} \rightarrow \mathrm{Na^+} + \mathrm{e^-}$$

The cation produced in this way, $\mathrm{Na^+}$, is called the sodium ion to distinguish it from the element. The outermost shell of the sodium ion is the second electron shell, which has eight electrons in it. The octet rule has been satisfied. Figure 1.2.5.1 is a graphical depiction of this process.

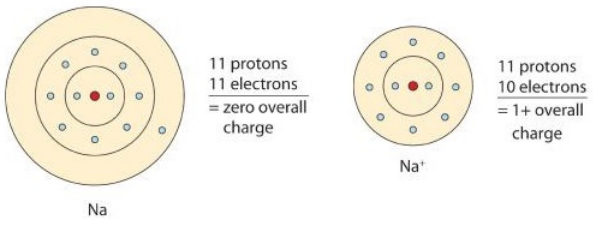


Figure 1.2.5.1: The Formation of a Sodium Ion. On the left, a sodium atom has 11 electrons. On the right, the sodium ion only has 10 electrons and a 1+ charge. Neutral sodium atom on left has 11 protons and 11 electrons. Sodium ion on right has 11 protons and 10 electrons, with a +1 overall charge.

Anions

Some atoms have nearly eight electrons in their valence shell and can gain additional valence electrons until they have an octet. When these atoms gain electrons, they acquire a negative charge because they now possess more electrons than protons. Negatively charged ions are called **anions**. Most nonmetals become anions when they make ionic compounds.

A neutral chlorine atom has seven electrons in its outermost shell. Only one more electron is needed to achieve an octet in chlorine's valence shell. (In table salt, this electron comes from the sodium atom.)

$$\mathrm{e^-} + \mathrm{Cl} \longrightarrow \mathrm{Cl^-}$$

In this case, the ion has the same outermost shell as the original atom, but now that shell has eight electrons in it. Once again, the octet rule has been satisfied. The resulting anion, $\mathrm{Cl^-}$, is called the chloride ion; note the slight change in the suffix (*-ide* instead of *-ine*) to create the name of this anion. Figure 1.2.5.2 is a graphical depiction of this process.

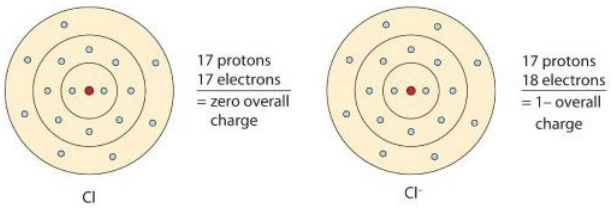


Figure 1.2.5.2: The Formation of a Chlorine Ion. On the left, the chlorine atom has 17 electrons. On the right, the chloride ion has 18 electrons and has a 1− charge. Neutral chlorine atom on

left has 17 protons and 17 electrons. Sodium ion on right has 17 protons and 18 electrons, with a -1 overall charge.

> *The names for positive and negative ions are pronounced CAT-eye-ons and ANN-eye-ons, respectively.*

In many cases, elements that belong to the same group (vertical column) on the periodic table form ions with the same charge because they have the same number of valence electrons. Thus, the periodic table becomes a tool for remembering the charges on many ions. For example, all ions made from alkali metals, the first column on the periodic table, have a 1+ charge. Ions made from alkaline earth metals, the second group on the periodic table, have a 2+ charge. On the other side of the periodic table, the next-to-last column, the halogens, form ions having a 1− charge. Figure 1.2.5.3 shows how the charge on many ions can be predicted by the location of an element on the periodic table. Note the convention of first writing the number and then the sign on a ion with multiple charges. The barium cation is written Ba^{2+}, not Ba^{+2}.

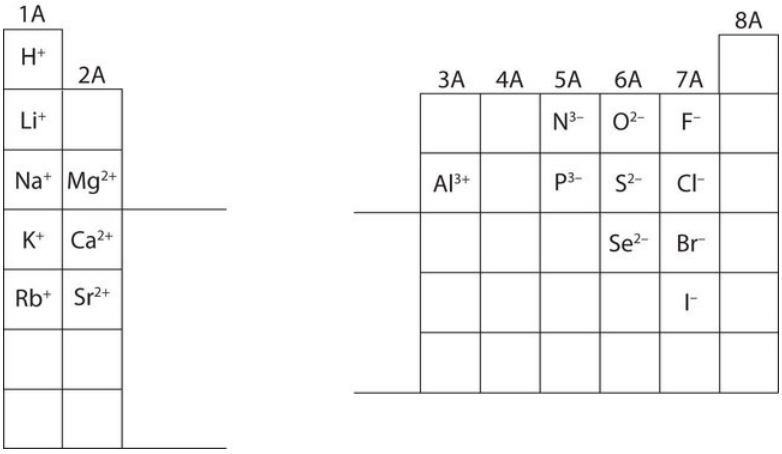


Figure 1.2.5.3: Predicting Ionic Charges. The charge that an atom acquires when it becomes an ion is related to the structure of the periodic table. Within a group (family) of elements, atoms form ions of a certain charge.

1.2.5: Ions - Losing and Gaining Electrons is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Henry Agnew & Marisa Alviar-Agnew.

- **4.7: Ions - Losing and Gaining Electrons** by Henry Agnew, Marisa Alviar-Agnew is licensed CC BY-NC-SA 3.0.

1.2.6: Isotopes - When the Number of Neutrons Varies

Learning Objectives

- Explain what isotopes are and how an isotope affects an element's atomic mass.
- Determine the number of protons, electrons, and neutrons of an element with a given mass number.

All atoms of the same element have the same number of protons, but some may have different numbers of neutrons. For example, all carbon atoms have six protons, and most have six neutrons as well. But some carbon atoms have seven or eight neutrons instead of the usual six. Atoms of the same element that differ in their numbers of neutrons are called **isotopes**. Many isotopes occur naturally. Usually one or two isotopes of an element are the most stable and common. Different isotopes of an element generally have the same physical and chemical properties because they have the same numbers of protons and electrons.

An Example: Hydrogen Isotopes

Hydrogen is an example of an element that has isotopes. Three isotopes of hydrogen are modeled in Figure 1.2.6.1. Most hydrogen atoms have just one proton, one electron, and lack a neutron. These atoms are just called hydrogen. Some hydrogen atoms have one neutron as well. These atoms are the isotope named deuterium. Other hydrogen atoms have two neutrons. These atoms are the isotope named tritium.

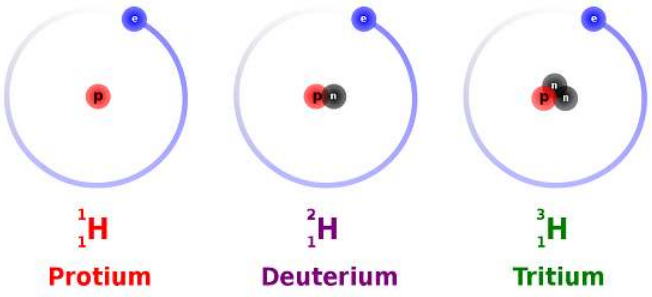


Figure 1.2.6.1: The three most stable isotopes of hydrogen: protium (A = 1), deuterium (A = 2), and tritium (A = 3). (CC SA-BY 3.0; Balajijagadesh via Wikipedia).

For most elements other than hydrogen, isotopes are named for their mass number. For example, carbon atoms with the usual 6 neutrons have a mass number of 12 (6 protons + 6 neutrons = 12), so they are called carbon-12. Carbon atoms with 7 neutrons have an atomic mass of 13 (6 protons + 7 neutrons = 13). These atoms are the isotope called carbon-13.

Example 1.2.6.1: Lithium Isotopes

a. What is the atomic number and the mass number of an isotope of lithium containing 3 neutrons?
b. What is the atomic number and the mass number of an isotope of lithium containing 4 neutrons?

Solution

A lithium atom contains 3 protons in its nucleus irrespective of the number of neutrons or electrons.

a.

$$\begin{aligned}\text{atomic number} &= (\text{number of protons}) = 3\\ &\quad (\text{number of neutrons}) = 3\end{aligned}$$

$$\begin{aligned}\text{mass number} &= (\text{number of protons}) + (\text{number of neutrons})\\ \text{mass number} &= 3 + 3\\ &= 6\end{aligned}$$

b.

$$\begin{aligned}\text{atomic number} &= (\text{number of protons}) = 3\\ &\quad (\text{number of neutrons}) = 4\end{aligned}$$

$$\begin{aligned}\text{mass number} &= (\text{number of protons}) + (\text{number of neutrons}) \\ \text{mass number} &= 3 + 4 \\ &= 7\end{aligned}$$

Notice that because the lithium atom always has 3 protons, the atomic number for lithium is always 3. The mass number, however, is 6 in the isotope with 3 neutrons, and 7 in the isotope with 4 neutrons. In nature, only certain isotopes exist. For instance, lithium exists as an isotope with 3 neutrons, and as an isotope with 4 neutrons, but it doesn't exist as an isotope with 2 neutrons or as an isotope with 5 neutrons.

Stability of Isotopes

Atoms need a certain ratio of neutrons to protons to have a stable nucleus. Having too many or too few neutrons relative to protons results in an unstable, or radioactive, nucleus that will sooner or later break down to a more stable form. This process is called radioactive decay. Many isotopes have radioactive nuclei, and these isotopes are referred to as radioisotopes. When they decay, they release particles that may be harmful. This is why radioactive isotopes are dangerous and why working with them requires special suits for protection. The isotope of carbon known as carbon-14 is an example of a radioisotope. In contrast, the carbon isotopes called carbon-12 and carbon-13 are stable.

This whole discussion of isotopes brings us back to Dalton's Atomic Theory. According to Dalton, atoms of a given element are identical. But if atoms of a given element can have different numbers of neutrons, then they can have different masses as well! How did Dalton miss this? It turns out that elements found in nature exist as constant uniform mixtures of their naturally occurring isotopes. In other words, a piece of lithium always contains both types of naturally occurring lithium (the type with 3 neutrons and the type with 4 neutrons). Moreover, it always contains the two in the same relative amounts (or "relative abundance"). In a chunk of lithium, 93% will always be lithium with 4 neutrons, while the remaining 7% will always be lithium with 3 neutrons.

Dalton always experimented with large chunks of an element—chunks that contained all of the naturally occurring isotopes of that element. As a result, when he performed his measurements, he was actually observing the averaged properties of all the different isotopes in the sample. For most of our purposes in chemistry, we will do the same thing and deal with the average mass of the atoms. Luckily, aside from having different masses, most other properties of different isotopes are similar.

There are two main ways in which scientists frequently show the mass number of an atom they are interested in. It is important to note that the mass number is *not* given on the periodic table. These two ways include writing a nuclear symbol or by giving the name of the element with the mass number written.

To write a **nuclear symbol**, the mass number is placed at the upper left (superscript) of the chemical symbol and the atomic number is placed at the lower left (subscript) of the symbol. The complete nuclear symbol for helium-4 is drawn below:



Nuclear symbol for helium-4: The element symbol is He, the mass number to the top left is 4, and the atomic number to the bottom left is 2

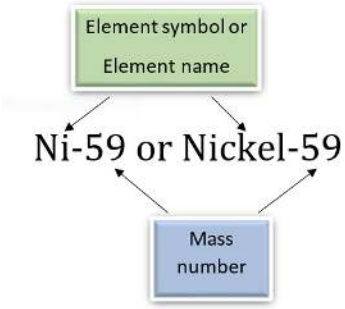
The following nuclear symbols are for a nickel nucleus with 31 neutrons and a uranium nucleus with 146 neutrons.

$$^{59}_{28}\text{Ni}$$

$$^{238}_{92}\text{U}$$

In the nickel nucleus represented above, the atomic number 28 indicates that the nucleus contains 28 protons, and therefore, it must contain 31 neutrons in order to have a mass number of 59. The uranium nucleus has 92 protons, as all uranium nuclei do; and this particular uranium nucleus has 146 neutrons.

Another way of representing isotopes is by adding a hyphen and the mass number to the chemical name or symbol. Thus the two nuclei would be Nickel-59 or Ni-59 and Uranium-238 or U-238, where 59 and 238 are the mass numbers of the two atoms, respectively. Note that the mass numbers (not the number of neutrons) are given to the side of the name.



Example 1.2.6.2: Potassium-40

How many protons, electrons, and neutrons are in an atom of $^{40}_{19}\text{K}$?

Solution

$$\text{atomic number} = (\text{number of protons}) = 19$$

For all atoms with no charge, the number of electrons is equal to the number of protons.

$$\text{number of electrons} = 19$$

The mass number, 40, is the sum of the protons and the neutrons.

To find the number of neutrons, subtract the number of protons from the mass number.

$$\text{number of neutrons} = 40 - 19 = 21.$$

Example 1.2.6.3: Zinc-65

How many protons, electrons, and neutrons are in an atom of zinc-65?

Solution

$$\text{number of protons} = 30$$

For all atoms with no charge, the number of electrons is equal to the number of protons.

$$\text{number of electrons} = 30$$

The mass number, 65, is the sum of the protons and the neutrons.

To find the number of neutrons, subtract the number of protons from the mass number.

$$\text{number of neutrons} = 65 - 30 = 35$$

Exercise 1.2.6.3

How many protons, electrons, and neutrons are in each atom?

a. $^{60}_{27}\text{Co}$
b. Na-24
c. $^{45}_{20}\text{Ca}$
d. Sr-90

Answer a:

27 protons, 27 electrons, 33 neutrons

Answer b:

11 protons, 11 electrons, 13 neutrons

Answer c:

20 protons, 20 electrons, 25 neutrons

Answer d:

38 protons, 38 electrons, 52 neutrons

Summary

- The number of protons is always the same in atoms of the same element.
- The number of neutrons can be different, even in atoms of the same element.
- Atoms of the same element that contain the same number of protons, but different numbers of neutrons, are known as **isotopes**.
- Isotopes of any given element all contain the same number of protons, so they have the same atomic number (for example, the atomic number of helium is always 2).
- Isotopes of a given element contain different numbers of neutrons, therefore, different isotopes have different mass numbers.

1.2.6: Isotopes - When the Number of Neutrons Varies is shared under a CK-12 license and was authored, remixed, and/or curated by Henry Agnew & Marisa Alviar-Agnew.

- **4.8: Isotopes - When the Number of Neutrons Varies** by Henry Agnew, Marisa Alviar-Agnew is licensed CK-12. Original source: https://www.ck12.org/c/chemistry/.

1.2.7: The Discovery of Radioactivity

Learning Objectives

- List the most common emissions from naturally radioactive nuclei.
- Compare the energy released per gram of matter in nuclear reactions to that in chemical reactions.
- Express the relationship between nuclear stability and the nuclei's binding energy per nucleon ratio.

No one could have known in the 1800's that the discovery of the fascinating science and art form of photography would eventually lead to the splitting of the atom. The basis of photography is the fact that visible light causes certain chemical reactions. If the chemicals are spread thinly on a surface but protected from light by a covering, no reaction occurs. When the covering is removed, however, light acting on the chemicals causes them to darken. With millions of cameras in use today, we do not think of it as a strange phenomenon—but at the time of its discovery, photography was a strange and wonderful thing.

Even stranger was the discovery by Wilhelm Roentgen—that radiation other than visible light could expose photographic film. He found that film wrapped in dark paper would react when x-rays went through the paper and struck the film.



Figure 1.2.7.1: Image of Becquerel's photographic plate, which has been fogged by exposure to radiation from a uranium salt. The shadow of a metal Maltese Cross placed between the plate and the uranium salt is clearly visible (Public Domain).

When Henri Becquerel heard about Roentgen's discovery, he wondered if his fluorescent minerals would give the same x-rays. Becquerel placed some of his rock crystals on top of a well-covered photographic plate and sat them in the sunlight. The sunlight made the crystals glow with a bright fluorescent light, but when Becquerel developed the film he was very disappointed. He found that only one of his minerals, a uranium salt, had fogged the photographic plate. He decided to try again, and this time, to leave them out in the sun for a longer period of time. Fortunately, the weather didn't cooperate, and Becquerel had to leave the crystals and film stored in a drawer for several cloudy days. Before continuing his experiments, Becquerel decided to check one of the photographic plates to make sure the chemicals were still good. To his amazement, he found that the plate had been exposed in spots where it had been near the uranium containing rocks, and some of these rocks had not been exposed to sunlight at all. In later experiments, Becquerel confirmed that the radiation from the uranium had no connection with light or fluorescence, but the amount of radiation was directly proportional to the concentration of uranium in the rock. Becquerel had discovered **radioactivity**.

The Curies and Radium

One of Becquerel's assistants, a young Polish scientist named Maria Sklowdowska (to become Marie Curie after she married Pierre Curie), became interested in the phenomenon of radioactivity. With her husband, she decided to find out if chemicals other than uranium were radioactive. The Austrian government was happy to send the Curies a ton of pitchblende from the mining region of Joachimstahl, because it was waste material that had to be disposed of anyway. The Curies wanted the pitchblende because it was the residue of uranium mining. From the ton of pitchblende, the Curies separated $0.10\,\mathrm{g}$ of a previously unknown element, radium, in the form of the compound radium chloride. This radium was many times more radioactive than uranium.

By 1902, the world was aware of a new phenomenon called radioactivity and of new elements which exhibited natural radioactivity. For this work, Becquerel and the Curies shared the 1903 Nobel Prize and for subsequent work; Marie Cure received a second Nobel Prize in 1911. She is the only person ever to receive two Nobel Prizes in science.

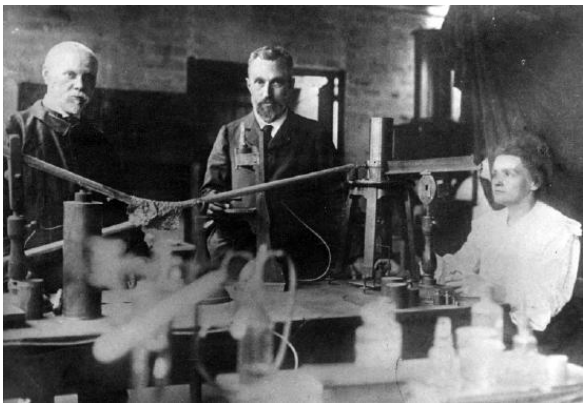


Figure 1.2.7.2: Marie Curie (right) and Pierre Curie (middle) with Henri Becquerel (left) shared the 1903 Nobel Prize.

Further experiments provided information about the characteristics of the penetrating emissions from radioactive substances. It was soon discovered that there were three common types of radioactive emissions. Some of the radiation could pass easily through aluminum foil while some of the radiation was stopped by the foil. Some of the radiation could even pass through foil up to a centimeter thick. The three basic types of radiation were named alpha, beta, and gamma radiation. The actual composition of the three types of radiation was still not known.

Eventually, scientists were able to demonstrate experimentally that the alpha particle, α, was a helium nucleus (a particle containing two protons and two neutrons), a beta particle, β, was a high speed electron, and gamma rays, γ, were a very high energy form of light (even higher energy than x-rays).

Unstable Nuclei May Disintegrate

A nucleus (with one exception, hydrogen-1) consists of some number of protons and neutrons pulled together in an extremely tiny volume. Since protons are positively charged and like charges repel, it is clear that protons cannot remain together in the nucleus unless there is a powerful force holding them there. The force which holds the nucleus together is generated by **nuclear binding energy**.

A nucleus with a large amount of binding energy per nucleon (proton or neutron) will be held together tightly and is referred to as stable. These nuclei do not break apart. When there is too little binding energy per nucleon, the nucleus will be less stable and may disintegrate (come apart). Such disintegration is referred to as **natural radioactivity**. It is also possible for scientists to smash nuclear particles together and cause nuclear reactions between normally stable nuclei. This disintegration is referred to as **artificial radioactivity**. None of the elements above #92 on the periodic table occur on earth naturally—they are all products of artificial (manmade) radioactivity.

When nuclei come apart, they come apart violently accompanied by a tremendous release of energy in the form of heat, light, and radiation. This energy comes from some of the nuclear binding energy. In nuclear changes, the energy involved comes from the nuclear binding energy. However, in chemical reactions, the energy comes from electrons moving energy levels. A typical nuclear change (such as fission) may involve millions of times more energy per atom changing compared to a chemical change (such as burning)!

Summary

- Henri Becquerel, Marie Curie, and Pierre Curie shared the discovery of radioactivity.

Contributions & Attributions

- Wikibooks

1.2.7: The Discovery of Radioactivity is shared under a CK-12 license and was authored, remixed, and/or curated by Henry Agnew & Marisa Alviar-Agnew.

- **17.2: The Discovery of Radioactivity** by Henry Agnew, Marisa Alviar-Agnew is licensed CK-12. Original source: https://www.ck12.org/c/chemistry/.

1.2.8: Types of Radioactivity- Alpha, Beta, and Gamma Decay

Learning Objectives

- Compare qualitatively the ionizing and penetration power of alpha particles (α), beta particles (β), and gamma rays (γ).
- Express the changes in the atomic number and mass number of a radioactive nuclei when an alpha, beta, or gamma particle is emitted.
- Write nuclear equations for alpha and beta decay reactions.

Many nuclei are radioactive; that is, they decompose by emitting particles and in doing so, become a different nucleus. In our studies up to this point, atoms of one element were unable to change into different elements. That is because in all other types of changes discussed, only the electrons were changing. In these changes, the nucleus, which contains the protons that dictate which element an atom is, is changing. All nuclei with 84 or more protons are radioactive, and elements with less than 84 protons have both stable and unstable isotopes. All of these elements can go through nuclear changes and turn into different elements.

In natural radioactive decay, three common emissions occur. When these emissions were originally observed, scientists were unable to identify them as some already known particles and so named them:

- alpha particles (α)
- beta particles (β)
- gamma rays (γ)

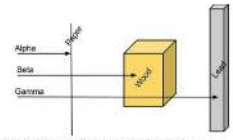
These particles were named using the first three letters of the Greek alphabet. Some later time, alpha particles were identified as helium-4 nuclei, beta particles were identified as electrons, and gamma rays as a form of electromagnetic radiation like x-rays, except much higher in energy and even more dangerous to living systems.

The Ionizing and Penetration Power of Radiation

With all the radiation from natural and man-made sources, we should quite reasonably be concerned about how all the radiation might affect our health. The damage to living systems is done by radioactive emissions when the particles or rays strike tissue, cells, or molecules and alter them. These interactions can alter molecular structure and function; cells no longer carry out their proper function and molecules, such as DNA, no longer carry the appropriate information. Large amounts of radiation are very dangerous, even deadly. In most cases, radiation will damage a single (or very small number) of cells by breaking the cell wall or otherwise preventing a cell from reproducing.

The ability of radiation to damage molecules is analyzed in terms of what is called **ionizing power**. When a radiation particle interacts with atoms, the interaction can cause the atom to lose electrons and thus become ionized. The greater the likelihood that damage will occur by an interaction is the ionizing power of the radiation.

Much of the threat from radiation is involved with the ease or difficulty of protecting oneself from the particles. How thick of a wall do you need to hide behind to be safe? The ability of each type of radiation to pass through matter is expressed in terms of **penetration power.** The more material the radiation can pass through, the greater the penetration power and the more dangerous it is. In general, the greater mass present, the greater the ionizing power, and the lower the penetration power.



Alpha particles can be blocked with paper. Beta particles penetrate through paper, but are blocked by wood. Gamma particles are capable of penetrating paper and wood but can be blocked by a few inches of lead. CC Tracy Poulsen

Comparing only the three common types of ionizing radiation, alpha particles have the greatest mass. Alpha particles have approximately four times the mass of a proton or neutron and approximately 8,000 times the mass of a beta particle. Because of the large mass of the alpha particle, it has the highest ionizing power and the greatest ability to damage tissue. That same large size of alpha particles, however, makes them less able to penetrate matter. They collide with molecules very quickly when striking matter, add two electrons, and become a harmless helium atom. Alpha particles have the least penetration power and can be stopped by a thick sheet of paper or even a layer of clothes. They are also stopped by the outer layer of dead skin on people. This may seem to remove the threat from alpha particles, but it is only from external sources. In a nuclear explosion or some sort of nuclear accident, where radioactive emitters are spread around in the environment, the emitters can be inhaled or taken in with food or water and once the alpha emitter is inside you, you have no protection at all.

Beta particles are much smaller than alpha particles and therefore, have much less ionizing power (less ability to damage tissue), but their small size gives them much greater penetration power. Most resources say that beta particles can be stopped by a one-quarter inch thick sheet of aluminum. Once again, however, the greatest danger occurs when the beta emitting source gets inside of you.

Gamma rays are not particles, but a high energy form of electromagnetic radiation (like x-rays, except more powerful). Gamma rays are energy that has no mass or charge. Gamma rays have tremendous penetration power and require several inches of dense material (like lead) to shield them. Gamma rays may pass all the way through a human body without striking anything. They are considered to have the least ionizing power and the greatest penetration power.

Table 1.2.8.1 Comparison of Penetrating Power,Ionizing Power and Shielding of Alpha and Beta Particles, and Gamma Rays.

Particle	Symbol	Mass	Penetrating Power	Ionizing Power	Shielding
Alpha	α	4amu	Very Low	Very High	Paper Skin
Beta	β	1/2000amu	Intermediate	Intermediate	Aluminum
Gamma	γ	0 (energy only)	Very High	Very Low	2 inches lead

The safest amount of radiation to the human body is zero. It is impossible to completely avoid ionizing radiation, so the next best goal is to be exposed to as little as possible. The two best ways to minimize exposure are to limit time of exposure, and to increase distance from the source.

Alpha Decay

The nuclear disintegration process that emits alpha particles is called alpha decay. An example of a nucleus that undergoes alpha decay is uranium-238. The alpha decay of U-238 is

$$^{238}_{92}\mathrm{U} \rightarrow {}^{4}_{2}\mathrm{He} + {}^{234}_{90}\mathrm{Th} \tag{1.2.8.1}$$

In this nuclear change, the uranium atom $(^{238}_{92}\mathrm{U})$ transmuted into an atom of thorium $(^{234}_{90}\mathrm{Th})$ and, in the process, gave off an alpha particle. Look at the symbol for the alpha particle: $^{4}_{2}\mathrm{He}$. Where does an alpha particle get this symbol? The bottom number in a nuclear symbol is the number of protons. That means that the alpha particle has two protons in it that were lost by the uranium atom. The two protons also have a charge of $+2$. The top number, 4, is the mass number or the total of the protons and neutrons in the particle. Because it has two protons, and a total of four protons and neutrons, alpha particles must also have two neutrons. Alpha particles always have this same composition: two protons and two neutrons.

Another alpha particle producer is thorium-230.

$$^{230}_{90}\mathrm{Th} \rightarrow {}^{4}_{2}\mathrm{He} + {}^{226}_{88}\mathrm{Ra} \tag{1.2.8.2}$$

These types of equations are called nuclear equations and are similar to the chemical equivalent discussed through the previous chapters.

Beta Decay

Another common decay process is beta particle emission, or beta decay. A beta particle is simply a high energy electron that is emitted from the nucleus. It may occur to you that we have a logically difficult situation here. Nuclei do not contain electrons and yet during beta decay, an electron is emitted from a nucleus. At the same time that the electron is being ejected from the nucleus, a neutron is becoming a proton. It is tempting to picture this as a neutron breaking into two pieces with the pieces being a proton and an electron. That would be convenient for simplicity, but unfortunately that is not what happens (more on this subject will be explained at the end of this section). For convenience, we will treat beta decay as a neutron splitting into a proton and an electron. The proton stays in the nucleus, increasing the atomic number of the atom by one. The electron is ejected from the nucleus and is the particle of radiation called beta.

To insert an electron into a nuclear equation and have the numbers add up properly, an atomic number and a mass number had to be assigned to an electron. The mass number assigned to an electron is zero (0), which is reasonable since the mass number is the number of protons plus neutrons, and an electron contains no protons and no neutrons. The atomic number assigned to an electron is

negative one (-1), because that allows a nuclear equation containing an electron to balance atomic numbers. Therefore, the nuclear symbol representing an electron (beta particle) is

$$ {}^{0}_{-1}\text{e} \text{ or } {}^{0}_{-1}\beta $$

Thorium-234 is a nucleus that undergoes beta decay. Here is the nuclear equation for this beta decay:

$$ {}^{234}_{90}\text{Th} \rightarrow {}^{0}_{-1}\text{e} + {}^{234}_{91}\text{Pa} \tag{1.2.8.3} $$

Gamma Radiation

Frequently, gamma ray production accompanies nuclear reactions of all types. In the alpha decay of U-238, two gamma rays of different energies are emitted in addition to the alpha particle.

$$ {}^{238}_{92}\text{U} \rightarrow {}^{4}_{2}\text{He} + {}^{234}_{90}\text{Th} + 2{}^{0}_{0}\gamma $$

Virtually all of the nuclear reactions in this chapter also emit gamma rays, but for simplicity the gamma rays are generally not shown. Nuclear reactions produce a great deal more energy than chemical reactions. Chemical reactions release the difference between the chemical bond energy of the reactants and products, and the energies released have an order of magnitude of $1 \times 10^{3}\ \text{kJ/mol}$. Nuclear reactions release some of the binding energy and may convert tiny amounts of matter into energy. The energy released in a nuclear reaction has an order of magnitude of $1 \times 10^{18}\ \text{kJ/mol}$. That means that nuclear changes involve almost **one million times more energy** per atom than chemical changes!

Note

Virtually all of the nuclear reactions in this chapter also emit gamma rays, but for simplicity the gamma rays are generally not shown.

The essential features of each reaction are shown in Figure 17.3.2

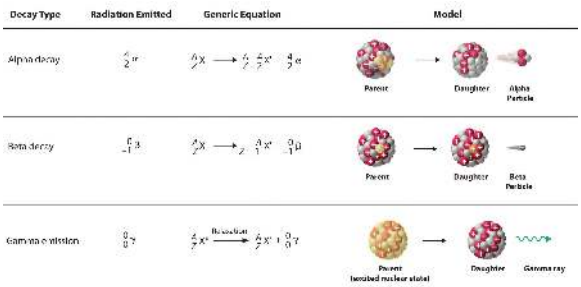


***Figure 17.3.2**: Three most common modes of nuclear decay.*

"Nuclear Accounting"

When writing nuclear equations, there are some general rules that will help you:

- The sum of the mass numbers (top numbers) on the reactant side equal the sum of the mass numbers on the product side.
- The atomic numbers (bottom numbers) on the two sides of the reaction will also be equal.

In the alpha decay of ${}^{238}\text{U}$ (Equation 1.2.8.1), both atomic and mass numbers are conserved:

- mass number: $238 = 4 + 234$
- atomic number: $92 = 2 + 90$

Confirm that this equation is correctly balanced by adding up the reactants' and products' atomic and mass numbers. Also, note that because this was an alpha reaction, one of the products is the alpha particle, ${}^{4}_{2}\text{He}$.

Note that both the mass numbers and the atomic numbers add up properly for the beta decay of thorium-234 (Equation 1.2.8.3):

- mass number: $234 = 0 + 234$
- atomic number: $90 = -1 + 91$

The mass numbers of the original nucleus and the new nucleus are the same because a neutron has been lost, but a proton has been gained, and so the sum of protons plus neutrons remains the same. The atomic number in the process has been increased by one since the new nucleus has one more proton than the original nucleus. In this beta decay, a thorium-234 nucleus has one more proton than the original nucleus. In this beta decay, a thorium-234 nucleus has become a protactinium-234 nucleus. Protactinium-234 is also a beta emitter and produces uranium-234.

$$ {}^{234}_{91}\text{Pa} \rightarrow {}^{0}_{-1}\text{e} + {}^{234}_{92}\text{U} \tag{1.2.8.4} $$

Once again, the atomic number increases by one and the mass number remains the same; this confirms that the equation is correctly balanced.

What About Balancing Charge?

Both alpha and beta particles are charged, but nuclear reactions in Equations 1.2.8.1, 1.2.8.3, and most of the other nuclear reactions above, are not balanced with respect to charge, as discussed when balancing redox reactions. When studying nuclear reactions in general, there is typically little information or concern about the chemical state of the radioactive isotopes, because the electrons from the electron cloud are not directly involved in the nuclear reaction (in contrast to chemical reactions).

So it is acceptable to **ignore** charge in balancing nuclear reactions, and concentrate on balancing mass and atomic numbers only.

Example 1.2.8.1

Complete the following nuclear reaction by filling in the missing particle.

$$ {}^{210}_{86}\text{Rn} \rightarrow {}^{4}_{2}\text{He} + ? $$

Solution

This reaction is an alpha decay. We can solve this problem one of two ways:

Solution 1: When an atom gives off an alpha particle, its atomic number drops by 2 and its mass number drops by 4, leaving: ${}^{206}_{84}\text{Po}$. We know the symbol is Po, for polonium, because this is the element with 84 protons on the periodic table.

Solution 2: Remember that the mass numbers on each side must total up to the same amount. The same is true of the atomic numbers.

- Mass numbers: $210 = 4 + ?$
- Atomic numbers: $86 = 2 + ?$

We are left with ${}^{206}_{84}\text{Po}$.

Example 1.2.8.2

Write each of the following nuclear reactions.

a) Carbon-14, used in carbon dating, decays by beta emission.

b) Uranium-238 decays by alpha emission.

Solution

a) Beta particles have the symbol ${}^{0}_{-1}\mathrm{e}$. Emitting a beta particle causes the atomic number to increase by 1 and the mass number to not change. We get atomic numbers and symbols for elements using our periodic table. We are left with the following reaction:

$$ {}^{14}_{6}\mathrm{C} \rightarrow {}^{0}_{-1}\mathrm{e} + {}^{14}_{7}\mathrm{N} $$

b) Alpha particles have the symbol ${}^{4}_{2}\mathrm{He}$. Emitting an alpha particle causes the atomic number to decrease by 2 and the mass number to decrease by 4. We are left with:

$$ {}^{238}_{92}\mathrm{U} \rightarrow {}^{4}_{2}\mathrm{He} + {}^{234}_{90}\mathrm{Th} $$

Decay Series

The decay of a radioactive nucleus is a move toward becoming stable. Often, a radioactive nucleus cannot reach a stable state through a single decay. In such cases, a series of decays will occur until a stable nucleus is formed. The decay of U-238 is an example of this. The U-238 decay series starts with U-238 and goes through fourteen separate decays to finally reach a stable nucleus, Pb-206 (Figure 17.3.3). There are similar decay series for U-235 and Th-232. The U-235 series ends with Pb-207 and the Th-232 series ends with Pb-208.

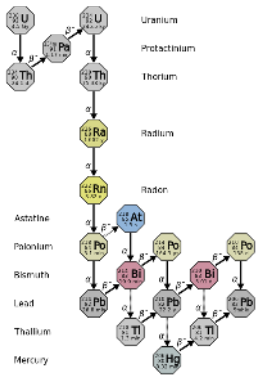


Figure 17.3.3: Uranium-238 decay chain. (CC-BY-3.0 Tosaka)

Several of the radioactive nuclei that are found in nature are present there because they are produced in one of the radioactive decay series. For example, there may have been radon on the earth at the time of its formation, but that original radon would have all decayed by this time. The radon that is present now is present because it was formed in a decay series (mostly by U-238).

Summary

A nuclear reaction is one that changes the structure of the nucleus of an atom. The atomic numbers and mass numbers in a nuclear equation must be balanced. Protons and neutrons are made up of quarks. The two most common modes of natural radioactivity are alpha decay and beta decay. Most nuclear reactions emit energy in the form of gamma rays.

Vocabulary

- **Alpha decay** - A common mode of radioactive decay in which a nucleus emits an alpha particle (a helium-4 nucleus).
- **Beta decay** - A common mode of radioactive decay in which a nucleus emits beta particles. The daughter nucleus will have a higher atomic number than the original nucleus.
- **Quark** - Particles that form one of the two basic constituents of matter. Various species of quarks combine in specific ways to form protons and neutrons, in each case taking exactly three quarks to make the composite particle.

1.2.8: Types of Radioactivity- Alpha, Beta, and Gamma Decay is shared under a CK-12 license and was authored, remixed, and/or curated by Henry Agnew & Marisa Alviar-Agnew.

- **17.3: Types of Radioactivity- Alpha, Beta, and Gamma Decay** by Henry Agnew, Marisa Alviar-Agnew is licensed CK-12. Original source: https://www.ck12.org/c/chemistry/.

1.2.9: Detecting Radioactivity

Learning Objectives

- Understand how the Geiger counter can be used to quantify the rate of ionization radiation.

When alpha, beta or gamma particles collide with a target, some of the energy in the particle is transferred to the target, typically resulting in the promotion of an electron to an “excited state”. In many “targets”, especially gasses, this results in *ionization*. Alpha, beta and gamma radiation are broadly referred to as **ionizing radiation**. A **Geiger counter** (or Geiger-Müller counter) takes advantage of this in order to detect these particles. In a Geiger tube, the electron produced by ionization of a captive gas travels to the anode and the change in voltage is detected by the attached circuitry. Most counters of this type are designed to emit an audible “click” in response to the change in voltage, and to also show it on a digital or analog meter. A simple schematic of a Geiger counter is shown in Figure 1.2.9.1.

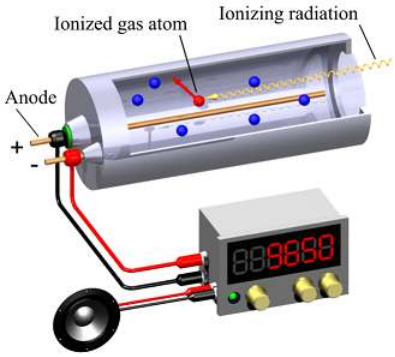


Figure 1.2.9.1: Schematic of a Geiger-Müller counter using an "end window" tube for low penetration radiation. A loudspeaker is also used for indication. (CC-BY-SA-3.0 Svjo-2 vai Wikipedia(opens in new window)).

Although scientists were not aware at the time of the Geiger counter's invention, all of us are subjected to a certain amount of radiation every day. This radiation is called **background radiation** and comes from a variety of natural and artificial radiation sources. Approximately 82% of background radiation comes from natural sources. These natural sources include:

1. Sources in the earth—including naturally occurring radioactive elements—which are incorporated in building materials, and also in the human body.
2. Sources from space in the form of cosmic rays.
3. Sources in the atmosphere, such as radioactive radon gas released from the earth; and radioactive atoms like carbon-14, produced in the atmosphere by bombardment from high-energy cosmic rays.

1.2.9: Detecting Radioactivity is shared under a CK-12 license and was authored, remixed, and/or curated by Henry Agnew & Marisa Alviar-Agnew.

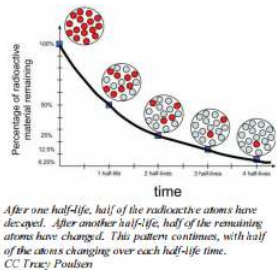
- **17.4: Detecting Radioactivity** by Henry Agnew, Marisa Alviar-Agnew is licensed CK-12. Original source: https://www.ck12.org/c/chemistry/.

1.2.10: Natural Radioactivity and Half-Life

Learning Objectives

- Describe what is meant by the term half-life and what factors affect half-life.
- Calculate the amount of radioactive material that will remain after an integral number of half-lives.
- Find the half-life of an isotope given graphical or other data.
- Describe how carbon-14 is used to determine the age of carbon-containing objects.

Rate of Radioactive Decay



After one half-life, half of the radioactive atoms have decayed. After another half-life, half of the remaining atoms have changed. This pattern continues, with half of the atoms changing over each half-life time.
CC Tracy Poulsen

During natural radioactive decay, not all atoms of an element are instantaneously changed to atoms of another element. The decay process takes time and there is value in being able to express the rate at which a process occurs. A useful concept is **half-life**, which is the time required for half of the starting material to change or decay. Half-lives can be calculated from measurements on the change in mass of a nuclide and the time it takes to occur. The only thing we know is that in the time of that substance's half-life, half of the original nuclei will disintegrate. Although chemical changes are sped up or slowed down by changing factors such as temperature and concentration, these factors have no effect on half-life. Each radioactive isotope will have its own unique half-life that is independent of any of these factors.

The half-lives of many radioactive isotopes have been determined; they have been found to range from extremely long half-lives of 10 billion years, to extremely short half-lives of fractions of a second.

Table 17.5.1: Table of Selected Half-lives

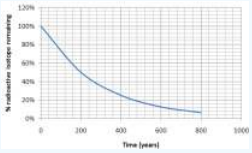
Element	Mass Number (A)	Half-life	Element	Mass Number (A)	Half Life
Uranium	238	4.5 Billion years	Californium	251	800 years
Neptunium	240	1 hour	Nobelium	254	3 seconds
Plutonium	243	5 hours	Carbon	14	5730 years
Americium	245	25 minutes	Carbon	16	740 milliseconds

The quantity of radioactive nuclei at any given time will decrease to half as much in one half-life. For example, if there are $100\ \text{g}$ of Cf-251 in a sample at some time, after 800 years, there will be $50\ \text{g}$ of Cf-251 remaining; after another 800 years (1600 years total), there will only be $25\ \text{g}$ remaining.

Remember, the half-life is the time it takes for half of your sample—no matter how much you have—to remain. Each half-life will follow the same general pattern as Cf-251. The only difference is the length of time it takes for half of a sample to decay.

Example 1.2.10.3

What is the half-life of an isotope that produces the following graph of decay over time?



Solution

We know that the half-life is the time it takes for half of a sample to change. How long did it take for half of our isotope to change? It took approximately 200 years for 100% of our sample to leave only 50% (half of the original amount) remaining. **The half-life is 200 years.**

*Notice that after another 200 years (400 years total), 25% remains (half of 50%).

Look carefully at the graph in the previous example. All types of radioactive decay make a graph of the same general shape. The only difference is the scale and units of the x-axis, as the half-life time will be different.

Example 1.2.10.2

If there are 60 grams of Np-240 present, how much Np-240 will remain after 4 hours? (Np-240 has a half-life of 1 hour.)

Solution

We create a table based on Np-240's half-life of 1 hour.

Amount of Np-240 present	Amount of time passed
60 g	0 (this is the amount before any time has passed)
30 g	1 hour (1 half-life)
15 g	2 hours (2 half-lives)
7.5 g	3 hours
3.75 g	4 hours

After 4 hours, only $3.75\ \text{g}$ of our original $60\ \text{g}$ sample would remain of the radioactive isotope Np-240.

Example 1.2.10.3

A sample of Ac-225 originally contained 80 grams, and after 50 days only 2.55 grams of the original Ac-225 remain. What is the half life of Ac-225?

Solution

We will tackle this problem similarly to the last problem. The difference is that we are looking for the half-life time. Let's set up a similar table, though:

Amount of Ac-225 present	Amount of time passed
80 g	0
40 g	1 half-life
20 g	2 half-lives
10 g	3 half-lives
5 g	4 half-lives
2.5 g	5 half-lives

We know that 50 days is the same as 5 half-lives. Therefore, 1 half-life is 10 days. The half-life of Ac-225 is 10 days.

Radioactive Dating

An ingenious application of half-life studies established a new science of determining ages of materials by half-life calculations. For geological dating, the decay of U-238 can be used. The half-life of U-238 is 4.5×10^9 years. The end product of the decay of U-238 is Pb-206. After one half-life, a 1.00 gram sample of uranium will have decayed to 0.50 grams of U-238 and 0.43 grams of Pb-206. By comparing the amount of U-238 to the amount of Pb-206 in a sample of uranium mineral, the age of the mineral can be estimated. Present day estimates for the age of the Earth's crust from this method is at 4 billion years.

Organic material (material made from things that were once living, such as paper and fabric) is radioactively dated using the long-lived nuclide of carbon, carbon-14. This method of determining the age of organic material (or once living materials) was given the name radiocarbon dating. The carbon dioxide consumed by living systems contains a certain concentration of $^{14}CO_2$. When an organism dies, the acquisition of carbon-14 stops, but the decay of the C-14 in the body continues. As time goes by, the ratio of C-14 to C-12 decreases at a rate determined by the half-life of C-14. Using half-life equations, the time since the organism died can be calculated. These procedures have been used to determine the age of organic artifacts and determine, for instance, whether art works are real or fake.

Summary and Vocabulary

The half-life of an isotope is used to describe the rate at which the isotope will decay and give off radiation. Using the half-life, it is possible to predict the amount of radioactive material that will remain after a given amount of time. C-14 dating procedures have been used to determine the age of organic artifacts. Its half-life is approximately 5700 years.

- **Background radiation** - Radiation that comes from environmental sources including the earth's crust, the atmosphere, cosmic rays, and radioisotopes. These natural sources of radiation account for the largest amount of radiation received by most people.
- **Half-life** - The time interval required for a quantity of (radioactive) material to decay to half of its original value.

1.2.10: Natural Radioactivity and Half-Life is shared under a CK-12 license and was authored, remixed, and/or curated by Henry Agnew & Marisa Alviar-Agnew.

- **17.5: Natural Radioactivity and Half-Life** by Henry Agnew, Marisa Alviar-Agnew is licensed CK-12. Original source: https://www.ck12.org/c/chemistry/.

SECTION OVERVIEW

1.3: Fluids

1.3.1: What Is a Fluid?

1.3.2: Pressure

1.3.3: Pressure Due to the Weight of Fluid

1.3.4: Archimedes' Principle

1.3.5: Bernoulli's Equation

1.3.6: Pascal's Principle

1.3: Fluids is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

1.3.1: What Is a Fluid?

Learning Objectives

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

Matter most commonly exists as a solid, liquid, or gas; these states are known as the three common *phases of matter*. Solids have a definite shape and a specific volume, liquids have a definite volume but their shape changes depending on the container in which they are held, and gases have neither a definite shape nor a specific volume as their molecules move to fill the container in which they are held. (See Figure 1.3.1.1.) Liquids and gases are considered to be fluids because they yield to shearing forces, whereas solids resist them. Note that the extent to which fluids yield to shearing forces (and hence flow easily and quickly) depends on a quantity called the viscosity. We can understand the phases of matter and what constitutes a fluid by considering the forces between atoms that make up matter in the three phases.

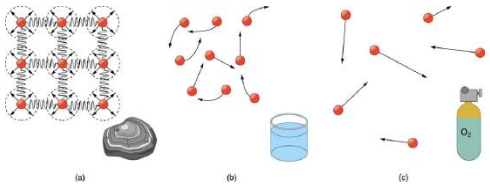


Figure 1.3.1.1: (a) Atoms in a solid always have the same neighbors, held near home by forces represented here by springs. These atoms are essentially in contact with one another. A rock is an example of a solid. This rock retains its shape because of the forces holding its atoms together. (b) Atoms in a liquid are also in close contact but can slide over one another. Forces between them strongly resist attempts to push them closer together and also hold them in close contact. Water is an example of a liquid. Water can flow, but it also remains in an open container because of the forces between its atoms. (c) Atoms in a gas are separated by distances that are considerably larger than the size of the atoms themselves, and they move about freely. A gas must be held in a closed container to prevent it from moving out freely.

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

CONNECTIONS: SUBMICROSCOPIC EXPLANATION OF SOLIDS AND LIQUIDS

Atomic and molecular characteristics explain and underlie the macroscopic characteristics of solids and fluids. This present section is devoted entirely to the submicroscopic explanation of solids and liquids.

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

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

Section Summary

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

Glossary

fluids

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

---

This page titled 1.3.1: What Is a Fluid? is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- **7.2: What Is a Fluid?** by OpenStax is licensed CC BY 4.0.

1.3.2: Pressure

Learning Objectives

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

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

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

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

Definition: PRESSURE

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

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

A given force can have a significantly different effect depending on the area over which the force is exerted, as shown in Figure 1.3.2.1. The SI unit for pressure is the pascal, where

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

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

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

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

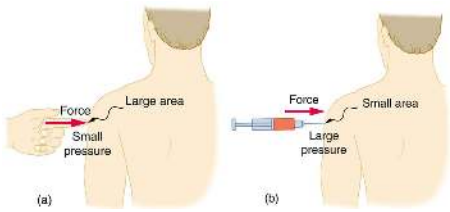


Figure 1.3.2.1: (a) While the person being poked with the finger might be irritated, the force has little lasting effect. (b) In contrast, the same force applied to an area the size of the sharp end of a needle is great enough to break the skin.

Example 1.3.2.1: Calculating Force Exerted by the Air: What Force Does a Pressure Exert?

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

Strategy

We can find the force exerted from the definition of pressure given in $P = \frac{F}{A}$, provided we can find the area A acted upon.

Solution

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

$$F = PA.$$

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

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

Discussion

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

The force exerted on the end of the tank is perpendicular to its inside surface. This direction is because the force is exerted by a static or stationary fluid. We have already seen that fluids cannot *withstand* shearing (sideways) forces; they cannot *exert* shearing forces, either. Fluid pressure has no direction, being a scalar quantity. The forces due to pressure have well-defined directions: they are always exerted perpendicular to any surface. (See the tire in Figure 1.3.2.2, for example.) Finally, note that pressure is exerted on all surfaces. Swimmers, as well as the tire, feel pressure on all sides. (See Figure 1.3.2.3.)

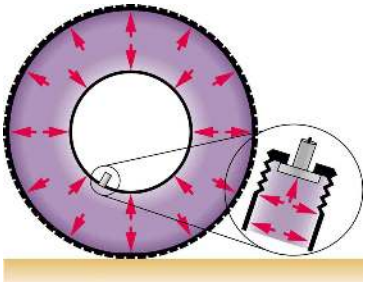


Figure 1.3.2.2: Pressure inside this tire exerts forces perpendicular to all surfaces it contacts. The arrows give representative directions and magnitudes of the forces exerted at various points. Note that static fluids do not exert shearing forces.

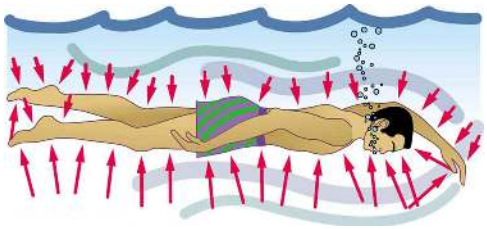


Figure 1.3.2.3: Pressure is exerted on all sides of this swimmer, since the water would flow into the space he occupies if he were not there. The arrows represent the directions and magnitudes of the forces exerted at various points on the swimmer. Note that the forces are larger underneath, due to greater depth, giving a net upward or buoyant force that is balanced by the weight of the swimmer.

Section Summary

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

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

- The SI unit of pressure is pascal and $1 \text{ Pa} = 1 \text{ N/m}^2$.

Glossary

pressure

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

This page titled 1.3.2: Pressure is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- **7.4: Pressure** by OpenStax is licensed CC BY 4.0.

1.3.3: Pressure Due to the Weight of Fluid

Learning Objectives

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

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

Consider the container in Figure 1.3.3.1. Its bottom supports the weight of the fluid in it. Let us calculate the pressure exerted on the bottom by the weight of the fluid. That **pressure** is the weight of the fluid mg divided by the area A supporting it (the area of the bottom of the container):

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

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

$$m = \rho V.$$

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

$$V = Ah,$$

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

$$m = \rho Ah.$$

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

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

The area cancels, and rearranging the variables yields

$$P = \rho gh.$$

This value is the **pressure due to the weight of a fluid**. The equation has general validity beyond the special conditions under which it is derived here. Even if the container were not there, the surrounding fluid would still exert this pressure, keeping the fluid static. Thus the equation $P = \rho gh$ represents the pressure due to the weight of any fluid of *average density* ρ at any depth h below its surface. For liquids, which are nearly incompressible, this equation holds to great depths.

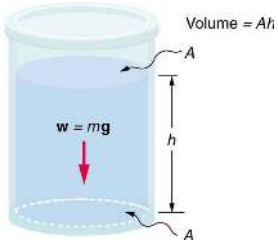


Figure 1.3.3.1: The bottom of this container supports the entire weight of the fluid in it. The vertical sides cannot exert an upward force on the fluid (since it cannot withstand a shearing force), and so the bottom must support it all.

Example 1.3.3.1: Calculating the Average Pressure and Force Exerted: What Force Must a Dam Withstand?

In Example 8.2.1, we calculated the mass of water in a large reservoir. We will now consider the pressure and force acting on the dam retaining water. (See Figure 1.3.3.2.) The dam is 500 m wide, and the water is 80.0 m deep at the dam. (a) What is the average pressure on the dam due to the water? (b) Calculate the force exerted against the dam and compare it with the weight of water in the dam (previously found to be 1.96×10^{13} N).

Strategy for (a)

The average pressure $\overline{P}$ due to the weight of the water is the pressure at the average depth $\bar{h}$ of 40.0 m, since pressure increases linearly with depth.

Solution for (a)

The average pressure due to the weight of a fluid is

$$\overline{P} = \rho g \bar{h}.$$

Entering the density of water from Table 8.2.1 and taking $\bar{h}$ to be the average depth of 40.0 m, we obtain

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

Strategy for (b)

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

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

Solution for (b)

We have already found the value for $\overline{P}$. The area of the dam is $A = 80.0 \text{ m} \times 500 \text{ m} = 4.00 \times 10^4 \text{ m}^2$, so that

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

Discussion

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

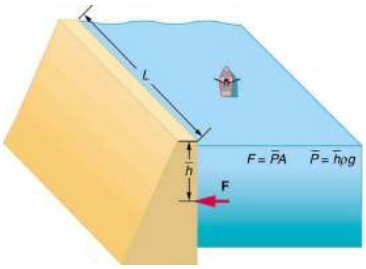


Figure 1.3.3.2: The dam must withstand the force exerted against it by the water it retains. This force is small compared with the weight of the water behind the dam.

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

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

This relationship means that, on average, at sea level, a column of air above $1.00\ \text{m}^2$ of the Earth's surface has a weight of 1.01×10^5 N, equivalent to 1 atm. (See Figure 1.3.3.3.)

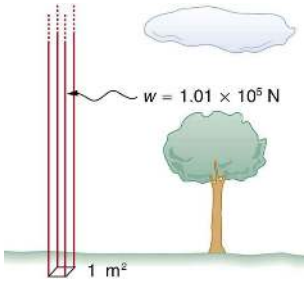


Figure 1.3.3.3: Atmospheric pressure at sea level averages 1.01×10^5 Pa (equivalent to 1 atm), since the column of air over this $1\ \text{m}^2$, extending to the top of the atmosphere, weighs 1.01×10^5 N.

Example 1.3.3.2: Calculating Depth Below the Surface of Water: What Depth of Water Creates the Same Pressure as the Entire Atmosphere?

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

Strategy

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

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

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

Solution

Entering the known values into the expression for h gives

$$h = \frac{1.01 \times 10^5\ \text{N/m}^2}{\left(1.00 \times 10^3\ \text{kg/m}^3\right)\left(9.80\ \text{m/s}^2\right)} = 10.3\ \text{m}.$$

Discussion

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

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

Section Summary

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

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

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

$$P = \rho gh,$$

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

Glossary

pressure due to the weight of fluid

pressure at a depth below a fluid surface due to its weight; given by $P = \rho gh$

This page titled 1.3.3: Pressure Due to the Weight of Fluid is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- **7.5: Pressure Due to the Weight of Fluid** by OpenStax is licensed CC BY 4.0.

1.3.4: Archimedes' Principle

Learning Objectives

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

When you rise from lounging in a warm bath, your arms feel strangely heavy. This is because you no longer have the buoyant support of the water. Where does this buoyant force come from? Why is it that some things float and others do not? Do objects that sink get any support at all from the fluid? Is your body buoyed by the atmosphere, or are only helium balloons affected? (See Figure 1.3.4.1.)

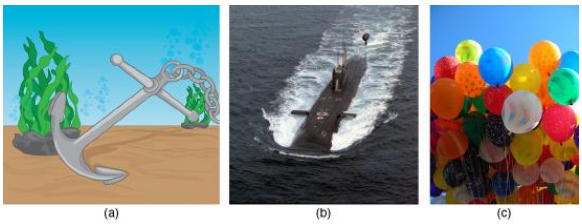


Figure 1.3.4.1: (a) Even objects that sink, like this anchor, are partly supported by water when submerged. (b) Submarines have adjustable density (ballast tanks) so that they may float or sink as desired. (credit: Allied Navy) (c) Helium-filled balloons tug upward on their strings, demonstrating air's buoyant effect. (credit: Crystl)

Answers to all these questions, and many others, are based on the fact that pressure increases with depth in a fluid. This means that the upward force on the bottom of an object in a fluid is greater than the downward force on the top of the object. There is a net upward, or **buoyant force** on any object in any fluid. (See Figure 1.3.4.2.) If the buoyant force is greater than the object's weight, the object will rise to the surface and float. If the buoyant force is less than the object's weight, the object will sink. If the buoyant force equals the object's weight, the object will remain suspended at that depth. The buoyant force is always present whether the object floats, sinks, or is suspended in a fluid.

Definition: BUOYANT FORCE

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

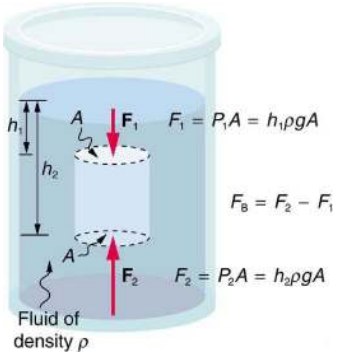


Figure 1.3.4.2: Pressure due to the weight of a fluid increases with depth since $P = \rho gh$. This pressure and associated upward force on the bottom of the cylinder are greater than the downward force on the top of the cylinder. Their difference is the buoyant force $\mathbf{F}_B$. (Horizontal forces cancel.)

Just how great is this buoyant force? To answer this question, think about what happens when a submerged object is removed from a fluid, as in Figure 1.3.4.3.

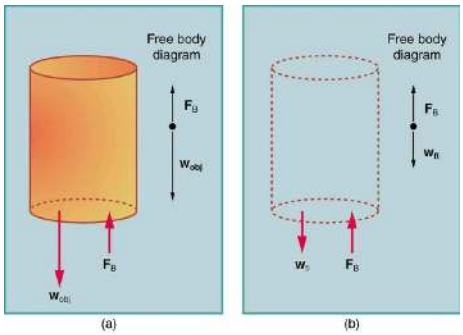


Figure 1.3.4.3: (a) An object submerged in a fluid experiences a buoyant force $\mathbf{F}_\mathrm{B}$. If $\mathbf{F}_\mathrm{B}$ is greater than the weight of the object, the object will rise. If $\mathbf{F}_\mathrm{B}$ is less than the weight of the object, the object will sink. (b) If the object is removed, it is replaced by fluid having weight w_fl. Since this weight is supported by surrounding fluid, the buoyant force must equal the weight of the fluid displaced. That is, $F_\mathrm{B} = w_\mathrm{fl}$,a statement of Archimedes' principle.

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

$$F_\mathrm{B} = w_\mathrm{fl},$$

where $\mathbf{F}_\mathrm{B}$ is the buoyant force and w_fl is the weight of the fluid displaced by the object. Archimedes' principle is valid in general, for any object in any fluid, whether partially or totally submerged.

Definition: ARCHIMEDES' PRINCIPLE

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

$$F_\mathrm{B} = w_\mathrm{fl},$$

where $\mathbf{F}_\mathrm{B}$ is the buoyant force and w_fl is the weight of the fluid displaced by the object.

Humm … High-tech body swimsuits were introduced in 2008 in preparation for the Beijing Olympics. One concern (and international rule) was that these suits should not provide any buoyancy advantage. How do you think that this rule could be verified?

MAKING CONNECTIONS: TAKE-HOME INVESTIGATION

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

Floating and Sinking

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

Example 1.3.4.1: Calculating buoyant force: dependency on shape

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

Strategy for (a)

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

Solution for (a)

First, we use the definition of density $\rho = \frac{m}{V}$ to find the steel's volume, and then we substitute values for mass and density. This gives

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

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

$$\begin{aligned} m_{\text{w}} &= \rho_{\text{w}} V_{\text{w}} = \left(1.000 \times 10^3 \text{ kg/m}^3\right)\left(1.28 \times 10^3 \text{ m}^3\right) \\ &= 1.28 \times 10^6 \text{ kg} \end{aligned}.$$

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

$$\begin{aligned} F_{\text{B}} &= w_{\text{w}} = m_{\text{w}} g = \left(1.28 \times 10^6 \text{ kg}\right)\left(9.80 \text{ m/s}^2\right) \\ &= 1.3 \times 10^7 \text{ N} \end{aligned}.$$

The steel's weight is $m_{\text{w}}g = 9.80 \times 10^7$ N, which is much greater than the buoyant force, so the steel will remain submerged. Note that the buoyant force is rounded to two digits because the density of steel is given to only two digits.

Strategy for (b)

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

Solution for (b)

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

$$\begin{aligned} m_{\text{w}} &= \rho_{\text{w}} V_{\text{w}} = \left(1.000 \times 10^3 \text{ kg/m}^3\right)\left(1.00 \times 10^5 \text{ m}^3\right) \\ &= 1.00 \times 10^8 \text{ kg} \end{aligned}.$$

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

$$\begin{aligned} F_{\text{B}} &= w_{\text{w}} = m_{\text{w}} g = \left(1.00 \times 10^8 \text{ kg}\right)\left(9.80 \text{ m/s}^2\right) \\ &= 9.80 \times 10^8 \text{ N} \end{aligned}.$$

Discussion

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

MAKING CONNECTIONS: TAKE-HOME INVESTIGATION

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

Density and Archimedes' Principle

Density plays a crucial role in Archimedes' principle. The average density of an object is what ultimately determines whether it floats. If its average density is less than that of the surrounding fluid, it will float. This is because the fluid, having a higher density, contains more mass and hence more weight in the same volume. The buoyant force, which equals the weight of the fluid displaced, is thus greater than the weight of the object. Likewise, an object denser than the fluid will sink.

The extent to which a floating object is submerged depends on how the object's density is related to that of the fluid. In Figure 1.3.4.4, for example, the unloaded ship has a lower density and less of it is submerged compared with the same ship loaded. We can

derive a quantitative expression for the fraction submerged by considering density. The fraction submerged is the ratio of the volume submerged to the volume of the object, or

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

The volume submerged equals the volume of fluid displaced, which we call V_{fl}. Now we can obtain the relationship between the densities by substituting $\rho = \frac{m}{V}$ into the expression. This gives

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

where $\bar{\rho}_{\text{obj}}$ is the average density of the object and ρ_{fl} is the density of the fluid. Since the object floats, its mass and that of the displaced fluid are equal, and so they cancel from the equation, leaving

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

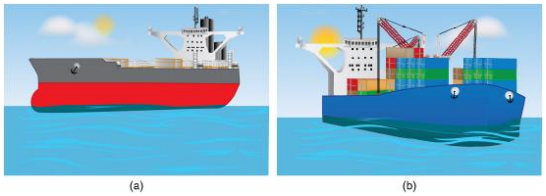


Figure 1.3.4.4: An unloaded ship (a) floats higher in the water than a loaded ship (b).

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

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

where $\bar{\rho}$ is the average density of the object or substance and ρ_{w} is the density of water at 4.00°C. Specific gravity is dimensionless, independent of whatever units are used for ρ. If an object floats, its specific gravity is less than one. If it sinks, its specific gravity is greater than one. Moreover, the fraction of a floating object that is submerged equals its specific gravity. If an object's specific gravity is exactly 1, then it will remain suspended in the fluid, neither sinking nor floating. Scuba divers try to obtain this state so that they can hover in the water. We measure the specific gravity of fluids, such as battery acid, radiator fluid, and urine, as an indicator of their condition. One device for measuring specific gravity is shown in Figure 1.3.4.5.

Definition: SPECIFIC GRAVITY

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

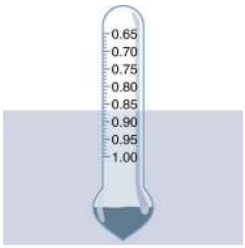


Figure 1.3.4.5: This hydrometer is floating in a fluid of specific gravity 0.87. The glass hydrometer is filled with air and weighted with lead at the bottom. It floats highest in the densest fluids and has been calibrated and labeled so that specific gravity can be read from it directly.

Example 1.3.4.2: Calculating Average Density: Floating Woman

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

Strategy

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

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

for the density of the object. This yields

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

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

Solution

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

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

Discussion

Her density is less than the fluid density. We expect this because she floats. Body density is one indicator of a person's percent body fat, of interest in medical diagnostics and athletic training. (See Figure 1.3.4.6.)



Figure 1.3.4.6: Subject in a "fat tank," where he is weighed while completely submerged as part of a body density determination. The subject must completely empty his lungs and hold a metal weight in order to sink. Corrections are made for the residual air in his lungs (measured separately) and the metal weight. His corrected submerged weight, his weight in air, and pinch tests of strategic fatty areas are used to calculate his percent body fat.

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

More Density Measurements

One of the most common techniques for determining density is shown in Figure 1.3.4.7.

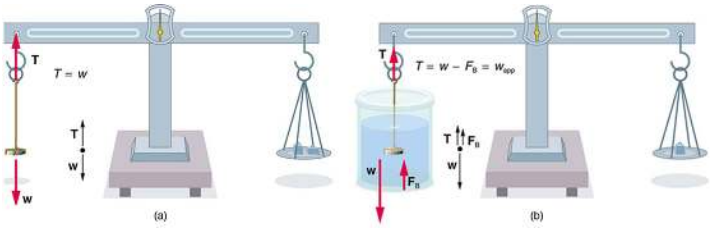


Figure 1.3.4.7: (a) A coin is weighed in air. (b) The apparent weight of the coin is determined while it is completely submerged in a fluid of known density. These two measurements are used to calculate the density of the coin.

An object, here a coin, is weighed in air and then weighed again while submerged in a liquid. The density of the coin, an indication of its authenticity, can be calculated if the fluid density is known. This same technique can also be used to determine the density of the fluid if the density of the coin is known. All of these calculations are based on Archimedes' principle.

Archimedes' principle states that the buoyant force on the object equals the weight of the fluid displaced. This, in turn, means that the object *appears* to weigh less when submerged; we call this measurement the object's *apparent weight*. The object suffers an *apparent weight loss* equal to the weight of the fluid displaced. Alternatively, on balances that measure mass, the object suffers an *apparent mass loss* equal to the mass of fluid displaced. That is

$$\text{apparent weight loss} = \text{weight of fluid displaced}$$

or

$$\text{apparent mass loss} = \text{mass of fluid displaced.}$$

The next example illustrates the use of this technique.

Example 1.3.4.3: Calculating Density: Is the Coin Authentic?

The mass of an ancient Greek coin is determined in air to be 8.630 g. When the coin is submerged in water as shown in Figure 1.3.4.7, its apparent mass is 7.800 g. Calculate its density, given that water has a density of $1.000\ \text{g/cm}^3$ and that effects caused by the wire suspending the coin are negligible.

Strategy

To calculate the coin's density, we need its mass (which is given) and its volume. The volume of the coin equals the volume of water displaced. The volume of water displaced V_w can be found by solving the equation for density $\rho = \frac{m}{V}$ for V.

Solution

The volume of water is $V_\text{w} = \frac{m_\text{w}}{\rho_\text{w}}$ where m_w is the mass of water displaced. As noted, the mass of the water displaced equals the apparent mass loss, which is $m_\text{w} = 8.630\ \text{g} - 7.800\ \text{g} = 0.830\ \text{g}$. Thus the volume of water is $V_\text{w} = \frac{0.830\ \text{g}}{1.000\ \text{g/cm}^3} = 0.830\ \text{cm}^3$. This is also the volume of the coin, since it is completely submerged. We can now find the density of the coin using the definition of density:

$$\rho_\text{c} = \frac{m_\text{c}}{V_\text{c}} = \frac{8.630\ \text{g}}{0.830\ \text{cm}^3} = 10.4\ \text{g/cm}^3.$$

Discussion

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

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

Section Summary

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

Glossary

Archimedes' principle
the buoyant force on an object equals the weight of the fluid it displaces

buoyant force

the net upward force on any object in any fluid

specific gravity

the ratio of the density of an object to a fluid (usually water)

This page titled 1.3.4: Archimedes' Principle is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- **7.6: Archimedes' Principle** by OpenStax is licensed CC BY 4.0.

1.3.5: Bernoulli's Equation

Learning Objectives

- Explain the terms in Bernoulli's equation.
- Explain how Bernoulli's equation is related to conservation of energy.
- Explain how to derive Bernoulli's principle from Bernoulli's equation.
- Calculate with Bernoulli's principle.
- List some applications of Bernoulli's principle.

When a fluid flows into a narrower channel, its speed increases. That means its kinetic energy also increases. Where does that change in kinetic energy come from? The increased kinetic energy comes from the net work done on the fluid to push it into the channel and the work done on the fluid by the gravitational force, if the fluid changes vertical position. Recall the work-energy theorem,

$$W_{\text{net}} = \frac{1}{2}mv^2 - \frac{1}{2}mv_0^2.$$

There is a pressure difference when the channel narrows. This pressure difference results in a net force on the fluid: recall that pressure times area equals force. The net work done increases the fluid's kinetic energy. As a result, the *pressure will drop in a rapidly-moving fluid*, whether or not the fluid is confined to a tube.

There are a number of common examples of pressure dropping in rapidly-moving fluids. Shower curtains have a disagreeable habit of bulging into the shower stall when the shower is on. The high-velocity stream of water and air creates a region of lower pressure inside the shower, and standard atmospheric pressure on the other side. The pressure difference results in a net force inward pushing the curtain in. You may also have noticed that when passing a truck on the highway, your car tends to veer toward it. The reason is the same—the high velocity of the air between the car and the truck creates a region of lower pressure, and the vehicles are pushed together by greater pressure on the outside. (See Figure 1.3.5.1.) This effect was observed as far back as the mid-1800s, when it was found that trains passing in opposite directions tipped precariously toward one another.

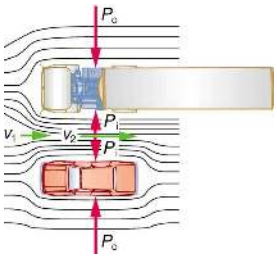


Figure 1.3.5.1: An overhead view of a car passing a truck on a highway. Air passing between the vehicles flows in a narrower channel and must increase its speed (v_2 is greater than v_1), causing the pressure between them to drop (P_i is less than P_o). Greater pressure on the outside pushes the car and truck together.

MAKING CONNECTIONS: TAKE-HOME INVESTIGATION WITH A SHEET OF PAPER

Hold the short edge of a sheet of paper parallel to your mouth with one hand on each side of your mouth. The page should slant downward over your hands. Blow over the top of the page. Describe what happens and explain the reason for this behavior.

Bernoulli's Equation

The relationship between pressure and velocity in fluids is described quantitatively by **Bernoulli's equation**, named after its discoverer, the Swiss scientist Daniel Bernoulli (1700–1782). Bernoulli's equation states that for an incompressible, frictionless fluid, the following sum is constant:

$$P + \frac{1}{2}\rho v^2 + \rho g h = \text{ constant },$$

where P is the absolute pressure, ρ is the fluid density, v is the velocity of the fluid, h is the height above some reference point, and g is the acceleration due to gravity. If we follow a small volume of fluid along its path, various quantities in the sum may change,

but the total remains constant. Let the subscripts 1 and 2 refer to any two points along the path that the bit of fluid follows; Bernoulli's equation becomes

$$P_1 + \frac{1}{2}\rho v_1^2 + \rho g h_1 = P_2 + \frac{1}{2}\rho v_2^2 + \rho g h_2.$$

Bernoulli's equation is a form of the conservation of energy principle. Note that the second and third terms are the kinetic and potential energy with m replaced by ρ. In fact, each term in the equation has units of energy per unit volume. We can prove this for the second term by substituting $\rho = m/V$ into it and gathering terms:

$$\frac{1}{2}\rho v^2 = \frac{\frac{1}{2}mv^2}{V} = \frac{\text{KE}}{V}.$$

So $\frac{1}{2}\rho v^2$ is the kinetic energy per unit volume. Making the same substitution into the third term in the equation, we find

$$\rho g h = \frac{mgh}{V} = \frac{\text{PE}_\text{g}}{V},$$

so $\rho g h$ is the gravitational potential energy per unit volume. Note that pressure P has units of energy per unit volume, too. Since $P = F/A$, its units are N/m^2. If we multiply these by m/m, we obtain $\text{N} \cdot \text{m/m}^3 = \text{J/m}^3$, or energy per unit volume. Bernoulli's equation is, in fact, just a convenient statement of conservation of energy for an incompressible fluid in the absence of friction.

MAKING CONNECTIONS: CONSERVATION OF ENERGY

Conservation of energy applied to fluid flow produces Bernoulli's equation. The net work done by the fluid's pressure results in changes in the fluid's KE and PE_g per unit volume. If other forms of energy are involved in fluid flow, Bernoulli's equation can be modified to take these forms into account. Such forms of energy include thermal energy dissipated because of fluid viscosity.

The general form of Bernoulli's equation has three terms in it, and it is broadly applicable. To understand it better, we will look at a number of specific situations that simplify and illustrate its use and meaning.

Bernoulli's Equation for Static Fluids

Let us first consider the very simple situation where the fluid is static—that is, $v_1 = v_2 = 0$. Bernoulli's equation in that case is

$$P_1 + \rho g h_1 = P_2 + \rho g h_2.$$

We can further simplify the equation by taking $h_2 = 0$ (we can always choose some height to be zero, just as we often have done for other situations involving the gravitational force, and take all other heights to be relative to this). In that case, we get

$$P_2 = P_1 + \rho g h_1.$$

This equation tells us that, in static fluids, pressure increases with depth. As we go from point 1 to point 2 in the fluid, the depth increases by h_1, and consequently, P_2 is greater than P_1 by an amount $\rho g h_1$. In the very simplest case, P_1 is zero at the top of the fluid, and we get the familiar relationship $P = \rho g h$. (Recall that $P = \rho g h$ and $\Delta\text{PE}_\text{g} = mgh$) Bernoulli's equation includes the fact that the pressure due to the weight of a fluid is $\rho g h$. Although we introduce Bernoulli's equation for fluid flow, it includes much of what we studied for static fluids in the preceding chapter.

Bernoulli's Principle—Bernoulli's Equation at Constant Depth

Another important situation is one in which the fluid moves but its depth is constant—that is, $h_1 = h_2$. Under that condition, Bernoulli's equation becomes

$$P_1 + \frac{1}{2}\rho v_1^2 = P_2 + \frac{1}{2}\rho v_2^2.$$

Situations in which fluid flows at a constant depth are so important that this equation is often called **Bernoulli's principle**. It is Bernoulli's equation for fluids at constant depth. (Note again that this applies to a small volume of fluid as we follow it along its path.) As we have just discussed, pressure drops as speed increases in a moving fluid. We can see this from Bernoulli's principle. For example, if v_2 is greater than v_1 in the equation, then P_2 must be less than P_1 for the equality to hold.

Example 1.3.5.1: Calculating Pressure: Pressure Drops as a Fluid Speeds Up

In Example 8.6.2, we found that the speed of water in a hose increased from 1.96 m/s to 25.5 m/s going from the hose to the nozzle. Calculate the pressure in the hose, given that the absolute pressure in the nozzle is $1.01 \times 10^5\ \mathrm{N/m^2}$ (atmospheric, as it must be) and assuming level, frictionless flow.

Strategy

Level flow means constant depth, so Bernoulli's principle applies. We use the subscript 1 for values in the hose and 2 for those in the nozzle. We are thus asked to find P_1.

Solution

Solving Bernoulli's principle for P_1 yields

$$P_1 = P_2 + \frac{1}{2}\rho v_2^2 - \frac{1}{2}\rho v_1^2 = P_2 + \frac{1}{2}\rho\left(v_2^2 - v_1^2\right).$$

Substituting known values,

$$\begin{aligned} P_1 =& 1.01 \times 10^5\ \mathrm{N/m^2} \\ &+ \frac{1}{2}\left(10^3\ \mathrm{kg/m^3}\right)\left[(25.5\ \mathrm{m/s})^2 - (1.96\ \mathrm{m/s})^2\right] \\ =& 4.24 \times 10^5\ \mathrm{N/m^2}. \end{aligned}$$

Discussion

This absolute pressure in the hose is greater than in the nozzle, as expected since v is greater in the nozzle. The pressure P_2 in the nozzle must be atmospheric since it emerges into the atmosphere without other changes in conditions.

Applications of Bernoulli's Principle

There are a number of devices and situations in which fluid flows at a constant height and, thus, can be analyzed with Bernoulli's principle.

Entrainment

People have long put the Bernoulli principle to work by using reduced pressure in high-velocity fluids to move things about. With a higher pressure on the outside, the high-velocity fluid forces other fluids into the stream. This process is called *entrainment*. Entrainment devices have been in use since ancient times, particularly as pumps to raise water small heights, as in draining swamps, fields, or other low-lying areas. Some other devices that use the concept of entrainment are shown in Figure 1.3.5.2.

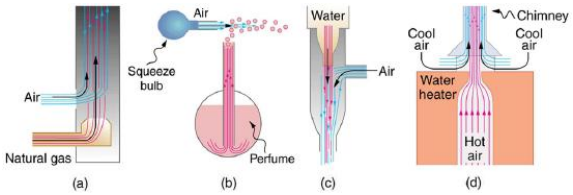


Figure 1.3.5.2: Examples of entrainment devices that use increased fluid speed to create low pressures, which then entrain one fluid into another. (a) A Bunsen burner uses an adjustable gas nozzle, entraining air for proper combustion. (b) An atomizer uses a squeeze bulb to create a jet of air that entrains drops of perfume. Paint sprayers and carburetors use very similar techniques to move their respective liquids. (c) A common aspirator uses a high-speed stream of water to create a region of lower pressure. Aspirators may be used as suction pumps in dental and surgical situations or for draining a flooded basement or producing a reduced pressure in a vessel. (d) The chimney of a water heater is designed to entrain air into the pipe leading through the ceiling.

Wings and Sails

The airplane wing is a beautiful example of Bernoulli's principle in action. Figure 1.3.5.3(a) shows the characteristic shape of a wing. The wing is tilted upward at a small angle and the upper surface is longer, causing air to flow faster over it. The pressure on top of the wing is therefore reduced, creating a net upward force or lift. (Wings can also gain lift by pushing air downward, utilizing the conservation of momentum principle. The deflected air molecules result in an upward force on the wing — Newton's third law.) Sails also have the characteristic shape of a wing. (See Figure 1.3.5.3(b).) The pressure on the front side of the sail, P_{front}, is lower than the pressure on the back of the sail, P_{back}. This results in a forward force and even allows you to sail into the wind.

MAKING CONNECTIONS: TAKE-HOME INVESTIGATION WITH TWO STRIPS OF PAPER

For a good illustration of Bernoulli's principle, make two strips of paper, each about 15 cm long and 4 cm wide. Hold the small end of one strip up to your lips and let it drape over your finger. Blow across the paper. What happens? Now hold two strips of paper up to your lips, separated by your fingers. Blow between the strips. What happens?

Velocity measurement

Figure 1.3.5.4 shows two devices that measure fluid velocity based on Bernoulli's principle. The manometer in Figure 1.3.5.4(a) is connected to two tubes that are small enough not to appreciably disturb the flow. The tube facing the oncoming fluid creates a dead spot having zero velocity ($v_1 = 0$) in front of it, while fluid passing the other tube has velocity v_2. This means that Bernoulli's principle as stated in $P_1 + \frac{1}{2}\rho v_1^2 = P_2 + \frac{1}{2}\rho v_2^2$ becomes

$$P_1 = P_2 + \frac{1}{2}\rho v_2^2.$$

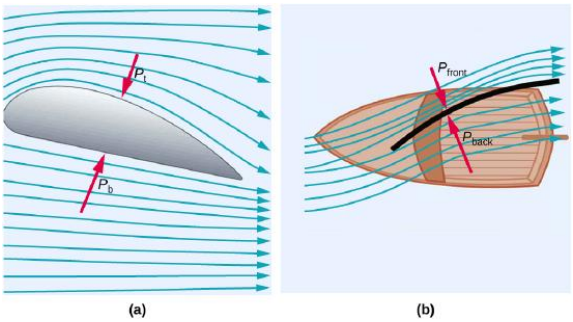


Figure 1.3.5.3: (a) The Bernoulli principle helps explain lift generated by a wing. (b) Sails use the same technique to generate part of their thrust.

Thus pressure P_2 over the second opening is reduced by $\frac{1}{2}\rho v_2^2$, and so the fluid in the manometer rises by h on the side connected to the second opening, where

$$h \propto \frac{1}{2}\rho v_2^2.$$

(Recall that the symbol $\propto$ means "proportional to.") Solving for v_2, we see that

$$v_2 \propto \sqrt{h}.$$

Figure 1.3.5.4(b) shows a version of this device that is in common use for measuring various fluid velocities; such devices are frequently used as air speed indicators in aircraft.

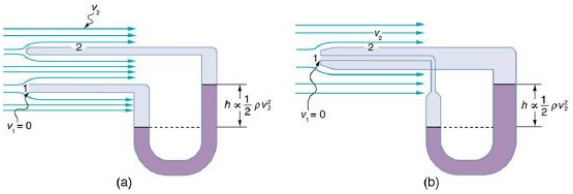


Figure 1.3.5.4: Measurement of fluid speed based on Bernoulli's principle. (a) A manometer is connected to two tubes that are close together and small enough not to disturb the flow. Tube 1 is open at the end facing the flow. A dead spot having zero speed is created there. Tube 2 has an opening on the side, and so the fluid has a speed v across the opening; thus, pressure there drops. The difference in pressure at the manometer is $\frac{1}{2}\rho v_2^2$, and so h is proportional to $\frac{1}{2}\rho v_2^2$. (b) This type of velocity measuring device is a Prandtl tube, also known as a pitot tube.

Summary

- Bernoulli's equation states that the sum on each side of the following equation is constant, or the same at any two points in an incompressible frictionless fluid:

$$P_1 + \frac{1}{2}\rho v_1^2 + \rho g h_1 = P_2 + \frac{1}{2}\rho v_2^2 + \rho g h_2.$$

- Bernoulli’s principle is Bernoulli’s equation applied to situations in which depth is constant. The terms involving depth (or height h) subtract out, yielding

$$P_1 + \frac{1}{2}\rho v_1^2 = P_2 + \frac{1}{2}\rho v_2^2.$$

- Bernoulli’s principle has many applications, including entrainment, wings and sails, and velocity measurement.

Glossary

Bernoulli’s equation

the equation resulting from applying conservation of energy to an incompressible frictionless fluid: $P + 1/2pv^2 + pgh$ = constant , through the fluid

Bernoulli’s principle

Bernoulli’s equation applied at constant depth: $P_1 + 1/2pv_1{}^2 = P_2 + 1/2pv_2{}^2$

This page titled 1.3.5: Bernoulli’s Equation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- **7.8: Bernoulli’s Equation** by OpenStax is licensed CC BY 4.0.

1.3.6: Pascal's Principle

Learning Objectives

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

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

Pressure is defined as force per unit area. Can pressure be increased in a fluid by pushing directly on the fluid? Yes, but it is much easier if the fluid is enclosed. The heart, for example, increases blood pressure by pushing directly on the blood in an enclosed system (valves closed in a chamber). If you try to push on a fluid in an open system, such as a river, the fluid flows away. An enclosed fluid cannot flow away, and so pressure is more easily increased by an applied force.

What happens to a pressure in an enclosed fluid? Since atoms in a fluid are free to move about, they transmit the pressure to all parts of the fluid and to the walls of the container. Remarkably, the pressure is transmitted *undiminished*. This phenomenon is called Pascal's principle, because it was first clearly stated by the French philosopher and scientist Blaise Pascal (1623–1662): A change in pressure applied to an enclosed fluid is transmitted undiminished to all portions of the fluid and to the walls of its container.

Definition: Pascal's Principle

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

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

Blaise Pascal had an interesting life in that he was home-schooled by his father who removed all of the mathematics textbooks from his house and forbade him to study mathematics until the age of 15. This, of course, raised the boy's curiosity, and by the age of 12, he started to teach himself geometry. Despite this early deprivation, Pascal went on to make major contributions in the mathematical fields of probability theory, number theory, and geometry. He is also well known for being the inventor of the first mechanical digital calculator, in addition to his contributions in the field of fluid statics.

Application of Pascal's Principle

One of the most important technological applications of Pascal's principle is found in a *hydraulic system*, which is an enclosed fluid system used to exert forces. The most common hydraulic systems are those that operate car brakes. Let us first consider the simple hydraulic system shown in Figure 1.3.6.1.

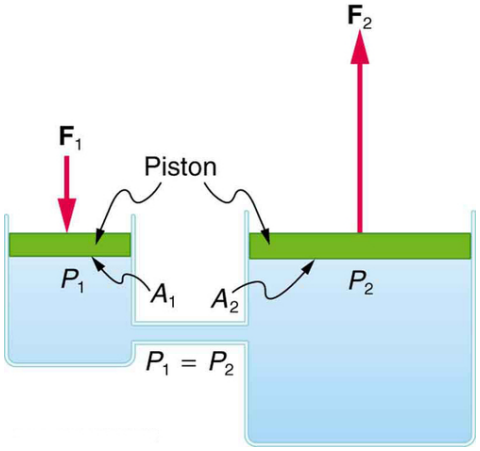


Figure 1.3.6.1: A typical hydraulic system with two fluid-filled cylinders, capped with pistons and connected by a tube called a hydraulic line. A downward force F_1 on the left piston creates a pressure that is transmitted undiminished to all parts of the enclosed fluid. This results in an upward force F_2 on the right piston that is larger than F_1 because the right piston has a larger area.

Relationship Between Forces in a Hydraulic System

We can derive a relationship between the forces in the simple hydraulic system shown in Figure 1.3.6.1 by applying Pascal's principle. Note first that the two pistons in the system are at the same height, and so there will be no difference in pressure due to a difference in depth. Now the pressure due to F_1 acting on area A_1 is simply $P_1 = \frac{F_1}{A_1}$, as defined by $P = \frac{F}{A}$. According to Pascal's principle, this pressure is transmitted undiminished throughout the fluid and to all walls of the container. Thus, a pressure P_2 is felt at the other piston that is equal to P_1. That is

$$P_1 = P_2 \tag{1.3.6.1}$$

But since $P_2 - \frac{F_2}{A_2}$, we see that

$$\frac{F_1}{A_1} = \frac{F_2}{A_2}. \tag{1.3.6.2}$$

This equation relates the ratios of force to area in any hydraulic system, providing the pistons are at the same vertical height and that friction in the system is negligible. Hydraulic systems can increase or decrease the force applied to them. To make the force larger, the pressure is applied to a larger area. For example, if a 100-N force is applied to the left cylinder in Figure 1.3.6.1 and the right one has an area five times greater, then the force out is 500 N. Hydraulic systems are analogous to simple levers, but they have the advantage that pressure can be sent through tortuously curved lines to several places at once.

Example 1.3.6.1: Calculating Force of Slave Cylinders: Pascal Puts on the Brakes

Consider the automobile hydraulic system shown in Figure 1.3.6.2.

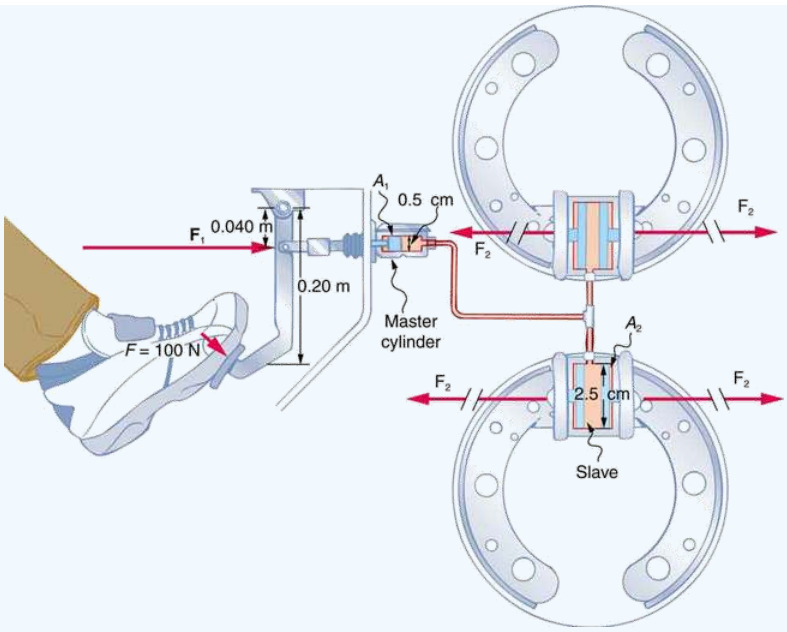


Figure 1.3.6.2: Hydraulic brakes use Pascal's principle. The driver exerts a force of 100 N on the brake pedal. This force is increased by the simple lever and again by the hydraulic system. Each of the identical slave cylinders receives the same pressure and, therefore, creates the same force output F_2. The circular cross-sectional areas of the master and slave cylinders are represented by A_1 and A_2, respectively

A force of 100 N is applied to the brake pedal, which acts on the cylinder—called the master—through a lever. A force of 500 N is exerted on the master cylinder. (The reader can verify that the force is 500 N using techniques of statics from Applications of Statics, Including Problem-Solving Strategies.) Pressure created in the master cylinder is transmitted to four so-called slave cylinders. The master cylinder has a diameter of 0.500 cm, and each slave cylinder has a diameter of 2.50 cm. Calculate the force F_2 created at each of the slave cylinders.

Strategy

We are given the force F_1 that is applied to the master cylinder. The cross-sectional areas A_1 and A_2 can be calculated from their given diameters. Then Equation 1.3.6.2 can be used to find the force F_2. Manipulate this algebraically to get F_2

on one side and substitute known values:

Solution

Pascal's principle applied to hydraulic systems is given by Equation 1.3.6.2:

$$\begin{aligned} F_2 &= \frac{A_2}{A_1} F_1 \\ &= \frac{\pi r_2^2}{\pi r_1^2} F_1 \\ &= \frac{(1.25\,cm)^2}{(0.250\,cm)^2} \times 500\,N \\ &= 1.25 \times 10^4\,N. \end{aligned}$$

Discussion

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

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

Definition: Conservation of Energy

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

Summary

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

Glossary

Pascal's Principle

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

This page titled 1.3.6: Pascal's Principle is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **11.5: Pascal's Principle** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/college-physics.

CHAPTER OVERVIEW

2: Motion



Under Construction(Public Domain ; Public Domain Pictures)

2.1: Introduction to One-Dimensional Kinematics

2.1.1: Displacement

2.1.2: Time, Velocity, and Speed

2.1.3: Acceleration

2.2: Motion Equations for Constant Acceleration in One Dimension

2.3: Falling Objects

2.3.1: Projectile Motion

2.3.E: Kinematics (Exercise)

2.4: Introduction to Dynamics- Newton’s Laws of Motion

2.4.1: Development of Force Concept

2.4.2: Newton’s First Law of Motion- Inertia

2.4.3: Newton's Second Law of Motion- Force and Acceleration

2.4.4: Newton’s Third Law of Motion- Symmetry in Forces

2.4.5: Friction

2.4.E: Dynamics (Exercise)

2: Motion is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

2.1: Introduction to One-Dimensional Kinematics



Figure 2.1.1: The motion of an American kestrel through the air can be described by the bird's displacement, speed, velocity, and acceleration. When it flies in a straight line without any change in direction, its motion is said to be one dimensional. (credit: Vince Maidens, Wikimedia Commons)

Objects are in motion everywhere we look. Everything from a tennis game to a space-probe flyby of the planet Neptune involves motion. When you are resting, your heart moves blood through your veins. And even in inanimate objects, there is continuous motion in the vibrations of atoms and molecules. Questions about motion are interesting in and of themselves: *How long will it take for a space probe to get to Mars?* But an understanding of motion is also key to understanding other concepts in physics. An understanding of acceleration, for example, is crucial to the study of force.

Our formal study of physics begins with **kinematics** which is defined as the *study of motion without considering its causes*. The word "kinematics" comes from a Greek term meaning motion and is related to other English words such as "cinema" (movies) and "kinesiology" (the study of human motion). In one-dimensional kinematics we will study only the *motion* of a football, for example, without worrying about what forces cause or change its motion. Such considerations come in other chapters. In this chapter, we examine the simplest type of motion—namely, motion along a straight line, or one-dimensional motion.

This page titled 2.1: Introduction to One-Dimensional Kinematics is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- **1.1: Introduction to One-Dimensional Kinematics** by OpenStax is licensed CC BY 4.0.

2.1.1: Displacement

Learning Objectives

- Define position, displacement, distance, and distance traveled.
- Explain the relationship between position and displacement.
- Distinguish between displacement and distance traveled.
- Calculate displacement and distance given initial position, final position, and the path between the two.



Figure 2.1.1.1: These cyclists in Vietnam can be described by their position relative to buildings and a canal. Their motion can be described by their change in position, or displacement, in the frame of reference. (credit: Suzan Black, Fotopedia)

Position

In order to describe the motion of an object, you must first be able to describe its **position**—where it is at any particular time. More precisely, you need to specify its position relative to a convenient reference frame. Earth is often used as a reference frame, and we often describe the position of an object as it relates to stationary objects in that reference frame. For example, a rocket launch would be described in terms of the position of the rocket with respect to the Earth as a whole, while a professor's position could be described in terms of where she is in relation to the nearby white board. (See Figure 2.1.1.2.) In other cases, we use reference frames that are not stationary but are in motion relative to the Earth. To describe the position of a person in an airplane, for example, we use the airplane, not the Earth, as the reference frame. (See Figure 2.1.1.3.)

Displacement

If an object moves relative to a reference frame (for example, if a professor moves to the right relative to a white board or a passenger moves toward the rear of an airplane), then the object's position changes. This change in position is known as **displacement**. The word "displacement" implies that an object has moved, or has been displaced.

Definition: DISPLACEMENT

Displacement is the *change in position* of an object:

$$\Delta x = x_{\mathrm{f}} - x_0,$$

where Δx is displacement, x_{f} is the final position, and x_0 is the initial position.

In this text the upper case Greek letter Δ always means "change in" whatever quantity follows it; thus, Δx means *change in position*. Always solve for displacement by subtracting initial position x_0 from final position x_f.

Note that the SI unit for displacement is the meter (m), but sometimes kilometers, miles, feet, and other units of length are used. Keep in mind that when units other than the meter are used in a problem, you may need to convert them into meters to complete the calculation.

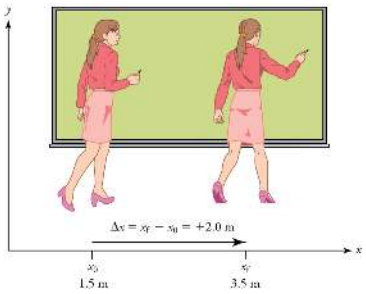


Figure 2.1.1.2: A professor paces left and right while lecturing. Her position relative to Earth is given by x. The **+2.0 m** displacement of the professor relative to Earth is represented by an arrow pointing to the right.

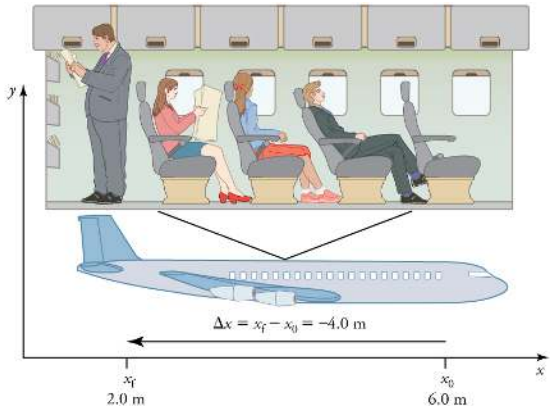


Figure 2.1.1.3: A passenger moves from his seat to the back of the plane. His location relative to the airplane is given by x. The **−4.0-m** displacement of the passenger relative to the plane is represented by an arrow toward the rear of the plane. Notice that the arrow representing his displacement is twice as long as the arrow representing the displacement of the professor (he moves twice as far) in Figure 2.1.1.2.

Note that displacement has a direction as well as a magnitude. The professor's displacement is 2.0 m to the right, and the airline passenger's displacement is 4.0 m toward the rear. In one-dimensional motion, direction can be specified with a plus or minus sign. When you begin a problem, you should select which direction is positive (usually that will be to the right or up, but you are free to select positive as being any direction). The professor's initial position is $x_0 = 1.5$ m and her final position is $x_f = 3.5$ m. Thus her displacement is

$$\Delta x = x_f - x_0 = 3.5 \text{ m} - 1.5 \text{ m} = +2.0 \text{ m}.$$

In this coordinate system, motion to the right is positive, whereas motion to the left is negative. Similarly, the airplane passenger's initial position is $x_0 = 6.0$ m and his final position is $x_f = 2.0$ m, so his displacement is

$$\Delta x = x_f - x_0 = 2.0 \text{ m} - 6.0 \text{ m} = -4.0 \text{ m}.$$

His displacement is negative because his motion is toward the rear of the plane, or in the negative x direction in our coordinate system.

Distance

Although displacement is described in terms of direction, distance is not. **Distance** is defined to be *the magnitude or size of displacement between two positions*. Note that the distance between two positions is not the same as the distance traveled between

them. **Distance traveled** is *the total length of the path traveled between two positions*. Distance has no direction and, thus, no sign. For example, the distance the professor walks is 2.0 m. The distance the airplane passenger walks is 4.0 m.

MISCONCEPTION ALERT: DISTANCE TRAVELED VS. MAGNITUDE OF DISPLACEMENT

It is important to note that the *distance traveled*, however, can be greater than the magnitude of the displacement (by magnitude, we mean just the size of the displacement without regard to its direction; that is, just a number with a unit). For example, the professor could pace back and forth many times, perhaps walking a distance of 150 m during a lecture, yet still end up only 2.0 m to the right of her starting point. In this case her displacement would be +2.0 m, the magnitude of her displacement would be 2.0 m, but the distance she traveled would be 150 m. In kinematics we nearly always deal with displacement and magnitude of displacement, and almost never with distance traveled. One way to think about this is to assume you marked the start of the motion and the end of the motion. The displacement is simply the difference in the position of the two marks and is independent of the path taken in traveling between the two marks. The distance traveled, however, is the total length of the path taken between the two marks.

Exercise 2.1.1.1

A cyclist rides 3 km west and then turns around and rides 2 km east. (a) What is her displacement? (b) What distance does she ride? (c) What is the magnitude of her displacement?

Answer

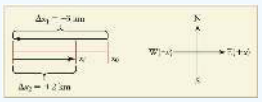


Figure 2.1.1.4

(a) The rider's displacement is $\Delta x = x_{\mathrm{f}} - x_0 = -1$ km. (The displacement is negative because we take east to be positive and west to be negative.)

(b) The distance traveled is **3 km + 2 km = 5 km**.

(c) The magnitude of the displacement is **1 km**.

Section Summary

- Kinematics is the study of motion without considering its causes. In this chapter, it is limited to motion along a straight line, called one-dimensional motion.
- Displacement is the change in position of an object.
- In symbols, displacement Δx is defined to be

$$\Delta x = x_{\mathrm{f}} - x_0,$$

  where x_0 is the initial position and x_f is the final position. In this text, the Greek letter Δ (delta) always means "change in" whatever quantity follows it. The SI unit for displacement is the meter (m). Displacement has a direction as well as a magnitude.
- When you start a problem, assign which direction will be positive.
- Distance is the magnitude of displacement between two positions.
- Distance traveled is the total length of the path traveled between two positions.

Glossary

kinematics

the study of motion without considering its causes

position

the location of an object at a particular time

displacement
the change in position of an object

distance
the magnitude of displacement between two positions

distance traveled
the total length of the path traveled between two positions

This page titled 2.1.1: Displacement is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- **1.2: Displacement** by OpenStax is licensed CC BY 4.0.

2.1.2: Time, Velocity, and Speed

Learning Objectives

- Explain the relationships between instantaneous velocity, average velocity, instantaneous speed, average speed, displacement, and time.
- Calculate velocity and speed given initial position, initial time, final position, and final time.
- Interpret graphs of position vs. time and velocity vs. time.



Figure 2.1.2.1: The motion of these racing snails can be described by their speeds and their velocities. (credit: tobitasflickr, Flickr)

There is more to motion than distance and displacement. Questions such as, “How long does a foot race take?” and “What was the runner’s speed?” cannot be answered without an understanding of other concepts. In this section we add definitions of time, velocity, and speed to expand our description of motion.

Time

The most fundamental physical quantities are defined by how they are measured. This is the case with time. Every measurement of time involves measuring a change in some physical quantity. It may be a number on a digital clock, a heartbeat, or the position of the Sun in the sky. In physics, the definition of time is simple—**time** is *change*, or the interval over which change occurs. It is impossible to know that time has passed unless something changes.

The amount of time or change is calibrated by comparison with a standard. The SI unit for time is the second, abbreviated s. We might, for example, observe that a certain pendulum makes one full swing every 0.75 s. We could then use the pendulum to measure time by counting its swings or, of course, by connecting the pendulum to a clock mechanism that registers time on a dial. This allows us to not only measure the amount of time, but also to determine a sequence of events.

How does time relate to motion? We are usually interested in elapsed time for a particular motion, such as how long it takes an airplane passenger to get from his seat to the back of the plane. To find elapsed time, we note the time at the beginning and end of the motion and subtract the two. For example, a lecture may start at 11:00 A.M. and end at 11:50 A.M., so that the elapsed time would be 50 min. **Elapsed time Δt** is the difference between the ending time and beginning time,

$$\Delta t = t_{\mathrm{f}} - t_0,$$

where Δt is the change in time or elapsed time, t_{f} is the time at the end of the motion, and t_0 is the time at the beginning of the motion. (As usual, the delta symbol, Δ, means the change in the quantity that follows it.)

Velocity

Your notion of velocity is probably the same as its scientific definition. You know that if you have a large displacement in a small amount of time you have a large velocity, and that velocity has units of distance divided by time, such as miles per hour or kilometers per hour.

Definition: AVERAGE VELOCITY

Average velocity is *displacement (change in position) divided by the time of travel*,

$$\bar{v} = \frac{\Delta x}{\Delta t} = \frac{x_{\mathrm{f}} - x_0}{t_{\mathrm{f}} - t_0},$$

where $\bar{v}$ is the *average* (indicated by the bar over the v) velocity,Δx is the change in position (or displacement), and x_{f} and x_0 are the final and beginning positions at times t_{f} and t_0, respectively.

Notice that this definition indicates that *velocity is a vector because displacement is a vector*. It has both magnitude and direction. The SI unit for velocity is meters per second or m/s, but many other units, such as km/h, mi/h (also written as mph), and cm/s, are in common use. Suppose, for example, an airplane passenger took 5 seconds to move −4 m (the negative sign indicates that displacement is toward the back of the plane). His average velocity would be

$$\bar{v} = \frac{\Delta x}{\Delta t} = \frac{-4 \text{ m}}{5 \text{ s}} = -0.8 \text{ m/s}.$$

The minus sign indicates the average velocity is also toward the rear of the plane.

The average velocity of an object does not tell us anything about what happens to it between the starting point and ending point, however. For example, we cannot tell from average velocity whether the airplane passenger stops momentarily or backs up before he goes to the back of the plane. To get more details, we must consider smaller segments of the trip over smaller time intervals.

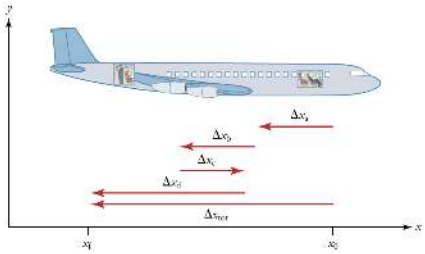


Figure 2.1.2.2: A more detailed record of an airplane passenger heading toward the back of the plane, showing smaller segments of his trip.

The smaller the time intervals considered in a motion, the more detailed the information. When we carry this process to its logical conclusion, we are left with an infinitesimally small interval. Over such an interval, the average velocity becomes the *instantaneous velocity* or the *velocity at a specific instant*. A car's speedometer, for example, shows the magnitude (but not the direction) of the instantaneous velocity of the car. (Police give tickets based on instantaneous velocity, but when calculating how long it will take to get from one place to another on a road trip, you need to use average velocity.) **Instantaneous velocity** v is the average velocity at a specific instant in time (or over an infinitesimally small time interval).

Mathematically, finding instantaneous velocity, v, at a precise instant t can involve taking a limit, a calculus operation beyond the scope of this text. However, under many circumstances, we can find precise values for instantaneous velocity without calculus.

Speed

In everyday language, most people use the terms "speed" and "velocity" interchangeably. In physics, however, they do not have the same meaning and they are distinct concepts. One major difference is that speed has no direction. Thus *speed is a scalar*. Just as we need to distinguish between instantaneous velocity and average velocity, we also need to distinguish between instantaneous speed and average speed.

Instantaneous speed is the magnitude of instantaneous velocity. For example, suppose the airplane passenger at one instant had an instantaneous velocity of −3.0 m/s (the minus meaning toward the rear of the plane). At that same time his instantaneous speed was 3.0 m/s. Or suppose that at one time during a shopping trip your instantaneous velocity is 40 km/h due north. Your instantaneous speed at that instant would be 40 km/h—the same magnitude but without a direction. Average speed, however, is very different from average velocity. **Average speed** is the distance traveled divided by elapsed time.

We have noted that distance traveled can be greater than displacement. So average speed can be greater than average velocity, which is displacement divided by time. For example, if you drive to a store and return home in half an hour, and your car's odometer shows the total distance traveled was 6 km, then your average speed was 12 km/h. Your average velocity, however, was zero, because your displacement for the round trip is zero. (Displacement is change in position and, thus, is zero for a round trip.) Thus average speed is *not* simply the magnitude of average velocity.

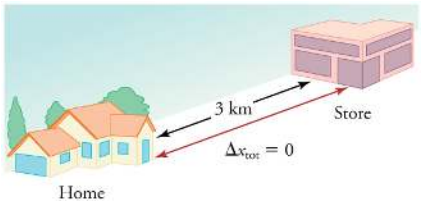


Figure 2.1.2.3: During a 30-minute round trip to the store, the total distance traveled is 6 km. The average speed is 12 km/h. The displacement for the round trip is zero, since there was no net change in position. Thus the average velocity is zero.

Another way of visualizing the motion of an object is to use a graph. A plot of position or of velocity as a function of time can be very useful. For example, for this trip to the store, the position, velocity, and speed-vs.-time graphs are displayed in Figure 2.1.2.4. (Note that these graphs depict a very simplified **model** of the trip. We are assuming that speed is constant during the trip, which is unrealistic given that we'll probably stop at the store. But for simplicity's sake, we will model it with no stops or changes in speed. We are also assuming that the route between the store and the house is a perfectly straight line.)

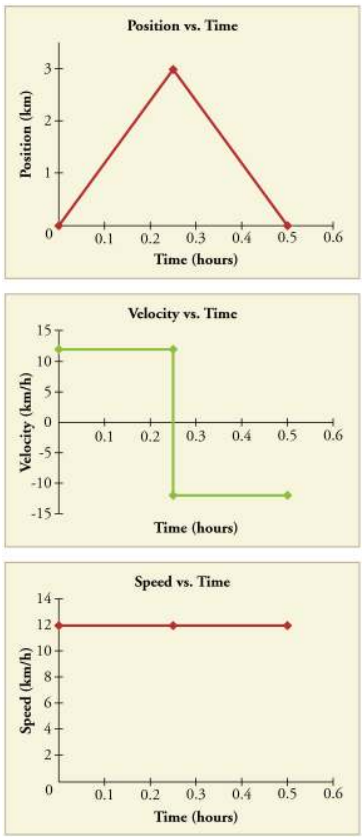


Figure 2.1.2.4: Position vs. time, velocity vs. time, and speed vs. time on a trip. Note that the velocity for the return trip is negative.

MAKING CONNECTIONS: TAKE-HOME INVESTIGATION—GETTING A SENSE OF SPEED

If you have spent much time driving, you probably have a good sense of speeds between about 10 and 70 miles per hour. But what are these in meters per second? What do we mean when we say that something is moving at 10 m/s? To get a better sense of what these values really mean, do some observations and calculations on your own:

- calculate typical car speeds in meters per second
- estimate jogging and walking speed by timing yourself; convert the measurements into both m/s and mi/h
- determine the speed of an ant, snail, or falling leaf

Exercise 2.1.2.1

A commuter train travels from Baltimore to Washington, DC, and back in 1 hour and 45 minutes. The distance between the two stations is approximately 40 miles. What is (a) the average velocity of the train, and (b) the average speed of the train in m/s?

Answer

(a) The average velocity of the train is zero because $x_f = x_0$; the train ends up at the same place it starts.

(b) The average speed of the train is calculated below. Note that the train travels 40 miles one way and 40 miles back, for a total distance of 80 miles.

$$\frac{\text{distance}}{\text{time}} = \frac{80 \text{ miles}}{105 \text{ minutes}} \tag{2.1.2.1}$$

$$\frac{80 \text{ miles}}{105 \text{ minutes}} \times \frac{5280 \text{ feet}}{1\text{mile}} \times \frac{1 \text{ meter}}{3.28 \text{ feet}} \times \frac{1 \text{ minute}}{60 \text{ seconds}} = 20 \text{ m/s} \tag{2.1.2.2}$$

Section Summary

- Time is measured in terms of change, and its SI unit is the second (s). Elapsed time for an event is

$$\Delta t = t_f - t_0,$$

  where t_f is the final time and t_0 is the initial time.
- Average velocity $\bar{v}$ is defined as displacement divided by the travel time. In symbols, average velocity is

$$\bar{v} = \frac{\Delta x}{\Delta t} = \frac{x_f - x_0}{t_f - t_0}.$$

- The SI unit for velocity is m/s.
- Velocity is a vector and thus has a direction.
- Instantaneous velocity v is the velocity at a specific instant or the average velocity for an infinitesimal interval.
- Instantaneous speed is the magnitude of the instantaneous velocity.
- Instantaneous speed is a scalar quantity, as it has no direction specified.
- Average speed is the total distance traveled divided by the elapsed time. (Average speed is *not* the magnitude of the average velocity.) Speed is a scalar quantity; it has no direction associated with it.

Glossary

average speed
distance traveled divided by time during which motion occurs

average velocity
displacement divided by time over which displacement occurs

instantaneous velocity
velocity at a specific instant, or the average velocity over an infinitesimal time interval

instantaneous speed
magnitude of the instantaneous velocity

time
change, or the interval over which change occurs

model
simplified description that contains only those elements necessary to describe the physics of a physical situation

elapsed time

the difference between the ending time and beginning time

This page titled 2.1.2: Time, Velocity, and Speed is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- **1.4: Time, Velocity, and Speed** by OpenStax is licensed CC BY 4.0.

2.1.3: Acceleration

Learning Objectives

- Define and distinguish between velocity and acceleration, and between instantaneous and average acceleration.
- Calculate acceleration given initial time, initial velocity, final time, and final velocity.



Figure 2.1.3.1: A plane slows down as it comes in for landing in St. Maarten. It is accelerating in a direction opposite to its velocity. (credit: Steve Conry, Flickr)

In everyday conversation, to accelerate means to speed up. The accelerator in a car can in fact cause it to speed up. The greater the **acceleration**, the greater the change in velocity over a given time. The formal definition of acceleration is consistent with these notions, but more inclusive.

Definition: AVERAGE ACCELERATION

Average Acceleration is *the rate at which velocity changes*,

$$\bar{a} = \frac{\Delta v}{\Delta t} = \frac{v_{\mathrm{f}} - v_0}{t_{\mathrm{f}} - t_0},$$

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

Because acceleration is velocity in m/s divided by time in s, the SI units for acceleration are $\mathrm{m/s^2}$, meters per second squared or meters per second per second, which literally means by how many meters per second the velocity changes every second.

Recall that velocity is a vector—it has both magnitude and direction. This means that a change in velocity can be a change in magnitude (or speed), but it can also be a change in *direction*. For example, if a car turns a corner at constant speed, it is accelerating because its direction is changing. The quicker you turn, the greater the acceleration. So there is an acceleration when velocity changes either in magnitude (an increase or decrease in speed) or in direction, or both.

ACCELERATION AS A VECTOR

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

Keep in mind that although acceleration is in the direction of the *change* in velocity, it is not always in the direction of *motion*. If acceleration is in a direction opposite to the direction of motion, the object slows down.



Figure 2.1.3.2: A subway train in Sao Paulo, Brazil, slows down as it comes into a station. It is accelerating in a direction opposite to its direction of motion. (credit: Yusuke Kawasaki, Flickr)

Example 2.1.3.1: Calculating Acceleration: A Racehorse Leaves the Gate

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



Figure 2.1.3.3

Strategy

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

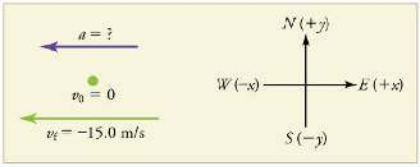


Figure 2.1.3.4

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

Solution

1. Identify the knowns. $v_0 = 0$, $v_f = -15.0$ m/s (the negative sign indicates direction toward the west), $\Delta t = 1.80$ s.

2. Find the change in velocity. Since the horse is going from zero to **−15.0 m/s**, its change in velocity equals its final velocity: $\Delta v = v_f = -15.0$ m/s.

3. Plug in the known values (Δv and Δt) and solve for the unknown $\bar{a}$.

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

Discussion

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

Instantaneous Acceleration

Instantaneous acceleration a, or the *acceleration at a specific instant in time*, is obtained by the same process as discussed for instantaneous velocity—that is, by considering an infinitesimally small interval of time. How do we find instantaneous acceleration using only algebra? The answer is that we choose an average acceleration that is representative of the motion. Figure 2.1.3.5 shows graphs of instantaneous acceleration versus time for two very different motions. In Figure 2.1.3.5 (a), the acceleration varies slightly and the average over the entire interval is nearly the same as the instantaneous acceleration at any time. In this case, we should treat this motion as if it had a constant acceleration equal to the average (in this case about 1.8 m/s^2). In Figure 2.1.3.5 (b), the acceleration varies drastically over time. In such situations it is best to consider smaller time intervals and choose an average acceleration for each. For example, we could consider motion over the time intervals from 0 to 1.0 s and from 1.0 to 3.0 s as separate motions with accelerations of $+3.0 \text{ m/s}^2$ and -2.0 m/s^2, respectively.

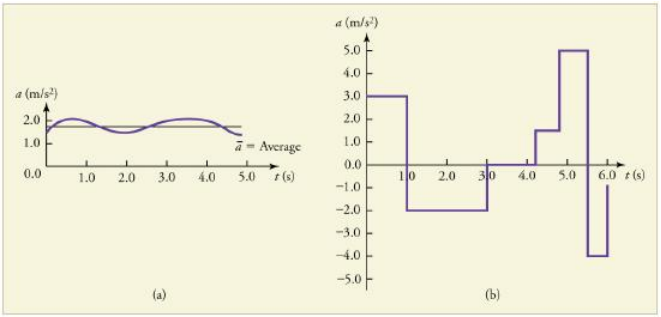


Figure 2.1.3.5: Graphs of instantaneous acceleration versus time for two different one-dimensional motions. (a) Here acceleration varies only slightly and is always in the same direction, since it is positive. The average over the interval is nearly the same as the acceleration at any given time. (b) Here the acceleration varies greatly, perhaps representing a package on a post office conveyor belt that is accelerated forward and backward as it bumps along. It is necessary to consider small time intervals (such as from 0 to 1.0 s) with constant or nearly constant acceleration in such a situation.

The next several examples consider the motion of the subway train shown in Figure 2.1.3.6. In (a) the shuttle moves to the right, and in (b) it moves to the left. The examples are designed to further illustrate aspects of motion and to illustrate some of the reasoning that goes into solving problems.

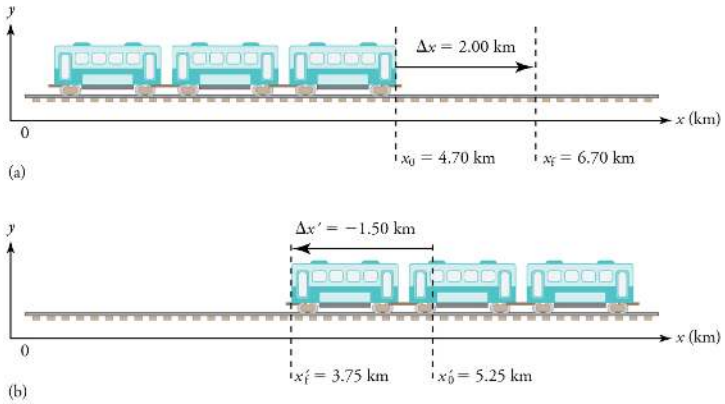


Figure 2.1.3.6: One-dimensional motion of a subway train considered in Example 2.1.3.2, Example 2.1.3.3, Example 2.1.3.4, Example 2.1.3.5, Example 2.1.3.6, and Example 2.1.3.7. Here we have chosen the x-axis so that + means to the right and − means to the left for displacements, velocities, and accelerations. (a) The subway train moves to the right from x_0 to x_f. Its displacement Δx is +2.0 km. (b) The train moves to the left from x'_0 to x'_f. Its displacement $\Delta x'$ is −1.5 km. (Note that the prime symbol (′) is used simply to distinguish between displacement in the two different situations. The distances of travel and the size of the cars are on different scales to fit everything into the diagram.)

Example 2.1.3.2: Calculating Displacement: A Subway Train

What are the magnitude and sign of displacements for the motions of the subway train shown in parts (a) and (b) of Figure 2.1.3.6?

Strategy

A drawing with a coordinate system is already provided, so we don't need to make a sketch, but we should analyze it to make sure we understand what it is showing. Pay particular attention to the coordinate system. To find displacement, we use the equation $\Delta x = x_f - x_0$. This is straightforward since the initial and final positions are given.

Solution

1. Identify the knowns. In the figure we see that $x_f = 6.70$ km and $x_0 = 4.70$ km for part (a), and $x'_f = 3.75$ km and $x'_0 = 5.25$ km for part (b).

2. Solve for displacement in part (a).

$$\Delta x = x_f - x_0 = 6.70 \text{ km} - 4.70 \text{ km} = +2.00 \text{ km}$$

3. Solve for displacement in part (b).

$$\Delta x' = x'_f - x'_0 = 3.75 \text{ km} - 5.25 \text{ km} = -1.50 \text{ km}$$

Discussion

The direction of the motion in (a) is to the right and therefore its displacement has a positive sign, whereas motion in (b) is to the left and thus has a negative sign.

Example 2.1.3.3: Comparing Distance Traveled with Displacement: A Subway Train

What are the distances traveled for the motions shown in parts (a) and (b) of the subway train in Figure 2.1.3.6?

Strategy

To answer this question, think about the definitions of distance and distance traveled, and how they are related to displacement. Distance between two positions is defined to be the magnitude of displacement, which was found in Example 2.1.3.2. Distance traveled is the total length of the path traveled between the two positions. In the case of the subway train shown in Figure 2.1.3.6, the distance traveled is the same as the distance between the initial and final positions of the train.

Solution

1. The displacement for part (a) was +2.00 km. Therefore, the distance between the initial and final positions was 2.00 km, and the distance traveled was 2.00 km.

2. The displacement for part (b) was **−1.5 km**. Therefore, the distance between the initial and final positions was 1.50 km, and the distance traveled was 1.50 km.

Discussion

Distance is a scalar. It has magnitude but no sign to indicate direction.

Example 2.1.3.4: Calculating Acceleration: A Subway Train Speeding Up

Suppose the train in Figure 2.1.3.6 (a) accelerates from rest to 30.0 km/h in the first 20.0 s of its motion. What is its average acceleration during that time interval?

Strategy

It is worth it at this point to make a simple sketch:

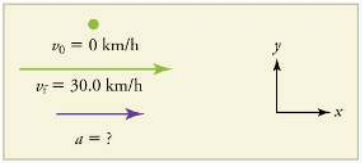


Figure 2.1.3.7

This problem involves three steps. First we must determine the change in velocity, then we must determine the change in time, and finally we use these values to calculate the acceleration.

Solution

1. Identify the knowns. $v_0 = 0$ (the trains starts at rest), $v_f = 30.0$ km/h, and $\Delta t = 20.0$ s.

2. Calculate Δv. Since the train starts from rest, its change in velocity is $\Delta v = +30.0$ km/h, where the plus sign means velocity to the right.

3. Plug in known values and solve for the unknown, $\bar{a}$.

$$\bar{a} = \frac{\Delta v}{\Delta t} = \frac{+30.0 \text{ km/h}}{20.0 \text{ s}}$$

4. Since the units are mixed (we have both hours and seconds for time), we need to convert everything into SI units of meters and seconds.

$$\bar{a} = \left(\frac{+30 \text{ km/h}}{20.0 \text{ s}}\right)\left(\frac{10^3 \text{ m}}{1 \text{ km}}\right)\left(\frac{1 \text{ h}}{3600 \text{ s}}\right) = 0.417 \text{ m/s}^2$$

Discussion

The plus sign means that acceleration is to the right. This is reasonable because the train starts from rest and ends up with a velocity to the right (also positive). So acceleration is in the same direction as the *change* in velocity, as is always the case.

Example 2.1.3.5: Calculate Acceleration: A Subway Train Slowing Down

Now suppose that at the end of its trip, the train in Figure 2.1.3.6 (a) slows to a stop from a speed of 30.0 km/h in 8.00 s. What is its average acceleration while stopping?

Strategy

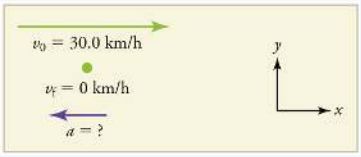


Figure 2.1.3.8

In this case, the train is slowing down and its acceleration is negative because it is toward the left. As in the previous example, we must find the change in velocity and the change in time and then solve for acceleration.

Solution

1. Identify the knowns. $v_0 = 30.0$ km/h, $v_f = 0$ km/h (the train is stopped, so its velocity is 0), and $\Delta t = 8.00$ s.

2. Solve for the change in velocity, Δv.

$$\Delta v = v_f - v_0 = 0 - 30.0 \text{ km/h} = -30.0 \text{ km/h}$$

3. Plug in the knowns, Δv and Δt, and solve for $\bar{a}$.

$$\bar{a} = \frac{\Delta v}{\Delta t} = \frac{-30.0 \text{ km/h}}{8.00 \text{ s}}$$

4. Convert the units to meters and seconds.

$$\bar{a} = \frac{\Delta v}{\Delta t} = \left(\frac{-30.0 \text{ km/h}}{8.00 \text{ s}}\right)\left(\frac{10^3 \text{ m}}{1 \text{ km}}\right)\left(\frac{1 \text{ h}}{3600 \text{ s}}\right) = -1.04 \text{ m/s}^2.$$

Discussion

The minus sign indicates that acceleration is to the left. This sign is reasonable because the train initially has a positive velocity in this problem, and a negative acceleration would oppose the motion.

The graphs of position, velocity, and acceleration vs. time for the trains in Example 2.1.3.4 and Example 2.1.3.5 are displayed in Figure 2.1.3.9. (We have taken the velocity to remain constant from 20 to 40 s, after which the train slows down.)

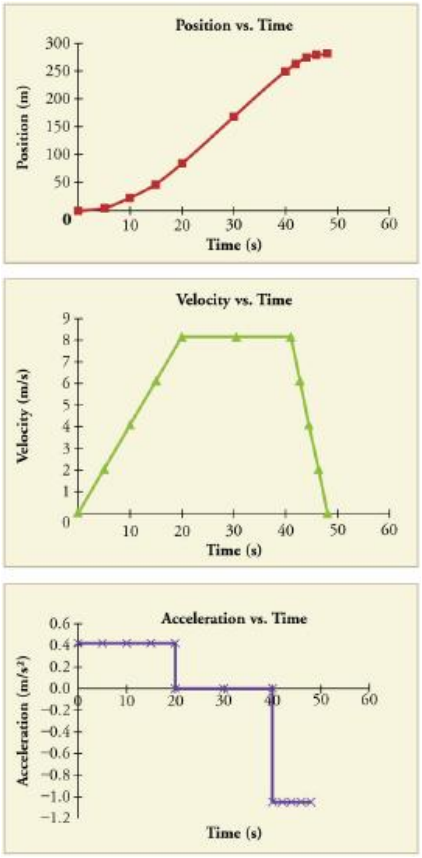


Figure 2.1.3.9: (a) Position of the train over time. Notice that the train's position changes slowly at the beginning of the journey, then more and more quickly as it picks up speed. Its position then changes more slowly as it slows down at the end of the journey. In the middle of the journey, while the velocity remains constant, the position changes at a constant rate. (b) Velocity of the train over time. The train's velocity increases as it accelerates at the beginning of the journey. It remains the same in the middle of the journey (where there is no acceleration). It decreases as the train slows down at the end of the journey. (c) The acceleration of the train over time. The train has positive acceleration as it speeds up at the beginning of the journey. It has no acceleration as it travels at constant velocity in the middle of the journey. Its acceleration is negative as it slows down at the end of the journey.

Example 2.1.3.6: Calculating Average Velocity: The Subway Train

What is the average velocity of the train in part b of Example 2.1.3.2, and shown again below, if it takes 5.00 min to make its trip?

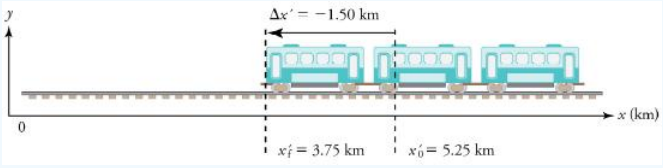


Figure 2.1.3.10

Strategy

Average velocity is displacement divided by time. It will be negative here, since the train moves to the left and has a negative displacement.

Solution

1. Identify the knowns. $x'_f = 3.75\ \text{km},\ x'_0 = 5.25\ \text{km},\ \Delta t = 5.00\ \text{min}.$

2. Determine displacement, $\Delta x'$. We found $\Delta x'$ to be **−1.5 km** in Example 2.1.3.2.

3. Solve for average velocity.

$$\bar{v} = \frac{\Delta x'}{} = \frac{-1.50\ \text{km}}{}$$

$$\Delta t \quad 5.00 \text{ min}$$

4. Convert units.

$$\bar{v} = \frac{\Delta x'}{\Delta t} = \left(\frac{-1.50 \text{ km}}{5.00 \text{ min}}\right)\left(\frac{60 \text{ min}}{1 \text{ h}}\right) = -18.0 \text{ km/h}$$

Discussion

The negative velocity indicates motion to the left.

Example 2.1.3.7: Calculating Deceleration: The Subway Train

Finally, suppose the train in Figure 2.1.3.10 slows to a stop from a velocity of 20.0 km/h in 10.0 s. What is its average acceleration?

Strategy

Once again, let's draw a sketch:

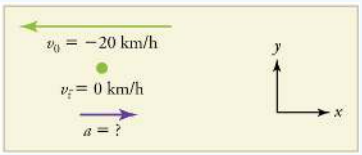


Figure 2.1.3.11

As before, we must find the change in velocity and the change in time to calculate average acceleration.

Solution

1. Identify the knowns. $v_0 = -20 \text{ km/h},\ v_f = 0 \text{ km/h},\ \Delta t = 10.0 \text{ s}$.

2. Calculate Δv. The change in velocity here is actually positive, since

$$\Delta v = v_f - v_0 = 0 - (-20 \text{ km/h}) = +20 \text{ km/h}$$

3. Solve for $\bar{a}$.

$$\bar{a} = \frac{\Delta v}{\Delta t} = \frac{+20.0 \text{ km/h}}{10.0 \text{ s}}$$

4. Convert units.

$$\bar{a} = \left(\frac{+20.0 \text{ km/h}}{10.0 \text{ s}}\right)\left(\frac{10^3 \text{ m}}{1 \text{ km}}\right)\left(\frac{1 \text{ h}}{3600 \text{ s}}\right) = +0.556 \text{ m/s}^2$$

Discussion

The plus sign means that acceleration is to the right. This is reasonable because the train initially has a negative velocity (to the left) in this problem and a positive acceleration opposes the motion (and so it is to the right).

Sign and Direction

Perhaps the most important thing to note about these examples is the signs of the answers. In our chosen coordinate system, plus means the quantity is to the right and minus means it is to the left. This is easy to imagine for displacement and velocity. But it is a little less obvious for acceleration. Most people interpret negative acceleration as the slowing of an object. This was not the case in Example 2.1.3.7, where a positive acceleration slowed a negative velocity. The crucial distinction was that the acceleration was in the opposite direction from the velocity. In fact, a negative acceleration will *increase* a negative velocity. For example, the train moving to the left in Figure 2.1.3.10 is sped up by an acceleration to the left. In that case, both v and a are negative. The plus and minus signs give the directions of the accelerations. If acceleration has the same sign as the change in velocity, the object is speeding up. If acceleration has the opposite sign of the change in velocity, the object is slowing down.

Exercise 2.1.3.1

An airplane lands on a runway traveling east. Describe its acceleration.

Answer

If we take east to be positive, then the airplane has negative acceleration, as it is accelerating toward the west. It is also slowing down: its acceleration is opposite in direction to its velocity.

Section Summary

- Acceleration is the rate at which velocity changes. In symbols, **average acceleration** $\bar{a}$ is

$$\bar{a} = \frac{\Delta v}{\Delta t} = \frac{v_{\mathrm{f}} - v_0}{t_{\mathrm{f}} - t_0}$$

- The SI unit for acceleration is $\mathrm{m/s^2}$.
- Acceleration is a vector, and thus has a both a magnitude and direction.
- Acceleration can be caused by either a change in the magnitude or the direction of the velocity.
- Instantaneous acceleration a is the acceleration at a specific instant in time.
- When an acceleration is in a direction opposite to that of the velocity of an object, the object slows down.

Glossary

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

average acceleration
the change in velocity divided by the time over which it changes

instantaneous acceleration
acceleration at a specific point in time

This page titled 2.1.3: Acceleration is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- **1.5: Acceleration** by OpenStax is licensed CC BY 4.0.

2.2: Motion Equations for Constant Acceleration in One Dimension

Learning Objectives

- Calculate displacement of an object that is not accelerating, given initial position and velocity.
- Calculate final velocity of an accelerating object, given initial velocity, acceleration, and time.
- Calculate displacement and final position of an accelerating object, given initial position, initial velocity, time, and acceleration.



Figure 2.2.1: Kinematic equations can help us describe and predict the motion of moving objects such as these kayaks racing in Newbury, England. (credit: Barry Skeates, Flickr)

We might know that the greater the acceleration of, say, a car moving away from a stop sign, the greater the displacement in a given time. But we have not developed a specific equation that relates acceleration and displacement. In this section, we develop some convenient equations for kinematic relationships, starting from the definitions of displacement, velocity, and acceleration already covered.

Notation: *t*, *x*, *v*, *a*

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

$$\left.\begin{aligned} \Delta t &= t \\ \Delta x &= x - x_0 \\ \Delta v &= v - v_0 \end{aligned}\right\}$$

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

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

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

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

SOLVING FOR DISPLACEMENT (Δx) AND FINAL POSITION (x) FROM AVERAGE VELOCITY WHEN ACCELERATION (a) IS CONSTANT

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

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

Substituting the simplified notation for Δx and Δt yields

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

Solving for x yields

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

where the average velocity is

$$\bar{v} = \frac{v_0 + v}{2} (\text{constant } a).$$

The equation $\bar{v} = \frac{v_0+v}{2}$ reflects the fact that, when acceleration is constant, v is just the simple average of the initial and final velocities. For example, if you steadily increase your velocity (that is, with constant acceleration) from 30 to 60 km/h, then your average velocity during this steady increase is 45 km/h. Using the equation $\bar{v} = \frac{v_0+v}{2}$ to check this, we see that

$$\bar{v} = \frac{v_0 + v}{2} = \frac{30 \text{ km/h} + 60 \text{ km/h}}{2} = 45 \text{ km/h},$$

which seems reasonable.

Example 2.2.1: Calculating Displacement: How Far does the Jogger Run?

A jogger runs down a straight stretch of road with an average velocity of 4.00 m/s for 2.00 min. What is his final position, taking his initial position to be zero?

Strategy

Draw a sketch.

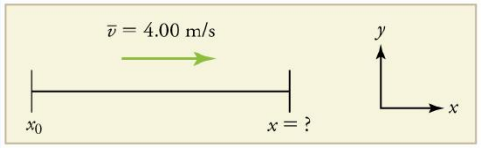


Figure 2.2.2

The final position x is given by the equation

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

To find x, we identify the values of x_0, $\bar{v}$, and t from the statement of the problem and substitute them into the equation.

Solution

1. Identify the knowns. $\bar{v} = 4.00$ m/s, $\Delta t = 2.00$ min, and $x_0 = 0$ m.

2. Enter the known values into the equation.

$$x = x_0 + \bar{v}t = 0 + (4.00 \text{ m/s})(120 \text{ s}) = 480 \text{ m}$$

Discussion

Velocity and final displacement are both positive, which means they are in the same direction.

The equation $x = x_0 + \bar{v}t$ gives insight into the relationship between displacement, average velocity, and time. It shows, for example, that displacement is a linear function of average velocity. (By linear function, we mean that displacement depends on $\bar{v}$ rather than on $\bar{v}$ raised to some other power, such as $\bar{v}^2$. When graphed, linear functions look like straight lines with a constant slope.) On a car trip, for example, we will get twice as far in a given time if we average 90 km/h than if we average 45 km/h.

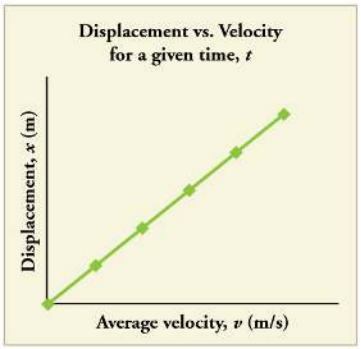


Figure 2.2.3: There is a linear relationship between displacement and average velocity. For a given time t, an object moving twice as fast as another object will move twice as far as the other object.

SOLVING FOR FINAL VELOCITY

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

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

Substituting the simplified notation for Δv and Δt gives us

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

Solving for v yields

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

Example 2.2.2: Calculating Final Velocity: An Airplane Slowing Down after Landing

An airplane lands with an initial velocity of 70.0 m/s and then slows down at $1.50\ \text{m/s}^2$ for 40.0 s. What is its final velocity?

Strategy

Draw a sketch. We draw the acceleration vector in the direction opposite the velocity vector because the plane is slowing down.

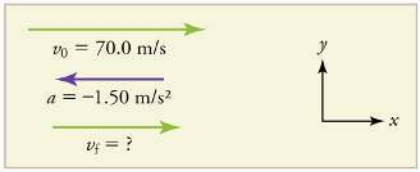


Figure 2.2.4

Solution

1. Identify the knowns. $v_0 = 70.0\ \text{m/s},\ a = -1.50\ \text{m/s}^2,\ t = 40.0\ \text{s}$.

2. Identify the unknown. In this case, it is final velocity, v_f.

3. Determine which equation to use. We can calculate the final velocity using the equation $v = v_0 + at$.

4. Plug in the known values and solve.

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

Discussion

The final velocity is much less than the initial velocity, as desired when slowing down, but still positive. With jet engines, reverse thrust could be maintained long enough to stop the plane and start moving it backward. That would be indicated by a

negative final velocity, which is not the case here.



Figure 2.2.5: The airplane lands with an initial velocity of 70.0 m/s and slows to a final velocity of 10.0 m/s before heading for the terminal. Note that the acceleration is negative because its direction is opposite to its velocity, which is positive.

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

- final velocity depends on how large the acceleration is and how long it lasts
- if the acceleration is zero, then the final velocity equals the initial velocity $(v = v_0)$, as expected (i.e., velocity is constant)
- if a is negative, then the final velocity is less than the initial velocity

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

MAKING CONNECTIONS: REAL-WORLD CONNECTION



Figure 2.2.6: The Space Shuttle *Endeavor* blasts off from the Kennedy Space Center in February 2010. (credit: Matthew Simantov, Flickr)

An intercontinental ballistic missile (ICBM) has a larger average acceleration than the Space Shuttle and achieves a greater velocity in the first minute or two of flight (actual ICBM burn times are classified—short-burn-time missiles are more difficult for an enemy to destroy). But the Space Shuttle obtains a greater final velocity, so that it can orbit the earth rather than come directly back down as an ICBM does. The Space Shuttle does this by accelerating for a longer time.

SOLVING FOR FINAL POSITION WHEN VELOCITY IS NOT CONSTANT (a≠0)

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

$$v = v_0 + at.$$

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

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

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

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

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

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

Example 2.2.3: Calculating Displacement of an Accelerating Object: Dragsters

Dragsters can achieve average accelerations of $26.0\ \text{m/s}^2$. Suppose such a dragster accelerates from rest at this rate for 5.56 s. How far does it travel in this time?



Figure 2.2.7: U.S. Army Top Fuel pilot Tony "The Sarge" Schumacher begins a race with a controlled burnout. (credit: Lt. Col. William Thurmond. Photo Courtesy of U.S. Army.)

Strategy

Draw a sketch.

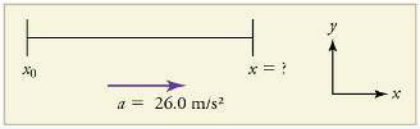


Figure 2.2.8

We are asked to find displacement, which is x if we take x_0 to be zero. (Think about it like the starting line of a race. It can be anywhere, but we call it 0 and measure all other positions relative to it.) We can use the equation $x = x_0 + v_0 t + \frac{1}{2}at^2$ once we identify v_0, a, and t from the statement of the problem.

Solution

1. Identify the knowns. Starting from rest means that $v_0 = 0$, a is given as $26.0\ \text{m/s}^2$ and t is given as 5.56 s.

2. Plug the known values into the equation to solve for the unknown x:

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

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

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

Substituting the identified values of a and t gives

$$x = \frac{1}{2}\left(26.0\ \text{m/s}^2\right)(5.56\ \text{s})^2,$$

yielding

$$x = 402\ \text{m}.$$

Discussion

If we convert 402 m to miles, we find that the distance covered is very close to one quarter of a mile, the standard distance for drag racing. So the answer is reasonable. This is an impressive displacement in only 5.56 s, but top-notch dragsters can do a quarter mile in even less time than this.

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

- displacement depends on the square of the elapsed time when acceleration is not zero. In Example 2.2.3, the dragster covers only one fourth of the total distance in the first half of the elapsed time
- if acceleration is zero, then the initial velocity equals average velocity $(v_0 = \bar{v})$ and $x = x_0 + v_0 t + \frac{1}{2}at^2$ becomes $x = x_0 + v_0 t$

SOLVING FOR FINAL VELOCITY WHEN VELOCITY IS NOT CONSTANT (a≠0)

A fourth useful equation can be obtained from another algebraic manipulation of previous equations.

If we solve $v = v_0 + at$ for t, we get

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

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

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

Example 2.2.4: Calculating Final Velocity: Dragsters

Calculate the final velocity of the dragster in Example 2.2.3 without using information about time.

Strategy

Draw a sketch.

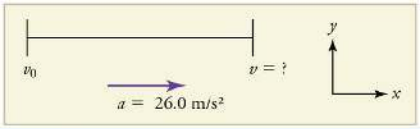


Figure 2.2.9

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

Solution

1. Identify the known values. We know that $v_0 = 0$, since the dragster starts from rest. Then we note that $x - x_0 = 402$ m (this was the answer in Example 2.2.3). Finally, the average acceleration was given to be $a = 26.0\ \text{m/s}^2$.

2. Plug the knowns into the equation $v^2 = v_0^2 + 2a\,(x - x_0)$ and solve for v.

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

Thus

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

To get v, we take the square root:

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

Discussion

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

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

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

Putting Equations Together

In the following examples, we further explore one-dimensional motion, but in situations requiring slightly more algebraic manipulation. The examples also give insight into problem-solving techniques. The box below provides easy reference to the

equations needed.

SUMMARY OF KINEMATIC EQUATIONS (CONSTANT a)

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

Example 2.2.5: Calculating Displacement: How Far Does a Car Go When Coming to a Halt?

On dry concrete, suppose a car can slow down at a rate of $7.00\ \text{m/s}^2$, whereas on wet concrete the maximum magnitude of acceleration is only $5.00\ \text{m/s}^2$. Find the distances necessary to stop a car moving at 30.0 m/s (about 110 km/h) (a) on dry concrete and (b) on wet concrete. (c) Repeat both calculations, finding the displacement from the point where the driver sees a traffic light turn red, taking into account his reaction time of 0.500 s to get his foot on the brake.

Strategy

Draw a sketch.

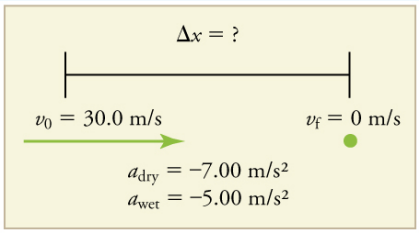


Figure 2.2.10

In order to determine which equations are best to use, we need to list all of the known values and identify exactly what we need to solve for. We shall do this explicitly in the next several examples, using tables to set them off.

Solution for (a)

1. Identify the knowns and what we want to solve for. We know that $v_0 = 30.0\ \text{m/s};\ v = 0;\ a = -7.00\ \text{m/s}^2$ (a is negative because it is in a direction opposite to velocity). We take x_0 to be 0. We are looking for displacement Δx, or $x - x_0$.

2. Identify the equation that will help up solve the problem. The best equation to use is

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

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

3. Rearrange the equation to solve for x.

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

4. Enter known values.

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

Thus,

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

Solution for (b)

This part can be solved in exactly the same manner as Part A. The only difference is that the acceleration is $-5.00\ \text{m/s}^2$. The result is

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

Solution for (c)

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

1. Identify the knowns and what we want to solve for. We know that $\bar{v} = 30.0\ \text{m/s}$; $t_{\text{reaction}} = 0.500\ \text{s}$; $a_{\text{reaction}} = 0$. We take $x_{\text{0-reaction}}$ to be 0. We are looking for $x_{reaction}$.

2. Identify the best equation to use.

$x = x_0 + \bar{v}t$ works well because the only unknown value is x, which is what we want to solve for.

3. Plug in the knowns to solve the equation.

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

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

4. Add the displacement during the reaction time to the displacement when braking.

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

a. 64.3 m + 15.0 m = 79.3 m when dry
b. 90.0 m + 15.0 m = 105 m when wet

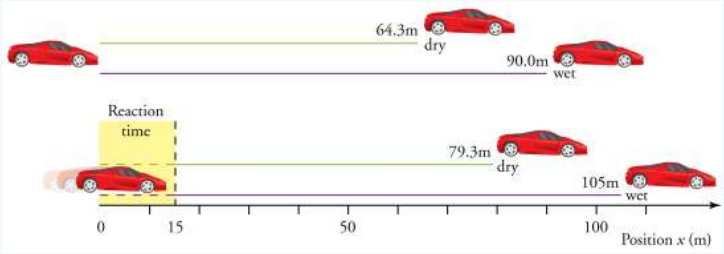


Figure 2.2.11: The distance necessary to stop a car varies greatly, depending on road conditions and driver reaction time. Shown here are the braking distances for dry and wet pavement, as calculated in this example, for a car initially traveling at 30.0 m/s. Also shown are the total distances traveled from the point where the driver first sees a light turn red, assuming a 0.500 s reaction time.

Discussion

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

With the basics of kinematics established, we can go on to many other interesting examples and applications. In the process of developing kinematics, we have also glimpsed a general approach to problem solving that produces both correct answers and insights into physical relationships.

MAKING CONNECTIONS: TAKE-HOME EXPERIMENT—BREAKING NEWS

We have been using SI units of meters per second squared to describe some examples of acceleration of cars, runners, and trains. To achieve a better feel for these numbers, one can measure the braking acceleration of a car doing a slow (and safe) stop. Recall that, for average acceleration, $\bar{a} = \Delta v/\Delta t$. While traveling in a car, slowly apply the brakes as you come up to a stop sign. Have a passenger note the initial speed in miles per hour and the time taken (in seconds) to stop. From this, calculate

the acceleration in miles per hour per second. Convert this to meters per second squared and compare with other accelerations mentioned in this chapter. Calculate the distance traveled in braking.

Exercise 2.2.1

A manned rocket accelerates at a rate of $20\ \text{m/s}^2$ during launch. How long does it take the rocket reach a velocity of 400 m/s?

Answer

To answer this, choose an equation that allows you to solve for time t, given only a, v_0, and v.

$$v = v_0 + at$$

Rearrange to solve for t.

$$t = \frac{v - v_0}{a} = \frac{400\ \text{m/s} - 0\ \text{m/s}}{20\ \text{m/s}^2} = 20\ \text{s}$$

Section Summary

- To simplify calculations we take acceleration to be constant, so that $\bar{a} = a$ at all times.
- We also take initial time to be zero.
- Initial position and velocity are given a subscript 0; final values have no subscript. Thus,

$$\left.\begin{aligned} \Delta t &= t \\ \Delta x &= x - x_0 \\ \Delta v &= v - v_0 \end{aligned}\right\}$$

- The following kinematic equations for motion with constant a are useful:

$$\begin{aligned} x &= x_0 + \bar{v}t \\ \bar{v} &= \frac{v_0 + v}{2} \\ v &= v_0 + at \\ x &= x_0 + v_0 t + \frac{1}{2}at^2 \\ v^2 &= v_0^2 + 2a\left(x - x_0\right) \end{aligned}$$

- In vertical motion, y is substituted for x.

This page titled 2.2: Motion Equations for Constant Acceleration in One Dimension is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- **1.6: Motion Equations for Constant Acceleration in One Dimension** by OpenStax is licensed CC BY 4.0.

2.3: Falling Objects

Learning Objectives

- Describe the effects of gravity on objects in motion.
- Describe the motion of objects that are in free fall.
- Calculate the position and velocity of objects in free fall.

Falling objects form an interesting class of motion problems. For example, we can estimate the depth of a vertical mine shaft by dropping a rock into it and listening for the rock to hit the bottom. By applying the kinematics developed so far to falling objects, we can examine some interesting situations and learn much about gravity in the process.

Gravity

The most remarkable and unexpected fact about falling objects is that, if air resistance and friction are negligible, then in a given location all objects fall toward the center of Earth with the *same constant acceleration, independent of their mass*. This experimentally determined fact is unexpected, because we are so accustomed to the effects of air resistance and friction that we expect light objects to fall slower than heavy ones.

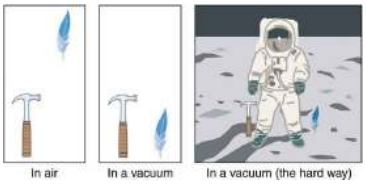


Figure 2.3.1: A hammer and a feather will fall with the same constant acceleration if air resistance is considered negligible. This is a general characteristic of gravity not unique to Earth, as astronaut David R. Scott demonstrated on the Moon in 1971, where the acceleration due to gravity is only $1.67\ \mathrm{m/s^2}$.

In the real world, air resistance can cause a lighter object to fall slower than a heavier object of the same size. A tennis ball will reach the ground after a hard baseball dropped at the same time. (It might be difficult to observe the difference if the height is not large.) Air resistance opposes the motion of an object through the air, while friction between objects—such as between clothes and a laundry chute or between a stone and a pool into which it is dropped—also opposes motion between them. For the ideal situations of these first few chapters, an object *falling without air resistance or friction* is defined to be in **free-fall**.

The force of gravity causes objects to fall toward the center of Earth. The acceleration of free-falling objects is therefore called the **acceleration due to gravity**. The acceleration due to gravity is *constant*, which means we can apply the kinematics equations to any falling object where air resistance and friction are negligible. This opens a broad class of interesting situations to us. The acceleration due to gravity is so important that its magnitude is given its own symbol, g. It is constant at any given location on Earth and has the average value

$$g = 9.80\ \mathrm{m/s^2}.$$

Although g varies from $9.78\ \mathrm{m/s^2}$ to $9.83\ \mathrm{m/s^2}$, depending on latitude, altitude, underlying geological formations, and local topography, the average value of $9.80\ \mathrm{m/s^2}$ will be used in this text unless otherwise specified. The direction of the acceleration due to gravity is *downward (towards the center of Earth)*. In fact, its direction *defines* what we call vertical. Note that whether the acceleration a in the kinematic equations has the value $+g$ or $-g$ depends on how we define our coordinate system. If we define the upward direction as positive, then $a = -g = -9.80\ \mathrm{m/s^2}$, and if we define the downward direction as positive, then $a = g = 9.80\ \mathrm{m/s^2}$. In certain examples and homework exercises, slightly inaccurate but more convenient value of $g \simeq 10\ \mathrm{m/s^2}$ will be used with appropriate specifications.

One-Dimensional Motion Involving Gravity

The best way to see the basic features of motion involving gravity is to start with the simplest situations and then progress toward more complex ones. So we start by considering straight up and down motion with no air resistance or friction. These assumptions mean that the velocity (if there is any) is vertical. If the object is dropped, we know the initial velocity is zero. Once the object has left contact with whatever held or threw it, the object is in free-fall. Under these circumstances, the motion is one-dimensional and has constant acceleration of magnitude g. We will also represent vertical displacement with the symbol y and use x for horizontal displacement.

KINEMATIC EQUATIONS FOR OBJECTS IN FREE-FALL

$$v = v_0 - gt$$

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

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

Example 2.3.1: Calculating Position and Velocity of a Falling Object: A Rock Thrown Upward

A person standing on the edge of a high cliff throws a rock straight up with an initial velocity of 13.0 m/s. The rock misses the edge of the cliff as it falls back to earth. Calculate the position and velocity of the rock 1.00 s, 2.00 s, and 3.00 s after it is thrown, neglecting the effects of air resistance.

Strategy

Draw a sketch.

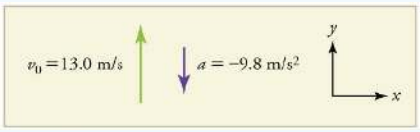


Figure 2.3.2

We are asked to determine the position y at various times. It is reasonable to take the initial position y_0 to be zero. This problem involves one-dimensional motion in the vertical direction. We use plus and minus signs to indicate direction, with up being positive and down negative. Since up is positive, and the rock is thrown upward, the initial velocity must be positive too. The acceleration due to gravity is downward, so a is negative. It is crucial that the initial velocity and the acceleration due to gravity have opposite signs. Opposite signs indicate that the acceleration due to gravity opposes the initial motion and will slow and eventually reverse it.

Since we are asked for values of position and velocity at three times, we will refer to these as y_1 and v_1; y_2 and v_2; and y_3 and v_3.

Solution for Position y_1

1. Identify the knowns. We know that $y_0 = 0;\ v_0 = 13.0\ \text{m/s};\ a = -g = -9.80\ \text{m/s}^2$; and $t = 1.00$ s.

2. Identify the best equation to use. We will use $y = y_0 + v_0 t + \frac{1}{2} a t^2$ because it includes only one unknown, y (or y_1, here), which is the value we want to find.

3. Plug in the known values and solve for y_1.

$$y_1 = 0 + (13.0\ \text{m/s})(1.00\ \text{s}) + \frac{1}{2}\left(-9.80\ \text{m/s}^2\right)(1.00\ \text{s})^2 = 8.10\ \text{m}$$

Discussion

The rock is 8.10 m above its starting point at $t = 1.00$ s, since $y_1 > y_0$. It could be *moving* up or down; the only way to tell is to calculate v_1 and find out if it is positive or negative.

Solution for Velocity v_1

1. Identify the knowns. We know that $y_0 = 0;\ v_0 = 13.0\ \text{m/s};\ a = -g = -9.80\ \text{m/s}^2$; and $t = 1.00s$. We also know from the solution above that $y_1 = 8.10$ m.

2. Identify the best equation to use. The most straightforward is $v = v_0 - gt$ (from $v = v_0 + at$, where a = gravitational acceleration $= -g$).

3. Plug in the knowns and solve.

$$v_1 = v_0 - gt = 13.0\ \text{m/s} - \left(9.80\ \text{m/s}^2\right)(1.00\ \text{s}) = 3.20\ \text{m/s}$$

Discussion

The positive value for v_1 means that the rock is still heading upward at t=**1.00s**. However, it has slowed from its original 13.0 m/s, as expected.

Solution for Remaining Times

The procedures for calculating the position and velocity at t = **2.00 s** and **3.00 s** are the same as those above. The results are summarized in Table 2.3.1 and illustrated in Figure 2.3.3.

Table 2.3.1: Results

Time, *t*	Position, *y*	Velocity, *v*	Acceleration, *a*
1.00 s	8.10 m	3.20 m	$-9.80\ \mathrm{m/s^2}$
2.00 s	6.40 m	−6.60 m	$-9.80\ \mathrm{m/s^2}$
3.00 s	−5.10 m	−16.4 m	$-9.80\ \mathrm{m/s^2}$

Graphing the data helps us understand it more clearly.

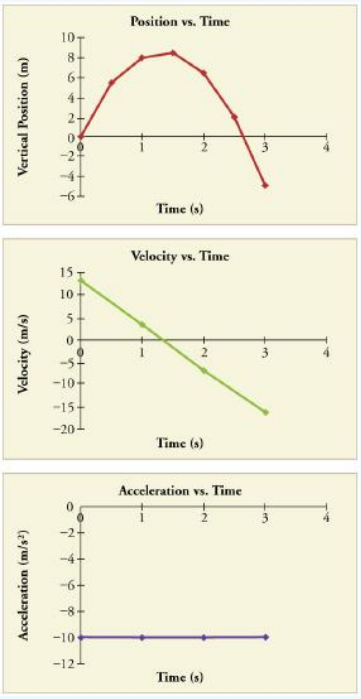


Figure 2.3.3: Vertical position, vertical velocity, and vertical acceleration vs. time for a rock thrown vertically up at the edge of a cliff. Notice that velocity changes linearly with time and that acceleration is constant. *Misconception Alert!* Notice that the position vs. time graph shows vertical position only. It is easy to get the impression that the graph shows some horizontal motion—the shape of the graph looks like the path of a projectile. But this is not the case; the horizontal axis is *time*, not space. The actual path of the rock in space is straight up, and straight down.

Discussion

The interpretation of these results is important. At 1.00 s the rock is above its starting point and heading upward, since y_1 and v_1 are both positive. At 2.00 s, the rock is still above its starting point, but the negative velocity means it is moving downward. At 3.00 s, both y_3 and v_3 are negative, meaning the rock is below its starting point and continuing to move downward. Notice that when the rock is at its highest point (at 1.5 s), its velocity is zero, but its acceleration is still $-9.80\ \mathrm{m/s^2}$. Its acceleration is $-9.80\ \mathrm{m/s^2}$ for the whole trip—while it is moving up and while it is moving down. Note that the values for y are the positions (or displacements) of the rock, not the total distances traveled. Finally, note that free-fall applies to upward motion as well as downward. Both have the same acceleration—the acceleration due to gravity, which remains constant the entire time. Astronauts training in the famous Vomit Comet, for example, experience free-fall while arcing up as well as down, as we will discuss in more detail later.

MAKING CONNECTIONS: TAKE-HOME EXPERIMENT—REACTION TIME

A simple experiment can be done to determine your reaction time. Have a friend hold a ruler between your thumb and index finger, separated by about 1 cm. Note the mark on the ruler that is right between your fingers. Have your friend drop the ruler unexpectedly, and try to catch it between your two fingers. Note the new reading on the ruler. Assuming acceleration is that due to gravity, calculate your reaction time. How far would you travel in a car (moving at 30 m/s) if the time it took your foot to go from the gas pedal to the brake was twice this reaction time?

Example 2.3.2: Calculating Velocity of a Falling Object: A Rock Thrown Down

What happens if the person on the cliff throws the rock straight down, instead of straight up? To explore this question, calculate the velocity of the rock when it is 5.10 m below the starting point, and has been thrown downward with an initial speed of 13.0 m/s.

Strategy

Draw a sketch.

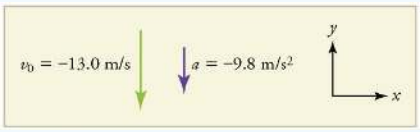


Figure 2.3.4

Since up is positive, the final position of the rock will be negative because it finishes below the starting point at $y_0 = 0$. Similarly, the initial velocity is downward and therefore negative, as is the acceleration due to gravity. We expect the final velocity to be negative since the rock will continue to move downward.

Solution

1. Identify the knowns. $y_0 = 0;\ y_1 = -5.10\ \text{m};\ v_0 = -13.0\ \text{m/s};\ a = -g = -9.80\ \text{m/s}^2$.

2. Choose the kinematic equation that makes it easiest to solve the problem. The equation $v^2 = v_0^2 + 2a\,(y - y_0)$ works well because the only unknown in it is v. (We will plug y_1 in for y.)

3. Enter the known values

$$v^2 = (-13.0\ \text{m/s})^2 + 2\,(-9.80\ \text{m/s}^2)\,(-5.10\ \text{m} - 0\ \text{m}) = 268.96\ \text{m}^2/\text{s}^2,$$

where we have retained extra significant figures because this is an intermediate result.

Taking the square root, and noting that a square root can be positive or negative, gives

$$v = \pm 16.4\ \text{m/s}.$$

The negative root is chosen to indicate that the rock is still heading down. Thus,

$$v = -16.4\ \text{m/s}.$$

Discussion

Note that *this is exactly the same velocity the rock had at this position when it was thrown straight upward with the same initial speed*. (See Example 2.3.1 and Figure 2.3.5 (a).) This is not a coincidental result. Because we only consider the acceleration due to gravity in this problem, the *speed* of a falling object depends only on its initial speed and its vertical position relative to the starting point. For example, if the velocity of the rock is calculated at a height of 8.10 m above the starting point (using the method from Example 2.3.1) when the initial velocity is 13.0 m/s straight up, a result of $\pm 3.20\ \text{m/s}$ is obtained. Here both signs are meaningful; the positive value occurs when the rock is at 8.10 m and heading up, and the negative value occurs when the rock is at 8.10 m and heading back down. It has the same *speed* but the opposite direction.

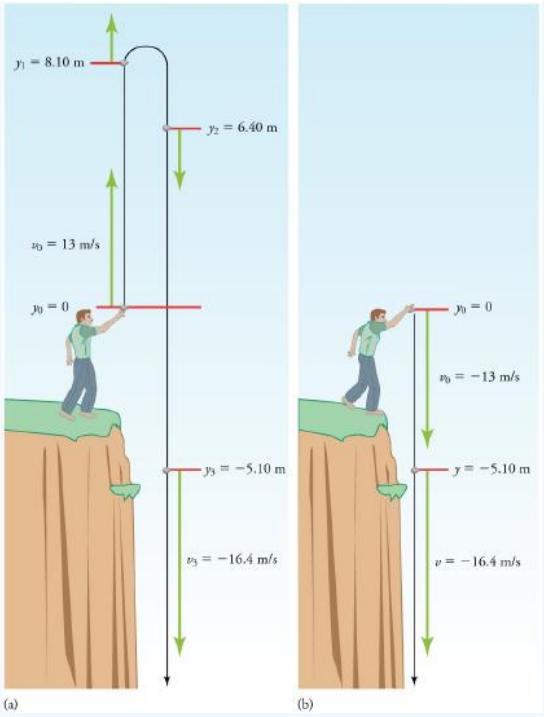


Figure 2.3.5: (a) A person throws a rock straight up, as explored in Example 2.3.1. The arrows are velocity vectors at 0, 1.00, 2.00, and 3.00 s. (b) A person throws a rock straight down from a cliff with the same initial speed as before, as in Example 2.3.2. Note that at the same distance below the point of release, the rock has the same velocity in both cases.

Another way to look at it is this: In Example 2.3.1, the rock is thrown up with an initial velocity of **13.0 m/s**. It rises and then falls back down. When its position is $y = 0$ on its way back down, its velocity is **−13.0 m/s**. That is, it has the same speed on its way down as on its way up. We would then expect its velocity at a position of $y = -5.10\text{m}$ to be the same whether we have thrown it upwards at **+13.0 m/s** or thrown it downwards at **−13.0 m/s**. The velocity of the rock on its way down from $y = 0$ is the same whether we have thrown it up or down to start with, as long as the speed with which it was initially thrown is the same.

Example 2.3.3: Find g from Data on a Falling Object

The acceleration due to gravity on Earth differs slightly from place to place, depending on topography (e.g., whether you are on a hill or in a valley) and subsurface geology (whether there is dense rock like iron ore as opposed to light rock like salt beneath you.) The precise acceleration due to gravity can be calculated from data taken in an introductory physics laboratory course. An object, usually a metal ball for which air resistance is negligible, is dropped and the time it takes to fall a known distance is measured. See, for example, Figure 2.3.6. Very precise results can be produced with this method if sufficient care is taken in measuring the distance fallen and the elapsed time.

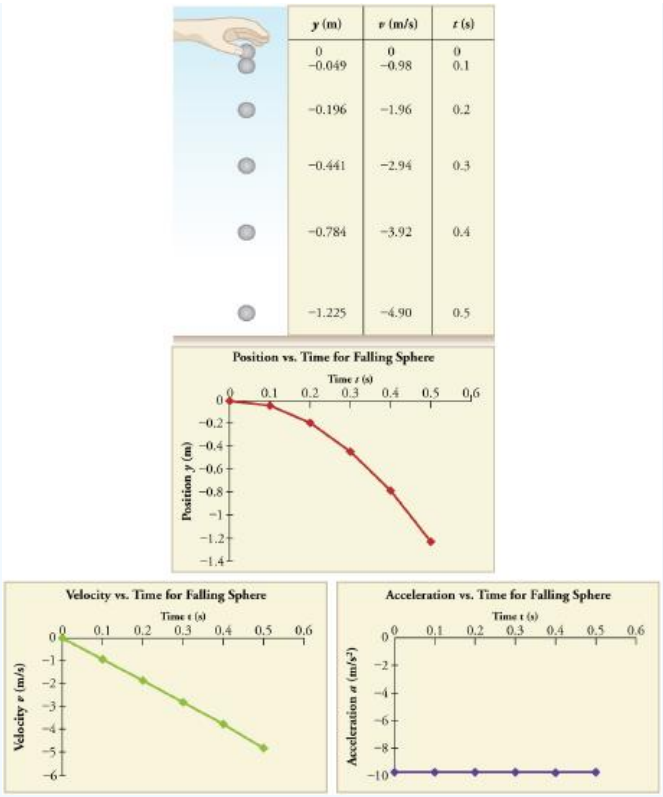


Figure 2.3.6: Positions and velocities of a metal ball released from rest when air resistance is negligible. Velocity is seen to increase linearly with time while displacement increases with time squared. Acceleration is a constant and is equal to gravitational acceleration.

Suppose the ball falls 1.0000 m in 0.45173 s. Assuming the ball is not affected by air resistance, what is the precise acceleration due to gravity at this location?

Strategy

Draw a sketch.

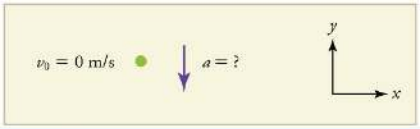


Figure 2.3.7

We need to solve for acceleration a. Note that in this case, displacement is downward and therefore negative, as is acceleration.

Solution

1. Identify the knowns. $y_0 = 0;\ y = -1.0000$ m; $t = 0.45173;\ v_0 = 0$..

2. Choose the equation that allows you to solve for a using the known values.

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

3. Substitute 0 for v_0 and rearrange the equation to solve for a. Substituting 0 for v_0 yields

$$y = y_0 + \frac{1}{2} a t^2.$$

Solving for a gives

$$a = \frac{2\,(y - y_0)}{t^2}.$$

4. Substitute known values yields

$$a = \frac{2(-1.0000\ \text{m} - 0)}{(0.45173\ \text{s})^2} = -9.8010\ \text{m/s}^2,$$

so, because $a = -g$ with the directions we have chosen,

$$g = 9.8010\ \text{m/s}^2.$$

Discussion

The negative value for a indicates that the gravitational acceleration is downward, as expected. We expect the value to be somewhere around the average value of $9.80\ \text{m/s}^2$, so $9.8010\ \text{m/s}^2$ makes sense. Since the data going into the calculation are relatively precise, this value for g is more precise than the average value of $9.80\ \text{m/s}^2$; it represents the local value for the acceleration due to gravity.

Exercise 2.3.1

A chunk of ice breaks off a glacier and falls 30.0 meters before it hits the water. Assuming it falls freely (there is no air resistance), how long does it take to hit the water?

Answer

We know that initial position $y_0 = 0$, final position $y = -30.0\ \text{m}$, and $a = -g = -9.80\ \text{m/s}^2$. We can then use the equation $y = y_0 + v_0 t + \frac{1}{2}at^2$ to solve for t. Inserting $a = -g$, we obtain

$$\begin{aligned} y &= 0 + 0 - \frac{1}{2}gt^2 \\ t^2 &= \frac{2y}{-g} \\ t &= \pm\sqrt{\frac{2y}{-g}} = \pm\sqrt{\frac{2(-30.0\ \text{m})}{-9.80\ \text{m/s}^2}} = \pm\sqrt{6.12\ \text{s}^2} = 2.47\ \text{s} \approx 2.5\ \text{s} \end{aligned}$$

where we take the positive value as the physically relevant answer. Thus, it takes about 2.5 seconds for the piece of ice to hit the water.

Section Summary

- An object in free-fall experiences constant acceleration if air resistance is negligible.
- On Earth, all free-falling objects have an acceleration due to gravity g, which averages

$$g = 9.80\ \text{m/s}^2.$$

- Whether the acceleration a should be taken as $+g$ or $-g$ is determined by your choice of coordinate system. If you choose the upward direction as positive,$a = -g = -9.80\ \text{m/s}^2$ is negative. In the opposite case, $a = +\text{g} = 9.80\ \text{m/s}^2$ is positive. Since acceleration is constant, the kinematic equations above can be applied with the appropriate $+g$ or $-g$ substituted for a.
- For objects in free-fall, up is normally taken as positive for displacement, velocity, and acceleration.

Glossary

free-fall

the state of movement that results from gravitational force only

acceleration due to gravity

acceleration an object undergoes while in free-fall

This page titled 2.3: Falling Objects is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- **1.7: Falling Objects** by OpenStax is licensed CC BY 4.0.

2.3.1: Projectile Motion

Learning Objectives

- Identify and explain the properties of a projectile, such as acceleration due to gravity, range, and trajectory.
- Apply the principle of independence of motion to solve projectile motion problems.

Projectile motion is the **motion** of an object thrown or projected into the air, subject to only the acceleration of gravity. The object is called a **projectile**, and its path is called its **trajectory**. The motion of falling objects is a simple one-dimensional type of projectile motion in which there is no horizontal movement. In this section, we consider two-dimensional projectile motion, such as that of a football or other object for which **air resistance** *is negligible*.

The most important fact to know here is that *motions along perpendicular axes are independent* and thus can be considered separately. In particular, it is often convenient to consider vertical and horizontal motions separately, or independently. The key to analyzing two-dimensional projectile motion is to break it into two motions, one along the horizontal axis and the other along the vertical. And we will assume all forces except gravity are negligible (that is, we ignore forces due to air resistance, wind, friction, etc.). The vertical component of acceleration is, then, $a_y = -g = -9.80\ \mathrm{m/s^2}$, where we take upward direction as the positive y direction. And because there is no force in the horizontal direction, $a_x = 0$. Figure 2.3.1.1 illustrates this approach to analyzing projectile motion.

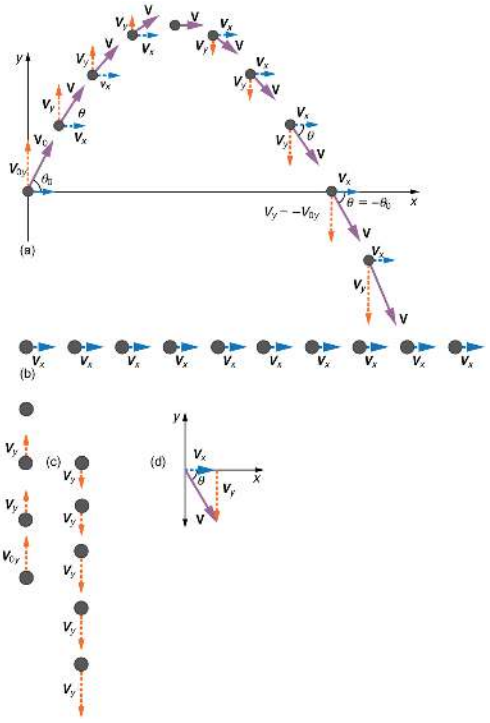


Figure 2.3.1.1: (a) We analyze two-dimensional projectile motion by breaking it into two independent one-dimensional motions along the vertical and horizontal axes. (b) The horizontal motion is simple, because $a_x = 0$ and v_x is thus constant. (c) The velocity in the vertical direction begins to decrease as the object rises; at its highest point, the vertical velocity is zero. As the object falls towards the Earth again, the vertical velocity increases again in magnitude. This is exactly same as in the case of falling objects (or objects thrown directly upward). (d) The *x* - and *y* -motions can be recombined to give the total velocity at any given point on the trajectory.

One of the conceptual aspects of projectile motion we can discuss without a detailed analysis is the range. On level ground, we define **range** to be the horizontal distance R traveled by a projectile. The range of projectiles describes everyday phenomena such as how far a football can be thrown or kicked. Also, investigating the range of projectiles can shed light on other interesting phenomena, such as the orbits of satellites around the Earth. Let us consider projectile range further.

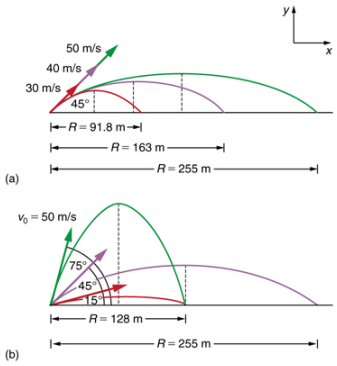


Figure 2.3.1.2: Trajectories of projectiles on level ground. (a) The greater the initial speed v_0, the greater the range for a given initial angle. (b) The effect of initial angle θ_0 on the range of a projectile with a given initial speed. Note that the range is the same for 15° and 75°, although the maximum heights of those paths are different.

How does the initial velocity of a projectile affect its range? It is intuitive to guess that, the greater the initial speed v_0, the greater the range, as shown in Figure 2.3.1.2 (a). The initial angle θ_0 also has a dramatic effect on the range, as illustrated in Figure 2.3.1.2 (b). For a fixed initial speed, such as might be produced by a cannon, the maximum range is obtained with $\theta_0 = 45^\circ$. Interestingly, for every initial angle except 45°, there are two angles that give the same range. The range also depends on the value of the acceleration of gravity g. The lunar astronaut Alan Shepherd was able to drive a golf ball a great distance on the Moon because gravity is weaker there. The range R of a projectile on *level ground* for which air resistance is negligible is given by

$$R = \frac{v_0^2 \sin 2\theta_0}{g},$$

where v_0 is the initial speed and θ_0 is the initial angle relative to the horizontal. This formula (which can be derived using algebra and trigonometry) fits the major features of projectile range as described.

When we speak of the range of a projectile on level ground, we assume that R is very small compared with the circumference of the Earth. If, however, the range is large, the Earth curves away below the projectile and acceleration of gravity changes direction along the path. The range is larger than predicted by the range equation given above because the projectile has farther to fall than it would on level ground. (See Figure 2.3.1.3.) If the initial speed is great enough, the projectile goes into orbit. This possibility was recognized centuries before it could be accomplished. When an object is in orbit, the Earth curves away from underneath the object at the same rate as it falls. The object thus falls continuously but never hits the surface.



Figure 2.3.1.3: Projectile to satellite. In each case shown here, a projectile is launched from a very high tower to avoid air resistance. With increasing initial speed, the range increases and becomes longer than it would be on level ground because the Earth curves away underneath its path. With a large enough initial speed, orbit is achieved.

Summary

- Projectile motion is the motion of an object through the air that is subject only to the acceleration of gravity.
- The most important fact regarding projectile motion is that motions along vertical direction and the horizontal direction are independent.
- The maximum horizontal distance traveled by a projectile is called the **range**. The range R of a projectile on level ground launched at an angle θ_0 above the horizontal with initial speed v_0 is given by

$$R = \frac{v_0^2 \sin 2\theta_0}{g}.$$

Glossary

centripetal acceleration
the acceleration of an object moving in a circle, directed toward the center

ultracentrifuge
a centrifuge optimized for spinning a rotor at very high speeds

uniform circular motion
an object moving in a circular path at constant speed

This page titled 2.3.1: Projectile Motion is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- **1.8: Projectile Motion** by OpenStax is licensed CC BY 4.0.

2.3.E: Kinematics (Exercise)

Conceptual Questions

1.1: Displacement

1. Give an example in which there are clear distinctions among distance traveled, displacement, and magnitude of displacement. Specifically identify each quantity in your example.

2. Under what circumstances does distance traveled equal magnitude of displacement? What is the only case in which magnitude of displacement and displacement are exactly the same?

3. Bacteria move back and forth by using their flagella (structures that look like little tails). Speeds of up to $50\ \mu\text{m/s}\ \left(50 \times 10^{-6}\ \text{m/s}\right)$ have been observed. The total distance traveled by a bacterium is large for its size, while its displacement is small. Why is this?

1.2: Vectors, Scalars, and Coordinate Systems

4. A student writes, "*A bird that is diving for prey has a speed of* −10 m/s." What is wrong with the student's statement? What has the student actually described? Explain.

5. What is the speed of the bird in Exercise 2.2.4?

6. Acceleration is the change in velocity over time. Given this information, is acceleration a vector or a scalar quantity? Explain.

7. A weather forecast states that the temperature is predicted to be −5ºC the following day. Is this temperature a vector or a scalar quantity? Explain.

1.3: Time, Velocity, and Speed

8. Give an example (but not one from the text) of a device used to measure time and identify what change in that device indicates a change in time.

9. There is a distinction between average speed and the magnitude of average velocity. Give an example that illustrates the difference between these two quantities.

10. Does a car's odometer measure position or displacement? Does its speedometer measure speed or velocity?

11. If you divide the total distance traveled on a car trip (as determined by the odometer) by the time for the trip, are you calculating the average speed or the magnitude of the average velocity? Under what circumstances are these two quantities the same?

12. How are instantaneous velocity and instantaneous speed related to one another? How do they differ?

1.4: Acceleration

13. Is it possible for speed to be constant while acceleration is not zero? Give an example of such a situation.

14. Is it possible for velocity to be constant while acceleration is not zero? Explain.

15. Give an example in which velocity is zero yet acceleration is not.

16. If a subway train is moving to the left (has a negative velocity) and then comes to a stop, what is the direction of its acceleration? Is the acceleration positive or negative?

17. Plus and minus signs are used in one-dimensional motion to indicate direction. What is the sign of an acceleration that reduces the magnitude of a negative velocity? Of a positive velocity?

1.5: Motion Equations for Constant Acceleration in One Dimension

18. In a constant-acceleration motion, how does the position change? How does the velocity change?

19. Describe the relationship between the position and the velocity in a constant-acceleration motion.

20. Describe the relationship between the velocity and the acceleration in a constant-acceleration motion.

21. In a constant-acceleration motion starting from rest, how does the distance traveled change with the duration of motion? If the motion continues for double the time (total), how much does the distance traveled increase (total)?

1.6: Falling Objects

22. What is the acceleration of a rock thrown straight upward on the way up? At the top of its flight? On the way down?

23. An object that is thrown straight up falls back to Earth. This is one-dimensional motion. (a) When is its velocity zero? (b) Does its velocity change direction? (c) Does the acceleration due to gravity have the same sign on the way up as on the way down?

24. Suppose you throw a rock nearly straight up at a coconut in a palm tree, and the rock misses on the way up but hits the coconut on the way down. Neglecting air resistance, how does the speed of the rock when it hits the coconut on the way down compare with what it would have been if it had hit the coconut on the way up? Is it more likely to dislodge the coconut on the way up or down? Explain.

25. If an object is thrown straight up and air resistance is negligible, then its speed when it returns to the starting point is the same as when it was released. If air resistance were not negligible, how would its speed upon return compare with its initial speed? How would the maximum height to which it rises be affected?

26. The severity of a fall depends on your speed when you strike the ground. All factors but the acceleration due to gravity being the same, how many times higher could a safe fall on the Moon be than on Earth (gravitational acceleration on the Moon is about 1/6 that of the Earth)?

27. How many times higher could an astronaut jump on the Moon than on Earth if his takeoff speed is the same in both locations (gravitational acceleration on the Moon is about 1/6 of g on Earth)?

1.7: Projectile Motion

28. Answer the following questions for projectile motion on level ground assuming negligible air resistance (the initial angle being neither 0° nor 90°): (a) Is the velocity ever zero? (b) When is the velocity a minimum? A maximum? (c) Can the velocity ever be the same as the initial velocity at a time other than at $t = 0$? (d) Can the speed ever be the same as the initial speed at a time other than at $t = 0$?

29. Answer the following questions for projectile motion on level ground assuming negligible air resistance (the initial angle being neither 0° nor 90°): (a) Is the acceleration ever zero? (b) Is the acceleration ever in the same direction as a component of velocity? (c) Is the acceleration ever opposite in direction to a component of velocity?

30. For a fixed initial speed, the range of a projectile is determined by the angle at which it is fired. For all but the maximum, there are two angles that give the same range. Considering factors that might affect the ability of an archer to hit a target, such as wind, explain why the smaller angle (closer to the horizontal) is preferable. When would it be necessary for the archer to use the larger angle? Why does the punter in a football game use the higher trajectory?

31. During a lecture demonstration, a professor places two coins on the edge of a table. She then flicks one of the coins horizontally off the table, simultaneously nudging the other over the edge. Describe the subsequent motion of the two coins, in particular discussing whether they hit the floor at the same time.

1.8: Centripetal Acceleration

32. Give examples of centripetal acceleration from everyday experience.

33. A satellite in a circular orbit around the Earth is undergoing a centripetal acceleration. Explain how this description is consistent with the description that the satellite in free-fall.

34. Can centripetal acceleration change the speed of circular motion? Explain.

Problems & Exercises

1.1: Displacement

35.

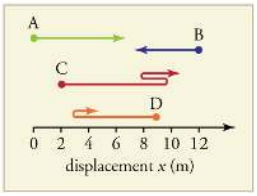


Figure 2.3.E.1

Find the following for path A in Figure 2.3.E.1: (a) The distance traveled. (b) The magnitude of the displacement from start to finish. (c) The displacement from start to finish.

Answer

(a) 7 m

(b) 7 m

(c) +7 m

36. Find the following for path B in Figure 2.3.E.1: (a) The distance traveled. (b) The magnitude of the displacement from start to finish. (c) The displacement from start to finish.

37. Find the following for path C in Figure 2.3.E.1: (a) The distance traveled. (b) The magnitude of the displacement from start to finish. (c) The displacement from start to finish.

Answer

(a) 13 m

(b) 9 m

(c) +9 m

38. Find the following for path D in Figure 2.3.E.1: (a) The distance traveled. (b) The magnitude of the displacement from start to finish. (c) The displacement from start to finish.

This page titled 2.3.E: Kinematics (Exercise) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- **1.E: Kinematics (Exercise)** by OpenStax is licensed CC BY 4.0.

2.4: Introduction to Dynamics- Newton's Laws of Motion

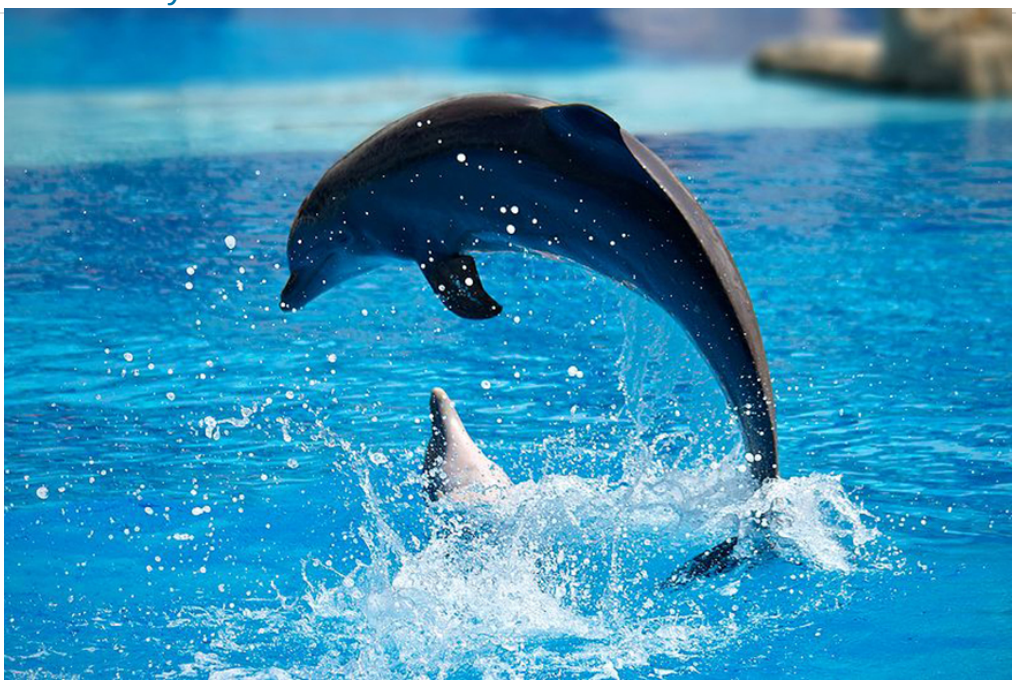
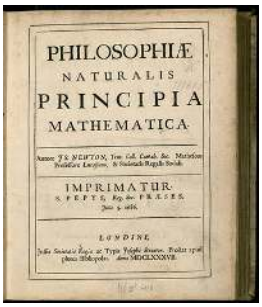


Figure 2.4.1: Newton's laws of motion describe the motion of the dolphin's path. (credit: Jin Jang)

Motion draws our attention. Motion itself can be beautiful, causing us to marvel at the forces needed to achieve spectacular motion, such as that of a dolphin jumping out of the water, or a pole vaulter, or the flight of a bird, or the orbit of a satellite. The study of motion is kinematics, but kinematics only *describes* the way objects move—their velocity and their acceleration. **Dynamics** considers the forces that affect the motion of moving objects and systems. Newton's laws of motion are the foundation of dynamics. These laws provide an example of the breadth and simplicity of principles under which nature functions. They are also universal laws in that they apply to similar situations on Earth as well as in space.

Issac Newton's (1642–1727) laws of motion were just one part of the monumental work that has made him legendary. The development of Newton's laws marks the transition from the Renaissance into the modern era. This transition was characterized by a revolutionary change in the way people thought about the physical universe. For many centuries natural philosophers had debated the nature of the universe based largely on certain rules of logic with great weight given to the thoughts of earlier classical philosophers such as Aristotle (384–322 BC). Among the many great thinkers who contributed to this change were Newton and Galileo.


PHILOSOPHIÆ
NATURALIS
PRINCIPIA
MATHEMATICA

IMPRIMATUR.

LONDINI

Figure 2.4.2: Issac Newton's monumental work, *Philosophiae Naturalis Principia Mathematica*, was published in 1687. It proposed scientific laws that are still used today to describe the motion of objects. (credit: Service commun de la documentation de l'Université de Strasbourg)

Galileo was instrumental in establishing *observation* as the absolute determinant of truth, rather than "logical" argument. Galileo's use of the telescope was his most notable achievement in demonstrating the importance of observation. He discovered moons orbiting Jupiter and made other observations that were inconsistent with certain ancient ideas and religious dogma. For this reason, and because of the manner in which he dealt with those in authority, Galileo was tried by the Inquisition and punished. He spent the final years of his life under a form of house arrest. Because others before Galileo had also made discoveries by *observing* the nature

of the universe, and because repeated observations verified those of Galileo, his work could not be suppressed or denied. After his death, his work was verified by others, and his ideas were eventually accepted by the church and scientific communities.

Galileo also contributed to the formation of what is now called Newton's first law of motion. Newton made use of the work of his predecessors, which enabled him to develop laws of motion, discover the law of gravity, invent calculus, and make great contributions to the theories of light and color. It is amazing that many of these developments were made with Newton working alone, without the benefit of the usual interactions that take place among scientists today.

It was not until the advent of modern physics early in the 20th century that it was discovered that Newton's laws of motion produce a good approximation to motion only when the objects are moving at speeds much, much less than the speed of light and when those objects are larger than the size of most molecules (about 10^{-9} m in diameter). These constraints define the realm of classical mechanics. At the beginning of the 20^{th} century, Albert Einstein (1879–1955) developed the theory of relativity and, along with many other scientists, developed quantum theory. This theory does not have the constraints present in classical physics. All of the situations we consider in this chapter are in the realm of classical physics.

MAKING CONNECTIONS: PAST AND PRESENT PHILOSOPHY

The importance of observation and the concept of *cause and effect* were not always so entrenched in human thinking. This realization was a part of the evolution of modern physics from natural philosophy. The achievements of Galileo, Newton, Einstein, and others were key milestones in the history of scientific thought. Most of the scientific theories that are described in this book descended from the work of these scientists.

This page titled 2.4: Introduction to Dynamics- Newton's Laws of Motion is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- **2.1: Introduction to Dynamics- Newton's Laws of Motion** by OpenStax is licensed CC BY 4.0.

2.4.1: Development of Force Concept

Learning Objectives

- Understand the definition of force.

Dynamics is the study of the forces that cause objects and systems to move. To understand this, we need a working definition of force. Our intuitive definition of **force**—that is, a push or a pull—is a good place to start. We know that a push or pull has both magnitude and direction (therefore, it is a vector quantity) and can vary considerably in each regard. For example, a cannon exerts a strong force on a cannonball that is launched into the air. In contrast, Earth exerts only a tiny downward pull on a flea.

A more quantitative definition of force can be based on some standard force, just as distance is measured in units relative to a standard distance. One possibility is to stretch a spring a certain fixed distance, as illustrated in Figure 2.4.1.1, and use the force it exerts to pull itself back to its relaxed shape—called a *restoring force*—as a standard. The magnitude of all other forces can be stated as multiples of this standard unit of force. Many other possibilities exist for standard forces. Some alternative definitions of force will be given later in this chapter.

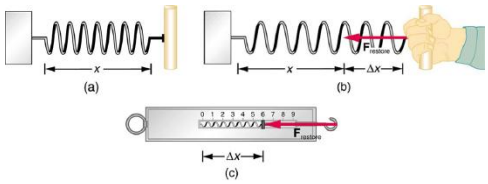


Figure 2.4.1.1: The force exerted by a stretched spring can be used as a standard unit of force. (a) This spring has a length x when undistorted. (b) When stretched a distance Δx, the spring exerts a restoring force, $\mathbf{F}_{\text{restore}}$, which is reproducible. (c) A spring scale is one device that uses a spring to measure force. The force $\mathbf{F}_{\text{restore}}$ is exerted on whatever is attached to the hook. Here $\mathbf{F}_{\text{restore}}$ has a magnitude of 6 units in the force standard being employed.

TAKE-HOME EXPERIMENT: FORCE STANDARDS

To investigate force standards and cause and effect, get two identical rubber bands. Hang one rubber band vertically on a hook. Find a small household item that could be attached to the rubber band using a paper clip, and use this item as a weight to investigate the stretch of the rubber band. Measure the amount of stretch produced in the rubber band with one, two, and four of these (identical) items suspended from the rubber band. What is the relationship between the number of items and the amount of stretch? How large a stretch would you expect for the same number of items suspended from two rubber bands? What happens to the amount of stretch of the rubber band (with the weights attached) if the weights are also pushed to the side with a pencil?

Section Summary

- **Dynamics** is the study of how forces affect the motion of objects.
- **Force** is a push or pull that can be defined in terms of various standards, and it is a vector having both magnitude and direction.

Glossary

dynamics

the study of how forces affect the motion of objects and systems

force

a push or pull on an object with a specific magnitude and direction; can be represented by vectors; can be expressed as a multiple of a standard force

This page titled 2.4.1: Development of Force Concept is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- **2.2: Development of Force Concept** by OpenStax is licensed CC BY 4.0.

2.4.2: Newton’s First Law of Motion- Inertia

Learning Objectives

- Define mass and inertia.
- Understand Newton's first law of motion.

Experience suggests that an object at rest will remain at rest if left alone, and that an object in motion tends to slow down and stop unless some effort is made to keep it moving. What **Newton’s first law of motion** states, however, is the following:

NEWTON’S FIRST LAW OF MOTION

A body at rest remains at rest, or, if in motion, remains in motion at a constant velocity unless acted on by a net external force.

Note the repeated use of the verb “remains.” We can think of this law as preserving the status quo of motion.

Rather than contradicting our experience, **Newton’s first law of motion** states that there must be a *cause* (which is a net external force) *for there to be any change in velocity (either a change in magnitude or direction)*. We will define *net external force* in the next section. An object sliding across a table or floor slows down due to the net force of friction acting on the object. If friction disappeared, would the object still slow down?

The idea of cause and effect is crucial in accurately describing what happens in various situations. For example, consider what happens to an object sliding along a rough horizontal surface. The object quickly grinds to a halt. If we spray the surface with talcum powder to make the surface smoother, the object slides farther. If we make the surface even smoother by rubbing lubricating oil on it, the object slides farther yet. Extrapolating to a frictionless surface, we can imagine the object sliding in a straight line indefinitely. Friction is thus the *cause* of the slowing (consistent with Newton’s first law). The object would not slow down at all if friction were completely eliminated. Consider an air hockey table. When the air is turned off, the puck slides only a short distance before friction slows it to a stop. However, when the air is turned on, it creates a nearly frictionless surface, and the puck glides long distances without slowing down. Additionally, if we know enough about the friction, we can accurately predict how quickly the object will slow down. Friction is an external force.

Newton’s first law is completely general and can be applied to anything from an object sliding on a table to a satellite in orbit to blood pumped from the heart. Experiments have thoroughly verified that any change in velocity (speed or direction) must be caused by an external force. The idea of *generally applicable or universal laws* is important not only here—it is a basic feature of all laws of physics. Identifying these laws is like recognizing patterns in nature from which further patterns can be discovered. The genius of Galileo, who first developed the idea for the first law, and Newton, who clarified it, was to ask the fundamental question, “What is the cause?” Thinking in terms of cause and effect is a worldview fundamentally different from the typical ancient Greek approach when questions such as “Why does a tiger have stripes?” would have been answered in Aristotelian fashion, “That is the nature of the beast.” True perhaps, but not a useful insight.

Mass

The property of a body to remain at rest or to remain in motion with constant velocity is called **inertia**. Newton’s first law is often called the **law of inertia**. As we know from experience, some objects have more inertia than others. It is obviously more difficult to change the motion of a large boulder than that of a basketball, for example. The inertia of an object is measured by its **mass**. Roughly speaking, mass is a measure of the amount of “stuff” (or matter) in something. The quantity or amount of matter in an object is determined by the numbers of atoms and molecules of various types it contains. Unlike weight, mass does not vary with location. The mass of an object is the same on Earth, in orbit, or on the surface of the Moon. In practice, it is very difficult to count and identify all of the atoms and molecules in an object, so masses are not often determined in this manner. Operationally, the masses of objects are determined by comparison with the standard kilogram.

Exercise 2.4.2.1

Which has more mass: a kilogram of cotton balls or a kilogram of gold?

Answer

> They are equal. A kilogram of one substance is equal in mass to a kilogram of another substance. The quantities that might differ between them are volume and density.

Section Summary

- **Newton's first law of motion** states that a body at rest remains at rest, or, if in motion, remains in motion at a constant velocity unless acted on by a net external force. This is also known as the **law of inertia**.
- **Inertia** is the tendency of an object to remain at rest or remain in motion. Inertia is related to an object's mass.
- **Mass** is the quantity of matter in a substance.

This page titled 2.4.2: Newton's First Law of Motion- Inertia is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- **2.3: Newton's First Law of Motion- Inertia** by OpenStax is licensed CC BY 4.0.

2.4.3: Newton's Second Law of Motion- Force and Acceleration

Learning Objectives

- Define net force, external force, and system.
- Understand Newton's second law of motion.
- Apply Newton's second law to determine the weight of an object.

Newton's second law of motion is closely related to Newton's first law of motion. It mathematically states the cause and effect relationship between force and changes in motion. Newton's second law of motion is more quantitative and is used extensively to calculate what happens in situations involving a force. Before we can write down Newton's second law as a simple equation giving the exact relationship of force, mass, and acceleration, we need to sharpen some ideas that have already been mentioned.

First, what do we mean by a change in motion? The answer is that a change in motion is equivalent to a change in velocity. A change in velocity means, by definition, that there is an **acceleration**. Newton's first law says that a net external force causes a change in motion; thus, we see that a *net external force causes acceleration*.

Another question immediately arises. What do we mean by an external force? An intuitive notion of external is correct—an **external force** acts from outside the **system** of interest. For example, in Figure 2.4.3.1(a) the system of interest is the wagon plus the child in it. The two forces exerted by the other children are external forces. An internal force acts between elements of the system. Again looking at Figure 2.4.3.1(a), the force the child in the wagon exerts to hang onto the wagon is an internal force between elements of the system of interest. Only external forces affect the motion of a system, according to Newton's first law. (The internal forces actually cancel, as we shall see in the next section.) *You must define the boundaries of the system before you can determine which forces are external*. Sometimes the system is obvious, whereas other times identifying the boundaries of a system is more subtle. The concept of a system is fundamental to many areas of physics, as is the correct application of Newton's laws. This concept will be revisited many times on our journey through physics.

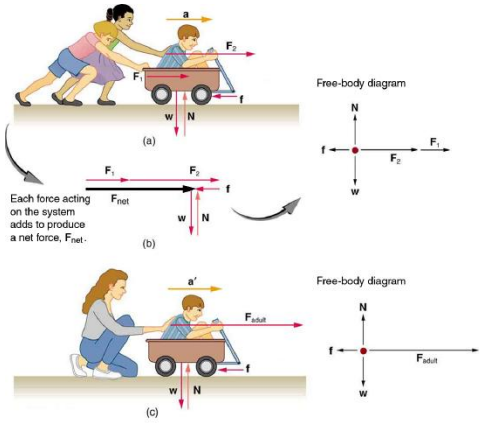


Figure 2.4.3.1: Different forces exerted on the same mass produce different accelerations. (a) Two children push a wagon with a child in it. Arrows representing all external forces are shown. The system of interest is the wagon and its rider. The weight w of the system and the support of the ground N are also shown for completeness and are assumed to cancel. The vector f represents the friction acting on the wagon, and it acts to the left, opposing the motion of the wagon. (b) All of the external forces acting on the system add together to produce a net force, $\boldsymbol{F}_{\text{net}}$. The free-body diagram shows all of the forces acting on the system of interest. The dot represents the center of mass of the system. Each force vector extends from this dot. Because there are two forces acting to the right, we draw the vectors collinearly. (c) A larger net external force produces a larger acceleration ($|\boldsymbol{a}'| > |\boldsymbol{a}|$) when an adult pushes the child.

Figure 2.4.3.1 is our first example of a **free-body diagram**, which is a technique used to illustrate all the external forces acting on a body. The body is represented by a single isolated point (or free body), and only those forces acting on the body from the outside (external forces) are shown. Free-body diagrams are very useful in analyzing forces acting on a system and are employed extensively in the study and application of Newton's laws of motion.

Now, it seems reasonable that acceleration should be directly proportional to and in the same direction as the net (total) external force acting on a system. This assumption has been verified experimentally and is illustrated in Figure 2.4.3.1. In part (a), a smaller force causes a smaller acceleration than the larger force illustrated in part (c). For completeness, the vertical forces are also shown; they are assumed to cancel since there is no acceleration in the vertical direction. The vertical forces are the weight w and the

support of the ground N, and the horizontal force f represents the force of friction. These will be discussed in more detail in later sections. For now, we will say that friction is a force that opposes the motion past each other of objects that are touching. Figure 2.4.3.1(b) shows how vectors representing the external forces add together to produce a net force, $\boldsymbol{F}_{\text{net}}$.

To obtain an equation for Newton's second law, we first write the relationship of acceleration and net external force as the proportionality

$$\boldsymbol{a} \propto \boldsymbol{F}_{\text{net}}$$

where the symbol $\propto$ means "proportional to," and $\boldsymbol{F}_{\text{net}}$ is the **net external force**, the vector sum of all external forces. This proportionality states what we have said in words—*acceleration is directly proportional to the net external force*. Once the system of interest is chosen, it is important to identify the external forces, so that the internal forces can be ignored (again, as we will see later, they exactly cancel out, allowing for a tremendous simplification).

Now, hopefully it also seems reasonable that acceleration should be inversely proportional to the mass of the system. That is, if the system has more "stuff", then for the same net external force applied, the acceleration is less. And indeed, as illustrated in Figure 2.4.3.2, the same net external force applied to a car produces a much smaller acceleration than when applied to a basketball. The proportionality is written as

$$a \propto \frac{1}{m}$$

where m is the mass of the system. In fact, this is how we will *define* mass. That is, for an object B to have double the mass of an object A, the same net external force applied on object B produces half the acceleration it does on object A.

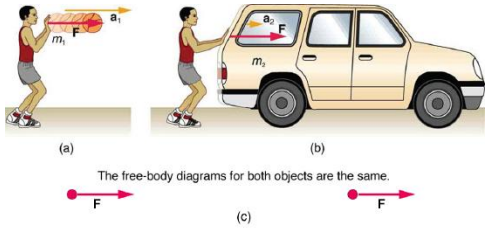


Figure 2.4.3.2: The same force exerted on systems of different masses produces different accelerations. (a) A basketball player pushes on a basketball to make a pass. (The effect of gravity on the ball is ignored.) (b) The same player exerts an identical force on a stalled SUV and produces a far smaller acceleration (even if friction is negligible). (c) The free-body diagrams are identical, permitting direct comparison of the two situations. A series of patterns for the free-body diagram will emerge as you do more problems.

As it turns out, the acceleration of an object depends only on the net external force and the mass of the object. Combining the two proportionalities just given yields Newton's second law of motion.

NEWTON'S SECOND LAW OF MOTION

The acceleration of a system is directly proportional to and in the same direction as the net external force acting on the system, and inversely proportional to its mass.

In equation form, Newton's second law of motion is

$$\boldsymbol{a} = \frac{\boldsymbol{F}_{\text{net}}}{m}$$

This is often written in the more familiar form

$$\boldsymbol{F}_{\text{net}} = m\boldsymbol{a},$$

with the vector notation indicating that the net external force is in the same direction as acceleration. When only the magnitude of force and acceleration are considered, this equation is simply (note the lack of vector notations)

$$F_{\text{net}} = ma.$$

Although these equations are really the same, the first gives more insight into what Newton's second law means. The law is a *cause and effect relationship* among three quantities. That is, acceleration is *caused* by a net external force, not vice versa, as the second equation might mistakenly imply.

Units of Force

$F_{\text{net}} = ma$ is used to define the units of force in terms of the three basic units for mass, length, and time. The SI unit of force is called the **newton** (abbreviated N) and 1 N is the force needed to accelerate a system of mass 1 kg at the rate of $1\ \text{m/s}^2$. Putting these together,

$$1\ \text{N} = 1\ \text{kg} \cdot \text{m/s}^2.$$

While almost the entire world uses the newton for the unit of force, in the United States the most familiar unit of force is the pound (lb), where 1 N = 0.225 lb.

Weight and the Gravitational Force

When an object is dropped, it accelerates toward the center of Earth. Newton's second law states that a net force on an object is responsible for its acceleration. If air resistance is negligible, the net force on a falling object is the gravitational force, commonly called its **weight**, w. Weight can be denoted as a vector w because it has a direction; *down* is, by definition, the direction of gravity, so w is directed downward. The magnitude of weight is denoted as w. Galileo was instrumental in showing that, in the absence of air resistance, all objects fall with the same acceleration g. Using Galileo's result and Newton's second law, we can derive an equation for the magnitude of weight.

Consider an object with mass m falling downward toward Earth. It experiences only the downward force of gravity, which has magnitude w. Newton's second law states that the magnitude of the net external force on an object is $F_{\text{net}} = ma$.

Since the object experiences only the downward force of gravity, $F_{\text{net}} = w$. We know that the acceleration of an object due to gravity is g, or $a = g$. Substituting these into Newton's second law gives

Definition: WEIGHT

This is the equation for *weight*—the gravitational force on a mass m:

$$w = mg.$$

Since $g = 9.80\ \text{m/s}^2$ on Earth, the weight of a 1.0 kg object on Earth is 9.8 N, as we see:

$$w = mg = (1.0\ \text{kg})\left(9.80\ \text{m/s}^2\right) = 9.8\ \text{N}.$$

When the net external force on an object is its weight, we say that it is in **free-fall**. That is, the only force acting on the object is the force of gravity. In the real world, when objects fall downward toward Earth, they are never truly in free-fall because there is always some upward force from the air acting on the object.

The acceleration due to gravity g varies slightly over the surface of Earth, so that the weight of an object depends on location and is not an intrinsic property of the object. Weight varies dramatically if one leaves Earth. On the Moon, for example, the acceleration due to gravity is only $1.67\ \text{m/s}^2$. A 1.0-kg mass thus has a weight of 9.8 N on Earth and only about 1.7 N on the Moon.

The broadest definition of weight in this sense is that *the weight of an object is the gravitational force on it from the nearest large body*, such as Earth, the Moon, the Sun, and so on. This is the most common and useful definition of weight in physics. It differs dramatically, however, from the definition of weight used by NASA and the popular media in relation to space travel and exploration. When they speak of "weightlessness" and "microgravity," they are really referring to the phenomenon we call "free-fall" in physics. We shall use the above definition of weight, and we will make careful distinctions between free-fall and actual weightlessness.

It is important to be aware that weight and mass are very different physical quantities, although they are closely related. Mass is the quantity of matter (how much "stuff") and does not vary, whereas weight is the gravitational force and does vary depending on gravity. It is tempting to equate the two, since most of our examples take place on Earth, where the weight of an object only varies a little with the location of the object. Furthermore, the terms *mass* and *weight* are used interchangeably in everyday language; for example, our medical records often show our "weight" in kilograms, but never in the correct units of newtons.

COMMON MISCONCEPTIONS: MASS VS. WEIGHT

Mass and weight are often used interchangeably in everyday language. However, in science, these terms are distinctly different from one another. Mass is a measure of how much matter is in an object. The typical measure of mass is the kilogram (or the "slug" in English units). Weight, on the other hand, is a measure of the force of gravity acting on an object. Weight is equal to the mass of an object (m) multiplied by the acceleration due to gravity (g). Like any other force, weight is measured in terms of newtons (or pounds in English units). This interchangeable use is why you might have heard an expression such as, "1 kilogram is 2.2 pounds." The *correct* statement is "1 kilogram mass weighs 2.2 pounds on Earth."

Assuming the mass of an object is kept intact, it will remain the same, regardless of its location. However, because weight depends on the acceleration due to gravity, the weight of an object *can change* when the object enters into a region with stronger or weaker gravity. For example, the acceleration due to gravity on the Moon is $1.67\ \mathrm{m/s^2}$ (which is much less than the acceleration due to gravity on Earth, $9.80\ \mathrm{m/s^2}$). If you measured your weight on Earth and then measured your weight on the Moon, you would find that you "weigh" much less, even though you do not look any skinnier. This is because the force of gravity is weaker on the Moon. In fact, when people say that they are "losing weight," they really mean that they are losing "mass" (which in turn causes them to weigh less).

TAKE-HOME EXPERIMENT: MASS AND WEIGHT

What do bathroom scales measure? When you stand on a bathroom scale, what happens to the scale? It depresses slightly. The scale contains springs that compress in proportion to your weight—similar to rubber bands expanding when pulled. The springs provide a measure of your weight (for an object which is not accelerating). This is a force in newtons (or pounds). In most countries, the measurement in newtons is divided by 9.80 to give a reading in mass units of kilograms. The scale measures weight but is calibrated to provide information about mass. While standing on a bathroom scale, push down on a table next to you. What happens to the reading? Why? Would your scale measure the same "mass" on Earth as on the Moon?

Example 2.4.3.1: What Acceleration Can a Person Produce when Pushing a Lawn Mower?

Suppose that the net external force (push minus friction) exerted on a lawn mower is 51 N (about 11 lb) parallel to the ground. The mass of the mower is 24 kg. What is its acceleration?



Figure 2.4.3.3: The net force on a lawn mower is 51 N to the right. At what rate does the lawn mower accelerate to the right?

Strategy

Since F_{net} and m are given, the acceleration can be calculated directly from Newton's second law as stated in $F_{\mathrm{net}} = ma$.

Solution

The magnitude of the acceleration a is $a = \frac{F_{\mathrm{net}}}{m}$. Entering known values gives

$$a = \frac{51\ \mathrm{N}}{24\ \mathrm{kg}}$$

Substituting the units $\mathrm{kg \cdot m/s^2}$ for N yields

$$a = \frac{51\ \mathrm{kg \cdot m/s^2}}{24\ \mathrm{kg}} = 2.1\ \mathrm{m/s^2}.$$

Discussion

The direction of the acceleration is the same direction as that of the net force, which is parallel to the ground. There is no information given in this example about the individual external forces acting on the system, but we can say something about

their relative magnitudes. For example, the force exerted by the person pushing the mower forward must be greater than the friction opposing the motion (directed backward) to result in a forward net force, and the vertical forces must cancel if there is to be no acceleration in the vertical direction. The acceleration found is small enough to be reasonable for a person pushing a mower. When the person reaches the top speed, the acceleration will be zero (no change of velocity) and the force exerted by the person pushing the mower will be equal to the friction opposing the motion.

Example 2.4.3.2: What Rocket Thrust Accelerates This Sled?

Prior to manned space flights, rocket sleds were used to test aircraft, missile equipment, and physiological effects on human subjects at high speeds. They consisted of a platform that was mounted on one or two rails and propelled by several rockets. Calculate the magnitude of force exerted by each rocket, called its thrust T, for the four-rocket propulsion system shown in Figure 2.4.3.4. The sled's initial acceleration is $49\ \mathrm{m/s^2}$ the mass of the system is 2100 kg, and the force of friction opposing the motion is known to be 650 N.

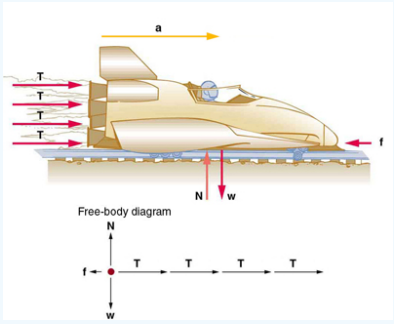


Figure 2.4.3.4: A sled experiences a rocket thrust that accelerates it to the right. Each rocket creates an identical thrust T. As in other situations where there is only horizontal acceleration, the vertical forces cancel. The ground exerts an upward force N on the system which balances out the weight force, w. The system here is the sled, its rockets, and rider, so none of the internal forces *between* these objects are considered. f represents the friction force (figure not drawn to scale).

Strategy

Although there are forces acting vertically and horizontally, we assume the vertical forces cancel since there is no vertical acceleration. This leaves us with only horizontal forces and a simpler one-dimensional problem. Directions are indicated with plus or minus signs, with right taken as the positive direction. See the free-body diagram in the figure.

Solution

Since acceleration, mass, and the force of friction are given, we start with Newton's second law and look for ways to find the thrust of the engines. Since we have defined the direction of the force and acceleration as acting "to the right," we need to consider only the magnitudes of these quantities in the calculations. Hence we begin with

$$F_{\text{net}} = ma,$$

where F_{net} is the net force along the horizontal direction. We can see from Figure 2.4.3.4 that the engine thrusts add, while friction opposes the thrust. In equation form, the net external force is

$$F_{\text{net}} = 4T - f.$$

Substituting this into Newton's second law gives

$$F_{\text{net}} = ma = 4T - f.$$

Using a little algebra, we solve for the total thrust $4T$:

$$4T = ma + f.$$

Substituting known values yields

$$4T = ma + f = (2100\ \text{kg})\left(49\ \text{m/s}^2\right) + 650\ \text{N}.$$

So the total thrust is

$$4T = 1.04 \times 10^5\ \text{N},$$

and the individual thrusts are

$$T = \frac{1.04 \times 10^5 \text{ N}}{4} = 2.6 \times 10^4 \text{ N}.$$

Discussion

The numbers are quite large, so to put the result in perspective, note that acceleration of 49 m/s^2 is approximately 5 times the gravitational acceleration on Earth. It's a large acceleration, which requires a large net external force to produce.

Experiments such as this were performed in the early 1960s to test the limits of human endurance and the setup designed to protect human subjects in jet fighter emergency ejections. Speeds of 1000 km/h were obtained, with accelerations of 45 g's. (Recall that g, the acceleration due to gravity, is 9.80 m/s^2. When we say that an acceleration is g's, it is $45 \times 9.80 \text{ m/s}^2$, which is approximately 440 m/s^2. While living subjects are not used any more, land speeds of 10,000 km/h have been obtained with rocket sleds. In this example, as in the preceding one, the system of interest is obvious. We will see in later examples that choosing the system of interest is crucial—and the choice is not always obvious.

Newton's second law of motion gives a relationship among acceleration, force, and mass. It can help us make predictions. Each of those physical quantities can be defined independently, so the second law tells us something basic and universal about nature. The next section introduces the third and final law of motion.

Section Summary

- Acceleration, a, is defined as a rate of change in velocity, resulting from a change in the magnitude and/or the direction of the velocity.
- An external force is one acting on a system from outside the system, as opposed to internal forces, which act between components within the system.
- Newton's second law of motion states that the acceleration of a system is directly proportional to and in the same direction as the net external force acting on the system, and inversely proportional to its mass.
- In equation form, Newton's second law of motion is $\boldsymbol{a} = \frac{\boldsymbol{F}_{\text{net}}}{m}$, often written in the more familiar form: $\boldsymbol{F}_{\text{net}} = m\boldsymbol{a}$.
- The weight w of an object is defined as the force of gravity acting on an object of mass mm. Given acceleration due to gravity g, the magnitude of weight is:

$$w = mg$$

Glossary

acceleration
the rate at which an object's velocity changes over a period of time

free-fall
a situation in which the only force acting on an object is the force due to gravity

external force
a force acting on an object or system that originates outside of the object or system

net external force
the vector sum of all external forces acting on an object or system; causes a mass to accelerate

free-body diagram
a sketch showing all of the external forces acting on an object or system; the system is represented by a dot, and the forces are represented by vectors extending outward from the dot

Newton's second law of motion
The acceleration of a system is directly proportional to and in the same direction as the net external force acting on the system, and inversely proportional to its mass.

system
the object or the group of objects under consideration

weight

the force due to gravity; $w = mg$ for objects on Earth

This page titled 2.4.3: Newton's Second Law of Motion- Force and Acceleration is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- **2.4: Newton's Second Law of Motion- Force and Acceleration** by OpenStax is licensed CC BY 4.0.

2.4.4: Newton's Third Law of Motion- Symmetry in Forces

Learning Objectives

- Understand Newton's third law of motion.
- Apply Newton's third law to define systems and solve problems of motion.

Whenever one body exerts a force on another—the first also experiences a force (equal in magnitude and opposite in direction). Numerous common experiences, such as stubbing a toe or throwing a ball, confirm this. It is precisely stated in **Newton's third law of motion**.

NEWTON'S THIRD LAW OF MOTION

Whenever one body exerts a force on a second body, the first body experiences a force by the second body that is equal in magnitude and opposite in direction to the force that it itself exerts.

This law represents a certain *symmetry in nature*: Forces always occur in pairs, and one body cannot exert a force on another without experiencing a force itself. We sometimes refer to this law loosely as "action-reaction," where the force exerted is the action and the force experienced as a consequence is the reaction. Newton's third law has practical uses in analyzing the origin of forces and understanding which forces are external to a system.

We can readily see Newton's third law at work by taking a look at how people move about. Consider a swimmer pushing off from the side of a pool, as illustrated in Figure 2.4.4.1. She pushes against the pool wall with her feet and accelerates in the direction *opposite* to that of her push. The wall has exerted an equal and opposite force back on the swimmer. You might think that two equal and opposite forces would cancel, but they do not *because they act on different bodies*. In this case, there are two bodies to consider: the swimmer and the wall. If we define the swimmer as the system of interest, as in the figure, then $\boldsymbol{F}_{\text{wall on feet}}$ is an external force on this system and affects its motion. The swimmer moves in the direction of $\boldsymbol{F}_{\text{wall on feet}}$. In contrast, the force $\boldsymbol{F}_{\text{feet on wall}}$ acts on the wall and not on our system of interest. Thus $\boldsymbol{F}_{\text{feet on wall}}$ does not affect the motion of the system and does not counteract $\boldsymbol{F}_{\text{wall on feet}}$. Note that the swimmer pushes the wall in the direction opposite to that in which she wishes to move. The reaction force by the wall on her pushes her in the desired direction.

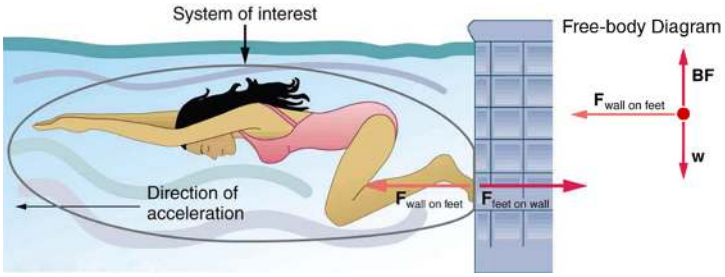


Figure 2.4.4.1: When the swimmer exerts a force $\boldsymbol{F}_{\text{feet on wall}}$ on the wall, she accelerates in the direction opposite to that of her push. This means the net external force on her is in the direction opposite to $\boldsymbol{F}_{\text{feet on wall}}$. This opposition occurs because, in accordance with Newton's third law of motion, the wall exerts a force $\boldsymbol{F}_{\text{wall on feet}}$ on her, equal in magnitude but in the direction opposite to the one she exerts on it. The line around the swimmer indicates the system of interest. Note that $\boldsymbol{F}_{\text{feet on wall}}$ does not act on this system (the swimmer) and, thus, does not cancel $\boldsymbol{F}_{\text{wall on feet}}$. Thus the free-body diagram shows only $\boldsymbol{F}_{\text{wall on feet}}$, w, the gravitational force, and **BF**, the buoyant force of the water supporting the swimmer's weight. The vertical forces w and **BF** cancel since there is no vertical motion.

Other examples of Newton's third law are easy to find. As a professor paces in front of a whiteboard, she exerts a force backward on the floor. The floor exerts a reaction force forward on the professor that causes her to accelerate forward. Similarly, a car accelerates because the ground pushes forward on the drive wheels in reaction to the drive wheels pushing backward on the ground. You can see evidence of the wheels pushing backward when tires spin on a gravel road and throw rocks backward. In another example, rockets move forward by expelling gas backward at high velocity. This means the rocket exerts a large backward force on the gas in the rocket combustion chamber, and the gas therefore exerts a large reaction force forward on the rocket. This reaction force is called **thrust**. It is a common misconception that rockets propel themselves by pushing on the ground or on the air behind them. The rockets propel themselves by pushing their own exhaust gases, and that is how they work even in a vacuum. Helicopters create lift by pushing air down, thereby experiencing an upward reaction force. Birds and airplanes also fly by exerting force on air

in a direction opposite to that of whatever force they need. For example, the wings of a bird force air downward and backward in order to get lift and move forward. An octopus propels itself in the water by ejecting water through a funnel from its body, similar to a jet ski. In a situation similar to Sancho's, professional cage fighters experience reaction forces when they punch, sometimes breaking their hand by hitting an opponent's body.

Example 2.4.4.1: Getting Up To Speed: Choosing the Correct System

A physics professor pushes a cart of demonstration equipment to a lecture hall, as seen in Figure 2.4.4.2. Her mass is 65.0 kg, the cart's is 12.0 kg, and the equipment's is 7.0 kg. Calculate the acceleration produced when the professor exerts a backward force of 150 N on the floor. All forces opposing the motion, such as friction on the cart's wheels and air resistance, total 24.0 N.

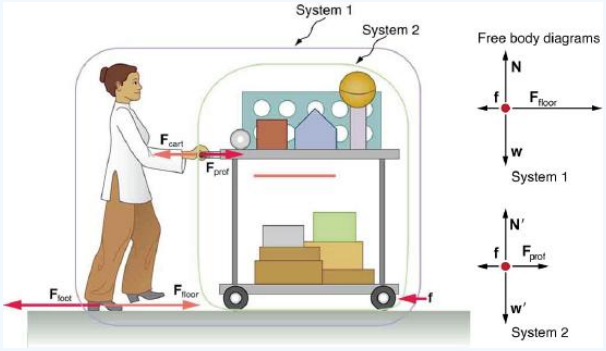


Figure 2.4.4.2: A professor pushes a cart of demonstration equipment. The lengths of the arrows are proportional to the magnitudes of the forces (except for **f**, since it is too small to draw to scale). Different questions are asked in each example; thus, the system of interest must be defined differently for each. System 1 is appropriate for this example, since it asks for the acceleration of the entire group of bodies. Only $\boldsymbol{F}_{\text{floor}}$ and f are external forces acting on System 1 along the line of motion. All other forces either cancel or act on the outside world. System 2 is appropriate for Example 2.4.4.2 so that $\boldsymbol{F}_{\text{prof}}$ will be an external force and enter into Newton's second law.

Strategy

Since they accelerate as a unit, we define the system to be the professor, cart, and equipment. This is System 1 in Figure 2.4.4.2. The professor pushes backward with a force $\boldsymbol{F}_{\text{foot}}$ of 150 N. According to Newton's third law, the floor exerts a forward reaction force $\boldsymbol{F}_{\text{floor}}$ of 150 N on System 1. Because all motion is horizontal, we can assume there is no net force in the vertical direction, and treat this as a one-dimensional problem in the horizontal direction. As noted, f opposes the motion and is thus in the opposite direction of $\boldsymbol{F}_{\text{floor}}$. Note that we do not include the forces $\boldsymbol{F}_{\text{prof}}$ or $\boldsymbol{F}_{\text{cart}}$ because these are internal forces, and we do not include $\boldsymbol{F}_{\text{foot}}$ because it acts on the floor, not on the system. There are no other significant forces acting on System 1. If the net external force can be found from all this information, we can use Newton's second law to find the acceleration as requested. See the free-body diagram in the figure.

Solution

Newton's second law is given by

$$a = \frac{F_{\text{net}}}{m}.$$

The net external force on System 1 is deduced from Figure 2.4.4.2 and the discussion above to be

$$F_{\text{net}} = F_{\text{floor}} - f = 150\ \text{N} - 24.0\ \text{N} = 126\ \text{N}.$$

The mass of System 1 is

$$m = (65.0 + 12.0 + 7.0)\text{kg} = 84\ \text{kg}.$$

These values of F_{net} and m produce an acceleration of

$$a = \frac{F_{\text{net}}}{m},$$
$$a = \frac{126\ \text{N}}{84\ \text{kg}} = 1.5\ \text{m/s}^2$$

Discussion

None of the forces between components of System 1, such as between the professor's hands and the cart, contribute to the net external force because they are internal to System 1. Another way to look at this is to note that forces between components of a system cancel because they are equal in magnitude and opposite in direction. For example, the force exerted by the professor on the cart results in an equal and opposite force back on her. In this case both forces act on the same system and, therefore, cancel. Thus internal forces (between components of a system) cancel. Choosing System 1 was crucial to solving this problem.

Example 2.4.4.2: Force on the Cart—Choosing a New System

Calculate the force the professor exerts on the cart in Figure 2.4.4.2 using data from the previous example if needed.

Strategy

If we now define the system of interest to be the cart plus equipment (System 2 in Figure 2.4.4.2), then the net external force on System 2 is the force the professor exerts on the cart minus friction. The force she exerts on the cart, $\mathbf{F}_{\text{prof}}$, is an external force acting on System 2. $\mathbf{F}_{\text{prof}}$ was internal to System 1, but it is external to System 2 and will enter Newton's second law for System 2.

Solution

Newton's second law can be used to find $\mathbf{F}_{\text{prof}}$. Starting with

$$a = \frac{F_{\text{net}}}{m}$$

and noting that the magnitude of the net external force on System 2 is

$$F_{\text{net}} = F_{\text{prof}} - f,$$

we solve for $\mathbf{F}_{\text{prof}}$, the desired quantity:

$$F_{\text{prof}} = F_{\text{net}} + f.$$

The value of f is given, so we must calculate net F_{net}. That can be done since both the acceleration and mass of System 2 are known. Using Newton's second law we see that

$$F_{\text{net}} = ma,$$

where the mass of System 2 is 19.0 kg (m= 12.0 kg + 7.0 kg) and its acceleration was found to be $a = 1.5\ \text{m/s}^2$ in the previous example. Thus,

$$F_{\text{net}} = ma,$$
$$F_{\text{net}} = (19.0\ \text{kg})\left(1.5\ \text{m/s}^2\right) = 29\ \text{N}.$$

Now we can find the desired force:

$$F_{\text{prof}} = F_{\text{net}} + f,$$
$$F_{\text{prof}} = 29\ \text{N} + 24.0\ \text{N} = 53\ \text{N}.$$

Discussion

It is interesting that this force is significantly less than the 150-N force the professor exerted backward on the floor. Not all of that 150-N force is transmitted to the cart; some of it accelerates the professor.

The choice of a system is an important analytical step both in solving problems and in thoroughly understanding the physics of the situation (which is not necessarily the same thing).

COMMON MISCONCEPTION: REACTION VS. RESULT

The phrases "action" and "reaction" in reference to the pair of forces described by Newton's third law often leads to unfortunate misunderstandings regarding the nature of the reaction force. Because of the common English usage of the word "reaction" (root word: react)—for example, in film, "reaction shot" means a shot of actors' faces as they react to an event that just

happened—many students intuitively guess that there is a cause-and-effect relationship between "action force" and "reaction force."

There is no such relationship. Newton's third law describes *symmetry* in forces. Action force and reaction force are symmetric—they occur simultaneously; they are two sides of one and the same interaction. Action force does not cause reaction force, no more than the head of a coin causes the tail of the coin (they merely occur together, always). Here's a quick check to see if you understood the pairing of action force and reaction force correctly: try swapping the labels. That is, if you originally called A "action force" and B "reaction force", try calling B "action force" and A "reaction force." Does this re-labeling make some sense? If not, chances are, you have mis-identified the pair of action force and reaction force. (Here, a little weirdness is O.K., but if it feels like complete nonsense to swap the labels—like reversal of a cause-and-effect relationship—that tells you you used "reaction" in the common English usage, not how the word is used in Newton's third law.)

Section Summary

- **Newton's third law of motion** represents a basic symmetry in nature. It states: Whenever one body exerts a force on a second body, the first body experiences a force that is equal in magnitude and opposite in direction to the force that the first body exerts.
- A **thrust** is a reaction force that pushes a body forward in response to a backward force. Rockets, airplanes, and cars are pushed forward by a thrust reaction force.

Glossary

Newton's third law of motion

Whenever one body exerts a force on a second body, the first body experiences a force by the second body that is equal in magnitude and opposite in direction to the force that it itself exerts.

thrust

a reaction force that pushes a body forward; rockets, airplanes, and cars are pushed forward by a thrust, a reaction force to propellants pushed backward

---

This page titled 2.4.4: Newton's Third Law of Motion- Symmetry in Forces is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- **2.5: Newton's Third Law of Motion- Symmetry in Forces** by OpenStax is licensed CC BY 4.0.

2.4.5: Friction

Learning Objectives

- Discuss the general characteristics of friction.
- Describe the various types of friction.
- Explain effect of friction on motion.

Friction is a force that is around us all the time that opposes relative motion between systems in contact but also allows us to move (which you have discovered if you have ever tried to walk on ice). While a common force, the behavior of friction is actually very complicated and is still not completely understood. We have to rely heavily on observations for whatever understandings we can gain. However, we can still deal with its more elementary general characteristics and understand the circumstances in which it behaves.

Definition: FRICTION

Friction is a force that opposes relative motion between systems in contact.

One of the simpler characteristics of friction is that it is parallel to the contact surface between systems and always in a direction that opposes motion or attempted motion of the systems relative to each other. If two systems are in contact and moving relative to one another, then the friction between them is called **kinetic friction**. For example, friction slows a hockey puck sliding on ice. But when objects are stationary, **static friction** can act between them; the static friction is usually greater than the kinetic friction between the objects.

Definition: KINETIC FRICTION

If two systems are in contact and moving relative to one another, then the friction between them is called kinetic friction.

Imagine, for example, trying to slide a heavy crate across a concrete floor—you may push harder and harder on the crate and not move it at all. This means that the static friction responds to what you do—it increases to be equal to and in the opposite direction of your push. But if you finally push hard enough, the crate seems to slip suddenly and starts to move. Once in motion it is easier to keep it in motion than it was to get it started, indicating that the kinetic friction force is less than the static friction force. If you add mass to the crate, say by placing a box on top of it, you need to push even harder to get it started and also to keep it moving. Furthermore, if you oiled the concrete you would find it to be easier to get the crate started and keep it going (as you might expect).

Figure 2.4.5.1 is a crude pictorial representation of how friction occurs at the interface between two objects. Close-up inspection of these surfaces shows them to be rough. So when you push to get an object moving (in this case, a crate), you must raise the object until it can skip along with just the tips of the surface hitting, break off the points, or do both. A considerable force can be resisted by friction with no apparent motion. The harder the surfaces are pushed together (such as if another box is placed on the crate), the more force is needed to move them. Part of the friction is due to adhesive forces between the surface molecules of the two objects, which explain the dependence of friction on the nature of the substances. Adhesion varies with substances in contact and is a complicated aspect of surface physics. Once an object is moving, there are fewer points of contact (fewer molecules adhering), so less force is required to keep the object moving. At small but nonzero speeds, friction is nearly independent of speed.

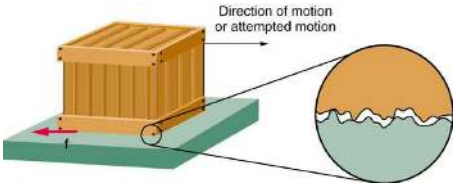


Figure 2.4.5.1: Frictional forces, such as f, always oppose motion or attempted motion between objects in contact. Friction arises in part because of the roughness of the surfaces in contact, as seen in the expanded view. In order for the object to move, it must rise to where the peaks can skip along the bottom surface. Thus a force is required just to set the object in motion. Some of the peaks will be broken off, also requiring a force to maintain motion. Much of the friction is actually due to attractive forces between molecules making up the two objects, so that even perfectly smooth surfaces are not friction-free. Such adhesive forces also depend on the substances the surfaces are made of, explaining, for example, why rubber-soled shoes slip less than those with leather soles.

The magnitude of the frictional force has two forms: one for static situations (static friction), the other for when there is motion (kinetic friction).

When there is no motion between the objects, the **magnitude of static friction $\mathbf{f_s}$** is

$$f_s \leq \mu_s N,$$

where μ_s is the coefficient of static friction and ***N*** is the magnitude of the normal force (the force perpendicular to the surface).

MAGNITUDE OF STATIC FRICTION

Magnitude of static friction f_s is

$$f_s \leq \mu_s N,$$

where μ_s is the coefficient of static friction and N is the magnitude of the normal force.

The symbol $\leq$ means *less than or equal to*, implying that static friction can have a minimum and a maximum value of $\mu_s N$. Static friction is a responsive force that increases to be equal and opposite to whatever force is exerted, up to its maximum limit. Once the applied force exceeds $f_{s(max)}$, the object will move. Thus

$$f_{s(max)} = \mu_s N.$$

Once an object is moving, the **magnitude of kinetic friction** f_k is given by

$$f_k = \mu_k N,$$

where μ_k is the coefficient of kinetic friction. A system in which $f_k = \mu_k N$ is described as a system in which *friction behaves simply*.

MAGNITUDE OF KINETIC FRICTION

The magnitude of kinetic friction f_k is given by

$$f_k = \mu_k N,$$

where μ_k is the coefficient of kinetic friction.

As seen in Table 2.4.5.1, the coefficients of kinetic friction are less than their static counterparts. That values of μ in Table 2.4.5.1 are stated to only one or, at most, two digits is an indication of the approximate description of friction given by the above two equations.

Table 2.4.5.1: Coefficients of Static and Kinetic Friction

System	Static friction μ_s	Kinetic friction μ_k
Rubber on dry concrete	1.0	0.7
Rubber on wet concrete	0.7	0.5
Wood on wood	0.5	0.3
Waxed wood on wet snow	0.14	0.1
Metal on wood	0.5	0.3
Steel on steel (dry)	0.6	0.3
Steel on steel (oiled)	0.05	0.03
Teflon on steel	0.04	0.04
Bone lubricated by synovial fluid	0.016	0.015

System	Static friction μ_s	Kinetic friction μ_k
Shoes on wood	0.9	0.7
Shoes on ice	0.1	0.05
Ice on ice	0.1	0.03
Steel on ice	0.4	0.02

The equations given earlier include the dependence of friction on materials and the normal force. The direction of friction is always opposite that of motion, parallel to the surface between objects, and perpendicular to the normal force. For example, if the crate you try to push (with a force parallel to the floor) has a mass of 100 kg, then the normal force would be equal to its weight, $W = mg = (100\ \text{kg})\,(9.80\ \text{m/s}^2) = 980\ \text{N}$, perpendicular to the floor. If the coefficient of static friction is 0.45, you would have to exert a force parallel to the floor greater than $f_{s(\max)} = \mu_s N = (0.45)(980\ \text{N}) = 440\ \text{N}$ to move the crate. Once there is motion, friction is less and the coefficient of kinetic friction might be 0.30, so that a force of only 290 N ($f_k = \mu_k N = (0.30)(980\ \text{N}) = 290\ \text{N}$) would keep it moving at a constant speed. If the floor is lubricated, both coefficients are considerably less than they would be without lubrication. Coefficient of friction is a unit less quantity with a magnitude usually between 0 and 1.0. The coefficient of the friction depends on the two surfaces that are in contact.

TAKE-HOME EXPERIMENT

Find a small plastic object (such as a food container) and slide it on a kitchen table by giving it a gentle tap. Now spray water on the table, simulating a light shower of rain. What happens now when you give the object the same-sized tap? Now add a few drops of (vegetable or olive) oil on the surface of the water and give the same tap. What happens now? This latter situation is particularly important for drivers to note, especially after a light rain shower. Why?

Many people have experienced the slipperiness of walking on ice. However, many parts of the body, especially the joints, have much smaller coefficients of friction—often three or four times less than ice. A joint is formed by the ends of two bones, which are connected by thick tissues. The knee joint is formed by the lower leg bone (the tibia) and the thighbone (the femur). The hip is a ball (at the end of the femur) and socket (part of the pelvis) joint. The ends of the bones in the joint are covered by cartilage, which provides a smooth, almost glassy surface. The joints also produce a fluid (synovial fluid) that reduces friction and wear. A damaged or arthritic joint can be replaced by an artificial joint (Figure 2.4.5.2). These replacements can be made of metals (stainless steel or titanium) or plastic (polyethylene), also with very small coefficients of friction.



Figure 2.4.5.2: Artificial knee replacement is a procedure that has been performed for more than 20 years. In this figure, we see the post-op X-rays of the right knee joint replacement. (credit: Mike Baird, Flickr)

Other natural lubricants include saliva produced in our mouths to aid in the swallowing process, and the slippery mucus found between organs in the body, allowing them to move freely past each other during heartbeats, during breathing, and when a person moves. Artificial lubricants are also common in hospitals and doctor's clinics. For example, when ultrasonic imaging is carried out,

the gel that couples the transducer to the skin also serves to lubricate the surface between the transducer and the skin—thereby reducing the coefficient of friction between the two surfaces. This allows the transducer to move freely over the skin.

We have discussed that when an object rests on a horizontal surface, there is a normal force supporting it equal in magnitude to its weight. Furthermore, simple friction is always proportional to the normal force.

MAKING CONNECTIONS: SUBMICROSCOPIC EXPLANATIONS OF FRICTION

The simpler aspects of friction dealt with so far are its macroscopic (large-scale) characteristics. Great strides have been made in the atomic-scale explanation of friction during the past several decades. Researchers are finding that the atomic nature of friction seems to have several fundamental characteristics. These characteristics not only explain some of the simpler aspects of friction—they also hold the potential for the development of nearly friction-free environments that could save hundreds of billions of dollars in energy which is currently being converted (unnecessarily) to heat.

Figure 2.4.5.3 illustrates one macroscopic characteristic of friction that is explained by microscopic (small-scale) research. We have noted that friction is proportional to the normal force, but not to the area in contact, a somewhat counterintuitive notion. When two rough surfaces are in contact, the actual contact area is a tiny fraction of the total area since only high spots touch. When a greater normal force is exerted, the actual contact area increases, and it is found that the friction is proportional to this area.

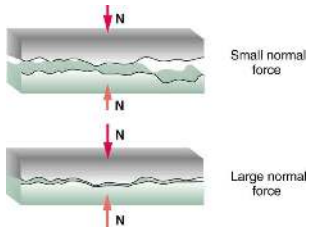


Figure 2.4.5.3: Two rough surfaces in contact have a much smaller area of actual contact than their total area. When there is a greater normal force as a result of a greater applied force, the area of actual contact increases as does friction.

But the atomic-scale view promises to explain far more than the simpler features of friction. The mechanism for how heat is generated is now being determined. In other words, why do surfaces get warmer when rubbed? Essentially, atoms are linked with one another to form lattices. When surfaces rub, the surface atoms adhere and cause atomic lattices to vibrate—essentially creating sound waves that penetrate the material. The sound waves diminish with distance and their energy is converted into heat. Chemical reactions that are related to frictional wear can also occur between atoms and molecules on the surfaces. Figure 2.4.5.4 shows how the tip of a probe drawn across another material is deformed by atomic-scale friction. The force needed to drag the tip can be measured and is found to be related to shear stress, which will be discussed later in this chapter. The variation in shear stress is remarkable (more than a factor of 10^{12}) and difficult to predict theoretically, but shear stress is yielding a fundamental understanding of a large-scale phenomenon known since ancient times—friction.

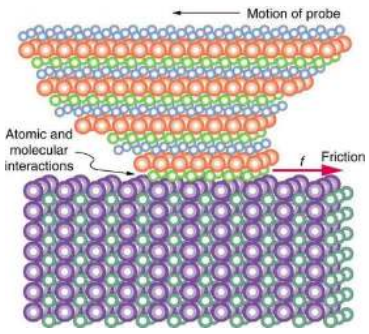


Figure 2.4.5.4: The tip of a probe is deformed sideways by frictional force as the probe is dragged across a surface. Measurements of how the force varies for different materials are yielding fundamental insights into the atomic nature of friction.

Section Summary

- Friction is a contact force between systems that opposes the motion or attempted motion between them. Simple friction is proportional to the normal force ***N*** pushing the systems together. (A normal force is always perpendicular to the contact surface between systems.) Friction depends on both of the materials involved. The magnitude of static friction f_s between systems stationary relative to one another is given by

$$f_s \leq \mu_s N,$$

where μ_s is the coefficient of static friction, which depends on both of the materials.

- The kinetic friction force f_k between systems moving relative to one another is given by

$$f_k = \mu_k N,$$

where μ_k is the coefficient of kinetic friction, which also depends on both materials.

Glossary

friction

a force that opposes relative motion or attempts at motion between systems in contact

kinetic friction

a force that opposes the motion of two systems that are in contact and moving relative to one another

static friction

a force that opposes the motion of two systems that are in contact and are not moving relative to one another

magnitude of static friction

$f_s \leq \mu_s N$, where μ_s is the coefficient of static friction and N is the magnitude of the normal force

magnitude of kinetic friction

$f_k = \mu_k N$, where μ_k is the coefficient of kinetic friction

---

This page titled 2.4.5: Friction is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- **2.8: Friction** by OpenStax is licensed CC BY 4.0.

2.4.E: Dynamics (Exercise)

Conceptual Questions

2.1: Development of Force Concept

1. Propose a force standard different from the example of a stretched spring discussed in the text. Your standard must be capable of producing the same force repeatedly.

2. What properties do forces have that allow us to classify them as vectors?

2.2: Newton's First Law of Motion- Inertia

3. How are inertia and mass related?

4. What is the relationship between weight and mass? Which is an intrinsic, unchanging property of a body?

2.3: Newton's Second Law of Motion- Force and Acceleration

5. Which statement is correct? (a) Net force causes motion. (b) Net force causes change in motion. Explain your answer and give an example.

6. Why can we neglect forces such as those holding a body together when we apply Newton's second law of motion?

7. Explain how the choice of the "system of interest" affects which forces must be considered when applying Newton's second law of motion.

8. Describe a situation in which the net external force on a system is not zero, yet its speed remains constant.

9. A system can have a nonzero velocity while the net external force on it *is* zero. Describe such a situation.

10. A rock is thrown straight up. What is the net external force acting on the rock when it is at the top of its trajectory?

11. (a) Give an example of different net external forces acting on the same system to produce different accelerations. (b) Give an example of the same net external force acting on systems of different masses, producing different accelerations. (c) What law accurately describes both effects? State it in words and as an equation.

12. If the acceleration of a system is zero, are no external forces acting on it? What about internal forces? Explain your answers.

13. If a constant, nonzero force is applied to an object, what can you say about the velocity and acceleration of the object?

14. The gravitational force on the basketball in Figure 2.4.*E*.1 is ignored. When gravity *is* taken into account, what is the direction of the net external force on the basketball—above horizontal, below horizontal, or still horizontal?

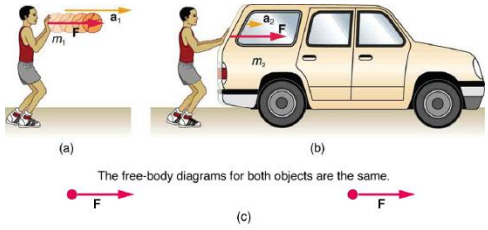


Figure 2.4.*E*.1: The same force exerted on systems of different masses produces different accelerations. (a) A basketball player pushes on a basketball to make a pass. (The effect of gravity on the ball is ignored.) (b) The same player exerts an identical force on a stalled SUV and produces a far smaller acceleration (even if friction is negligible). (c) The free-body diagrams are identical, permitting direct comparison of the two situations. A series of patterns for the free-body diagram will emerge as you do more problems.

2.4: Newton's Third Law of Motion- Symmetry in Forces

15. When you take off in a jet aircraft, there is a sensation of being pushed back into the seat. Explain why you move backward in the seat—is there really a force backward on you? (The same reasoning explains whiplash injuries, in which the head is apparently thrown backward.)

16. A device used since the 1940s to measure the kick or recoil of the body due to heart beats is the “ballistocardiograph.” What physics principle(s) are involved here to measure the force of cardiac contraction? How might we construct such a device?

17. Describe a situation in which one system exerts a force on another and, as a consequence, experiences a force that is equal in magnitude and opposite in direction. Which of Newton’s laws of motion apply?

18. Why does an ordinary rifle recoil (kick backward) when fired? The barrel of a recoilless rifle is open at both ends. Describe how Newton’s third law applies when one is fired. Can you safely stand close behind one when it is fired?

19. An American football lineman reasons that it is senseless to try to out-push the opposing player, since no matter how hard he pushes he will experience an equal and opposite force from the other player. Use Newton’s laws and draw a free-body diagram of an appropriate system to explain how he can still out-push the opposition if he is strong enough.

20. Newton’s third law of motion tells us that forces always occur in pairs of equal and opposite magnitude. Explain how the choice of the “system of interest” affects whether one such pair of forces cancels.

2.5: Normal Force and Tension

21. Find at least one example of each case from everyday experiences (find an example that is not in the textbook!): (a) a situation where normal force on a person is equal to the person's weight, (b) a situation where normal force on a person is greater than the person's weight, and (c) a situation where normal force on a person is less than the person's weight. Also, can you find an example (that is not in the textbook!) which covers each of all three cases, depending on changing circumstances?

22. As noted in the text, the normal force and the tension force can only act in one direction: the normal force can only push, and the tension force can only pull. The numerical value of the force in these given directions is determined by what force is necessary to keep a body in the observed state of motion (either not accelerating, in presence of other forces, or accelerating at a particular rate). Under certain conditions, the normal force would go to zero (but not negative, as direction can't reverse), as would the tension force.

23. Explain qualitatively what you would see as: (a) the normal force goes to zero, and (b) the tension force goes to zero.

2.6: Spring Force- Hooke’s Law

24. Describe a system which undergoes an oscillation under a Hooke's law force.

25. Imagine an object under a hypothetical force behaving opposite to Hooke's law force. That is, instead of restoring force ($F = -kx$, with the negative sign indicating that the force acts in the opposite direction of displacement), the force reinforces an initial displacement: $F = +kx$. What would happen to this object if it is placed at the equilibrium ($x = 0$) and then very slightly disturbed?

2.7: Friction

26. Define normal force. What is its relationship to friction when friction behaves simply?

27. The glue on a piece of tape can exert forces. Can these forces be a type of simple friction? Explain, considering especially that tape can stick to vertical walls and even to ceilings.

28. When you learn to drive, you discover that you need to let up slightly on the brake pedal as you come to a stop or the car will stop with a jerk. Explain this in terms of the relationship between static and kinetic friction.

29. When you push a piece of chalk across a chalkboard, it sometimes screeches because it rapidly alternates between slipping and sticking to the board. Describe this process in more detail, in particular explaining how it is related to the fact that kinetic friction is less than static friction. (The same slip-grab process occurs when tires screech on pavement.)

2.8: Newton’s Universal Law of Gravitation

30. Action at a distance, such as is the case for gravity, was once thought to be illogical and therefore untrue. What is the ultimate determinant of the truth in physics, and why was this action ultimately accepted?

31. Two friends are having a conversation. Anna says a satellite in orbit is in freefall because the satellite keeps falling toward Earth. Tom says a satellite in orbit is not in freefall because the acceleration due to gravity is not $9.80\ \mathrm{m/s^2}$. Who do you agree with and why?

32. Newton's laws of motion and gravity were among the first to convincingly demonstrate the underlying simplicity and unity in nature. Many other examples have since been discovered, and we now expect to find such underlying order in complex situations. Is there proof that such order will always be found in new explorations?

2.9: Centripetal Force

33. If you wish to reduce the stress (which is related to centripetal force) on high-speed tires, would you use large- or small-diameter tires? Explain.

34. Define centripetal force. Can any type of force (for example, tension, gravitational force, friction, and so on) be a centripetal force? Can any combination of forces be a centripetal force?

35. If centripetal force is directed toward the center, why do you feel that you are 'thrown' away from the center as a car goes around a curve? Explain.

36. Race car drivers routinely cut corners as shown in Figure 2.4.E.2. Explain how this allows the curve to be taken at the greatest speed.

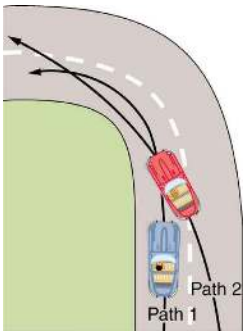


Figure 2.4.E.2: Two paths around a race track curve are shown. Race car drivers will take the inside path (called cutting the corner) whenever possible because it allows them to take the curve at the highest speed.

37. A number of amusement parks have rides that make vertical loops like the one shown in Figure 2.4.E.3. For safety, the cars are attached to the rails in such a way that they cannot fall off. If the car goes over the top at just the right speed, gravity alone will supply the centripetal force. What other force acts and what is its direction if:

(a) The car goes over the top at faster than this speed?

(b)The car goes over the top at slower than this speed?

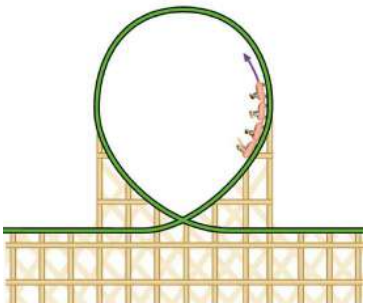


Figure 2.4.E.3: Amusement rides with a vertical loop are an example of a form of curved motion.

38. What is the direction of the force exerted by the car on the passenger as the car goes over the top of the amusement ride pictured in Figure 2.4.E.3 under the following circumstances:

(a) The car goes over the top at such a speed that the gravitational force is the only force acting?

(b) The car goes over the top faster than this speed?

(c) The car goes over the top slower than this speed?

39. Suppose a child is riding on a merry-go-round at a distance about halfway between its center and edge. She has a lunch box resting on wax paper, so that there is very little friction between it and the merry-go-round. Which path shown in Figure 2.4.E.4 will the lunch box take when she lets go? The lunch box leaves a trail in the dust on the merry-go-round. Is that trail straight, curved to the left, or curved to the right? Explain your answer.

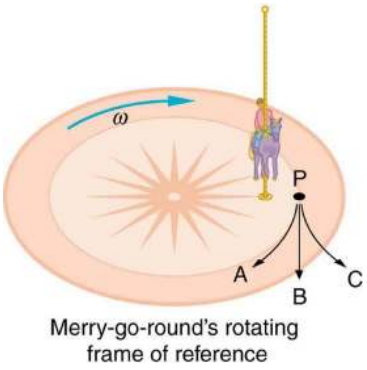


Figure 2.4.E.4: A child riding on a merry-go-round releases her lunch box at point P. This is a view from above the clockwise rotation. Assuming it slides with negligible friction, will it follow path A, B, or C, as viewed from Earth's frame of reference? What will be the shape of the path it leaves in the dust on the merry-go-round?

40. Do you feel yourself thrown to either side when you negotiate a curve that is ideally banked for your car's speed? What is the direction of the force exerted on you by the car seat?

41. Suppose a mass is moving in a circular path on a frictionless table as shown in figure. In the Earth's frame of reference, there is no centrifugal force pulling the mass away from the centre of rotation, yet there is a very real force stretching the string attaching the mass to the nail. Using concepts related to centripetal force and Newton's third law, explain what force stretches the string, identifying its physical origin.

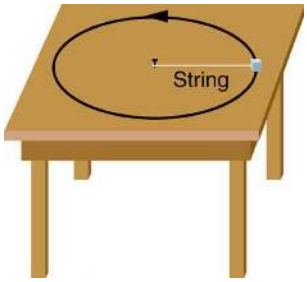


Figure 2.4.E.5: A mass attached to a nail on a frictionless table moves in a circular path. The force stretching the string is real and not fictional. What is the physical origin of the force on the string?

This page titled 2.4.E: Dynamics (Exercise) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- **2.E: Dynamics (Exercise)** by OpenStax is licensed CC BY 4.0.

CHAPTER OVERVIEW

3: Electricity and Magnetism



Under Construction(Public Domain ; Public Domain Pictures)

3.1: Static Electricity and Charge- Conservation of Charge

3.1.1: Ohm’s Law- Resistance and Simple Circuits

3.2: Resistors in Series and Parallel

3.3: Capacitors and Dielectrics

3.4: Magnets

3.5: Ferromagnets and Electromagnets

3: Electricity and Magnetism is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

3.1: Static Electricity and Charge- Conservation of Charge

Learning Objectives

- Define electric charge, and describe how the two types of charge interact.
- Describe three common situations that generate static electricity.
- State the law of conservation of charge.



Figure 3.1.1: Borneo amber was mined in Sabah, Malaysia, from shale-sandstone-mudstone veins. When a piece of amber is rubbed with a piece of silk, the amber gains more electrons, giving it a net negative charge. At the same time, the silk, having lost electrons, becomes positively charged. (credit: Sebakoamber, Wikimedia Commons)

What makes plastic wrap cling? Static electricity. Not only are applications of static electricity common these days, its existence has been known since ancient times. The first record of its effects dates to ancient Greeks who noted more than 500 years B.C. that polishing amber temporarily enabled it to attract bits of straw (see Figure 3.1.1). The very word *electric* derives from the Greek word for amber (*electron*).

Many of the characteristics of static electricity can be explored by rubbing things together. Rubbing creates the spark you get from walking across a wool carpet, for example. Static cling generated in a clothes dryer and the attraction of straw to recently polished amber also result from rubbing. Similarly, lightning results from air movements under certain weather conditions. You can also rub a balloon on your hair, and the static electricity created can then make the balloon cling to a wall. We also have to be cautious of static electricity, especially in dry climates. When we pump gasoline, we are warned to discharge ourselves (after sliding across the seat) on a metal surface before grabbing the gas nozzle. Attendants in hospital operating rooms must wear booties with aluminum foil on the bottoms to avoid creating sparks which may ignite the oxygen being used.

Some of the most basic characteristics of static electricity include:

- The effects of static electricity are explained by a physical quantity not previously introduced, called electric charge.
- There are only two types of charge, one called positive and the other called negative.
- Like charges repel, whereas unlike charges attract.
- The force between charges decreases with distance.

How do we know there are two types of **electric charge**? When various materials are rubbed together in controlled ways, certain combinations of materials always produce one type of charge on one material and the opposite type on the other. By convention, we call one type of charge “positive”, and the other type “negative.” For example, when glass is rubbed with silk, the glass becomes positively charged and the silk negatively charged. Since the glass and silk have opposite charges, they attract one another like clothes that have rubbed together in a dryer. Two glass rods rubbed with silk in this manner will repel one another, since each rod has positive charge on it. Similarly, two silk cloths so rubbed will repel, since both cloths have negative charge. Figure 3.1.2 shows how these simple materials can be used to explore the nature of the force between charges.

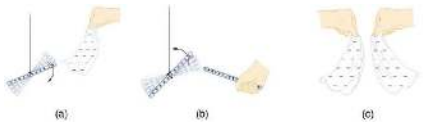


Figure 3.1.2: A glass rod becomes positively charged when rubbed with silk, while the silk becomes negatively charged. (a) The glass rod is attracted to the silk because their charges are opposite. (b) Two similarly charged glass rods repel. (c) Two similarly charged silk cloths repel.

More sophisticated questions arise. Where do these charges come from? Can you create or destroy charge? Is there a smallest unit of charge? Exactly how does the force depend on the amount of charge and the distance between charges? Such questions

obviously occurred to Benjamin Franklin and other early researchers, and they interest us even today.

Charge Carried by Electrons and Protons

Franklin wrote in his letters and books that he could see the effects of electric charge but did not understand what caused the phenomenon. Today we have the advantage of knowing that normal matter is made of atoms, and that atoms contain positive and negative charges, usually in equal amounts.

Figure 3.1.3 shows a simple model of an atom with negative **electrons** orbiting its positive nucleus. The nucleus is positive due to the presence of positively charged **protons**. Nearly all charge in nature is due to electrons and protons, which are two of the three building blocks of most matter. (The third is the neutron, which is neutral, carrying no charge.) Other charge-carrying particles are observed in cosmic rays and nuclear decay, and are created in particle accelerators. All but the electron and proton survive only a short time and are quite rare by comparison.

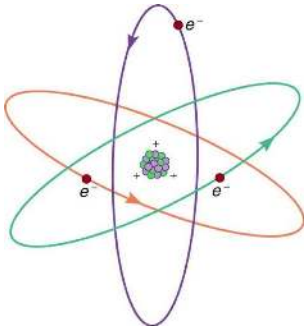


Figure 3.1.3: This simplified (and not to scale) view of an atom is called the planetary model of the atom. Negative electrons orbit a much heavier positive nucleus, as the planets orbit the much heavier sun. There the similarity ends, because forces in the atom are electromagnetic, whereas those in the planetary system are gravitational. Normal macroscopic amounts of matter contain immense numbers of atoms and molecules and, hence, even greater numbers of individual negative and positive charges.

The charges of electrons and protons are identical in magnitude but opposite in sign. Furthermore, all charged objects in nature are integral multiples of this basic quantity of charge, meaning that all charges are made of combinations of a basic unit of charge. Usually, charges are formed by combinations of electrons and protons. The magnitude of this basic charge is

$$|q_e| = 1.60 \times 10^{-19}\text{C}.$$

The symbol q is commonly used for charge and the subscript e indicates the charge of a single electron (or proton).

The SI unit of charge is the coulomb (C). The number of protons needed to make a charge of 1.00 C is

$$1.00\text{C} \times \frac{1 \text{ proton}}{1.60 \times 10^{-19}\text{C}} = 6.25 \times 10^{18} \text{ protons}.$$

Similarly, 6.25×10^{18} electrons have a combined charge of −1.00 coulomb. Just as there is a smallest bit of an element (an atom), there is a smallest bit of charge. There is no directly observed charge smaller than $|q_e|$, and all observed charges are integral multiples of $|q_e|$.

THINGS GREAT AND SMALL: THE SUBMICROSCOPIC ORIGIN OF CHARGE

With the exception of exotic, short-lived particles, all charge in nature is carried by electrons and protons. Electrons carry the charge we have named negative. Protons carry an equal-magnitude charge that we call positive. (See Figure 3.1.4.) Electron and proton charges are considered fundamental building blocks, since all other charges are integral multiples of those carried by electrons and protons. Electrons and protons are also two of the three fundamental building blocks of ordinary matter. The neutron is the third and has zero total charge.

Figure 3.1.4 shows a person touching a Van de Graaff generator and receiving excess positive charge. The expanded view of a hair shows the existence of both types of charges but an excess of positive. The repulsion of these positive like charges causes the strands of hair to repel other strands of hair and to stand up. The further blowup shows an artist's conception of an electron and a proton perhaps found in an atom in a strand of hair.

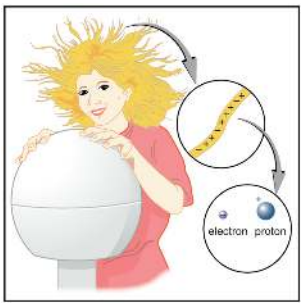


Figure 3.1.4: When this person touches a Van de Graaff generator, she receives an excess of positive charge, causing her hair to stand on end. The charges in one hair are shown. An artist's conception of an electron and a proton illustrate the particles carrying the negative and positive charges. We cannot really see these particles with visible light because they are so small (the electron seems to be an infinitesimal point), but we know a great deal about their measurable properties, such as the charges they carry.

The electron seems to have no substructure; in contrast, when the substructure of protons is explored by scattering extremely energetic electrons from them, it appears that there are point-like particles inside the proton. These sub-particles, named quarks, have never been directly observed, but they are believed to carry fractional charges as seen in Figure 3.1.5. Charges on electrons and protons and all other directly observable particles are unitary, but these quark substructures carry charges of either $-\frac{1}{3}$ or $+\frac{2}{3}$. There are continuing attempts to observe fractional charge directly and to learn of the properties of quarks, which are perhaps the ultimate substructure of matter.

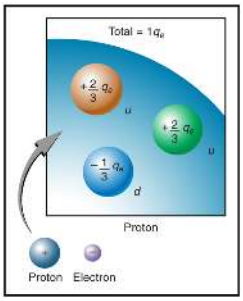


Figure 3.1.5: Artist's conception of fractional quark charges inside a proton. A group of three quark charges add up to the single positive charge on the proton: $-\frac{1}{3}q_e + \frac{2}{3}q_e + \frac{2}{3}q_e = +1q_e$.

Separation of Charge in Atoms

Charges in atoms and molecules can be separated—for example, by rubbing materials together. Some atoms and molecules have a greater affinity for electrons than others and will become negatively charged by close contact in rubbing, leaving the other material positively charged. (See Figure 3.1.6.) Positive charge can similarly be induced by rubbing. Methods other than rubbing can also separate charges. Batteries, for example, use combinations of substances that interact in such a way as to separate charges. Chemical interactions may transfer negative charge from one substance to the other, making one battery terminal negative and leaving the first one positive.

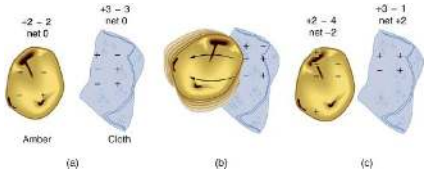


Figure 3.1.6: When materials are rubbed together, charges can be separated, particularly if one material has a greater affinity for electrons than another. (a) Both the amber and cloth are originally neutral, with equal positive and negative charges. Only a tiny fraction of the charges are involved, and only a few of them are shown here. (b) When rubbed together, some negative charge is transferred to the amber, leaving the cloth with a net positive charge. (c) When separated, the amber and cloth now have net charges, but the absolute value of the net positive and negative charges will be equal.

No charge is actually created or destroyed when charges are separated as we have been discussing. Rather, existing charges are moved about. In fact, in all situations the total amount of charge is always constant. This universally obeyed law of nature is called

the **law of conservation of charge**.

LAW OF CONSERVATION OF CHARGE

Total charge is constant in any process.

In more exotic situations, such as in particle accelerators, mass, Δm, can be created from energy in the amount $\Delta m = \frac{E}{c^2}$. Sometimes, the created mass is charged, such as when an electron is created. Whenever a charged particle is created, another having an opposite charge is always created along with it, so that the total charge created is zero. Usually, the two particles are "matter-antimatter" counterparts. For example, an antielectron would usually be created at the same time as an electron. The antielectron has a positive charge (it is called a positron), and so the total charge created is zero. (See Figure 3.1.7.) All particles have antimatter counterparts with opposite signs. When matter and antimatter counterparts are brought together, they completely annihilate one another. By annihilate, we mean that the mass of the two particles is converted to energy E, again obeying the relationship $\Delta m = \frac{E}{c^2}$. Since the two particles have equal and opposite charge, the total charge is zero before and after the annihilation; thus, total charge is conserved.

MAKING CONNECTIONS: CONSERVATION LAWS

Only a limited number of physical quantities are universally conserved. Charge is one—energy, momentum, and angular momentum are others. Because they are conserved, these physical quantities are used to explain more phenomena and form more connections than other, less basic quantities. We find that conserved quantities give us great insight into the rules followed by nature and hints to the organization of nature. Discoveries of conservation laws have led to further discoveries, such as the weak nuclear force and the quark substructure of protons and other particles.

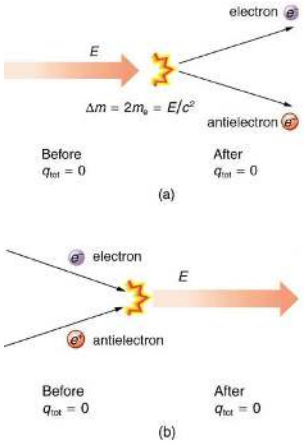


Figure 3.1.7: (a) When enough energy is present, it can be converted into matter. Here the matter created is an electron–antielectron pair. (m_e is the electron's mass.) The total charge before and after this event is zero. (b) When matter and antimatter collide, they annihilate each other; the total charge is conserved at zero before and after the annihilation.

The law of conservation of charge is absolute—it has never been observed to be violated. Charge, then, is a special physical quantity, joining a very short list of other quantities in nature that are always conserved. Other conserved quantities include energy, momentum, and angular momentum.

Section Summary

- There are only two types of charge, which we call positive and negative.
- Like charges repel, unlike charges attract, and the force between charges decreases with the square of the distance.
- The vast majority of positive charge in nature is carried by protons, while the vast majority of negative charge is carried by electrons.
- The electric charge of one electron is equal in magnitude and opposite in sign to the charge of one proton.
- An ion is an atom or molecule that has nonzero total charge due to having unequal numbers of electrons and protons.
- The SI unit for charge is the coulomb (C), with protons and electrons having charges of opposite sign but equal magnitude; the magnitude of this basic charge $|q_e|$ is

$$|q_e| = 1.60 \times 10^{-19}\text{C}.$$

- Whenever charge is created or destroyed, equal amounts of positive and negative are involved.
- Most often, existing charges are separated from neutral objects to obtain some net charge.
- Both positive and negative charges exist in neutral objects and can be separated by rubbing one object with another. For macroscopic objects, negatively charged means an excess of electrons and positively charged means a depletion of electrons.
- The law of conservation of charge ensures that whenever a charge is created, an equal charge of the opposite sign is created at the same time.

Glossary

electric charge

a physical property of an object that causes it to be attracted toward or repelled from another charged object; each charged object generates and is influenced by a force called an electromagnetic force

law of conservation of charge

states that whenever a charge is created, an equal amount of charge with the opposite sign is created simultaneously

electron

a particle orbiting the nucleus of an atom and carrying the smallest unit of negative charge

proton

a particle in the nucleus of an atom and carrying a positive charge equal in magnitude and opposite in sign to the amount of negative charge carried by an electron

---

This page titled 3.1: Static Electricity and Charge- Conservation of Charge is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- **9.2: Static Electricity and Charge- Conservation of Charge** by OpenStax is licensed CC BY 4.0.

3.1.1: Ohm's Law- Resistance and Simple Circuits

Learning Objectives

- Explain the origin of Ohm's law.
- Calculate voltages, currents, or resistances with Ohm's law.
- Explain what an ohmic material is.
- Describe a simple circuit.

What drives current? We can think of various devices—such as batteries, generators, wall outlets, and so on—which are necessary to maintain a current. All such devices create a potential difference and are loosely referred to as voltage sources. When a voltage source is connected to a conductor, it applies a potential difference V that creates an electric field. The electric field in turn exerts force on charges, causing current.

Ohm's Law

The current that flows through most substances is directly proportional to the voltage V applied to it. The German physicist Georg Simon Ohm (1787–1854) was the first to demonstrate experimentally that the current in a metal wire is *directly proportional to the voltage applied*:

$$I \propto V.$$

This important relationship is known as **Ohm's law**. It can be viewed as a cause-and-effect relationship, with voltage the cause and current the effect. This is an empirical law like that for friction—an experimentally observed phenomenon. Such a linear relationship doesn't always occur.

Resistance and Simple Circuits

If voltage drives current, what impedes it? The electric property that impedes current (crudely similar to friction and air resistance) is called **resistance** R. Collisions of moving charges with atoms and molecules in a substance transfer energy to the substance and limit current. Resistance is defined as inversely proportional to current, or

$$I \propto \frac{1}{R}.$$

Thus, for example, current is cut in half if resistance doubles. Combining the relationships of current to voltage and current to resistance gives

$$I = \frac{V}{R}.$$

This relationship is also called Ohm's law. Ohm's law in this form really defines resistance for certain materials. Ohm's law (like Hooke's law) is not universally valid. The many substances for which Ohm's law holds are called **ohmic**. These include good conductors like copper and aluminum, and some poor conductors under certain circumstances. Ohmic materials have a resistance R that is independent of voltage V and current I. An object that has simple resistance is called a *resistor*, even if its resistance is small. The unit for resistance is an **ohm** and is given the symbol Ω (upper case Greek omega). Rearranging $I = V/R$ gives $R = V/I$, and so the units of resistance are 1 ohm = 1 volt per ampere:

$$1\ \Omega = 1\frac{V}{A}.$$

Figure 3.1.1.1 shows the schematic for a simple circuit. A **simple circuit** has a single voltage source and a single resistor. The wires connecting the voltage source to the resistor can be assumed to have negligible resistance, or their resistance can be included in R.

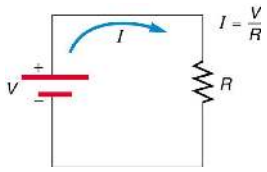


Figure 3.1.1.1: A simple electric circuit in which a closed path for current to flow is supplied by conductors (usually metal wires) connecting a load to the terminals of a battery, represented by the red parallel lines. The zigzag symbol represents the single resistor and includes any resistance in the connections to the voltage source.

Example 3.1.1.1: Calculating Resistance: An Automobile Headlight

What is the resistance of an automobile headlight through which 2.50 A flows when 12.0 V is applied to it?

Strategy

We can rearrange Ohm's law as stated by $I = V/R$ and use it to find the resistance.

Solution

Rearranging $I = V/R$ and substituting known values gives

$$R = \frac{V}{I} = \frac{12.0\ \text{V}}{2.50\ \text{A}} = 4.80\Omega.$$

Discussion

This is a relatively small resistance, but it is larger than the cold resistance of the headlight. Resistance usually increases with temperature, and so the bulb has a lower resistance when it is first switched on and will draw considerably more current during its brief warm-up period.

Resistances range over many orders of magnitude. Some ceramic insulators, such as those used to support power lines, have resistances of $10^{12}\ \Omega$ or more. A dry person may have a hand-to-foot resistance of $10^{5}\ \Omega$, whereas the resistance of the human heart is about $10^{3}\ \Omega$. A meter-long piece of large-diameter copper wire may have a resistance of $10^{-5}\ \Omega$, and superconductors have no resistance at all (they are non-ohmic). Resistance is related to the shape of an object and the material of which it is composed.

Additional insight is gained by solving $I = V/R$ for V yielding

$$V = IR.$$

This expression for V can be interpreted as the *voltage drop across a resistor produced by the flow of current* I. The phrase IR *drop* is often used for this voltage. For instance, the headlight in Example 3.1.1.1 has an IR drop of 12.0 V. If voltage is measured at various points in a circuit, it will be seen to increase at the voltage source and decrease at the resistor. Voltage is similar to fluid pressure. The voltage source is like a pump, creating a pressure difference, causing current—the flow of charge. The resistor is like a pipe that reduces pressure and limits flow because of its resistance. Conservation of energy has important consequences here. The voltage source supplies energy (causing an electric field and a current), and the resistor converts it to another form (such as thermal energy). In a simple circuit (one with a single simple resistor), the voltage supplied by the source equals the voltage drop across the resistor, since $\text{PE} = q\Delta V$, and the same q flows through each. Thus the energy supplied by the voltage source and the energy converted by the resistor are equal. (See Figure 3.1.1.2.)

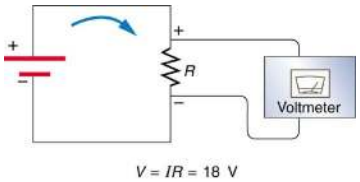


Figure 3.1.1.2: The voltage drop across a resistor in a simple circuit equals the voltage output of the battery.

MAKING CONNECTIONS: CONSERVATION OF ENERGY

In a simple electrical circuit, the sole resistor converts energy supplied by the source into another form. Conservation of energy is evidenced here by the fact that all of the energy supplied by the source is converted to another form by the resistor alone. We will find that conservation of energy has other important applications in circuits and is a powerful tool in circuit analysis.

Section Summary

- A simple circuit *is* one in which there is a single voltage source and a single resistance.
- One statement of Ohm's law gives the relationship between current I, voltage V, and resistance R in a simple circuit to be $I = \frac{V}{R}$
- Resistance has units of ohms (Ω), related to volts and amperes by $1\ \Omega = 1\ \mathrm{V/A}$.
- There is a voltage or IR drop across a resistor, caused by the current flowing through it, given by $V = IR$.

Glossary

Ohm's law

an empirical relation stating that the current *I* is proportional to the potential difference *V*. It is often written as *I* = *V*/*R*, where *R* is the resistance

resistance

the electric property that impedes current; for ohmic materials, it is the ratio of voltage to current, *R* = *V*/*I*

ohm

the unit of resistance, given by $1\ \Omega = 1\ \mathrm{V/A}$

ohmic

a type of a material for which Ohm's law is valid

simple circuit

a circuit with a single voltage source and a single resistor

This page titled 3.1.1: Ohm's Law- Resistance and Simple Circuits is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- **9.9: Ohm's Law- Resistance and Simple Circuits** by OpenStax is licensed CC BY 4.0.

3.2: Resistors in Series and Parallel

Learning Objectives

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

- Draw a circuit with resistors in parallel and in series.
- Calculate the voltage drop of a current across a resistor using Ohm's law.
- Contrast the way total resistance is calculated for resistors in series and in parallel.
- Explain why total resistance of a parallel circuit is less than the smallest resistance of any of the resistors in that circuit.
- Calculate total resistance of a circuit that contains a mixture of resistors connected in series and in parallel.

sMost circuits have more than one component, called a **resistor** that limits the flow of charge in the circuit. A measure of this limit on charge flow is called **resistance**. The simplest combinations of resistors are the series and parallel connections illustrated in Figure 3.2.1. The total resistance of a combination of resistors depends on both their individual values and how they are connected.

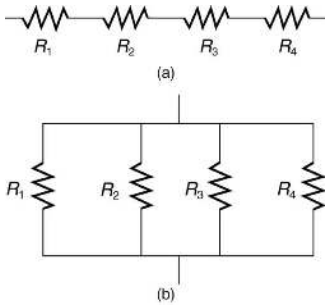


Figure 3.2.1: (a) A series connection of resistors. (b) A parallel connection of resistors.

Resistors in Series

When are resistors in **series**? Resistors are in series whenever the flow of charge, called the **current**, must flow through devices sequentially. For example, if current flows through a person holding a screwdriver and into the Earth, then R_1 in Figure 3.2.1(a) could be the resistance of the screwdriver's shaft, R_2 the resistance of its handle, R_3 the person's body resistance, and R_4 the resistance of her shoes.

Figure 3.2.2 shows resistors in series connected to a **voltage** source. It seems reasonable that the total resistance is the sum of the individual resistances, considering that the current has to pass through each resistor in sequence. (This fact would be an advantage to a person wishing to avoid an electrical shock, who could reduce the current by wearing high-resistance rubber-soled shoes. It could be a disadvantage if one of the resistances were a faulty high-resistance cord to an appliance that would reduce the operating current.)

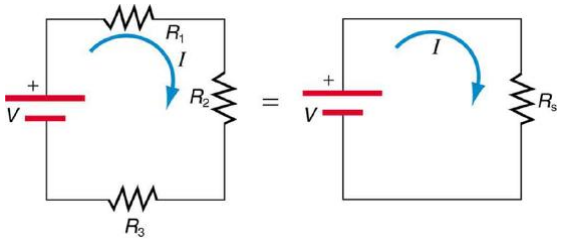


Figure 3.2.2: Three resistors connected in series to a battery (left) and the equivalent single or series resistance (right).

To verify that resistances in series do indeed add, let us consider the loss of electrical power, called a **voltage drop**, in each resistor in Figure 3.2.2.

According to **Ohm's law**, the voltage drop, V, across a resistor when a current flows through it is calculated using the equation $V = IR$, where I equals the current in amps (A) and R is the resistance in ohms(Ω). Another way to think of this is that V is the voltage necessary to make a current I flow through a resistance R.

So the voltage drop across R_1 is $V_1 = IR_1$, that across $V_2 = IR_2$, and that across R_3 is $V_3 = IR_3$. The sum of these voltages equals the voltage output of the source; that is,

$$V = V_1 + V_2 + V_3. \tag{3.2.1}$$

This equation is based on the conservation of energy and conservation of charge. Electrical potential energy can be described by the equation $\mathrm{PE} = qV$, where q is the electric charge and V is the voltage. Thus the energy supplied by the source is qV, while that dissipated by the resistors is

$$qV_1 + qV_2 + qV_3. \tag{3.2.2}$$

CONNECTIONS: CONSERVATION LAWS

The derivations of the expressions for series and parallel resistance are based on the laws of conservation of energy and conservation of charge, which state that total charge and total energy are constant in any process. These two laws are directly involved in all electrical phenomena and will be invoked repeatedly to explain both specific effects and the general behavior of electricity.

These energies must be equal, because there is no other source and no other destination for energy in the circuit. Thus, $qV = qV_1 + qV_2 + qV_3$. The charge q cancels, yielding $V = V_1 + V_2 + V_3$, as stated. (Note that the same amount of charge passes through the battery and each resistor in a given amount of time, since there is no capacitance to store charge, there is no place for charge to leak, and charge is conserved.)

Now substituting the values for the individual voltages gives

$$V = IR_1 + IR_2 + IR_3 = I(R_1 + R_2 + R_3). \tag{3.2.3}$$

Note that for the equivalent single series resistance R_S, we have

$$V = IR_\mathrm{S}. \tag{3.2.4}$$

This implies that the total or equivalent series resistance R_S of three resistors is $R_\mathrm{S} = R_1 + R_2 + R_3$.

This logic is valid in general for any number of resistors in series; thus, the total resistance R_S of a series connection is

$$R_\mathrm{S} = R_1 + R_2 + R_3 + \ldots, \tag{3.2.5}$$

as proposed. Since all of the current must pass through each resistor, it experiences the resistance of each, and resistances in series simply add up.

Example 3.2.1: Calculating Resistance, Current, Voltage Drop, and Power Dissipation: Analysis of a Series Circuit

Suppose the voltage output of the battery in Figure 3.2.2 is $12.0\mathrm{V}$, and the resistances are $R_1 = 1.00\Omega$, $R_2 = 6.00\Omega$, and $R_3 = 13.0\Omega$. (a) What is the total resistance? (b) Find the current. (c) Calculate the voltage drop in each resistor, and show these add to equal the voltage output of the source. (d) Calculate the power dissipated by each resistor. (e) Find the power output of the source, and show that it equals the total power dissipated by the resistors.

Strategy and Solution for (a)

The total resistance is simply the sum of the individual resistances, as given by this equation:

$$R_\mathrm{S} = R_1 + R_2 + R_3 \tag{3.2.6}$$

$$= 1.00\Omega + 6.00\Omega + 13.0\Omega \tag{3.2.7}$$

$$= 20.0\Omega. \tag{3.2.8}$$

Strategy and Solution for (b)

The current is found using Ohm's law, $V = IR$. Entering the value of the applied voltage and the total resistance yields the current for the circuit:

$$I = \frac{V}{R_\mathrm{S}} = \frac{12.0\Omega}{20.0\Omega} = 0.600\mathrm{A}. \tag{3.2.9}$$

Strategy and Solution for (c)

The voltage—or IR drop—in a resistor is given by Ohm's law. Entering the current and the value of the first resistance yields

$$V = IR = (0.600\mathrm{A})(1.0\Omega) = 0.600\mathrm{V} \tag{3.2.10}$$

$$V_1 = IR_1 = (0.600\text{A})(1.0\Omega) = 0.600\text{V}. \tag{3.2.10}$$

Similarly,

$$V_2 = IR_2 = (0.600\text{A})(6.0\Omega) = 3.60\text{V} \tag{3.2.11}$$

and

$$V_3 = IR_3 = (0.600\text{A})(13.0\Omega) = 7.80\text{V}. \tag{3.2.12}$$

Discussion for (c)

The three IR drops add to 12.0V, as predicted:

$$V_1 + V_2 + V_3 = (0.600 + 3.60 + 7.80)\text{V} = 12.0\text{V}. \tag{3.2.13}$$

Strategy and Solution for (d)

The easiest way to calculate power in watts (W) dissipated by a resistor in a DC circuit is to use **Joule's law**, $P = IV$, where P is electric power. In this case, each resistor has the same full current flowing through it. By substituting Ohm's law $V = IR$ into Joule's law, we get the power dissipated by the first resistor as

$$P_1 = I^2R_1 = (0.600\text{A})^2(1.00\Omega) = 0.360\text{W}. \tag{3.2.14}$$

Similarly,

$$P_2 = I^2R_2 = (0.600\text{A})^2(6.00\Omega) = 2.16\text{W} \tag{3.2.15}$$

and

$$P_3 = I^2R_3 = (0.600\text{A})^2(13.0\Omega) = 4.68\text{W}. \tag{3.2.16}$$

Discussion for (d)

Power can also be calculated using either $P = IV$ or $P = \dfrac{V^2}{R}$, where V is the voltage drop across the resistor (not the full voltage of the source). The same values will be obtained.

Strategy and Solution for (e)

The easiest way to calculate power output of the source is to use $P = IV$, where V is the source voltage. This gives

$$P = (0.600\text{A})(12.0\text{V}) = 7.20\text{W}. \tag{3.2.17}$$

Discussion for (e)

Note, coincidentally, that the total power dissipated by the resistors is also 7.20 W, the same as the power put out by the source. That is,

$$P_1 + P_2 + P_3 = (0.360 + 2.16 + 4.68)\text{W} = 7.20\text{W}. \tag{3.2.18}$$

Power is energy per unit time (watts), and so conservation of energy requires the power output of the source to be equal to the total power dissipated by the resistors.

MAJOR FEATURES OF RESISTORS IN SERIES

1. Series resistances add: $R_\text{S} = R_1 + R_2 + R_3 + ...$
2. The same current flows through each resistor in series.
3. Individual resistors in series do not get the total source voltage, but divide it.

Resistors in Parallel

Figure 3.2.3 shows resistors in **parallel**, wired to a voltage source. Resistors are in parallel when each resistor is connected directly to the voltage source by connecting wires having negligible resistance. Each resistor thus has the full voltage of the source applied to it.

Each resistor draws the same current it would if it alone were connected to the voltage source (provided the voltage source is not overloaded). For example, an automobile's headlights, radio, and so on, are wired in parallel, so that they utilize the full voltage of

the source and can operate completely independently. The same is true in your house, or any building. (See Figure 3.2.3(b).)

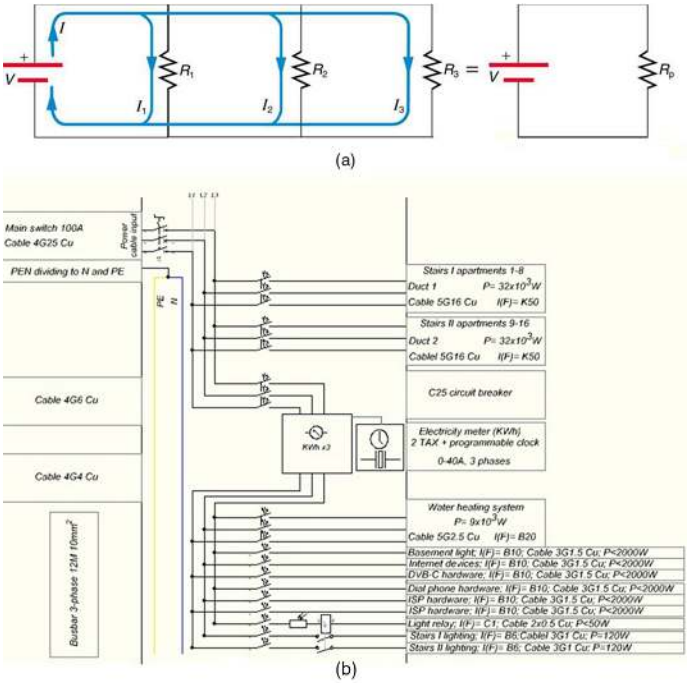


Figure 3.2.3: (a) Three resistors connected in parallel to a battery and the equivalent single or parallel resistance. (b) Electrical power setup in a house. (credit: Dmitry G, Wikimedia Commons)

To find an expression for the equivalent parallel resistance R_{p}, let us consider the currents that flow and how they are related to resistance. Since each resistor in the circuit has the full voltage, the currents flowing through the individual resistors are $I_1 = \frac{V}{R_1}$, $I_2 = \frac{V}{R_2}$, and $I_3 = \frac{V}{R_3}$. Conservation of charge implies that the total current I produced by the source is the sum of these currents:

$$I = I_1 + I_2 + I_3. \tag{3.2.19}$$

Substituting the expressions for the individual currents gives

$$I = \frac{V}{R_1} + \frac{V}{R_2} + \frac{V}{R_3} = V(\frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3}). \tag{3.2.20}$$

Note that Ohm's law for the equivalent single resistance gives

$$I = \frac{V}{R_{\text{p}}} = V(\frac{1}{R_{\text{p}}}). \tag{3.2.21}$$

The terms inside the parentheses in the last two equations must be equal. Generalizing to any number of resistors, the total resistance R_{p} of a parallel connection is related to the individual resistances by

$$\frac{1}{R_{\text{p}}} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} + ... \tag{3.2.22}$$

This relationship results in a total resistance R_{p} that is less than the smallest of the individual resistances. (This is seen in the next example.) When resistors are connected in parallel, more current flows from the source than would flow for any of them individually, and so the total resistance is lower.

Example 3.2.2: Calculating Resistance, Current, Power Dissipation, and Power Output: Analysis of a Parallel Circuit

Let the voltage output of the battery and resistances in the parallel connection in Figure 3.2.3 be the same as the previously considered series connection: $V = 12.0\text{V},\ R_1 = 1.00\Omega,\ R_2 = 6.00\Omega$, and $R_3 = 13.0\Omega$. (a) What is the total resistance? (b) Find the total current. (c) Calculate the currents in each resistor, and show these add to equal the total current output of the source. (d) Calculate the power dissipated by each resistor. (e) Find the power output of the source, and show that it equals the total power dissipated by the resistors.

Strategy and Solution for (a)

The total resistance for a parallel combination of resistors is found using the equation below. Entering known values gives

$$\frac{1}{R_{\text{p}}} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} = \frac{1}{1.00\Omega} + \frac{1}{6.00\Omega} + \frac{1}{13.0\Omega}. \tag{3.2.23}$$

Thus,

$$\frac{1}{R_{\text{p}}} = \frac{1.00}{\Omega} + \frac{0.1667}{\Omega} + \frac{0.07692}{\Omega} = \frac{1.2436}{\Omega}. \tag{3.2.24}$$

(Note that in these calculations, each intermediate answer is shown with an extra digit.)

We must invert this to find the total resistance R_{p}. This yields

$$R_{\text{p}} = \frac{1}{1.2436}\Omega = 0.8041\Omega. \tag{3.2.25}$$

The total resistance with the correct number of significant digits is $R_{\text{p}} = 0.804\Omega$.

Discussion for (a)

R_{p} is, as predicted, less than the smallest individual resistance.

Strategy and Solution for (b)

The total current can be found from Ohm's law, substituting R_{p} for the total resistance. This gives

$$I = \frac{V}{R_{\text{p}}} = \frac{12.0\text{V}}{0.8041\Omega} = 14.92\text{A}. \tag{3.2.26}$$

Discussion for (b)

Current i for each device is much larger than for the same devices connected in series (see the previous example). A circuit with parallel connections has a smaller total resistance than the resistors connected in series.

Strategy and Solution for (c)

The individual currents are easily calculated from Ohm's law, since each resistor gets the full voltage. Thus,

$$I_1 = \frac{V}{R_1} = \frac{12.0\text{V}}{1.00\Omega} = 12.0\text{A}. \tag{3.2.27}$$

Similarly,

$$I_2 = \frac{V}{R_2} = \frac{12.0\text{V}}{6.00\Omega} = 2.00\text{A} \tag{3.2.28}$$

and

$$I_3 = \frac{V}{R_3} = \frac{12.0\text{V}}{13.0\Omega} = 0.92\text{A}. \tag{3.2.29}$$

Discussion for (c)

The total current is the sum of the individual currents:

$$I_1 + I_2 + I_3 = 14.92\text{A}. \tag{3.2.30}$$

This is consistent with conservation of charge.

Strategy and Solution for (d)

The power dissipated by each resistor can be found using any of the equations relating power to current, voltage, and resistance, since all three are known. Let us use $P = \frac{V^2}{R}$, since each resistor gets full voltage. Thus,

$$P_1 = \frac{V^2}{R_1} = \frac{(12.0\text{V})^2}{1.00\Omega} = 144\text{W}. \tag{3.2.31}$$

Similarly,

$$P_2 = \frac{V^2}{R_2} = \frac{(12.0\text{V})^2}{6.00\Omega} = 24.0\text{W} \tag{3.2.32}$$

and

$$P_3 = \frac{V^2}{R_3} = \frac{(12.0\text{V})^2}{13.0\Omega} = 11.1\text{W}. \tag{3.2.33}$$

Discussion for (d)

The power dissipated by each resistor is considerably higher in parallel than when connected in series to the same voltage source.

Strategy and Solution for (e)

The total power can also be calculated in several ways. Choosing $P = IV$, and entering the total current, yields

$$P = IV = (14.92\text{A})(12.0\text{V}) = 179\text{W}. \tag{3.2.34}$$

Discussion for (e)

Total power dissipated by the resistors is also 179 W:

$$P_1 + P_2 + P_3 = 144\text{W} + 24.0\text{W} + 11.1\text{W} = 179\text{W}. \tag{3.2.35}$$

This is consistent with the law of conservation of energy.

Overall Discussion

Note that both the currents and powers in parallel connections are greater than for the same devices in series.

MAJOR FEATURES OF RESISTORS IN PARALLEL

1. Parallel resistance is found from $\frac{1}{R_\text{p}} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} + \ldots$, and it is smaller than any individual resistance in the combination.
2. Each resistor in parallel has the same full voltage of the source applied to it. (Power distribution systems most often use parallel connections to supply the myriad devices served with the same voltage and to allow them to operate independently.)
3. Parallel resistors do not each get the total current; they divide it.

Combinations of Series and Parallel

More complex connections of resistors are sometimes just combinations of series and parallel. These are commonly encountered, especially when wire resistance is considered. In that case, wire resistance is in series with other resistances that are in parallel.

Combinations of series and parallel can be reduced to a single equivalent resistance using the technique illustrated in Figure 3.2.4. Various parts are identified as either series or parallel, reduced to their equivalents, and further reduced until a single resistance is left. The process is more time consuming than difficult.

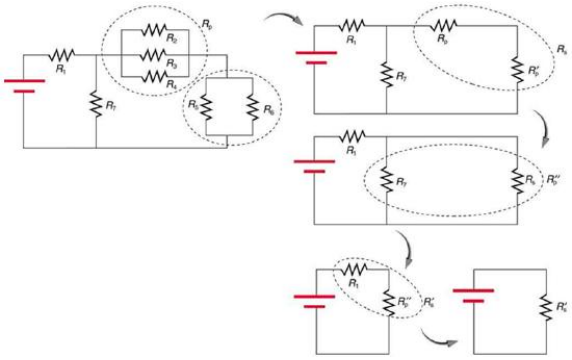


Figure 3.2.4: This combination of seven resistors has both series and parallel parts. Each is identified and reduced to an equivalent resistance, and these are further reduced until a single equivalent resistance is reached.

The simplest combination of series and parallel resistance, shown in Figure 3.2.5, is also the most instructive, since it is found in many applications. For example, R_1 could be the resistance of wires from a car battery to its electrical devices, which are in parallel. R_2 and R_3 could be the starter motor and a passenger compartment light. We have previously assumed that wire resistance is negligible, but, when it is not, it has important effects, as the next example indicates.

Example 3.2.3: Calculating Resistance, IR Drop, Current, and Power Dissipation: Combining Series and Parallel Circuits

Figure 3.2.5 shows the resistors from the previous two examples wired in a different way—a combination of series and parallel. We can consider R_1 to be the resistance of wires leading to R_2 and R_3. (a) Find the total resistance. (b) What is the IR drop in R_1? (c) Find the current I_2 through R_2. (d) What power is dissipated by R_2?

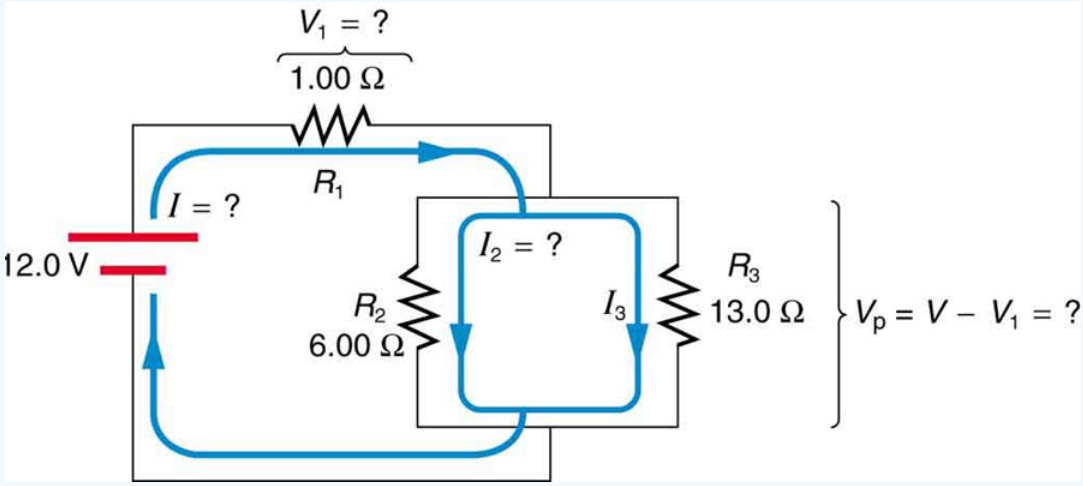


Figure 3.2.5. These three resistors are connected to a voltage source so that R_2 and R_3 are in parallel with one another and that combination is in series with R_1.

Strategy and Solution for (a)

To find the total resistance, we note that R_2 and R_3 are in parallel and their combination R_p is in series with R_1. Thus the total (equivalent) resistance of this combination is

$$R_{tot} = R_1 + R_2. \tag{3.2.36}$$

First, we find R_p using the equation for resistors in parallel and entering known values:

$$\frac{1}{R_p} = \frac{1}{R_2} + \frac{1}{6.00\,\Omega} + \frac{1}{13.0\,\Omega} = \frac{0.2436}{\Omega}. \tag{3.2.37}$$

Inverting gives

$$R_p = \frac{1}{0.2436}\Omega = 4.11\,\Omega. \tag{3.2.38}$$

So the total resistance is

$$R_{tot} = R_1 + R_p = 1.00\Omega + 4.11\Omega = 5.11\,\Omega. \tag{3.2.39}$$

Discussion for (a)

The total resistance of this combination is intermediate between the pure series and pure parallel values ($20.0\,\Omega$ and $0.804\,\Omega$, respectively) found for the same resistors in the two previous examples.

Strategy and Solution for (b)

To find the IR drop in R_1, we note that the full current I flows through R_1. Thus, its IR drop is

$$V_1 = IR_1. \tag{3.2.40}$$

We must find I before we can calculate V_1. The total current I is found using Ohm's law for the circuit. That is,

$$I = \frac{V}{R_{tot}} = \frac{12.0\,V}{5.11\,\Omega} = 2.35\,A. \tag{3.2.41}$$

Entering this into the expression above, we get

$$V_1 = IR_1 = (2.35\,A)(1.00\,\Omega) = 2.35\,V. \tag{3.2.42}$$

Discussion for (b)

The voltage applied to R_2 and R_3 is less than the total voltage by an amount V_1. When wire resistance is large, it can significantly affect the operation of the devices represented by R_2 and R_3.

Strategy and Solution for (c)

To find the current through R_2, we must first find the voltage applied to it. We call this voltage V_p, because it is applied to a parallel combination of resistors. The voltage applied to both R_2 and R_3 is reduced by the amount V_1, and so it is

$$V_p = V - V_1 = 12.0\,V - 2.35\,V = 9.65\,V. \tag{3.2.43}$$

Now the current I_2 through resistance R_2 is found using Ohm's law:

$$I_2 = \frac{V_p}{R_2} = \frac{9.65\,V}{6.00\,\Omega} = 1.61\,A. \tag{3.2.44}$$

Discussion for (c)

The current is less than the 2.00 A that flowed through R_2 when it was connected in parallel to the battery in the previous parallel circuit example.

Strategy and Solution for (d)

The power dissipated by R_2 is given by

$$P_2 = (I_2)^2 R_2 = (161\,A)^2(6.00\,\Omega) = 15.5\,W. \tag{3.2.45}$$

Discussion for (d)

The power is less than the 24.0 W this resistor dissipated when connected in parallel to the 12.0-V source.

Practical Implications

One implication of this last example is that resistance in wires reduces the current and power delivered to a resistor. If wire resistance is relatively large, as in a worn (or a very long) extension cord, then this loss can be significant. If a large current is drawn, the IR drop in the wires can also be significant.

For example, when you are rummaging in the refrigerator and the motor comes on, the refrigerator light dims momentarily. Similarly, you can see the passenger compartment light dim when you start the engine of your car (although this may be due to resistance inside the battery itself).

What is happening in these high-current situations is illustrated in Figure 3.2.6. The device represented by R_3 has a very low resistance, and so when it is switched on, a large current flows. This increased current causes a larger IR drop in the wires represented by R_1 reducing the voltage across the light bulb (which is R_2), which then dims noticeably.

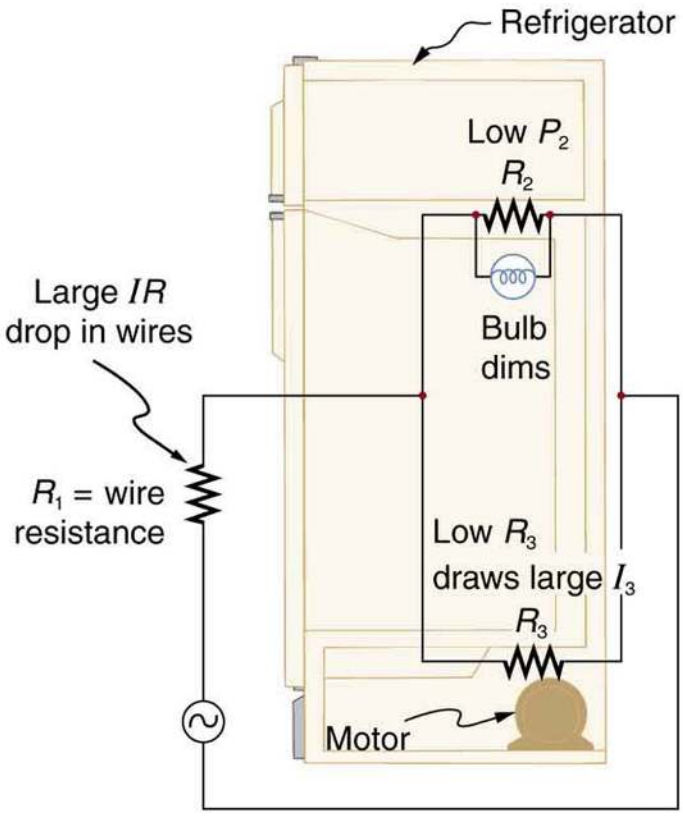


Figure 3.2.6: Why do lights dim when a large appliance is switched on? The answer is that the large current the appliance motor draws causes a significant IR drop in the wires and reduces the voltage across the light.

Exercise 3.2.1

Can any arbitrary combination of resistors be broken down into series and parallel combinations? See if you can draw a circuit diagram of resistors that cannot be broken down into combinations of series and parallel.

Answer

No, there are many ways to connect resistors that are not combinations of series and parallel, including loops and junctions. In such cases Kirchhoff's rules, to be introduced in Kirchhoff's Rules, will allow you to analyze the circuit.

Problem Solving Strategies For Series and Parallel Resistors

1. Draw a clear circuit diagram, labeling all resistors and voltage sources. This step includes a list of the knowns for the problem, since they are labeled in your circuit diagram.
2. Identify exactly what needs to be determined in the problem (identify the unknowns). A written list is useful.
3. Determine whether resistors are in series, parallel, or a combination of both series and parallel. Examine the circuit diagram to make this assessment. Resistors are in series if the same current must pass sequentially through them.
4. Use the appropriate list of major features for series or parallel connections to solve for the unknowns. There is one list for series and another for parallel. If your problem has a combination of series and parallel, reduce it in steps by considering individual groups of series or parallel connections, as done in this module and the examples. Special note: When finding R_p, the reciprocal must be taken with care.
5. Check to see whether the answers are reasonable and consistent. Units and numerical results must be reasonable. Total series resistance should be greater, whereas total parallel resistance should be smaller, for example. Power should be greater for the same devices in parallel compared with series, and so on.

Summary

- The total resistance of an electrical circuit with resistors wired in a series is the sum of the individual resistances: $R_s = R_1 + R_2 + R_3 + ...$
- Each resistor in a series circuit has the same amount of current flowing through it.
- The voltage drop, or power dissipation, across each individual resistor in a series is different, and their combined total adds up to the power source input.
- The total resistance of an electrical circuit with resistors wired in parallel is less than the lowest resistance of any of the components and can be determined using the formula: $\frac{1}{R_p} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} + ...$
- Each resistor in a parallel circuit has the same full voltage of the source applied to it.
- The current flowing through each resistor in a parallel circuit is different, depending on the resistance.
- If a more complex connection of resistors is a combination of series and parallel, it can be reduced to a single equivalent resistance by identifying its various parts as series or parallel, reducing each to its equivalent, and continuing until a single resistance is eventually reached.

Glossary

series
a sequence of resistors or other components wired into a circuit one after the other

resistor
a component that provides resistance to the current flowing through an electrical circuit

resistance
causing a loss of electrical power in a circuit

Ohm's law
the relationship between current, voltage, and resistance within an electrical circuit: V=IR

voltage
the electrical potential energy per unit charge; electric pressure created by a power source, such as a battery

voltage drop
the loss of electrical power as a current travels through a resistor, wire or other component

current
the flow of charge through an electric circuit past a given point of measurement

Joule's law
tthe relationship between potential electrical power, voltage, and resistance in an electrical circuit, given by: $P_e = IV$

parallel
the wiring of resistors or other components in an electrical circuit such that each component receives an equal voltage from the power source; often pictured in a ladder-shaped diagram, with each component on a rung of the ladder

---

This page titled 3.2: Resistors in Series and Parallel is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **21.1: Resistors in Series and Parallel** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/college-physics.

3.3: Capacitors and Dielectrics

Learning Objectives

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

- Describe the action of a capacitor and define capacitance.
- Explain parallel plate capacitors and their capacitances.
- Discuss the process of increasing the capacitance of a dielectric.
- Determine capacitance given charge and voltage.

A **capacitor** is a device used to store electric charge. Capacitors have applications ranging from filtering static out of radio reception to energy storage in heart defibrillators. Typically, commercial capacitors have two conducting parts close to one another, but not touching, such as those in Figure 3.3.1. (Most of the time an insulator is used between the two plates to provide separation —see the discussion on dielectrics below.)

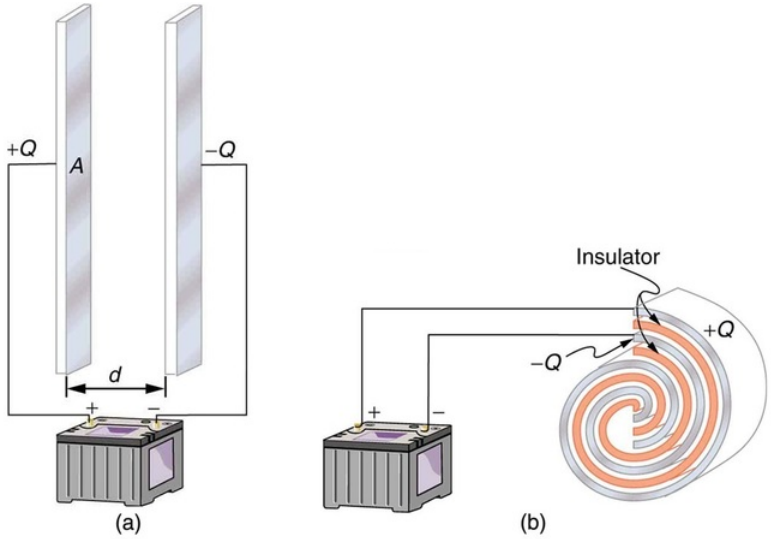


Figure 3.3.1: Both capacitors shown here were initially uncharged before being connected to a battery. They now have separated charges of $+Q$ and $-Q$ on their two halves. (a) A parallel plate capacitor. (b) A rolled capacitor with an insulating material between its two conducting sheets.

Definition: CAPACITOR

A capacitor is a device used to store electric charge.

When battery terminals are connected to an initially uncharged capacitor, equal amounts of positive and negative charge, $+Q$ and $-Q$, are separated into its two plates. The capacitor remains neutral overall, but we refer to it as storing a charge Q in this circumstance. The amount of charge Q a *capacitor* can store depends on two major factors—the voltage applied and the capacitor's physical characteristics, such as its size.

A system composed of two identical, parallel conducting plates separated by a distance, as in Figure 3.3.2, is called a **parallel plate capacitor**. It is easy to see the relationship between the voltage and the stored charge for a parallel plate capacitor, as shown in Figure 3.3.2. Each electric field line starts on an individual positive charge and ends on a negative one, so that there will be more field lines if there is more charge. (Drawing a single field line per charge is a convenience, only. We can draw many field lines for each charge, but the total number is proportional to the number of charges.) The electric field strength is, thus, directly proportional to Q.

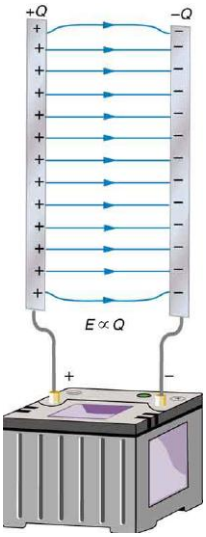


Figure 3.3.2: Electric field lines in this parallel plate capacitor, as always, start on positive charges and end on negative charges. Since the electric field strength is proportional to the density of field lines, it is also proportional to the amount of charge on the capacitor.

The field is proportional to the charge:

$$E \propto Q, \tag{3.3.1}$$

where the symbol $\propto$ means "proportional to." From the discussion in Electric Potential in a Uniform Electric Field, we know that the voltage across parallel plates is $V = Ed$. Thus,

$$V \propto E. \tag{3.3.2}$$

It follows, then, that $V \propto Q$, and conversely,

$$Q \propto V. \tag{3.3.3}$$

This is true in general: The greater the voltage applied to any capacitor, the greater the charge stored in it.

Different capacitors will store different amounts of charge for the same applied voltage, depending on their physical characteristics. We define their **capacitance** C to be such that the charge Q stored in a capacitor is proportional to C. The charge stored in a capacitor is given by

$$Q = CV. \tag{3.3.4}$$

This equation expresses the two major factors affecting the amount of charge stored. Those factors are the physical characteristics of the capacitor, C, and the voltage, V. Rearranging the equation, we define the *capacitance C of a capacitor.*

Definition: CAPACITANCE

Capacitance C is the amount of charge stored per volt, or

$$C = \frac{Q}{V}. \tag{3.3.5}$$

The unit of capacitance is the farad (F), named for Michael Faraday (1791–1867), an English scientist who contributed to the fields of electromagnetism and electrochemistry. Since capacitance is charge per unit voltage, we see that a farad is a coulomb per volt, or

$$1\ \mathrm{F} = \frac{1\ \mathrm{C}}{1\ \mathrm{V}}. \tag{3.3.6}$$

A 1-farad capacitor would be able to store 1 coulomb (a very large amount of charge) with the application of only 1 volt. One farad is, thus, a very large capacitance. Typical capacitors range from fractions of a picofarad $(1\ \mathrm{pF} = 10^{-12}\mathrm{F})$ to millifarads $(1\ \mathrm{mF} = 10^{-3}\mathrm{F})$.

Figure 3.3.3 shows some common capacitors. Capacitors are primarily made of ceramic, glass, or plastic, depending upon purpose and size. Insulating materials, called dielectrics, are commonly used in their construction, as discussed below.

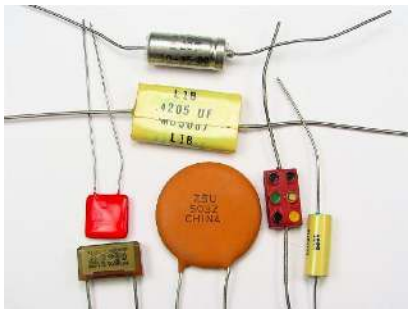


Figure 3.3.3: Some typical capacitors. Size and value of capacitance are not necessarily related. (credit: Windell Oskay)

Parallel Plate Capacitor

The parallel plate capacitor shown in Figure 3.3.4 has two identical conducting plates, each having a surface area A, separated by a distance d (with no material between the plates). When a voltage V is applied to the capacitor, it stores a charge Q, as shown. We can see how its capacitance depends on A and d by considering the characteristics of the Coulomb force. We know that like charges repel, unlike charges attract, and the force between charges decreases with distance. So it seems quite reasonable that the bigger the plates are, the more charge they can store—because the charges can spread out more. Thus C should be greater for larger A. Similarly, the closer the plates are together, the greater the attraction of the opposite charges on them. So C should be greater for smaller d.

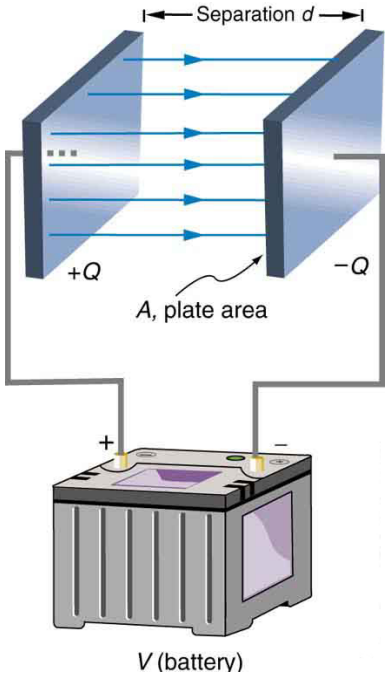


Figure 3.3.4: Parallel plate capacitor with plates separated by a distance d. Each plate has an area A.

It can be shown that for a parallel plate capacitor there are only two factors (A and d) that affect its capacitance C. The capacitance of a parallel plate capacitor in equation form can be defined:

Definition: CAPACITANCE OF A PARALLEL PLATE CAPACITOR

The capacitance of a parallel plate capacitor in equation form is given by

$$C = \varepsilon_0 \frac{A}{d}. \tag{3.3.7}$$

A is the area of one plate in square meters, and d is the distance between the plates in meters. The constant ε_0 is the permittivity of free space; its numerical value in SI units is $\varepsilon_0 = 8.85 \times 10^{-12}\text{F/m}$. The units of F/m are equivalent to $\text{C}^2/\text{N} \cdot \text{m}^2$. The small numerical value of ε_0 is related to the large size of the farad. A parallel plate capacitor must have a large area to have a capacitance approaching a farad. (Note that the above equation is valid when the parallel plates are separated by air or free space. When another material is placed between the plates, the equation is modified, as discussed below.)

Example 3.3.1: Capacitance and Charge Stored in a Parallel Plate Capacitor

a. What is the capacitance of a parallel plate capacitor with metal plates, each of area 1.00m^2, separated by 1.00 mm?
b. What charge is stored in this capacitor if a voltage of $3.00 \times 10^3\text{V}$ is applied to it?

Strategy

Finding the capacitance C is a straightforward application of the equation $C = \varepsilon_0 A/d$. Once C is found, the charge stored can be found using the equation $Q = CV$.

Solution for (a)

Entering the given values into the equation for the capacitance of a parallel plate capacitor yields

$$\begin{aligned} C &= \varepsilon \frac{A}{d} \\ &= (8.85 \times 10^{-12} \frac{\text{F}}{\text{m}}) \frac{1.00\text{m}^2}{1.00 \times 10^{-3}\text{m}} \\ &= 8.85 \times 10^{-9}\text{F} = 8.85\text{nF}. \end{aligned}$$

Discussion for (a)

This small value for the capacitance indicates how difficult it is to make a device with a large capacitance. Special techniques help, such as using very large area thin foils placed close together.

Solution for (b)

The charge stored in any capacitor is given by the equation $Q = CV$. Entering the known values into this equation gives

$$\begin{aligned} Q &= CV \\ &= (8.85 \times 10^{-9}\text{F})(3.00 \times 10^3\text{V}) \\ &= 26.6\mu\text{C}. \end{aligned}$$

Discussion for (b)

This charge is only slightly greater than those found in typical static electricity. Since air breaks down at about $3.00 \times 10^6\text{V/m}$, more charge cannot be stored on this capacitor by increasing the voltage.

Membrane Potential

Another interesting biological example dealing with electric potential is found in the cell's plasma membrane. The membrane sets a cell off from its surroundings and also allows ions to selectively pass in and out of the cell. There is a potential difference across the membrane of about -70mV. This is due to the mainly negatively charged ions in the cell and the predominance of positively charged sodium (Na^+) ions outside. Things change when a nerve cell is stimulated. Na^+ ions are allowed to pass through the membrane into the cell, producing a positive membrane potential—the nerve signal. The cell membrane is about 7 to 10 nm thick. An approximate value of the electric field across it is given by

$$\begin{aligned} E &= \frac{V}{d} \\ &= \frac{-70 \times 10^{-3}\text{V}}{8 \times 10^{-9}\text{m}} \\ &= -9 \times 10^6\text{V/m}. \end{aligned}$$

This electric field is enough to cause a breakdown in air.

Dielectric

The previous example highlights the difficulty of storing a large amount of charge in capacitors. If d is made smaller to produce a larger capacitance, then the maximum voltage must be reduced proportionally to avoid breakdown (since $E = V/d$). An important solution to this difficulty is to put an insulating material, called a **dielectric**, between the plates of a capacitor and allow d to be as small as possible. Not only does the smaller d make the capacitance greater, but many insulators can withstand greater electric fields than air before breaking down.

There is another benefit to using a dielectric in a capacitor. Depending on the material used, the capacitance is greater than that given by the equation $C = \varepsilon \frac{A}{d}$ by a factor κ, called the *dielectric constant*. A parallel plate capacitor with a dielectric between its plates has a capacitance given by

$$C = \kappa \varepsilon_0 \frac{A}{d} (\text{parallel plate capacitor with dielectric}). \tag{3.3.8}$$

Values of the dielectric constant κ for various materials are given in Table 3.3.1. Note that κ for vacuum is exactly 1, and so the above equation is valid in that case, too. If a dielectric is used, perhaps by placing Teflon between the plates of the capacitor in Example 3.3.1, then the capacitance is greater by the factor κ, which for Teflon is 2.1.

TAKE-HOME EXPERIMENT: BUILDING A CAPACITOR

How large a capacitor can you make using a chewing gum wrapper? The plates will be the aluminum foil, and the separation (dielectric) in between will be the paper.

Table 3.3.1: Dielectric Constants and Dielectric Strengths for Various Materials at 20ºC

Material	Dielectric constant (κ)	Dielectric Strength (V/m)
Vacuum	1.00000	–
Air	1.00059	3×10^6
Bakelite	4.9	24×10^6
Fused quartz	3.78	8×10^6
Neoprene rubber	6.7	12×10^6
Nylon	3.4	14×10^6
Paper	3.7	16×10^6
Polystyrene	2.56	24×10^6
Pyrex glass	5.6	14×10^6
Silicon oil	2.5	15×10^6
Strontium titanate	233	8×10^6
Teflon	2.1	60×10^6
Water	80	–

Note also that the dielectric constant for air is very close to 1, so that air-filled capacitors act much like those with vacuum between their plates *except* that the air can become conductive if the electric field strength becomes too great. (Recall that $E = V/d$ for a parallel plate capacitor.) Also shown in Table 3.3.1 are maximum electric field strengths in V/m, called **dielectric strengths**, for several materials. These are the fields above which the material begins to break down and conduct. The dielectric strength imposes a limit on the voltage that can be applied for a given plate separation. For instance, in Example, the separation is 1.00 mm, and so the voltage limit for air is

$$\begin{aligned} V &= E \cdot V \\ &= (3 \times 10^6 \text{V/m})(1.00 \times 10^{-3} \text{m}) \\ &= 3000 \text{V}. \end{aligned}$$

However, the limit for a 1.00 mm separation filled with Teflon is 60,000 V, since the dielectric strength of Teflon is $60 \times 10^6 \text{V/m}$ V/m. So the same capacitor filled with Teflon has a greater capacitance and can be subjected to a much greater voltage. Using the capacitance we calculated in the above example for the air-filled parallel plate capacitor, we find that the Teflon-filled capacitor can store a maximum charge of

$$\begin{aligned} Q &= CV \\ &= \kappa C_{air} V \\ &= (2.1)(8.85 \text{nF})(6.0 \times 10^4 \text{V}) \\ &= 1.1 \text{mC}. \end{aligned}$$

This is 42 times the charge of the same air-filled capacitor.

DIELECTRIC STRENGTH

The maximum electric field strength above which an insulating material begins to break down and conduct is called its dielectric strength.

Microscopically, how does a dielectric increase capacitance? Polarization of the insulator is responsible. The more easily it is polarized, the greater its dielectric constant κ. Water, for example, is a **polar molecule** because one end of the molecule has a slight positive charge and the other end has a slight negative charge. The polarity of water causes it to have a relatively large dielectric constant of 80. The effect of polarization can be best explained in terms of the characteristics of the Coulomb force. Figure 3.3.5 shows the separation of charge schematically in the molecules of a dielectric material placed between the charged plates of a capacitor. The Coulomb force between the closest ends of the molecules and the charge on the plates is attractive and very strong, since they are very close together. This attracts more charge onto the plates than if the space were empty and the opposite charges were a distance d away.

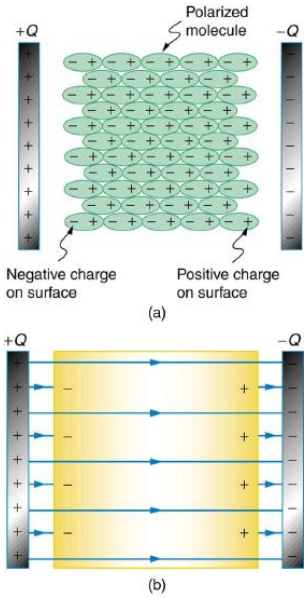


Figure 3.3.5: (a) The molecules in the insulating material between the plates of a capacitor are polarized by the charged plates. This produces a layer of opposite charge on the surface of the dielectric that attracts more charge onto the plate, increasing its capacitance. (b) The dielectric reduces the electric field strength inside the capacitor, resulting in a smaller voltage between the plates for the same charge. The capacitor stores the same charge for a smaller voltage, implying that it has a larger capacitance because of the dielectric.

Another way to understand how a dielectric increases capacitance is to consider its effect on the electric field inside the capacitor. Figure 3.3.5(b) shows the electric field lines with a dielectric in place. Since the field lines end on charges in the dielectric, there are fewer of them going from one side of the capacitor to the other. So the electric field strength is less than if there were a vacuum

between the plates, even though the same charge is on the plates. The voltage between the plates is $V = Ed$, so it too is reduced by the dielectric. Thus there is a smaller voltage V for the same charge Q; since $C = Q/V$, the capacitance C is greater.

The dielectric constant is generally defined to be $\kappa = E_0/E$, or the ratio of the electric field in a vacuum to that in the dielectric material, and is intimately related to the polarizability of the material.

THINGS GREAT AND SMALL: The Submicroscopic Origin of Polarization

Polarization is a separation of charge within an atom or molecule. As has been noted, the planetary model of the atom pictures it as having a positive nucleus orbited by negative electrons, analogous to the planets orbiting the Sun. Although this model is not completely accurate, it is very helpful in explaining a vast range of phenomena and will be refined elsewhere, such as in the Chapter on Atomic Physics. The submicroscopic origin of polarization can be modeled as shown in Figure 3.3.6.

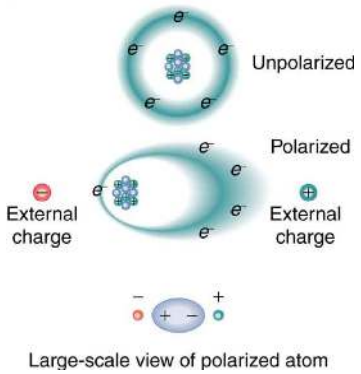


Figure 3.3.6: Artist's conception of a polarized atom. The orbits of electrons around the nucleus are shifted slightly by the external charges (shown exaggerated). The resulting separation of charge within the atom means that it is polarized. Note that the unlike charge is now closer to the external charges, causing the polarization.

We will find in Atomic Physics that the orbits of electrons are more properly viewed as electron clouds with the density of the cloud related to the probability of finding an electron in that location (as opposed to the definite locations and paths of planets in their orbits around the Sun). This cloud is shifted by the Coulomb force so that the atom on average has a separation of charge. Although the atom remains neutral, it can now be the source of a Coulomb force, since a charge brought near the atom will be closer to one type of charge than the other.

Some molecules, such as those of water, have an inherent separation of charge and are thus called polar molecules.Figure 3.3.7 illustrates the separation of charge in a water molecule, which has two hydrogen atoms and one oxygen atom (H_2O). The water molecule is not symmetric—the hydrogen atoms are repelled to one side, giving the molecule a boomerang shape. The electrons in a water molecule are more concentrated around the more highly charged oxygen nucleus than around the hydrogen nuclei. This makes the oxygen end of the molecule slightly negative and leaves the hydrogen ends slightly positive. The inherent separation of charge in polar molecules makes it easier to align them with external fields and charges. Polar molecules therefore exhibit greater polarization effects and have greater dielectric constants. Those who study chemistry will find that the polar nature of water has many effects. For example, water molecules gather ions much more effectively because they have an electric field and a separation of charge to attract charges of both signs. Also, as brought out in the previous chapter, polar water provides a shield or screening of the electric fields in the highly charged molecules of interest in biological systems.

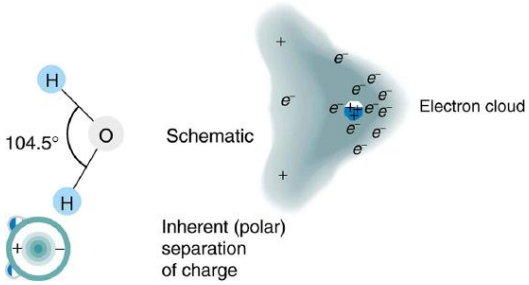


Figure 3.3.7: Artist's conception of a water molecule. There is an inherent separation of charge, and so water is a polar molecule. Electrons in the molecule are attracted to the oxygen nucleus and leave an excess of positive charge near the two hydrogen nuclei. (Note that the schematic on the right is a rough illustration of the distribution of electrons in the water molecule. It does not show the actual numbers of protons and electrons involved in the structure.)

PHET EXPLORATIONS: CAPACITOR LAB

Explore how a capacitor works! Change the size of the plates and add a dielectric to see the effect on capacitance. Change the voltage and see charges built up on the plates. Observe the electric field in the capacitor. Measure the voltage and the electric field.

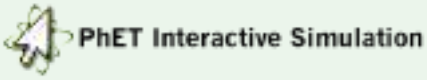


Figure 3.3.8: Capacitor Lab

Summary

- A capacitor is a device used to store charge.
- The amount of charge Q a capacitor can store depends on two major factors—the voltage applied and the capacitor's physical characteristics, such as its size.
- The capacitance C is the amount of charge stored per volt, or $C = \frac{Q}{V}$.
- The capacitance of a parallel plate capacitor is $C = \varepsilon_0 \frac{A}{d}$, when the plates are separated by air or free space. ε_0 is called the permittivity of free space.
- A parallel plate capacitor with a dielectric between its plates has a capacitance given by $C = \kappa \varepsilon_0 \frac{A}{d}$, where κ is the dielectric constant of the material.
- The maximum electric field strength above which an insulating material begins to break down and conduct is called dielectric strength.

Glossary

capacitor
a device that stores electric charge

capacitance
amount of charge stored per unit volt

dielectric
an insulating material

dielectric strength
the maximum electric field above which an insulating material begins to break down and conduct

parallel plate capacitor
two identical conducting plates separated by a distance

polar molecule
a molecule with inherent separation of charge

This page titled 3.3: Capacitors and Dielectrics is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **19.5: Capacitors and Dielectrics** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/college-physics.

3.4: Magnets

Learning Objectives

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

- Describe the difference between the north and south poles of a magnet.
- Describe how magnetic poles interact with each other.

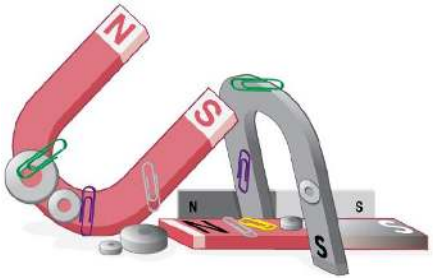


Figure 3.4.1: Magnets come in various shapes, sizes, and strengths. All have both a north pole and a south pole. There is never an isolated pole (a monopole).

All magnets attract iron, such as that in a refrigerator door. However, magnets may attract or repel other magnets. Experimentation shows that all magnets have two poles. If freely suspended, one pole will point toward the north. The two poles are thus named the north magnetic pole and the south magnetic pole (or more properly, north-seeking and south-seeking poles, for the attractions in those directions).

UNIVERSAL CHARACTERISTICS OF MAGNETS AND MAGNET POLES

It is a universal characteristic of all magnets that *like poles repel and unlike poles attract*. (Note the similarity with electrostatics: unlike charges attract and like charges repel.)

Further experimentation shows that it is *impossible to separate north and south poles* in the manner that + and − charges can be separated.

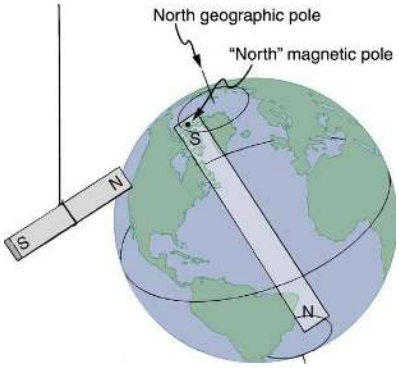


Figure 3.4.2: One end of a bar magnet is suspended from a thread that points toward north. The magnet's two poles are labeled N and S for north-seeking and south-seeking poles, respectively.

MISCONCEPTION ALERT: EARTH'S GEOGRAPHIC NORTH POLE HIDES AN S

The Earth acts like a very large bar magnet with its south-seeking pole near the geographic North Pole. That is why the north pole of your compass is attracted toward the geographic north pole of the Earth—because the magnetic pole that is near the geographic North Pole is actually a south magnetic pole! Confusion arises because the geographic term "North Pole" has come to be used (incorrectly) for the magnetic pole that is near the North Pole. Thus, "North magnetic pole" is actually a misnomer—it should be called the South magnetic pole.

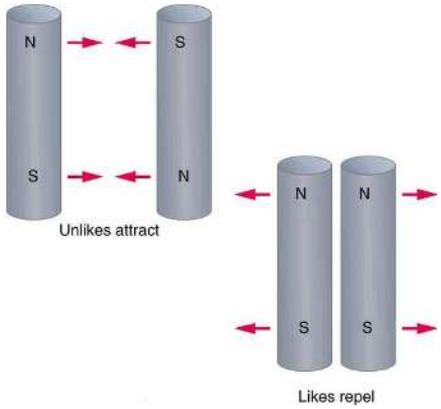


Figure 3.4.3: Unlike poles attract, whereas like poles repel.

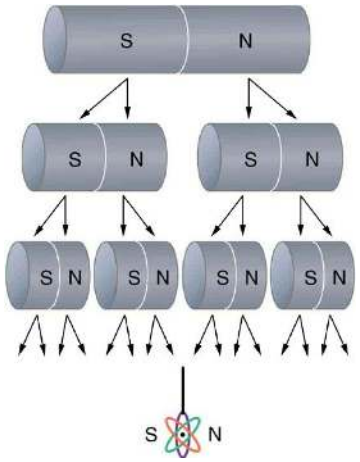


Figure 3.4.4: North and south poles always occur in pairs. Attempts to separate them result in more pairs of poles. If we continue to split the magnet, we will eventually get down to an iron atom with a north pole and a south pole—these, too, cannot be separated.

The fact that magnetic poles always occur in pairs of north and south is true from the very large scale—for example, sunspots always occur in pairs that are north and south magnetic poles—all the way down to the very small scale. Magnetic atoms have both a north pole and a south pole, as do many types of subatomic particles, such as electrons, protons, and neutrons.

MAKING CONNECTIONS: TAKE-HOME EXPERIMENT -- REFRIGERATOR MAGNETS

We know that like magnetic poles repel and unlike poles attract. See if you can show this for two refrigerator magnets. Will the magnets stick if you turn them over? Why do they stick to the door anyway? What can you say about the magnetic properties of the door next to the magnet? Do refrigerator magnets stick to metal or plastic spoons? Do they stick to all types of metal?

Summary

- Magnetism is a subject that includes the properties of magnets, the effect of the magnetic force on moving charges and currents, and the creation of magnetic fields by currents.
- There are two types of magnetic poles, called the north magnetic pole and south magnetic pole.
- North magnetic poles are those that are attracted toward the Earth's geographic north pole.
- Like poles repel and unlike poles attract.
- Magnetic poles always occur in pairs of north and south—it is not possible to isolate north and south poles.

Glossary

north magnetic pole

the end or the side of a magnet that is attracted toward Earth's geographic north pole

south magnetic pole

the end or the side of a magnet that is attracted toward Earth's geographic south pole

This page titled 3.4: Magnets is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **22.1: Magnets** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/college-physics.

3.5: Ferromagnets and Electromagnets

Learning Objectives

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

- Define ferromagnet.
- Describe the role of magnetic domains in magnetization.
- Explain the significance of the Curie temperature.
- Describe the relationship between electricity and magnetism.

Ferromagnets

Only certain materials, such as iron, cobalt, nickel, and gadolinium, exhibit strong magnetic effects. Such materials are called ferromagnetic, after the Latin word for iron, *ferrum*. A group of materials made from the alloys of the rare earth elements are also used as strong and permanent magnets; a popular one is neodymium. Other materials exhibit weak magnetic effects, which are detectable only with sensitive instruments. Not only do ferromagnetic materials respond strongly to magnets (the way iron is attracted to magnets), they can also be magnetized themselves—that is, they can be induced to be magnetic or made into permanent magnets.

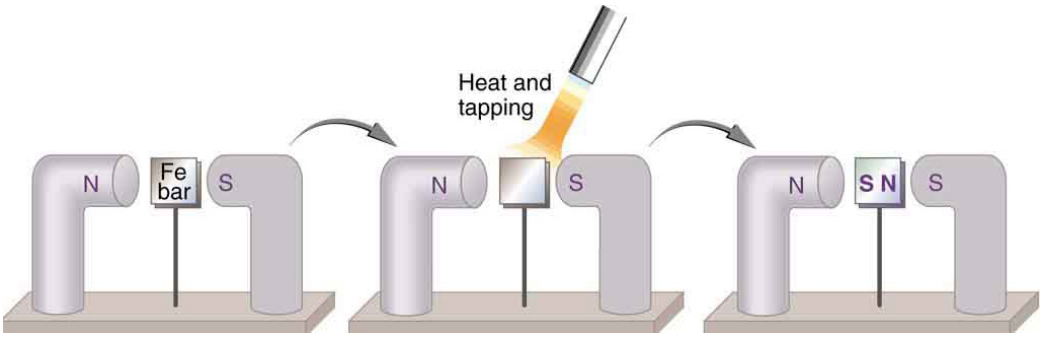


Figure 3.5.1: An unmagnetized piece of iron is placed between two magnets, heated, and then cooled, or simply tapped when cold. The iron becomes a permanent magnet with the poles aligned as shown: its south pole is adjacent to the north pole of the original magnet, and its north pole is adjacent to the south pole of the original magnet. Note that there are attractive forces between the magnets.

When a magnet is brought near a previously unmagnetized ferromagnetic material, it causes local magnetization of the material with unlike poles closest, as in Figure 3.5.1. (This results in the attraction of the previously unmagnetized material to the magnet.) What happens on a microscopic scale is illustrated in Figure 3.5.2. The regions within the material called domains act like small bar magnets. Within domains, the poles of individual atoms are aligned. Each atom acts like a tiny bar magnet. Domains are small and randomly oriented in an unmagnetized ferromagnetic object. In response to an external magnetic field, the domains may grow to millimeter size, aligning themselves as shown in Figure 2b. This induced magnetization can be made permanent if the material is heated and then cooled, or simply tapped in the presence of other magnets.

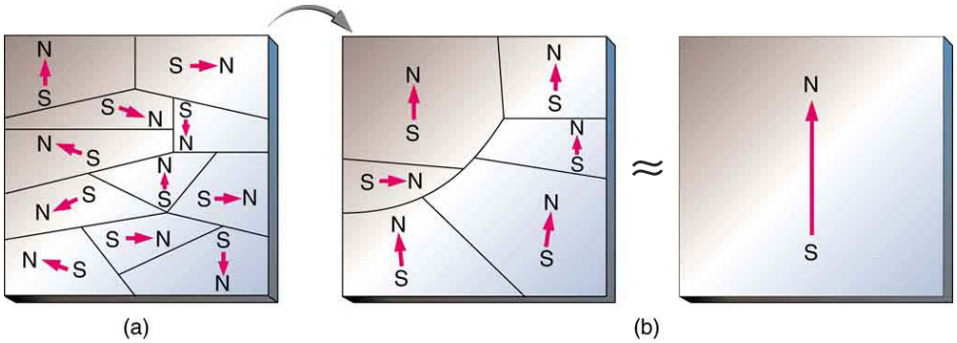


Figure 3.5.2: (a) An unmagnetized piece of iron (or other ferromagnetic material) has randomly oriented domains. (b) When magnetized by an external field, the domains show greater alignment, and some grow at the expense of others. Individual atoms are aligned within domains; each atom acts like a tiny bar magnet.

Conversely, a permanent magnet can be demagnetized by hard blows or by heating it in the absence of another magnet. Increased thermal motion at higher temperature can disrupt and randomize the orientation and the size of the domains. There is a well-defined temperature for ferromagnetic materials, which is called the Curie temperature, above which they cannot be magnetized. The Curie temperature for iron is 1043 K $(770^{\circ}C)$, which is well above room temperature. There are several elements and alloys that have Curie temperatures much lower than room temperature and are ferromagnetic only below those temperatures.

Electromagnets

Early in the 19th century, it was discovered that electrical currents cause magnetic effects. The first significant observation was by the Danish scientist Hans Christian Oersted (1777–1851), who found that a compass needle was deflected by a current-carrying wire. This was the first significant evidence that the movement of charges had any connection with magnets. Electromagnetism is the use of electric current to make magnets. These temporarily induced magnets are called electromagnets. Electromagnets are employed for everything from a wrecking yard crane that lifts scrapped cars to controlling the beam of a 90-km-circumference particle accelerator to the magnets in medical imaging machines (Figure 3.5.3).



Figure 3.5.3: Instrument for magnetic resonance imaging (MRI). The device uses a superconducting cylindrical coil for the main magnetic field. The patient goes into this “tunnel” on the gurney. (credit: Bill McChesney, Flickr)

Figure 3.5.4 shows that the response of iron filings to a current-carrying coil and to a permanent bar magnet. The patterns are similar. In fact, electromagnets and ferromagnets have the same basic characteristics—for example, they have north and south poles that cannot be separated and for which like poles repel and unlike poles attract.

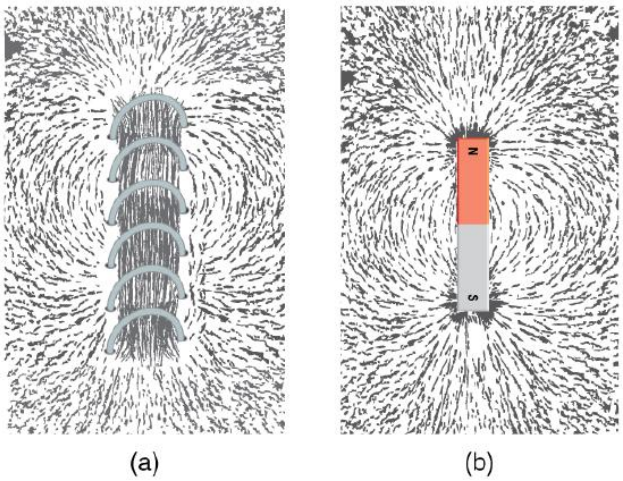


Figure 3.5.4: Iron filings near (a) a current-carrying coil and (b) a magnet act like tiny compass needles, showing the shape of their fields. Their response to a current-carrying coil and a permanent magnet is seen to be very similar, especially near the ends of the coil and the magnet.

Combining a ferromagnet with an electromagnet can produce particularly strong magnetic effects (Figure 3.5.5). Whenever strong magnetic effects are needed, such as lifting scrap metal, or in particle accelerators, electromagnets are enhanced by ferromagnetic materials. Limits to how strong the magnets can be made are imposed by coil resistance (it will overheat and melt at sufficiently high current), and so superconducting magnets may be employed. These are still limited, because superconducting properties are destroyed by too great a magnetic field.

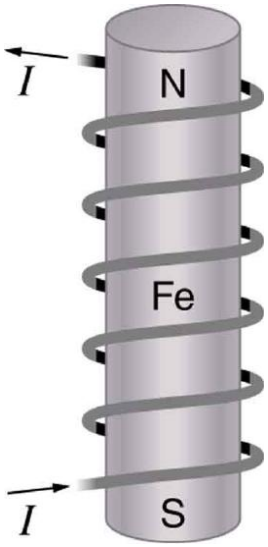


Figure 3.5.5: An electromagnet with a ferromagnetic core can produce very strong magnetic effects. Alignment of domains in the core produces a magnet, the poles of which are aligned with the electromagnet.

Figure 3.5.6 shows a few uses of combinations of electromagnets and ferromagnets. Ferromagnetic materials can act as memory devices, because the orientation of the magnetic fields of small domains can be reversed or erased. Magnetic information storage on videotapes and computer hard drives are among the most common applications. This property is vital in our digital world.

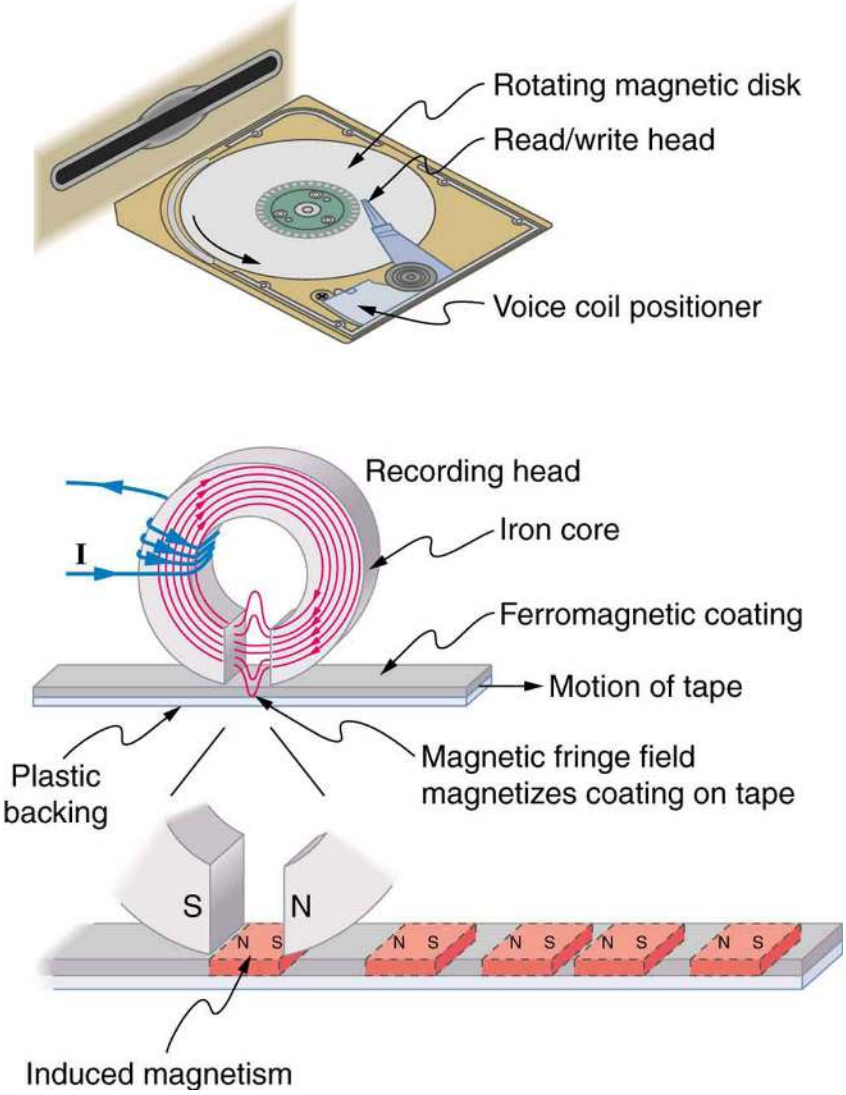


Figure 3.5.6: An electromagnet induces regions of permanent magnetism on a floppy disk coated with a ferromagnetic material. The information stored here is digital (a region is either magnetic or not); in other applications, it can be analog (with a varying strength), such as on audiotapes.

Current: The Source of All Magnetism

An electromagnet creates magnetism with an electric current. In later sections we explore this more quantitatively, finding the strength and direction of magnetic fields created by various currents. But what about ferromagnets? Figure 3.5.7 shows models of how electric currents create magnetism at the submicroscopic level. (Note that we cannot directly observe the paths of individual electrons about atoms, and so a model or visual image, consistent with all direct observations, is made. We can directly observe the electron's orbital angular momentum, its spin momentum, and subsequent magnetic moments, all of which are explained with electric-current-creating subatomic magnetism.) Currents, including those associated with other submicroscopic particles like protons, allow us to explain ferromagnetism and all other magnetic effects. Ferromagnetism, for example, results from an internal cooperative alignment of electron spins, possible in some materials but not in others.

Crucial to the statement that electric current is the source of all magnetism is the fact that it is impossible to separate north and south magnetic poles. (This is far different from the case of positive and negative charges, which are easily separated.) A current loop always produces a magnetic dipole—that is, a magnetic field that acts like a north pole and south pole pair. Since isolated north and south magnetic poles, called magnetic monopoles, are not observed, currents are used to explain all magnetic effects. If magnetic monopoles did exist, then we would have to modify this underlying connection that all magnetism is due to electrical current. There is no known reason that magnetic monopoles should not exist—they are simply never observed—and so searches at

the subnuclear level continue. If they do *not* exist, we would like to find out why not. If they *do* exist, we would like to see evidence of them.

ELECTRIC CURRENTS AND MAGNETISM

Electric current is the source of all magnetism.

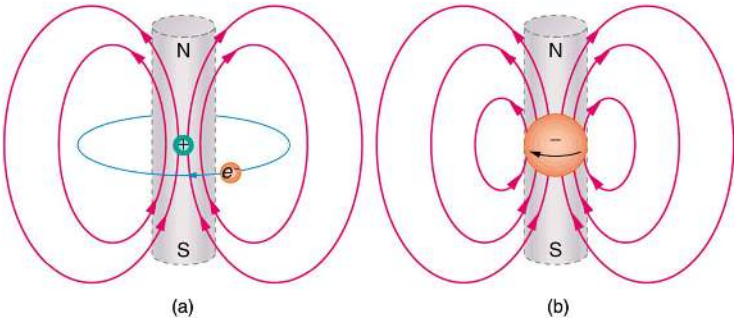


Figure 3.5.7: (a) In the planetary model of the atom, an electron orbits a nucleus, forming a closed-current loop and producing a magnetic field with a north pole and a south pole. (b) Electrons have spin and can be crudely pictured as rotating charge, forming a current that produces a magnetic field with a north pole and a south pole. Neither the planetary model nor the image of a spinning electron is completely consistent with modern physics. However, they do provide a useful way of understanding phenomena.

PHET EXPLORATIONS: MAGNETS AND ELECTROMAGNETS

Explore the interactions between a compass and bar magnet. Discover how you can use a battery and wire to make a magnet! Can you make it a stronger magnet? Can you make the magnetic field reverse?

Summary

- Magnetic poles always occur in pairs of north and south—it is not possible to isolate north and south poles.
- All magnetism is created by electric current.
- Ferromagnetic materials, such as iron, are those that exhibit strong magnetic effects.
- The atoms in ferromagnetic materials act like small magnets (due to currents within the atoms) and can be aligned, usually in millimeter-sized regions called domains.
- Domains can grow and align on a larger scale, producing permanent magnets. Such a material is magnetized, or induced to be magnetic.
- Above a material's Curie temperature, thermal agitation destroys the alignment of atoms, and ferromagnetism disappears.
- Electromagnets employ electric currents to make magnetic fields, often aided by induced fields in ferromagnetic materials.

Glossary

ferromagnetic

materials, such as iron, cobalt, nickel, and gadolinium, that exhibit strong magnetic effects

magnetized

to be turned into a magnet; to be induced to be magnetic

domains

regions within a material that behave like small bar magnets

Curie temperature

the temperature above which a ferromagnetic material cannot be magnetized

electromagnetism

the use of electrical currents to induce magnetism

electromagnet

an object that is temporarily magnetic when an electrical current is passed through it

magnetic monopoles

an isolated magnetic pole; a south pole without a north pole, or vice versa (no magnetic monopole has ever been observed)

This page titled 3.5: Ferromagnets and Electromagnets is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **22.2: Ferromagnets and Electromagnets** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/college-physics.

CHAPTER OVERVIEW

4: Waves



Under Construction(Public Domain ; Public Domain Pictures)

4.1: Oscillations and Waves

4.1.1: Introduction to Oscillatory Motion and Waves

4.1.2: Period and Frequency in Oscillations

4.1.3: Simple Harmonic Motion- A Special Periodic Motion

4.1.4: Forced Oscillations and Resonance

4.1.5: Waves

4.1.6: Wave Interference- Standing Waves and Beats

4.1.7: Sound

4.1.8: Speed of Sound, Frequency, and Wavelength

4.1.9: Doppler Effect and Sonic Booms

4.1.E: Oscillations and Waves (Exercise)

4.2: Light

4.2.1: Introduction to Light

4.2.2: Maxwell's Equations- Electromagnetic Waves Predicted and Observed

4.2.3: Production and Properties of Electromagnetic Waves

4.2.4: The Electromagnetic Spectrum- an Overview

4.2.5: The Electromagnetic Spectrum- Application Notes

4.2.6: Reflection

4.2.7: Refraction

4.2.8: Dispersion- The Rainbow and Prisms

4.2.9: Image Formation by Mirrors

4.2.10: Polarization

4.3: Image Formation by Lenses

4: Waves is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

SECTION OVERVIEW

4.1: Oscillations and Waves

4.1.1: Introduction to Oscillatory Motion and Waves

4.1.2: Period and Frequency in Oscillations

4.1.3: Simple Harmonic Motion- A Special Periodic Motion

4.1.4: Forced Oscillations and Resonance

4.1.5: Waves

4.1.6: Wave Interference- Standing Waves and Beats

4.1.7: Sound

4.1.8: Speed of Sound, Frequency, and Wavelength

4.1.9: Doppler Effect and Sonic Booms

4.1.E: Oscillations and Waves (Exercise)

This page titled 4.1: Oscillations and Waves is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

4.1.1: Introduction to Oscillatory Motion and Waves



Figure 4.1.1.1: There are at least four types of waves in this picture—only the water waves are evident. There are also sound waves, light waves, and waves on the guitar strings. (credit: John Norton)

What do an ocean buoy, a child in a swing, the cone inside a speaker, a guitar, atoms in a crystal, the motion of chest cavities, and the beating of hearts all have in common? They all **oscillate**—-that is, they move back and forth between two points. Many systems oscillate, and they have certain characteristics in common. All oscillations involve force and energy. You push a child in a swing to get the motion started. The energy of atoms vibrating in a crystal can be increased with heat. You put energy into a guitar string when you pluck it.

Some oscillations create **waves**. A guitar creates sound waves. You can make water waves in a swimming pool by slapping the water with your hand. You can no doubt think of other types of waves. Some, such as water waves, are visible. Some, such as sound waves, are not. But *every wave is a disturbance that moves from its source and carries energy*. Other examples of waves include earthquakes and visible light. Even subatomic particles, such as electrons, can behave like waves.

By studying oscillatory motion and waves, we shall find that a small number of underlying principles describe all of them and that wave phenomena are more common than you have ever imagined.

Glossary

oscillate

moving back and forth regularly between two points

wave

a disturbance that moves from its source and carries energy

This page titled 4.1.1: Introduction to Oscillatory Motion and Waves is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- **5.1: Introduction to Oscillatory Motion and Waves** by OpenStax is licensed CC BY 4.0.

4.1.2: Period and Frequency in Oscillations

Learning Objectives

- Understand the relationship between the frequency and the period of oscillations.
- Determine the frequency of oscillations.

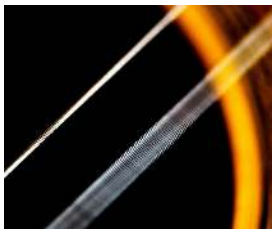


Figure 4.1.2.1: The strings on this guitar vibrate at regular time intervals. (credit: JAR)

When you pluck a guitar string, the resulting sound has a steady tone and lasts a long time. Each successive vibration of the string takes the same time as the previous one. We define **periodic motion** to be a motion that repeats itself at regular time intervals, such as exhibited by the guitar string or by an object on a spring moving up and down. The time to complete one oscillation remains constant and is called the **period** T. Its units are usually seconds, but may be any convenient unit of time. The word period refers to the time for some event whether repetitive or not; but we shall be primarily interested in periodic motion, which is by definition repetitive. A concept closely related to period is the frequency of an event. For example, if you get a paycheck twice a month, the frequency of payment is two per month and the period between checks is half a month. **Frequency** f is defined to be the number of events per unit time. For periodic motion, frequency is the number of oscillations per unit time. The relationship between frequency and period is

$$f = \frac{1}{T}.$$

The SI unit for frequency is the *cycle per second*, which is defined to be a *hertz* (Hz):

$$1 \text{ Hz} = 1\frac{\text{cycle}}{\text{sec}} \text{ or } 1 \text{ Hz} = \frac{1}{\text{s}}$$

A cycle is one complete oscillation. Note that a vibration can be a single or multiple event, whereas oscillations are usually repetitive for a significant number of cycles.

Example 4.1.2.1: Determine the Frequency of Two Oscillations: Medical Ultrasound and the Period of Middle C

We can use the formulas presented in this module to determine both the frequency based on known oscillations and the oscillation based on a known frequency. Let's try one example of each. (a) A medical imaging device produces ultrasound by oscillating with a period of 0.400 µs. What is the frequency of this oscillation? (b) The frequency of middle C on a typical musical instrument is 264 Hz. What is the time for one complete oscillation?

Strategy

Both questions (a) and (b) can be answered using the relationship between period and frequency. In question (a), the period T is given and we are asked to find frequency f. In question (b), the frequency f is given and we are asked to find the period T.

Solution a

1. Substitute 0.400 μs for T in $f = \frac{1}{T}$:

$$f = \frac{1}{T} = \frac{1}{0.400 \times 10^{-6} \text{ s}}.$$

Solve to find

$$f = 2.50 \times 10^{6} \text{ Hz}.$$

Discussion a

The frequency of sound found in (a) is much higher than the highest frequency that humans can hear and, therefore, is called ultrasound. Appropriate oscillations at this frequency generate ultrasound used for noninvasive medical diagnoses, such as observations of a fetus in the womb.

Solution b

1. Identify the known values:

   The time for one complete oscillation is the period T:

   $$f = \frac{1}{T}.$$

2. Solve for T:

   $$T = \frac{1}{f}.$$

3. Substitute the given value for the frequency into the resulting expression:

   $$T = \frac{1}{f} = \frac{1}{264\ \text{Hz}} = \frac{1}{264\text{cycles/s}} = 3.79 \times 10^{-3}\ \text{s} = 3.79\ \text{ms}.$$

Discussion

The period found in (b) is the time per cycle, but this value is often quoted as simply the time in convenient units (ms or milliseconds in this case).

Exercise 4.1.2.1

Identify an event in your life (such as receiving a paycheck) that occurs regularly. Identify both the period and frequency of this event.

Answer

I visit my parents for dinner every other Sunday. The frequency of my visits is 26 per calendar year. The period is two weeks.

Section Summary

- Periodic motion is a repetitious oscillation.
- The time for one oscillation is the period T.
- The number of oscillations per unit time is the frequency f.
- These quantities are related by

$$f = \frac{1}{T}.$$

Glossary

period
time it takes to complete one oscillation

periodic motion
motion that repeats itself at regular time intervals

frequency
number of events per unit of time

This page titled 4.1.2: Period and Frequency in Oscillations is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- **5.2: Period and Frequency in Oscillations** by OpenStax is licensed CC BY 4.0.

4.1.3: Simple Harmonic Motion- A Special Periodic Motion

Learning Objectives

- Describe a simple harmonic oscillator.
- Explain the link between simple harmonic motion and waves.

The oscillations of a system in which the net force can be described by Hooke's law are of special importance, because they are very common. They are also the simplest oscillatory systems. **Simple Harmonic Motion** (SHM) is the name given to oscillatory motion for a system where the net force can be described by Hooke's law, and such a system is called a **simple harmonic oscillator**. If the net force can be described by Hooke's law and there is no *damping* (by friction or other non-conservative forces), then a simple harmonic oscillator will oscillate with equal displacement on either side of the equilibrium position, as shown for an object on a spring in Figure 4.1.3.1. The maximum displacement from equilibrium is called the **amplitude** X. The units for amplitude and displacement are the same, but depend on the type of oscillation. For the object on the spring, the units of amplitude and displacement are meters; whereas for sound oscillations, they have units of pressure (and other types of oscillations have yet other units). Because amplitude is the maximum displacement, it is related to the energy in the oscillation.

TAKE-HOME EXPERIMENT: SHM AND THE MARBLE

Find a bowl or basin that is shaped like a hemisphere on the inside. Place a marble inside the bowl and tilt the bowl periodically so the marble rolls from the bottom of the bowl to equally high points on the sides of the bowl. Get a feel for the force required to maintain this periodic motion. What is the restoring force and what role does the force you apply play in the simple harmonic motion (SHM) of the marble?

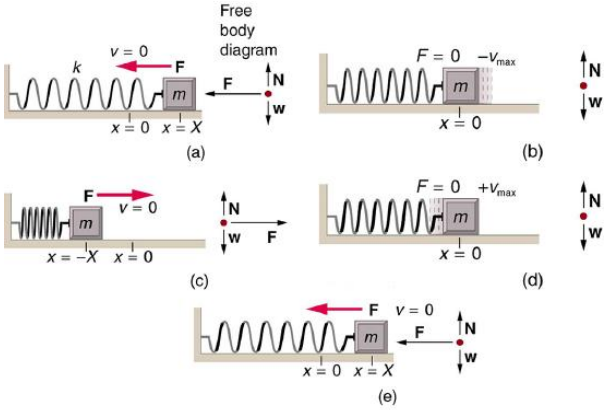


Figure 4.1.3.1: An object attached to a spring sliding on a frictionless surface is an uncomplicated simple harmonic oscillator. When displaced from equilibrium, the object performs simple harmonic motion that has an amplitude X and a period T. The object's maximum speed occurs as it passes through equilibrium. The stiffer the spring is, the smaller the period T. The greater the mass of the object is, the greater the period T.

What is so significant about simple harmonic motion? One special thing is that the period T and frequency f of a simple harmonic oscillator are independent of amplitude. The string of a guitar, for example, will oscillate with the same frequency whether plucked gently or hard. Because the period is constant, a simple harmonic oscillator can be used as a clock.

Two important factors do affect the period of a simple harmonic oscillator. The period is related to how stiff the system is. A very stiff object has a large spring constant k, which causes the system to have a smaller period. For example, you can adjust a diving board's stiffness—the stiffer it is, the faster it vibrates, and the shorter its period. Period also depends on the mass of the oscillating system. The more massive the system is, the longer the period. For example, a heavy person on a diving board bounces up and down more slowly than a light one.

In fact, the mass m and the spring constant k are the *only* factors that affect the period and frequency of simple harmonic motion of a mass on a spring.

PERIOD OF SIMPLE HARMONIC OSCILLATOR

The *period of a simple harmonic oscillator* is given by

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

and, because $f = 1/T$, the *frequency of a simple harmonic oscillator* is

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

Note that neither T nor f has any dependence on amplitude.

TAKE-HOME EXPERIMENT: MASS AND RULER OSCILLATIONS

Find two identical wooden or plastic rulers. Tape one end of each ruler firmly to the edge of a table so that the length of each ruler that protrudes from the table is the same. On the free end of one ruler tape a heavy object such as a few large coins. Pluck the ends of the rulers at the same time and observe which one undergoes more cycles in a time period, and measure the period of oscillation of each of the rulers.

Example 4.1.3.1: Calculate the Frequency and Period of Oscillations: Bad Shock Absorbers in a Car

If the shock absorbers in a car go bad, then the car will oscillate at the least provocation, such as when going over bumps in the road and after stopping (See Figure 4.1.3.2). Calculate the frequency and period of these oscillations for such a car if the car's mass (including its load) is 900 kg and the spring constant (k) of the suspension system is 6.53×10^4 N/m.

Strategy

The frequency of the car's oscillations will be that of a simple harmonic oscillator as given in the equation $f = \frac{1}{2\pi}\sqrt{\frac{k}{m}}$. The mass and the spring constant are both given.

Solution

1. Enter the known values of *k* and *m*:

$$f = \frac{1}{2\pi}\sqrt{\frac{k}{m}} = \frac{1}{2\pi}\sqrt{\frac{6.53 \times 10^4 \text{ N/m}}{900 \text{ kg}}}.$$

2. Calculate the frequency:

$$\frac{1}{2\pi}\sqrt{72.6/\text{s}^{-2}} = 1.3656/\text{s}^{-1} \approx 1.36/\text{s}^{-1} = 1.36 \text{ Hz}.$$

3. You could use $T = 2\pi\sqrt{\frac{m}{k}}$ to calculate the period, but it is simpler to use the relationship $T = 1/f$ and substitute the value just found for f:

$$T = \frac{1}{f} = \frac{1}{1.356 \text{ Hz}} = 0.738 \text{ s}.$$

Discussion

The values of T and f both seem about right for a bouncing car. You can observe these oscillations if you push down hard on the end of a car and let go.

The Link between Simple Harmonic Motion and Waves

If a time-exposure photograph of the bouncing car were taken as it drove by, the headlight would make a wavelike streak, as shown in Figure 4.1.3.2. Similarly, Figure 4.1.3.3 shows an object bouncing on a spring as it leaves a wavelike "trace of its position on a moving strip of paper. Both waves are sine functions. All simple harmonic motion is intimately related to sine and cosine waves.

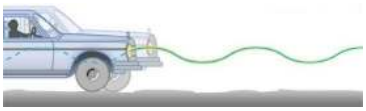


Figure 4.1.3.2: The bouncing car makes a wavelike motion. If the restoring force in the suspension system can be described only by Hooke's law, then the wave is a sine function. (The wave is the trace produced by the headlight as the car moves to the right.)

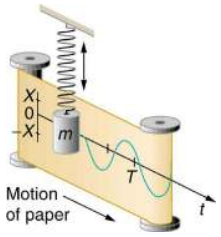


Figure 4.1.3.3: The vertical position of an object bouncing on a spring is recorded on a strip of moving paper, leaving a sine wave.

The displacement $x(t)$, the velocity $v(t)$, and the acceleration $a(t)$ of simple harmonic motion sketches out sinusoidal functions as a function of time. That is, if you sketch out $\sin(t)$ and $\cos(t)$, with t in seconds, for example, what you graph will appear similar to the actual graphs of $x(t)$,$v(t)$, and $a(t)$. You just need to scale them horizontally and vertically, so that they have correct period and amplitude, respectively.

Figure 4.1.3.4 shows the simple harmonic motion of an object on a spring and presents graphs of $x(t)$,$v(t)$, and $a(t)$ versus time.

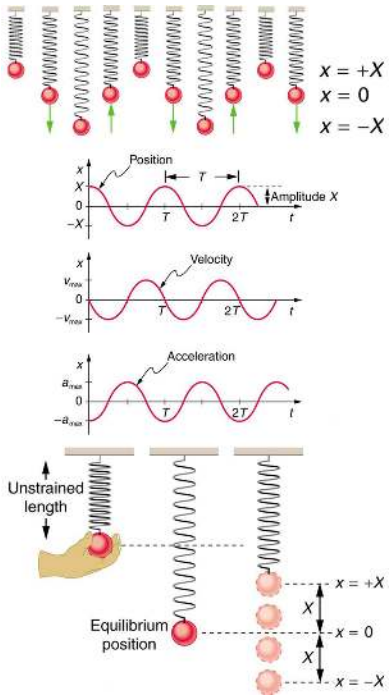


Figure 4.1.3.4: Graphs of $x(t)$,$v(t)$, and $a(t)$ versus t for the motion of an object on a spring. The net force on the object can be described by Hooke's law, and so the object undergoes simple harmonic motion. Note that the initial position has the vertical displacement at its maximum value X; v is initially zero and then negative as the object moves down; and the initial acceleration is negative, back toward the equilibrium position and becomes zero at that point.

The most important point here is that these relationships are valid for all simple harmonic motion. They are very useful in visualizing waves associated with simple harmonic motion, including visualizing how waves add with one another.

Exercise 4.1.3.1

Suppose you pluck a banjo string. You hear a single note that starts out loud and slowly quiets over time. Describe what happens to the sound waves in terms of period, frequency and amplitude as the sound decreases in volume.

Answer

Frequency and period remain essentially unchanged. Only amplitude decreases as volume decreases.

Exercise 4.1.3.2

A babysitter is pushing a child on a swing. At the point where the swing reaches x, where would the corresponding point on a wave of this motion be located?

Answer

x is the maximum deformation, which corresponds to the amplitude of the wave. The point on the wave would either be at the very top or the very bottom of the curve.

Section Summary

- Simple harmonic motion is oscillatory motion for a system that can be described only by Hooke's law. Such a system is also called a simple harmonic oscillator.
- Maximum displacement is the amplitude X. The period T and frequency f of a simple harmonic oscillator are given by $T = 2\pi\sqrt{\frac{m}{k}}$ and $f = \frac{1}{2\pi}\sqrt{\frac{k}{m}}$, where m is the mass of the system.
- Displacement, velocity, and acceleration in simple harmonic motion as a function of time are represented by sinusoidal functions in time, that is functions that are similar to $\sin(t)$ and $\cos(t)$.

Glossary

amplitude
the maximum displacement from the equilibrium position of an object oscillating around the equilibrium position

simple harmonic motion
the oscillatory motion in a system where the net force can be described by Hooke's law

simple harmonic oscillator
a device that implements Hooke's law, such as a mass that is attached to a spring, with the other end of the spring being connected to a rigid support such as a wall

This page titled 4.1.3: Simple Harmonic Motion- A Special Periodic Motion is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- **5.3: Simple Harmonic Motion- A Special Periodic Motion** by OpenStax is licensed CC BY 4.0.

4.1.4: Forced Oscillations and Resonance

Learning Objectives

- Observe the resonance phenomena in several examples.
- Observe amplitude of a damped harmonic oscillator.
- Understand the origin of damping of resonance.



Figure 4.1.4.1: You can cause the strings in a piano to vibrate simply by producing sound waves from your voice. (credit: Matt Billings, Flickr)

Sit in front of a piano sometime and sing a loud brief note at it with the dampers off its strings. It will sing the same note back at you—the strings, having the same frequencies as your voice, are resonating in response to the forces from the sound waves that you sent to them. Your voice and a piano's strings is a good example of the fact that objects—in this case, piano strings—can be forced to oscillate but oscillate best at their natural frequency. In this section, we shall briefly explore applying a *periodic driving force* acting on a simple harmonic oscillator. The driving force puts energy into the system at a certain frequency, not necessarily the same as the natural frequency of the system. The **natural frequency** is the frequency at which a simple harmonic oscillator oscillates, if there is no damping force, and other than an initial kick to get the motion started, there is no driving force. A **damping force** is a frictional force (friction, air resistance, and other forces that lead to decrease in mechanical energy) which converts the mechanical energy of the oscillatory motion into thermal energy.

Most of us have played with toys involving an object supported on an elastic band, something like the paddle ball suspended from a finger in Figure 4.1.4.2. Imagine the finger in the figure is your finger. At first you hold your finger steady, and the ball bounces up and down with a small amount of damping. If you move your finger up and down slowly, the ball will follow along without bouncing much on its own. As you increase the frequency at which you move your finger up and down, the ball will respond by oscillating with increasing amplitude. When you drive the ball at its natural frequency, the ball's oscillations increase in amplitude with each oscillation for as long as you drive it. The phenomenon of driving a system with a frequency equal to its natural frequency is called **resonance**. A system being driven at its natural frequency is said to **resonate**. As the driving frequency gets progressively higher than the resonant or natural frequency, the amplitude of the oscillations becomes smaller, until the oscillations nearly disappear and your finger simply moves up and down with little effect on the ball.

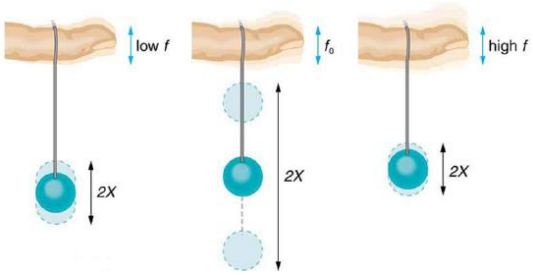


Figure 4.1.4.2: The paddle ball on its rubber band moves in response to the finger supporting it. If the finger moves with the natural frequency f_0 of the ball on the rubber band, then a resonance is achieved, and the amplitude of the ball's oscillations increases dramatically. At higher and lower driving frequencies, energy is transferred to the ball less efficiently, and it responds with lower-amplitude oscillations.

Figure 4.1.4.3 shows a graph of the amplitude of a damped harmonic oscillator as a function of the frequency of the periodic force driving it. There are three curves on the graph, each representing a different amount of damping. All three curves peak at the point where the frequency of the driving force equals the natural frequency of the harmonic oscillator. The highest peak, or greatest response, is for the least amount of damping, because less energy is removed by the damping force.

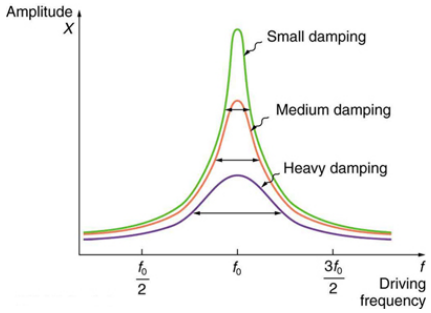


Figure 4.1.4.3: Amplitude of a harmonic oscillator as a function of the frequency of the driving force. The curves represent the same oscillator with the same natural frequency but with different amounts of damping. Resonance occurs when the driving frequency equals the natural frequency, and the greatest response is for the least amount of damping. The narrowest response is also for the least damping.

It is interesting that the widths of the resonance curves shown in Figure 4.1.4.3 depend on damping: the less the damping, the narrower the resonance. The message is that if you want a driven oscillator to resonate at a very specific frequency, you need as little damping as possible. Little damping is the case for piano strings and many other musical instruments. Conversely, if you want small-amplitude oscillations, such as in a car's suspension system, then you want heavy damping. Heavy damping reduces the amplitude, but the tradeoff is that the system responds at more frequencies.

These features of driven harmonic oscillators apply to a huge variety of systems. When you tune a radio, for example, you are adjusting its resonant frequency so that it only oscillates to the desired station's broadcast (driving) frequency. The more selective the radio is in discriminating between stations, the smaller its damping. Magnetic resonance imaging (MRI) is a widely used medical diagnostic tool in which atomic nuclei (mostly hydrogen nuclei) are made to resonate by incoming radio waves (on the order of 100 MHz). A child on a swing is driven by a parent at the swing's natural frequency to achieve maximum amplitude. In all of these cases, the efficiency of energy transfer from the driving force into the oscillator is best at resonance. Speed bumps and gravel roads prove that even a car's suspension system is not immune to resonance. In spite of finely engineered shock absorbers, which ordinarily convert mechanical energy to thermal energy almost as fast as it comes in, speed bumps still cause a large-amplitude oscillation. On gravel roads that are corrugated, you may have noticed that if you travel at the "wrong" speed, the bumps are very noticeable whereas at other speeds you may hardly feel the bumps at all. Figure 4.1.4.4 shows a photograph of a famous example (the Tacoma Narrows Bridge) of the destructive effects of a driven harmonic oscillation. The Millennium Bridge in London was closed for a short period of time for the same reason while inspections were carried out.

In our bodies, the chest cavity is a clear example of a system at resonance. The diaphragm and chest wall drive the oscillations of the chest cavity which result in the lungs inflating and deflating. The system is critically damped and the muscular diaphragm oscillates at the resonant value for the system, making it highly efficient.



Figure 4.1.4.4: In 1940, the Tacoma Narrows Bridge in Washington state collapsed. Heavy cross winds drove the bridge into oscillations at its resonant frequency. Damping decreased when support cables broke loose and started to slip over the towers, allowing increasingly greater amplitudes until the structure failed (credit: PRI's *Studio 360*, via Flickr)

Exercise 4.1.4.1

A famous magic trick involves a performer singing a note toward a crystal glass until the glass shatters. Explain why the trick works in terms of resonance and natural frequency.

Answer

> The performer must be singing a note that corresponds to the natural frequency of the glass. As the sound wave is directed at the glass, the glass responds by resonating at the same frequency as the sound wave. With enough energy introduced into the system, the glass begins to vibrate and eventually shatters.

Section Summary

- A system’s natural frequency is the frequency at which the system will oscillate if not affected by driving or damping forces.
- A periodic force driving a harmonic oscillator at its natural frequency produces resonance. The system is said to resonate.
- The less damping a system has, the higher the amplitude of the forced oscillations near resonance. The more damping a system has, the broader response it has to varying driving frequencies.

Glossary

natural frequency
the frequency at which a simple harmonic oscillator oscillates if it is set in motion without a driving force

damping force
a frictional force which converts the mechanical energy of the oscillatory motion into thermal energy

resonance
the phenomenon of driving a system with a frequency equal to the system's natural frequency

resonate
a system being driven at its natural frequency

This page titled 4.1.4: Forced Oscillations and Resonance is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- **5.4: Forced Oscillations and Resonance** by OpenStax is licensed CC BY 4.0.

4.1.5: Waves

Learning Objectives

- State the characteristics of a wave.
- Calculate the velocity of wave propagation, given frequency and wavelength



Figure 4.1.5.1: Waves in the ocean behave similarly to all other types of waves. (credit: Steve Jurveston, Flickr)

What do we mean when we say something is a wave? The most intuitive and easiest wave to imagine is the familiar water wave. More precisely, a **wave** is a disturbance that propagates, or moves from the place it was created. For water waves, the disturbance is in the surface of the water, perhaps created by a rock thrown into a pond or by a swimmer splashing the surface repeatedly. For sound waves, the disturbance is a change in air pressure, perhaps created by the oscillating cone inside a speaker. For earthquakes, there are several types of disturbances, including disturbance of Earth's surface and pressure disturbances under the surface. Even radio waves are most easily understood using an analogy with water waves. Visualizing water waves is useful because there is more to it than just a mental image. Water waves exhibit characteristics common to all waves, such as amplitude, period, frequency and energy. All wave characteristics can be described by a small set of underlying principles.

A wave is a disturbance that propagates, or moves from the place it was created. The simplest waves repeat themselves for several cycles and are associated with simple harmonic motion. Let us start by considering the simplified water wave in Figure 4.1.5.2. The wave is an up and down disturbance of the water surface. It causes a sea gull to move up and down in simple harmonic motion as the wave crests and troughs (peaks and valleys) pass under the bird. The time for one complete up and down motion is the wave's period T. The wave's frequency is $f = 1/T$, as usual. The wave itself moves to the right in the figure. This movement of the wave is actually the disturbance moving to the right, not the water itself (or the bird would move to the right). We define **wave velocity** v_{w} to be the speed at which the disturbance moves. Wave velocity is sometimes also called the *propagation velocity or propagation speed*, because the disturbance propagates from one location to another.

MISCONCEPTION ALERT

Many people think that water waves push water from one direction to another. In fact, the particles of water tend to stay in one location, save for moving up and down due to the energy in the wave. The energy moves forward through the water, but the water stays in one place. If you feel yourself pushed in an ocean, what you feel is the energy of the wave, not a rush of water.

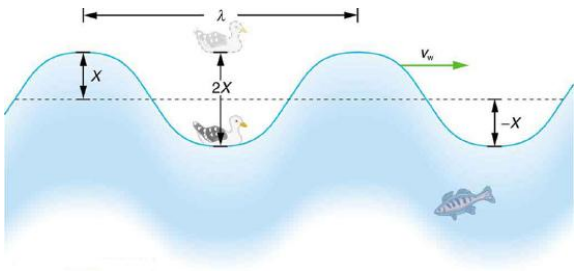


Figure 4.1.5.2: An idealized ocean wave passes under a sea gull that bobs up and down in simple harmonic motion. The wave has a wavelength λ, which is the distance between adjacent identical parts of the wave. The up and down disturbance of the surface propagates parallel to the surface at a speed v_{w}.

he water wave in the figure also has a length associated with it, called its **wavelength** λ, the distance between adjacent identical parts of a wave. (λ is the distance parallel to the direction of propagation.) The speed of propagation v_{w} is the distance the wave travels in a given time, which is one wavelength in the time of one period. In equation form, that is

$$v_{\text{w}} = \frac{\lambda}{}$$

$$v_w = \frac{\lambda}{T}$$

or

$$v_w = f\lambda. \tag{4.1.5.1}$$

This fundamental relationship holds for all types of waves. For water waves, v_w is the speed of a surface wave; for sound, v_w is the speed of sound; and for visible light, v_w is the speed of light, for example.

One note of caution with Equation (4.1.5.1): despite the appearance, wave speed v_w is a *not* a function of frequency f and wavelength λ. When f changes (or λ changes), v_w does not change (you will see this explicitly later with sound and light waves). Wave speed is a *property of the medium* the wave travels in. Unless the medium itself changes, the wave speed remains constant. This is true of sound and light waves, as well as almost all other types of waves you will see in this textbook (water waves, waves on a string, etc.). So, what happens when frequency or wavelength does change, then? The wavelength and frequency change to compensate. For example, if the frequency doubles for a given wave, instead of the wave speed doubling, the wavelength will decrease to half, so that the product $f\lambda$ remains constant.

TAKE-HOME EXPERIMENT: WAVES IN A BOWL

Fill a large bowl or basin with water and wait for the water to settle so there are no ripples. Gently drop a cork into the middle of the bowl. Estimate the wavelength and period of oscillation of the water wave that propagates away from the cork. Remove the cork from the bowl and wait for the water to settle again. Gently drop the cork at a height that is different from the first drop. Does the wavelength depend upon how high above the water the cork is dropped?

Calculate the Velocity of Wave Propagation: Gull in the Ocean

Calculate the wave velocity of the ocean wave in Figure 4.1.5.2 if the distance between wave crests is 10.0 m and the time for a sea gull to bob up and down is 5.00 s.

Strategy

We are asked to find v_w. The given information tells us that $\lambda = 10.0$ m and $T = 5.00$ s. Therefore, we can use $v_w = \frac{\lambda}{T}$ to find the wave velocity.

Solution

1. Enter the known values into $v_w = \frac{\lambda}{T}$:

$$v_w = \frac{10.0 \text{ m}}{5.00 \text{ s}}.$$

2. Solve for v_w to find $v_w = 2.00$ m/s.

Discussion

This slow speed seems reasonable for an ocean wave. Note that the wave moves to the right in the figure at this speed, not the varying speed at which the sea gull moves up and down.

Transverse and Longitudinal Waves

A simple wave consists of a periodic disturbance that propagates from one place to another. The wave in Figure 4.1.5.3 propagates in the horizontal direction while the surface is disturbed in the vertical direction. Such a wave is called a **transverse wave** or shear wave; in such a wave, the disturbance is perpendicular to the direction of propagation. In contrast, in a **longitudinal wave** or compressional wave, the disturbance is parallel to the direction of propagation. Figure 4.1.5.4 shows an example of a longitudinal wave. The size of the disturbance is its amplitude X and is completely independent of the speed of propagation v_w.

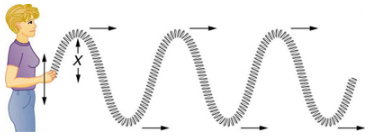


Figure 4.1.5.3: In this example of a transverse wave, the wave propagates horizontally, and the disturbance in the cord is in the vertical direction.

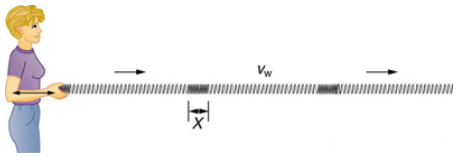


Figure 4.1.5.4: In this example of a longitudinal wave, the wave propagates horizontally, and the disturbance in the cord is also in the horizontal direction.

Waves may be transverse, longitudinal, or *a combination of the two*. (Water waves are actually a combination of transverse and longitudinal. The simplified water wave illustrated in Figure 4.1.5.2 shows no longitudinal motion of the bird.) The waves on the strings of musical instruments are transverse—so are electromagnetic waves, such as visible light.

Sound waves in air and water are longitudinal. Their disturbances are periodic variations in pressure that are transmitted in fluids. Fluids do not have appreciable shear strength, and thus the sound waves in them must be longitudinal or compressional. Sound in solids can be both longitudinal and transverse.

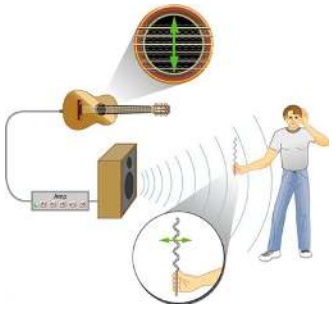


Figure 4.1.5.5: The wave on a guitar string is transverse. The sound wave rattles a sheet of paper in a direction that shows the sound wave is longitudinal.

Earthquake waves under Earth's surface also have both longitudinal and transverse components (called compressional or P-waves and shear or S-waves, respectively). These components have important individual characteristics—they propagate at different speeds, for example. Earthquakes also have surface waves that are similar to surface waves on water.

Exercise 4.1.5.1

Why is it important to differentiate between longitudinal and transverse waves?

Answer

In the different types of waves, energy can propagate in a different direction relative to the motion of the wave. This is important to understand how different types of waves affect the materials around them.

Section Summary

- A wave is a disturbance that moves from the point of creation with a wave velocity v_w.
- A wave has a wavelength λ, which is the distance between adjacent identical parts of the wave.
- Wave velocity and wavelength are related to the wave's frequency and period by $v_w = \frac{\lambda}{T}$ or $v_w = f\lambda$. However, wave velocity is a property of the medium and remains constant as wavelength or frequency changes.
- A transverse wave has a disturbance perpendicular to its direction of propagation, whereas a longitudinal wave has a disturbance parallel to its direction of propagation.

Glossary

longitudinal wave

a wave in which the disturbance is parallel to the direction of propagation

transverse wave

a wave in which the disturbance is perpendicular to the direction of propagation

wave velocity

the speed at which the disturbance moves; also called wave speed, propagation velocity, or propagation speed

wavelength

the distance between adjacent identical parts of a wave

This page titled 4.1.5: Waves is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- **5.5: Waves** by OpenStax is licensed CC BY 4.0.

4.1.6: Wave Interference- Standing Waves and Beats

Learning Objectives

- Distinguish between constructive interference and destructive interference.
- Explain standing waves in terms of interference, nodes, and antinodes.
- Explain the beat phenomenon in terms of interference.



Figure 4.1.6.1: These waves result from the superposition of several waves from different sources, producing a complex pattern. (credit: waterborough, Wikimedia Commons)

Most waves do not look very simple. They look more like the waves in Figure 4.1.6.1 than like the simple water wave considered in "Waves". (Simple waves may be created by a simple harmonic oscillation, and thus have a sinusoidal shape). Complex waves are more interesting, even beautiful, but they look formidable. Most waves appear complex because they result from several simple waves adding together. Luckily, the rules for adding waves are quite simple.

When two or more waves arrive at the same point, they superimpose themselves on one another. More specifically, the disturbances of waves are superimposed when they come together—a phenomenon called **superposition**. Each disturbance corresponds to a force, and forces add. If the disturbances are along the same line, then the resulting wave is a simple addition of the disturbances of the individual waves—that is, their amplitudes add. Figure 4.1.6.2 and Figure 4.1.6.3 illustrate superposition in two special cases, both of which produce simple results.

Figure 4.1.6.2 shows two identical waves that arrive at the same point exactly in phase. The crests of the two waves are precisely aligned, as are the troughs. This superposition produces pure **constructive interference**. Because the disturbances add, pure constructive interference produces a wave that has twice the amplitude of the individual waves, but has the same wavelength.

Figure 4.1.6.3 shows two identical waves that arrive exactly out of phase—that is, precisely aligned crest to trough—producing pure **destructive interference**. Because the disturbances are in the opposite direction for this superposition, the resulting amplitude is zero for pure destructive interference—the waves completely cancel.

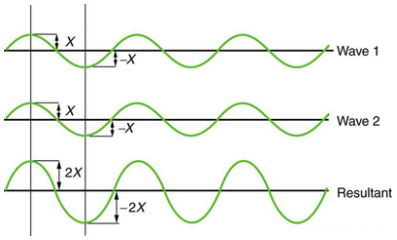


Figure 4.1.6.2: Pure constructive interference of two identical waves produces one with twice the amplitude, but the same wavelength.

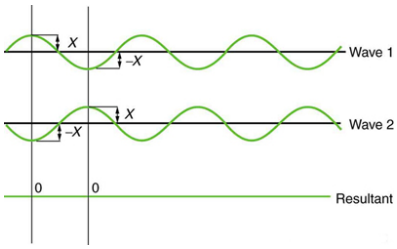


Figure 4.1.6.3: Pure destructive interference of two identical waves produces zero amplitude, or complete cancellation.

While pure constructive and pure destructive interference do occur, they require precisely aligned identical waves. The superposition of most waves produces a combination of constructive and destructive interference and can vary from place to place and time to time. Sound from a stereo, for example, can be loud in one spot and quiet in another. Varying loudness means the sound waves add partially constructively and partially destructively at different locations. A stereo has at least two speakers creating sound waves, and waves can reflect from walls. All these waves superimpose. An example of sounds that vary over time from constructive to destructive is found in the combined whine of airplane jets heard by a stationary passenger. The combined sound can fluctuate up and down in volume as the sound from the two engines varies in time from constructive to destructive. These examples are of waves that are similar.

An example of the superposition of two dissimilar waves is shown in Figure 4.1.6.4. Here again, the disturbances add and subtract, producing a more complicated looking wave.

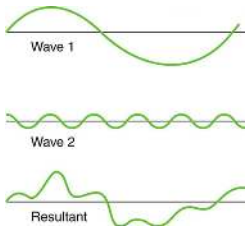


Figure 4.1.6.4: Superposition of non-identical waves exhibits both constructive and destructive interference.

Standing Waves

Sometimes waves do not seem to move; rather, they just vibrate in place. Unmoving waves can be seen on the surface of a glass of milk in a refrigerator, for example. Vibrations from the refrigerator motor create waves on the milk that oscillate up and down but do not seem to move across the surface. These waves are formed by the superposition of two or more moving waves, such as illustrated in Figure 4.1.6.5 for two identical waves moving in opposite directions. The waves move through each other with their disturbances adding as they go by. If the two waves have the same amplitude and wavelength, then they alternate between constructive and destructive interference. The resultant looks like a wave standing in place and, thus, is called a **standing wave**. Waves on the glass of milk are one example of standing waves. There are other standing waves, such as on guitar strings and in organ pipes. With the glass of milk, the two waves that produce standing waves may come from reflections from the side of the glass.

A closer look at earthquakes provides evidence for conditions appropriate for resonance, standing waves, and constructive and destructive interference. A building may be vibrated for several seconds with a driving frequency matching that of the natural frequency of vibration of the building—producing a resonance resulting in one building collapsing while neighboring buildings do not. Often buildings of a certain height are devastated while other taller buildings remain intact. The building height matches the condition for setting up a standing wave for that particular height. As the earthquake waves travel along the surface of Earth and reflect off denser rocks, constructive interference occurs at certain points. Often areas closer to the epicenter are not damaged while areas farther away are damaged.

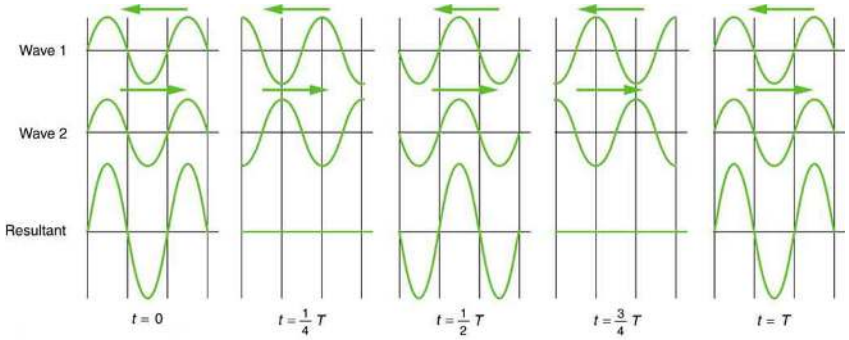


Figure 4.1.6.5: Standing wave created by the superposition of two identical waves moving in opposite directions. The oscillations are at fixed locations in space and result from alternately constructive and destructive interference.

Standing waves are also found on the strings of musical instruments and are due to reflections of waves from the ends of the string. Figure 4.1.6.6 and Figure 4.1.6.7 show three standing waves that can be created on a string that is fixed at both ends. **Nodes** are the points where the string does not move; more generally, nodes are where the wave disturbance is zero in a

standing wave. The fixed ends of strings must be nodes, too, because the string cannot move there. The word **antinode** is used to denote the location of maximum amplitude in standing waves. Standing waves on strings have a frequency that is related to the propagation speed v_w of the disturbance on the string. The wavelength λ is determined by the distance between the points where the string is fixed in place.

The lowest frequency, called the **fundamental frequency**, is thus for the longest wavelength, which is seen to be $\lambda_1 = 2L$. Therefore, the fundamental frequency is $f_1 = v_w/\lambda_1 = v_w/2L$. In this case, the **overtones** or harmonics are multiples of the fundamental frequency. As seen in Figure 4.1.6.7, the first harmonic can easily be calculated since $\lambda_2 = L$. Thus, $f_2 = v_w/\lambda_2 = v_w/L = 2f_1$. Similarly, $f_3 = 3f_1$, and so on. All of these frequencies can be changed by adjusting the tension in the string. The greater the tension, the greater v_w is and the higher the frequencies. This observation is familiar to anyone who has ever observed a string instrument being tuned. We will see in later chapters that standing waves are crucial to many resonance phenomena, such as in sounding boxes on string instruments.

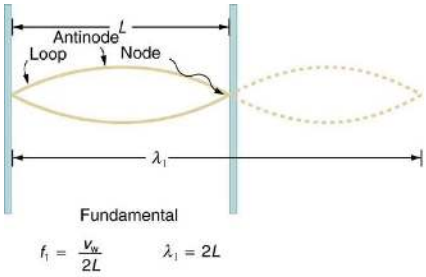


Figure 4.1.6.6: The figure shows a string oscillating at its fundamental frequency.

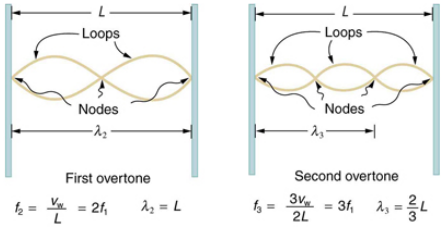


Figure 4.1.6.7: First and second harmonic frequencies are shown.

Beats

Striking two adjacent keys on a piano produces a warbling combination usually considered to be unpleasant. The superposition of two waves of similar but not identical frequencies is the culprit. Another example is often noticeable in jet aircraft, particularly the two-engine variety, while taxiing. The combined sound of the engines goes up and down in loudness. This varying loudness happens because the sound waves have similar but not identical frequencies. The discordant warbling of the piano and the fluctuating loudness of the jet engine noise are both due to alternately constructive and destructive interference as the two waves go in and out of phase. Figure 4.1.6.8 illustrates this graphically.

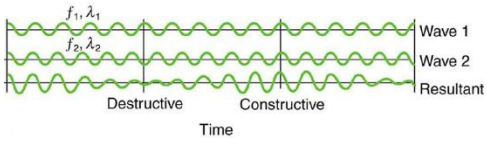


Figure 4.1.6.8: Beats are produced by the superposition of two waves of slightly different frequencies but identical amplitudes. The waves alternate in time between constructive interference and destructive interference, giving the resulting wave a time-varying amplitude.

The wave resulting from the superposition of two similar-frequency waves has a frequency that is the average of the two. This wave fluctuates in amplitude, or *beats*, with a frequency called the **beat frequency**. We can determine the beat frequency by adding two waves together mathematically. The result,

$$f_B = |f_1 - f_2|$$

is the beat frequency. The resultant wave has the average frequency of the two superimposed waves, but it also fluctuates in overall amplitude at the beat frequency f_B. This result is valid for all types of waves. However, if it is a sound wave, providing the two

frequencies are similar, then what we hear is an average frequency that gets louder and softer (or warbles) at the beat frequency.

MAKING CAREER CONNECTIONS

Piano tuners use beats routinely in their work. When comparing a note with a tuning fork, they listen for beats and adjust the string until the beats go away (to zero frequency). For example, if the tuning fork has a 256 Hz frequency and two beats per second are heard, then the other frequency is either 254 or 258 Hz. Most keys hit multiple strings, and these strings are actually adjusted until they have nearly the same frequency and give a slow beat for richness. Twelve-string guitars and mandolins are also tuned using beats.

While beats may sometimes be annoying in audible sounds, we will find that beats have many applications. Observing beats is a very useful way to compare similar frequencies. There are applications of beats as apparently disparate as in ultrasonic imaging and radar speed traps.

Exercise 4.1.6.1

Imagine you are holding one end of a jump rope, and your friend holds the other. If your friend holds her end still, you can move your end up and down, creating a transverse wave. If your friend then begins to move her end up and down, generating a wave in the opposite direction, what resultant wave forms would you expect to see in the jump rope?

Answer

The rope would alternate between having waves with amplitudes two times the original amplitude and reaching equilibrium with no amplitude at all. The wavelengths will result in both constructive and destructive interference.

Exercise 4.1.6.2

Define nodes and antinodes.

Answer

Nodes are areas of wave interference where there is no motion. Antinodes are areas of wave interference where the motion is at its maximum point.

Exercise 4.1.6.3

You hook up a stereo system. When you test the system, you notice that in one corner of the room, the sounds seem dull. In another area, the sounds seem excessively loud. Describe how the sound moving about the room could result in these effects.

Answer

With multiple speakers putting out sounds into the room, and these sounds bouncing off walls, there is bound to be some wave interference. In the dull areas, the interference is probably mostly destructive. In the louder areas, the interference is probably mostly constructive.

Section Summary

- Superposition is the combination of two waves at the same location.
- Constructive interference occurs when two waves are superimposed in phase (both disturbances in the same direction).
- Destructive interference occurs when two waves are superimposed exactly out of phase (two disturbances in opposite directions).
- A standing wave is one in which two waves superimpose to produce a wave that forms a standing pattern of nodes and antinodes, locations of no displacement and of maximum displacement.
- Standing waves form at specific resonance frequencies, with the lowest frequency being the fundamental frequency, and higher frequencies being called overtones or harmonics.
- Beats occur when waves of similar frequencies f_1 and f_2 are superimposed. The resulting amplitude oscillates with a beat frequency given by

$$f_{\mathrm{B}} = |f_1 - f_2| \, .$$

Glossary

antinode
the location of maximum amplitude in standing waves

beat frequency
the frequency of the amplitude fluctuations of a wave

constructive interference
when two waves arrive at the same point exactly in phase; that is, the crests of the two waves are precisely aligned, as are the troughs

destructive interference
when two identical waves arrive at the same point exactly out of phase; that is, precisely aligned crest to trough

fundamental frequency
the lowest frequency of a periodic waveform

nodes
the points where the string does not move; more generally, nodes are where the wave disturbance is zero in a standing wave

overtones
multiples of the fundamental frequency of a sound

superposition
the phenomenon that occurs when two or more waves arrive at the same point

This page titled 4.1.6: Wave Interference- Standing Waves and Beats is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- **5.6: Wave Interference- Standing Waves and Beats** by OpenStax is licensed CC BY 4.0.

4.1.7: Sound

Learning Objectives

- Define sound and hearing.
- Describe sound as a longitudinal wave.
- Distinguish between audible sound, ultrasound, and infrasound.



Figure 4.1.7.1: This glass has been shattered by a high-intensity sound wave of the same frequency as the resonant frequency of the glass. While the sound is not visible, the effects of the sound prove its existence. (credit: ||read||, Flickr)

Sound can be used as a familiar illustration of waves. Because hearing is one of our most important senses, it is interesting to see how the physical properties of sound correspond to our perceptions of it. **Hearing** is the perception of sound, just as vision is the perception of visible light. But sound has important applications beyond hearing. Ultrasound, for example, is not heard but can be employed to form medical images and is also used in treatment.

The physical phenomenon of **sound** is defined to be a disturbance of matter that is transmitted from its source outward. Sound is a wave. On the atomic scale, it is a disturbance of atoms that is far more ordered than their thermal motions. In many instances, sound is a periodic wave, and the atoms undergo simple harmonic motion. In this text, we shall explore such periodic sound waves.

A vibrating string produces a sound wave as illustrated in Figure 4.1.7.2, Figure 4.1.7.3, and Figure 4.1.7.4. As the string oscillates back and forth, it transfers energy to the air, mostly as thermal energy created by turbulence. But a small part of the string's energy goes into compressing and expanding the surrounding air, creating slightly higher and lower local pressures. These compressions (high pressure regions) and rarefactions (low pressure regions) move out as longitudinal pressure waves having the same frequency as the string—they are the disturbance that is a sound wave. (Sound waves in air and most fluids are longitudinal, because fluids have almost no shear strength. In solids, sound waves can be both transverse and longitudinal.) Figure 4.1.7.4 shows a graph of gauge pressure versus distance from the vibrating string.

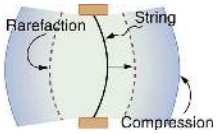


Figure 4.1.7.2: A vibrating string moving to the right compresses the air in front of it and expands the air behind it.

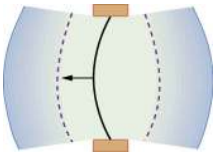


Figure 4.1.7.3: As the string moves to the left, it creates another compression and rarefaction as the ones on the right move away from the string.

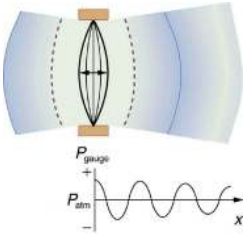


Figure 4.1.7.4: After many vibrations, there are a series of compressions and rarefactions moving out from the string as a sound wave. The graph shows gauge pressure versus distance from the source. Pressures vary only slightly from atmospheric for ordinary sounds.

The amplitude of a sound wave decreases with distance from its source, because the energy of the wave is spread over a larger and larger area. But it is also absorbed by objects, such as the eardrum in Figure 4.1.7.5, and converted to thermal energy by the viscosity of air. In addition, during each compression a little heat transfers to the air and during each rarefaction even less heat transfers from the air, so that the heat transfer reduces the organized disturbance into random thermal motions. Whether the heat transfer from compression to rarefaction is significant depends on how far apart they are—that is, it depends on wavelength. Wavelength, frequency, amplitude, and speed of propagation are important for sound, as they are for all waves.

When the frequency of the sound wave is in **audible range** of 20 to 20,000 Hz, the vibration of the eardrums at these frequencies can be sensed, which is what we hear. Sound waves of frequency above 20 kHz are called **ultrasound** waves. Some construction equipment produces **infrasound** waves of frequencies less than 20 Hz, which, although they can't be heard, can still affect the human body at large enough amplitudes.

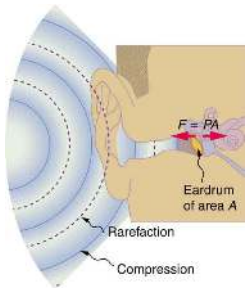


Figure 4.1.7.5: Sound wave compressions and rarefactions travel up the ear canal and force the eardrum to vibrate. There is a net force on the eardrum, since the sound wave pressures differ from the atmospheric pressure found behind the eardrum. A complicated mechanism converts the vibrations to nerve impulses, which are perceived by the person.

Section Summary

- Sound is a disturbance of matter that is transmitted from its source outward.
- Sound is one type of wave.
- Hearing is the perception of sound.

Glossary

audible range
frequencies between 20 Hz and 20,000 Hz

infrasound
sound waves below audible frequencies (less than 20 Hz)

hearing
the perception of sound

sound
a disturbance of matter that is transmitted from its source outward

ultrasound
sound waves above audible frequencies (greater than 20,000 Hz)

This page titled 4.1.7: Sound is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- **5.7: Sound** by OpenStax is licensed CC BY 4.0.

4.1.8: Speed of Sound, Frequency, and Wavelength

Learning Objectives

- Define pitch.
- Describe the relationship between the speed of sound, its frequency, and its wavelength.
- Describe the effects on the speed of sound as it travels through various media.
- Describe the effects of temperature on the speed of sound.



Figure 4.1.8.1: When a firework explodes, the light energy is perceived before the sound energy. Sound travels more slowly than light does. (credit: Dominic Alves, Flickr)

Sound, like all waves, travels at a certain speed and has the properties of frequency and wavelength. You can observe direct evidence of the speed of sound while watching a fireworks display. The flash of an explosion is seen well before its sound is heard, implying both that sound travels at a finite speed and that it is much slower than light. You can also directly sense the frequency of a sound. Perception of frequency is called **pitch**. The wavelength of sound is not directly sensed, but indirect evidence is found in the correlation of the size of musical instruments with their pitch. Small instruments, such as a piccolo, typically make high-pitch sounds, while large instruments, such as a tuba, typically make low-pitch sounds. High pitch means small wavelength, and the size of a musical instrument is directly related to the wavelengths of sound it produces. So a small instrument creates short-wavelength sounds. Similar arguments hold that a large instrument creates long-wavelength sounds.

The relationship of the speed of sound, its frequency, and wavelength is the same as for all waves:

$$v_{\text{w}} = f\lambda,$$

where v_{w} is the speed of sound, f is its frequency, and λ is its wavelength. The wavelength of a sound is the distance between adjacent identical parts of a wave—for example, between adjacent compressions as illustrated in Figure 4.1.8.2. The frequency is the same as that of the source and is the number of waves that pass a point per unit time.

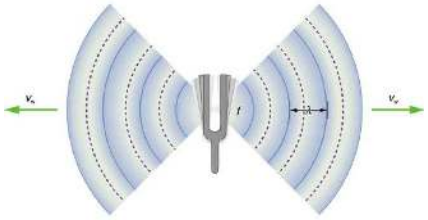


Figure 4.1.8.2: A sound wave emanates from a source vibrating at a frequency f, propagates at v_{w}, and has a wavelength λ.

Table 4.1.8.1 makes it apparent that the speed of sound varies greatly in different media. The speed of sound in a medium is determined by a combination of the medium's rigidity (or compressibility in gases) and its density. The more rigid (or less compressible) the medium, the faster the speed of sound. This observation is analogous to the fact that the frequency of a simple harmonic motion is directly proportional to the stiffness of the oscillating object. The greater the density of a medium, the slower the speed of sound. This observation is analogous to the fact that the frequency of a simple harmonic motion is inversely proportional to the mass of the oscillating object. The speed of sound in air is low, because air is compressible. Because liquids and solids are relatively rigid and very difficult to compress, the speed of sound in such media is generally greater than in gases.

Table 4.1.8.1: Speed of Sound in Various Media

Medium	v_{w}(m/s)
Gases at 0ºC	

Medium	v_w(m/s)
Air	331
Carbon dioxide	259
Oxygen	316
Helium	965
Hydrogen	1290
Liquids at 20ºC	
Ethanol	1160
Mercury	1450
Water, fresh	1480
Sea water	1540
Human tissue	1540
Solids (longitudinal or bulk)	
Vulcanized rubber	54
Polyethylene	920
Marble	3810
Glass, Pyrex	5640
Lead	1960
Aluminum	5120
Steel	5960

Earthquakes, essentially sound waves in Earth's crust, are an interesting example of how the speed of sound depends on the rigidity of the medium. Earthquakes have both longitudinal and transverse components, and these travel at different speeds. The bulk modulus of granite is greater than its shear modulus. For that reason, the speed of longitudinal or pressure waves (P-waves) in earthquakes in granite is significantly higher than the speed of transverse or shear waves (S-waves). Both components of earthquakes travel slower in less rigid material, such as sediments. P-waves have speeds of 4 to 7 km/s, and S-waves correspondingly range in speed from 2 to 5 km/s, both being faster in more rigid material. The P-wave gets progressively farther ahead of the S-wave as they travel through Earth's crust. The time between the P- and S-waves is routinely used to determine the distance to their source, the epicenter of the earthquake.

The speed of sound is affected by temperature in a given medium. For air at sea level, the speed of sound is given by

$$v_w = (331 \text{ m/s})\sqrt{\frac{T}{273 \text{ K}}},$$

where the temperature (denoted as T) is in units of kelvin. While not negligible, this is not a strong dependence. At 0ºC, the speed of sound is 331 m/s, whereas at 20.0ºC it is 343 m/s, less than a 4% increase. Figure 4.1.8.3 shows a use of the speed of sound by a bat to sense distances. Echoes are also used in medical imaging.



Figure 4.1.8.3: A bat uses sound echoes to find its way about and to catch prey. The time for the echo to return is directly proportional to the distance.

One of the more important properties of sound is that its speed is nearly independent of frequency. This independence is certainly true in open air for sounds in the audible range of 20 to 20,000 Hz. If this independence were not true, you would certainly notice it for music played by a marching band in a football stadium, for example. Suppose that high-frequency sounds traveled faster—then the farther you were from the band, the more the sound from the low-pitch instruments would lag that from the high-pitch ones. But the music from all instruments arrives in cadence independent of distance, and so all frequencies must travel at nearly the same speed. Recall that

$$v_{\text{w}} = f\lambda.$$

In a given medium under fixed conditions, v_{w} is constant, so that there is a relationship between f and λ; the higher the frequency, the smaller the wavelength. See Figure 4.1.8.4 and consider the following example.

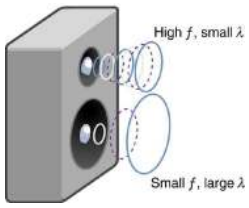


Figure 4.1.8.4: Because they travel at the same speed in a given medium, low-frequency sounds must have a greater wavelength than high-frequency sounds. Here, the lower-frequency sounds are emitted by the large speaker, called a woofer, while the higher-frequency sounds are emitted by the small speaker, called a tweeter.

Example 4.1.8.1: Calculating Wavelengths: What Are the Wavelengths of Audible Sounds?

Calculate the wavelengths of sounds at the extremes of the audible range, 20 and 20,000 Hz, in 30.0ºC air. (Assume that the frequency values are accurate to two significant figures.)

Strategy

To find wavelength from frequency, we can use $v_{\text{w}} = f\lambda$.

Solution

1. Identify knowns. The value for v_{w}, is given by

$$v_{\text{w}} = (331 \text{ m/s})\sqrt{\frac{T}{273 \text{ K}}}.$$

2. Convert the temperature into kelvin and then enter the temperature into the equation

$$v_{\text{w}} = (331 \text{ m/s})\sqrt{\frac{303 \text{ K}}{273 \text{ K}}} = 348.7 \text{ m/s}.$$

3. Solve the relationship between speed and wavelength for λ:

$$\lambda = \frac{v_{\text{w}}}{f}.$$

4. Enter the speed and the minimum frequency to give the maximum wavelength:

$$348.7 \text{ m/s}$$

$$\lambda_{\text{max}} = \frac{348.7 \text{ m/s}}{20 \text{ Hz}} = 17 \text{ m}.$$

5. Enter the speed and the maximum frequency to give the minimum wavelength:

$$\lambda_{\text{min}} = \frac{348.7 \text{ m/s}}{20,000 \text{ Hz}} = 0.017 \text{ m} = 1.7 \text{ cm}.$$

Discussion

Because the product of f multiplied by λ equals a constant, the smaller f is, the larger λ must be, and vice versa.

The speed of sound can change when sound travels from one medium to another. However, the frequency usually remains the same because it is like a driven oscillation and has the frequency of the original source. If v_{w} changes and f remains the same, then the wavelength λ must change. That is, because $v_{\text{w}} = f\lambda$, the higher the speed of a sound, the greater its wavelength for a given frequency.

MAKING CONNECTIONS: TAKE-HOME INVESTIGATION—VOICE AS A SOUND WAVE

Suspend a sheet of paper so that the top edge of the paper is fixed and the bottom edge is free to move. You could tape the top edge of the paper to the edge of a table. Gently blow near the edge of the bottom of the sheet and note how the sheet moves. Speak softly and then louder such that the sounds hit the edge of the bottom of the paper, and note how the sheet moves. Explain the effects.

Exercise 4.1.8.1

Imagine you observe two fireworks explode. You hear the explosion of one as soon as you see it. However, you see the other firework for several milliseconds before you hear the explosion. Explain why this is so.

Answer

Sound and light both travel at definite speeds. The speed of sound is slower than the speed of light. The first firework is probably very close by, so the speed difference is not noticeable. The second firework is farther away, so the light arrives at your eyes noticeably sooner than the sound wave arrives at your ears.

Exercise 4.1.8.2

You observe two musical instruments that you cannot identify. One plays high-pitch sounds and the other plays low-pitch sounds. How could you determine which is which without hearing either of them play?

Answer

Compare their sizes. High-pitch instruments are generally smaller than low-pitch instruments because they generate a smaller wavelength.

Section Summary

The relationship of the speed of sound v_{w}, its frequency f, and its wavelength λ is given by

$$v_{\text{w}} = f\lambda,$$

which is the same relationship given for all waves.

In air, the speed of sound is related to air temperature T by

$$v_{\text{w}} = (331 \text{ m/s})\sqrt{\frac{T}{273 \text{ K}}}.$$

v_{w} is the same for all frequencies and wavelengths.

Glossary

pitch

the perception of the frequency of a sound

This page titled 4.1.8: Speed of Sound, Frequency, and Wavelength is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- **5.8: Speed of Sound, Frequency, and Wavelength** by OpenStax is licensed CC BY 4.0.

4.1.9: Doppler Effect and Sonic Booms

Learning Objectives

- Define Doppler effect, Doppler shift, and sonic boom.
- Describe the sounds produced by objects moving faster than the speed of sound.

The characteristic sound of a motorcycle buzzing by is an example of the Doppler effect. The high-pitch scream shifts dramatically to a lower-pitch roar as the motorcycle passes by a stationary observer. The closer the motorcycle brushes by, the more abrupt the shift. The faster the motorcycle moves, the greater the shift. We also hear this characteristic shift in frequency for passing race cars, airplanes, and trains. It is so familiar that it is used to imply motion and children often mimic it in play.

The Doppler effect is an alteration in the observed frequency of a sound due to motion of either the source or the observer. Although less familiar, this effect is easily noticed for a stationary source and moving observer. For example, if you ride a train past a stationary warning bell, you will hear the bell's frequency shift from high to low as you pass by. This change in frequency due to *relative* motion of source and observer is called a **Doppler shift**. The Doppler effect is named for the Austrian physicist and mathematician Christian Johann Doppler (1803–1853), who did experiments with both moving sources and moving observers. Doppler, for example, had musicians play on a moving open train car and also play standing next to the train tracks as a train passed by. Their music was observed both on and off the train, and changes in frequency were measured.

What causes the Doppler shift? Figure 4.1.9.1, Figure 4.1.9.2, and Figure 4.1.9.3 compare sound waves emitted by stationary and moving sources in a stationary air mass. Each disturbance spreads out spherically from the point where the sound was emitted. If the source is stationary, then all of the spheres representing the air compressions in the sound wave centered on the same point, and the stationary observers on either side see the same wavelength and frequency as emitted by the source, as in Figure 4.1.9.1. If the source is moving, as in Figure 4.1.9.2, then the situation is different. Each compression of the air moves out in a sphere from the point where it was emitted, but the point of emission moves. This moving emission point causes the air compressions to be closer together on one side and farther apart on the other. Thus, the wavelength is shorter in the direction the source is moving (on the right in Figure 4.1.9.2), and longer in the opposite direction (on the left in Figure 4.1.9.2). Finally, if the observers move, as in Figure 4.1.9.3, the frequency at which they receive the compressions changes. The observer moving toward the source receives them at a higher frequency, and the person moving away from the source receives them at a lower frequency.



Figure 4.1.9.1: Sounds emitted by a source spread out in spherical waves. Because the source, observers, and air are stationary, the wavelength and frequency are the same in all directions and to all observers.

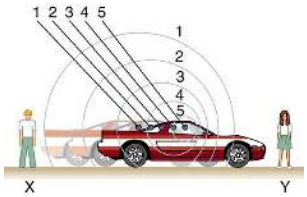


Figure 4.1.9.2: Sounds emitted by a source moving to the right spread out from the points at which they were emitted. The wavelength is reduced and, consequently, the frequency is increased in the direction of motion, so that the observer on the right hears a higher-pitch sound. The opposite is true for the observer on the left, where the wavelength is increased and the frequency is reduced.

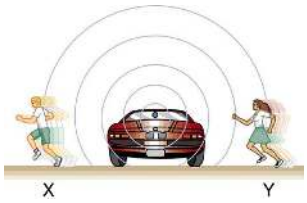


Figure 4.1.9.3: The same effect is produced when the observers move relative to the source. Motion toward the source increases frequency as the observer on the right passes through more wave crests than she would if stationary. Motion away from the source decreases frequency as the observer on the left passes through fewer wave crests than he would if stationary.

We know that wavelength and frequency are related by $v_w = f\lambda$, where v_w is the fixed speed of sound. The sound moves in a medium and has the same speed v_w in that medium whether the source is moving or not. Thus f multiplied by λ is a constant. Because the observer on the right in Figure 4.1.9.2 receives a shorter wavelength, the frequency she receives must be higher. Similarly, the observer on the left receives a longer wavelength, and hence he hears a lower frequency. The same thing happens in Figure 4.1.9.3. A higher frequency is received by the observer moving toward the source, and a lower frequency is received by an observer moving away from the source. In general, then, relative motion of source and observer toward one another increases the received frequency. Relative motion apart decreases frequency. The greater the relative speed is, the greater the effect.

THE DOPPLER EFFECT

The Doppler effect occurs not only for sound but for any wave when there is relative motion between the observer and the source. There are Doppler shifts in the frequency of sound, light, and water waves, for example. Doppler shifts can be used to determine velocity, such as when ultrasound is reflected from blood in a medical diagnostic. The recession of galaxies is determined by the observed frequency shift of light received from them measured against the frequency emitted by atoms in the laboratory. This implies much about the origins of the universe. Modern physics has been profoundly affected by observations of Doppler shifts.

Sonic Booms to Bow Wakes

What happens to the sound produced by a moving source, such as a jet airplane, that approaches or even exceeds the speed of sound? The answer to this question applies not only to sound but to all other waves as well.

Suppose a jet airplane is coming nearly straight at you, emitting a sound of frequency f_s. The greater the plane's speed v_s, the greater the Doppler shift and the greater the value observed for f_{obs} . Now, as v_s approaches the speed of sound, f_{obs} approaches infinity. At the speed of sound, this result means that in front of the source, each successive wave is superimposed on the previous one because the source moves forward at the speed of sound. The observer gets them all at the same instant, and so the frequency is infinite. (Before airplanes exceeded the speed of sound, some people argued it would be impossible because such constructive superposition would produce pressures great enough to destroy the airplane.) If the source exceeds the speed of sound, no sound is received by the observer until the source has passed, so that the sounds from the approaching source are mixed with those from it when receding. This mixing appears messy, but something interesting happens—a sonic boom is created. (See Figure 4.1.9.4.)

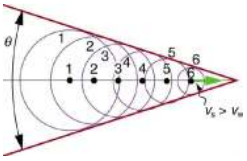


Figure 4.1.9.4: Sound waves from a source that moves faster than the speed of sound spread spherically from the point where they are emitted, but the source moves ahead of each. Constructive interference along the lines shown (actually a cone in three dimensions) creates a shock wave called a sonic boom. The faster the speed of the source, the smaller the angle θ.

There is constructive interference along the lines shown (a cone in three dimensions) from similar sound waves arriving there simultaneously. This superposition forms a disturbance called a **sonic boom**, a constructive interference of sound created by an object moving faster than sound. Inside the cone, the interference is mostly destructive, and so the sound intensity there is much less than on the shock wave. An aircraft creates two sonic booms, one from its nose and one from its tail. (See Figure 4.1.9.5.) During television coverage of space shuttle landings, two distinct booms could often be heard. These were separated by exactly the time it would take the shuttle to pass by a point. Observers on the ground often do not see the aircraft creating the sonic boom, because it has passed by before the shock wave reaches them, as seen in Figure 4.1.9.5. If the aircraft flies close by at low altitude,

pressures in the sonic boom can be destructive and break windows as well as rattle nerves. Because of how destructive sonic booms can be, supersonic flights are banned over populated areas of the United States.

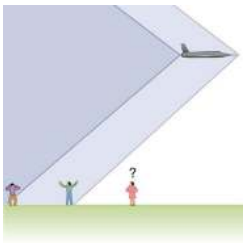


Figure 4.1.9.5: Two sonic booms, created by the nose and tail of an aircraft, are observed on the ground after the plane has passed by.

Sonic booms are one example of a broader phenomenon called bow wakes. A **bow wake**, such as the one in Figure 4.1.9.6, is created when the wave source moves faster than the wave propagation speed. Water waves spread out in circles from the point where created, and the bow wake is the familiar V-shaped wake trailing the source. A more exotic bow wake is created when a subatomic particle travels through a medium faster than the speed of light travels in that medium (the speed of light in vacuum is $c = 2.998 \times 10^8$ m/s, and this is the maximum speed anything can move, but in water, for example, speed of light slows to $0.75c$, and other particles can move faster). If the particle creates light in its passage, that light spreads on a cone with an angle indicative of the speed of the particle, as illustrated in Figure 4.1.9.7. Such a bow wake is called Cherenkov radiation and is commonly observed in particle physics.



Figure 4.1.9.6: Bow wake created by a duck. Constructive interference produces the rather structured wake, while there is relatively little wave action inside the wake, where interference is mostly destructive. (credit: Horia Varlan, Flickr)



Figure 4.1.9.7: The blue glow in this research reactor pool is Cherenkov radiation caused by subatomic particles traveling faster than the speed of light in water. (credit: U.S. Nuclear Regulatory Commission)

Doppler shifts and sonic booms are interesting sound phenomena that occur in all types of waves. They can be of considerable use. For example, the Doppler shift in ultrasound can be used to measure blood velocity, while police use the Doppler shift in radar (a microwave) to measure car velocities. In meteorology, the Doppler shift is used to track the motion of storm clouds; such "Doppler Radar" can give velocity and direction and rain or snow potential of imposing weather fronts. In astronomy, we can examine the light emitted from distant galaxies and determine their speed relative to ours. As galaxies move away from us, their light is shifted to a lower frequency, and so to a longer wavelength—the so-called red shift. Such information from galaxies far, far away has allowed us to estimate the age of the universe (from the Big Bang) as about 14 billion years.

Exercise 4.1.9.1

Why did scientist Christian Doppler observe musicians both on a moving train and also from a stationary point not on the train?

Answer

Doppler needed to compare the perception of sound when the observer is stationary and the sound source moves, as well as when the sound source and the observer are both in motion.

Exercise 4.1.9.2

Describe a situation in your life when you might rely on the Doppler shift to help you either while driving a car or walking near traffic.

Answer

If I am driving and I hear Doppler shift in an ambulance siren, I would be able to tell when it was getting closer and also if it has passed by. This would help me to know whether I needed to pull over and let the ambulance through

Section Summary

- The Doppler shift is change in observed frequency of a wave (for example, a sound wave) due to relative motion of the source and the observer, with approaching source and/or observer increasing the observed frequency and receding source and/or observer decreasing the observed frequency.
- A sonic boom is constructive interference of sound created by an object moving faster than sound.
- A shock wave is a type of bow wake created when any wave source moves faster than the wave propagation speed.

Glossary

Doppler shift
the change in wave frequency due to relative motion of source and observer

sonic boom
a constructive interference of sound created by an object moving faster than sound

bow wake
V-shaped disturbance created when the wave source moves faster than the wave propagation speed

This page titled 4.1.9: Doppler Effect and Sonic Booms is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- **5.9: Doppler Effect and Sonic Booms** by OpenStax is licensed CC BY 4.0.

4.1.E: Oscillations and Waves (Exercise)

Conceptual Questions

5.1: Period and Frequency in Oscillations

1. If an oscillation occurs more rapidly, how does the period of oscillation change? How does the frequency change?

2. Give examples of natural phenomena that occur with constant frequency (or period) over a length of time.

5.2: Simple Harmonic Motion- A Special Periodic Motion

3. What conditions must be met to produce simple harmonic motion?

4. (a) If frequency is not constant for some oscillation, can the oscillation be simple harmonic motion?

(b) Can you think of any examples of harmonic motion where the frequency may depend on the amplitude?

5. Give an example of a simple harmonic oscillator, specifically noting how its frequency is independent of amplitude.

6. Explain why you expect an object made of a stiff material to vibrate at a higher frequency than a similar object made of a spongy material.

7. As you pass a freight truck with a trailer on a highway, you notice that its trailer is bouncing up and down slowly. Is it more likely that the trailer is heavily loaded or nearly empty? Explain your answer.

8. Some people modify cars to be much closer to the ground than when manufactured. Should they install stiffer springs? Explain your answer.

5.3: Forced Oscillations and Resonance

9. Give additional examples of resonance phenomenon from everyday life.

10. What "damps" the resonance of a forced oscillation? What are the dynamics of "damping" and what physical quantity is involved?

5.4: Waves

11. Give one example of a transverse wave and another of a longitudinal wave, being careful to note the relative directions of the disturbance and wave propagation in each.

12. What is the difference between propagation speed and the frequency of a wave? Does one or both affect wavelength? If so, how?

13. Given a wave of frequency ff traveling with wave speed v_{w} , what happens to the wave speed if the frequency doubles? What happens to the wavelength?

5.5: Wave Interference- Standing Waves and Beats

14. How does destructive interference occur? What do you see in a destructive interference?

15. Explain how noise-cancelling headphones work in terms of wave interference phenomena.

16. Explain how nodes and antinodes form on standing waves.

17. In the beat phenomenon, the loudness of the sound combined from two sound waves "wobbles" (it increases and decreases over time). Explain how the loudness of the combined sound becomes louder then quieter, even though individual sound waves remain at the same loudness, or amplitude.

5.7: Speed of Sound, Frequency, and Wavelength

18. How do sound vibrations of atoms differ from thermal motion?

19. When sound passes from one medium to another where its propagation speed is different, does its frequency or wavelength change? Explain your answer briefly.

20. For a high-fidelity speaker containing a woofer and a tweeter (see Figure 4.1.E.1 for example), why is the woofer larger than the tweeter?

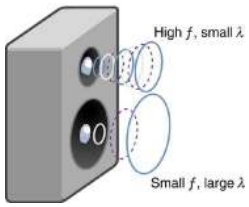


Figure 4.1.*E*.1: Because they travel at the same speed in a given medium, low-frequency sounds must have a greater wavelength than high-frequency sounds. Here, the lower-frequency sounds are emitted by the large speaker, called a woofer, while the higher-frequency sounds are emitted by the small speaker, called a tweeter.

5.8: Doppler Effect and Sonic Booms

21. Explain Doppler shift. How does the measured frequency increase or decrease?

22. Due to efficiency considerations related to its bow wake, the supersonic transport aircraft must maintain a cruising speed that is a constant ratio to the speed of sound (a constant Mach number). If the aircraft flies from warm air into colder air, should it increase or decrease its speed? Explain your answer.

23. When you hear a sonic boom, you often cannot see the plane that made it. Why is that?

This page titled 4.1.E: Oscillations and Waves (Exercise) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- **5.E: Oscillations and Waves (Exercise)** by OpenStax is licensed CC BY 4.0.

SECTION OVERVIEW

4.2: Light

4.2.1: Introduction to Light

4.2.2: Maxwell's Equations- Electromagnetic Waves Predicted and Observed

4.2.3: Production and Properties of Electromagnetic Waves

4.2.4: The Electromagnetic Spectrum- an Overview

4.2.5: The Electromagnetic Spectrum- Application Notes

4.2.6: Reflection

4.2.7: Refraction

4.2.8: Dispersion- The Rainbow and Prisms

4.2.9: Image Formation by Mirrors

4.2.10: Polarization

This page titled 4.2: Light is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

4.2.1: Introduction to Light

Light from this page or screen is formed into an image by the lens of your eye, much as the lens of the camera that made this photograph. Mirrors, like lenses, can also form images that in turn are captured by your eye.



Figure 4.2.1.1: Image seen as a result of reflection of light on a plane smooth surface. (credit: NASA Goddard Photo and Video, via Flickr)

Our lives are filled with light. Through vision, the most valued of our senses, light can evoke spiritual emotions, such as when we view a magnificent sunset or glimpse a rainbow breaking through the clouds. Light can also simply amuse us in a theater, or warn us to stop at an intersection. It has innumerable uses beyond vision. Light can carry telephone signals through glass fibers or cook a meal in a solar oven. Life itself could not exist without light's energy. From photosynthesis in plants to the sun warming a cold-blooded animal, its supply of energy is vital.



Figure 4.2.1.2: Double Rainbow over the bay of Pocitos in Montevideo, Uruguay. (credit: Madrax, Wikimedia Commons)

We will start our discussion of visible light as a type of electromagnetic wave. This knowledge will help us answer questions regarding the nature of light and vision. What is color, and how do our eyes detect it? Why do diamonds sparkle? How does light travel? How do lenses and mirrors form images? These are but a few of the questions that are answered by the study of optics. Optics is the branch of physics that deals with the behavior of visible light and other electromagnetic waves.

It is convenient to divide optics into two major parts based on the features we are interested in. The wave characteristics of light, such as frequency and wavelength, relates to the colors we perceive and to how we characterize different types of electromagnetic wave along the electromagnetic spectrum. The wave nature of light is also responsible for phenomena such as diffraction and interference. We call this part of optics "wave optics" or "physical optics." But when light interacts with an object that is several times as large as the light's wavelength, its observable behavior is like that of a ray; it does not prominently display its wave characteristics, and we can look at primarily how it refracts and reflects. We call this part of optics "geometric optics" or "ray optics."

This page titled 4.2.1: Introduction to Light is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- **11.1: Introduction to Light** by OpenStax is licensed CC BY 4.0.

4.2.2: Maxwell's Equations- Electromagnetic Waves Predicted and Observed

Learning Objectives

- Restate Maxwell's equations.

The Scotsman James Clerk Maxwell (1831–1879) is regarded as the greatest theoretical physicist of the 19th century. (See Figure 4.2.2.1). Although he died young, Maxwell not only formulated a complete electromagnetic theory, represented by **Maxwell's equations**, he also developed the kinetic theory of gases and made significant contributions to the understanding of color vision and the nature of Saturn's rings.



Figure 4.2.2.1: James Clerk Maxwell, a 19th-century physicist, developed a theory that explained the relationship between electricity and magnetism and correctly predicted that visible light is caused by electromagnetic waves. (credit: G. J. Stodart)

Maxwell brought together all the work that had been done by brilliant physicists such as Oersted, Coulomb, Gauss, and Faraday, and added his own insights to develop the overarching theory of electromagnetism. The results of Maxwell's equations are paraphrased here in words because their mathematical statement is beyond the level of this text. However, the equations illustrate how apparently simple mathematical statements can elegantly unite and express a multitude of concepts—why mathematics is the language of science.

THE RESULTS OF MAXWELL'S EQUATIONS

1. Electric field lines originate on positive charges and terminate on negative charges. The electric field is defined as the force per unit charge on a test charge.
2. Magnetic field lines are continuous, having no beginning or end. No magnetic monopoles are known to exist.
3. A changing magnetic field induces an electric field.
4. Magnetic fields are generated by moving charges or by changing electric fields.

Maxwell's equations encompass the major laws of electricity and magnetism. What is not so apparent is the symmetry that Maxwell introduced in his mathematical framework. Especially important is his addition of the hypothesis that changing electric fields create magnetic fields. This is exactly analogous (and symmetric) to Faraday's law of induction and had been suspected for some time, but fits beautifully into Maxwell's equations.

MAKING CONNECTIONS: UNIFICATION OF FORCES

Maxwell's complete and symmetric theory showed that electric and magnetic forces are not separate, but different manifestations of the same thing—the electromagnetic force. This classical unification of forces is one motivation for current attempts to unify the four basic forces in nature—the gravitational, electrical, strong, and weak nuclear forces.

Since changing electric fields create relatively weak magnetic fields, they could not be easily detected at the time of Maxwell's hypothesis. Maxwell realized, however, that oscillating charges, like those in AC circuits, produce changing electric fields. He predicted that these changing fields would propagate from the source like waves generated on a lake by a jumping fish.

The waves predicted by Maxwell would consist of oscillating electric and magnetic fields—defined to be an electromagnetic wave (EM wave). Electromagnetic waves would be capable of exerting forces on charges great distances from their source, and they might thus be detectable. Maxwell calculated that electromagnetic waves would propagate at the speed of light,

$$c = 3.00 \times 10^8 \text{ m/s}.$$

In fact, Maxwell concluded that light is an electromagnetic wave having such wavelengths that it can be detected by the eye.

Other wavelengths should exist—it remained to be seen if they did. If so, Maxwell's theory and remarkable predictions would be verified, the greatest triumph of physics since Newton. Experimental verification came within a few years, but not before Maxwell's death.

Hertz's Observations

The German physicist Heinrich Hertz (1857–1894) was the first to generate and detect certain types of electromagnetic waves in the laboratory. Starting in 1887, he performed a series of experiments that not only confirmed the existence of electromagnetic waves, but also verified that they travel at the speed of light.

Hertz used an AC circuit that resonates at a known frequency and connected it to a loop of wire as shown in Figure 4.2.2.2. High voltages induced across the gap in the loop produced sparks that were visible evidence of the current in the circuit and that helped generate electromagnetic waves.

Across the laboratory, Hertz had another loop attached to another circuit, which could be tuned (as the dial on a radio) to the same resonant frequency as the first and could, thus, be made to receive electromagnetic waves. This loop also had a gap across which sparks were generated, giving solid evidence that electromagnetic waves had been received.

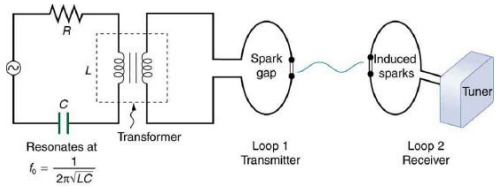


Figure 4.2.2.2: The apparatus used by Hertz in 1887 to generate and detect electromagnetic waves. An AC circuit connected to the first loop caused sparks across a gap in the wire loop and generated electromagnetic waves. Sparks across a gap in the second loop located across the laboratory gave evidence that the waves had been received.

Hertz also studied the reflection, refraction, and interference patterns of the electromagnetic waves he generated, verifying their wave character. He was able to determine wavelength from the interference patterns, and knowing their frequency, he could calculate the propagation speed using the equation $v = f\lambda$ (velocity—or speed—equals frequency times wavelength). Hertz was thus able to prove that electromagnetic waves travel at the speed of light. The SI unit for frequency, the hertz $(1\ \text{Hz} = 1\ \text{cycle}\ /\text{sec})$, is named in his honor.

Section Summary

- Electromagnetic waves consist of oscillating electric and magnetic fields and propagate at the speed of light c. They were predicted by Maxwell, who also showed that
- Maxwell's prediction of electromagnetic waves resulted from his formulation of a complete and symmetric theory of electricity and magnetism, known as Maxwell's equations.
- These four equations are paraphrased in this text, rather than presented numerically, and encompass the major laws of electricity and magnetism. First is Gauss's law for electricity, second is Gauss's law for magnetism, third is Faraday's law of induction, including Lenz's law, and fourth is Ampere's law in a symmetric formulation that adds another source of magnetism—changing electric fields.

Glossary

electromagnetic waves

radiation in the form of waves of electric and magnetic energy

Maxwell's equations

a set of four equations that comprise a complete, overarching theory of electromagnetism

hertz

an SI unit denoting the frequency of an electromagnetic wave, in cycles per second

speed of light

in a vacuum, such as space, the speed of light is a constant 3 x 10^8 m/s

electric field lines

a pattern of imaginary lines that extend between an electric source and charged objects in the surrounding area, with arrows pointed away from positively charged objects and toward negatively charged objects. The more lines in the pattern, the stronger the electric field in that region

magnetic field lines

a pattern of continuous, imaginary lines that emerge from and enter into opposite magnetic poles. The density of the lines indicates the magnitude of the magnetic field

---

This page titled 4.2.2: Maxwell's Equations- Electromagnetic Waves Predicted and Observed is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- **11.2: Maxwell's Equations- Electromagnetic Waves Predicted and Observed** by OpenStax is licensed CC BY 4.0.

4.2.3: Production and Properties of Electromagnetic Waves

Learning Objectives

- Describe the electric and magnetic waves as they move out from a source, such as an AC generator.
- Describe the momentum carried by electromagnetic waves and its relationship to radiation pressure.

We can get a good understanding of **electromagnetic waves** (EM) by considering how they are produced. Whenever a current varies, associated electric and magnetic fields vary, moving out from the source like waves. Perhaps the easiest situation to visualize is a varying current in a long straight wire, produced by an AC generator at its center, as illustrated in Figure 4.2.3.1.

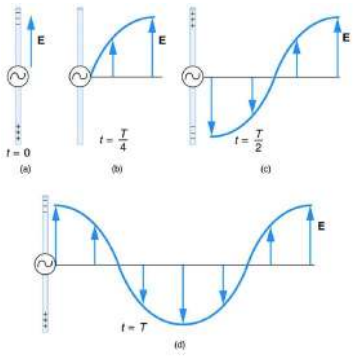


Figure 4.2.3.1: This long straight gray wire with an AC generator at its center becomes a broadcast antenna for electromagnetic waves. Shown here are the charge distributions at four different times. The electric field (**E**) propagates away from the antenna at the speed of light, forming part of an electromagnetic wave.

The **electric field** (**E**) shown surrounding the wire is produced by the charge distribution on the wire. Both the **E** and the charge distribution vary as the current changes. The changing field propagates outward at the speed of light.

There is an associated **magnetic field** (B) which propagates outward as well (see Figure 4.2.3.2). The electric and magnetic fields are closely related and propagate as an electromagnetic wave. This is what happens in broadcast antennae such as those in radio and TV stations.

Closer examination of the one complete cycle shown in Figure 4.2.3.1 reveals the periodic nature of the generator-driven charges oscillating up and down in the antenna and the electric field produced. At time $t = 0$, there is the maximum separation of charge, with negative charges at the top and positive charges at the bottom, producing the maximum magnitude of the electric field (or E-field) in the upward direction. One-fourth of a cycle later, there is no charge separation and the field next to the antenna is zero, while the maximum E-field has moved away at speed c.

As the process continues, the charge separation reverses and the field reaches its maximum downward value, returns to zero, and rises to its maximum upward value at the end of one complete cycle. The outgoing wave has an **amplitude** proportional to the maximum separation of charge. Its **wavelength** (λ) is proportional to the period of the oscillation and, hence, is smaller for short periods or high frequencies. (As usual, wavelength and **frequency** (f) are inversely proportional.)

Electric and Magnetic Waves: Moving Together

Following Ampere's law, current in the antenna produces a magnetic field, as shown in Figure 4.2.3.2. The relationship between **E** and B is shown at one instant in Figure 4.2.3.2(a). As the current varies, the magnetic field varies in magnitude and direction.

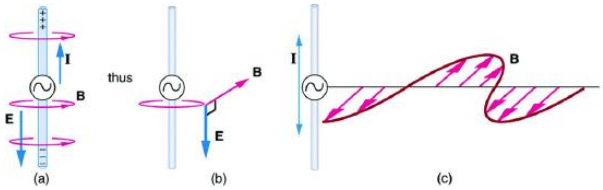


Figure 4.2.3.2: (a) The current in the antenna produces the circular magnetic field lines. The current (I) produces the separation of charge along the wire, which in turn creates the electric field as shown. (b) The electric and magnetic fields (**E** and B) near the wire are perpendicular; they are shown here for one point in space. (c) The magnetic field varies with current and propagates away from the antenna at the speed of light.

The magnetic field lines also propagate away from the antenna at the speed of light, forming the other part of the electromagnetic wave, as seen in Figure 4.2.3.2(b). The magnetic part of the wave has the same period and wavelength as the electric part, since they are both produced by the same movement and separation of charges in the antenna.

The electric and magnetic waves are shown together at one instant in time in Figure 4.2.3.3. The electric and magnetic fields produced by a long straight wire antenna are exactly in phase. Note that they are perpendicular to one another and to the direction of propagation, making this a **transverse wave**.

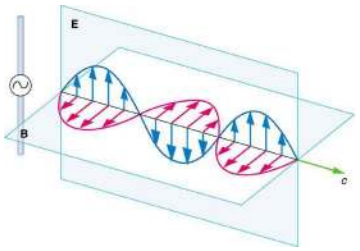


Figure 4.2.3.3: A part of the electromagnetic wave sent out from the antenna at one instant in time. The electric and magnetic fields (**E** and **B**) are in phase, and they are perpendicular to one another and the direction of propagation. For clarity, the waves are shown only along one direction, but they propagate out in other directions too.

Electromagnetic waves generally propagate out from a source in all directions, sometimes forming a complex radiation pattern. A linear antenna like this one will not radiate parallel to its length, for example. The wave is shown in one direction from the antenna in Figure 4.2.3.3 to illustrate its basic characteristics.

Instead of the AC generator, the antenna can also be driven by an AC circuit. In fact, charges radiate whenever they are accelerated. But while a current in a circuit needs a complete path, an antenna has a varying charge distribution forming a **standing wave**, driven by the AC. The dimensions of the antenna are critical for determining the frequency of the radiated electromagnetic waves. This is a **resonant** phenomenon and when we tune radios or TV, we vary electrical properties to achieve appropriate resonant conditions in the antenna.

Receiving Electromagnetic Waves

Electromagnetic waves carry energy away from their source, similar to a sound wave carrying energy away from a standing wave on a guitar string. An antenna for receiving EM signals works in reverse. And like antennas that produce EM waves, receiver antennas are specially designed to resonate at particular frequencies.

An incoming electromagnetic wave accelerates electrons in the antenna, setting up a standing wave. If the radio or TV is switched on, electrical components pick up and amplify the signal formed by the accelerating electrons. The signal is then converted to audio and/or video format. Sometimes big receiver dishes are used to focus the signal onto an antenna.

In fact, charges radiate whenever they are accelerated. When designing circuits, we often assume that energy does not quickly escape AC circuits, and mostly this is true. A broadcast antenna is specially designed to enhance the rate of electromagnetic radiation, and shielding is necessary to keep the radiation close to zero. Some familiar phenomena are based on the production of electromagnetic waves by varying currents. Your microwave oven, for example, sends electromagnetic waves, called microwaves, from a concealed antenna that has an oscillating current imposed on it.

Momentum Carried by Electromagnetic Waves

As electromagnetic waves are received, as described above, the forces exerted on the charged particles (see [link][1] for diagrams) do work on the particles and increase its energy. The energy that sunlight carries is a familiar part of every warm sunny day. A much less familiar feature of electromagnetic radiation is the extremely weak pressure that electromagnetic radiation produces by exerting a force in the direction of the wave. This force occurs because electromagnetic waves contain and transport momentum.

To understand the direction of the force for a very specific case, consider a plane electromagnetic wave incident on a metal. The electrons in the metal move with the velocity $\boldsymbol{v}$ that is proportional to the force on it due to applied electric force (in opposite direction to $\boldsymbol{E}$ because electron is negatively charged). There are other, frictional forces in the metal that make this motion happen (this is the origin of Ohm's law in electric circuits). Note that there is also a magnetic field in the electromagnetic wave (Figure 4.2.3.4). Using the right-hand rule (magnetic force is $\boldsymbol{F} = q(\boldsymbol{v} \times \boldsymbol{B})$), we can get the direction of force on the electron (remember that electron is negatively charged), which is shown in Figure 4.2.3.4. This force is in the same direction as the direction of wave

propagation, and it represents a transfer of momentum carried by the electromagnetic wave from the wave to the electron. This momentum is directly proportional to the energy carried by the electromagnetic wave and is given by,

$$p = \frac{E}{c},$$

where p is the magnitude of momentum of electromagnetic wave, E is the energy of electromagnetic wave, and c is the speed of light.

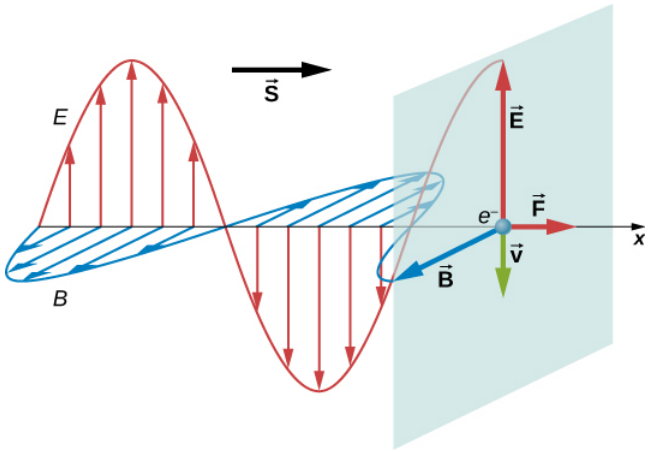


Figure 4.2.3.4: Electric and magnetic fields of an electromagnetic wave can combine to produce a force in the direction of propagation, as illustrated for the special case of electrons whose motion is highly damped by the resistance of a metal.

Maxwell predicted that an electromagnetic wave carries momentum. An object absorbing an electromagnetic wave would experience a force in the direction of propagation of the wave. The force corresponds to **radiation pressure** exerted on the object by the wave. The force would be twice as great if the radiation were reflected rather than absorbed, because the change in the momentum of electromagnetic wave itself is twice as great (and we enforce momentum conservation of the system as a whole).

Maxwell's prediction was confirmed in 1903 by Nichols and Hull by precisely measuring radiation pressures with a torsion balance. The schematic arrangement is shown in Figure 4.2.3.5. The mirrors suspended from a fiber were housed inside a glass container. Nichols and Hull were able to obtain a small measurable deflection of the mirrors from shining light on one of them. From the measured deflection, they could calculate the unbalanced force on the mirror, and obtained agreement with the predicted value of the force.

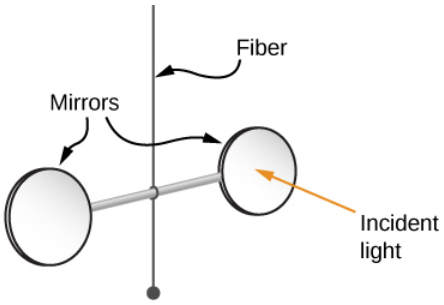


Figure 4.2.3.5: Simplified diagram of the central part of the apparatus Nichols and Hull used to precisely measure radiation pressure and confirm Maxwell's prediction.

Radiation pressure plays a role in explaining many observed astronomical phenomena, including the appearance of **comets**. Comets are basically chunks of icy material in which frozen gases and particles of rock and dust are embedded. When a comet approaches the Sun, it warms up and its surface begins to evaporate. The *coma* of the comet is the hazy area around it from the gases and dust. Some of the gases and dust form tails when they leave the comet. Notice in Figure 4.2.3.6 that a comet has *two* tails. The *ion tail* (or *gas tail* in Figure 4.2.3.6) is composed mainly of ionized gases. These ions interact electromagnetically with the solar wind, which is a continuous stream of charged particles emitted by the Sun. The force of the solar wind on the ionized gases is strong enough that the ion tail almost always points directly away from the Sun. The second tail is composed of dust particles. Because the *dust tail* is electrically neutral, it does not interact with the solar wind. However, this tail is affected by the radiation pressure

produced by the light from the Sun. Although quite small, this pressure is strong enough to cause the dust tail to be displaced from the path of the comet.

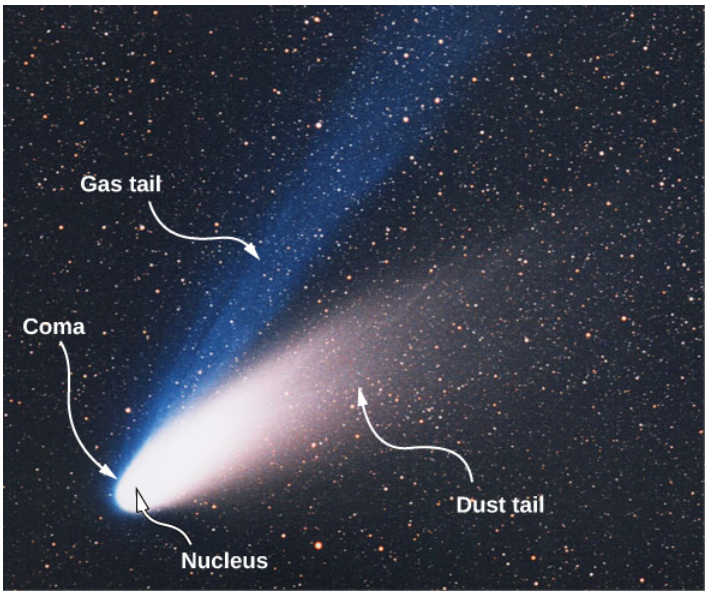


Figure 4.2.3.6: Evaporation of material being warmed by the Sun forms two tails, as shown in this photo of Comet Ison. (credit: modification of work by E. Slawik—ESO)

After Maxwell showed that light carried momentum as well as energy, a novel idea eventually emerged, initially only as science fiction. Perhaps a spacecraft with a large reflecting **solar sail** could use radiation pressure for propulsion. Such a vehicle would not have to carry fuel. It would experience a constant but small force from solar radiation, instead of the short bursts from rocket propulsion. It would accelerate slowly, but by being accelerated continuously, it would eventually reach great speeds. A spacecraft with small total mass and a sail with a large area would be necessary to obtain a usable acceleration.

The latest accomplishment on this concept is represented by the IKAROS (Interplanetary Kite-craft Accelerated by Radiation Of the Sun) project by Japan Aerospace Exploration Agency (JAXA), which was launched in May 2010. IKAROS solar sail is currently traveling through the solar system. During its six-month journey to Venus, acceleration and attitude control using the solar sail, rather than thrusters, were successfully tested. Last update from the spacecraft was in May 2014, on a ten-month orbit around the Sun. A scale model of the spacecraft is shown in Figure 4.2.3.7.

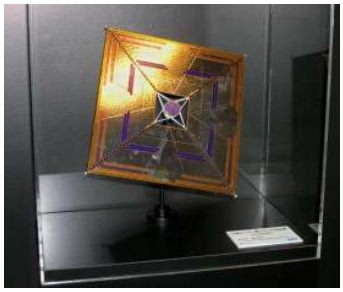


Figure 4.2.3.7: A 1:64 scale model of the IKAROS spacecraft at the 61st International Astronautical Congress in Prague, Czech Republic ("IKAROS model" By Pavel Hrdlička, Wikipedia, CC BY-SA 3.0)

Section Summary

- Electromagnetic waves are created by oscillating charges (which radiate whenever accelerated) and have the same frequency as the oscillation.
- Since the electric and magnetic fields in most electromagnetic waves are perpendicular to the direction in which the wave moves, it is ordinarily a transverse wave.
- Electromagnetic waves carry momentum and exert radiation pressure.
- The momentum of an electromagnetic wave is directly proportional to its energy.

Glossary

electric field
a vector quantity (**E**); the lines of electric force per unit charge, moving radially outward from a positive charge and in toward a negative charge

electric field strength
the magnitude of the electric field, denoted *E*-field

magnetic field
a vector quantity (**B**); can be used to determine the magnetic force on a moving charged particle

magnetic field strength
the magnitude of the magnetic field, denoted *B*-field

transverse wave
a wave, such as an electromagnetic wave, which oscillates perpendicular to the axis along the line of travel

standing wave
a wave that oscillates in place, with nodes where no motion happens

wavelength
the distance from one peak to the next in a wave

amplitude
the height, or magnitude, of an electromagnetic wave

frequency
the number of complete wave cycles (up-down-up) passing a given point within one second (cycles/second)

resonant
a system that displays enhanced oscillation when subjected to a periodic disturbance of the same frequency as its natural frequency

oscillate
to fluctuate back and forth in a steady beat

radiation pressure
pressure exerted by an electromagnetic wave on a surface

solar sail
a spacecraft that utilizes radiation pressure due to solar radiation in its propulsion

---

1 Section 16.2: Maxwell's Equations and Electromagnetic Waves
University Physics II - Thermodynamics, Electricity, and Magnetism

---

This page titled 4.2.3: Production and Properties of Electromagnetic Waves is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- **11.3: Production and Properties of Electromagnetic Waves** by OpenStax is licensed CC BY 4.0.

4.2.4: The Electromagnetic Spectrum- an Overview

Learning Objectives

- List three “rules of thumb” that apply to the different frequencies along the electromagnetic spectrum.
- Explain why the higher the frequency, the shorter the wavelength of an electromagnetic wave.
- Draw a simplified electromagnetic spectrum, indicating the relative positions, frequencies, and spacing of the different types of radiation bands.

In this module we examine how electromagnetic waves are classified into categories such as radio, infrared, ultraviolet, and so on, so that we can understand some of their similarities as well as some of their differences. We will also find that there are many connections with previously discussed topics, such as wavelength and resonance. A brief overview of the production and utilization of electromagnetic waves is found in Table 4.2.4.1. Note that the vast majority of the different types of electromagnetic waves originate from atomic and/or molecular electron transitions—that is, from electrons changing their energy levels within atoms or molecules.

Table 4.2.4.1: Electromagnetic Waves

Type of EM wave	Production	Applications	Life sciences aspect	Issues
Radio and TV	Accelerating charges	Communications, Remote controls	MRI	Requires controls for band use
Microwaves	Accelerating charges and thermal agitation	Communications, Ovens, Radar	Deep heating	Cell phone use
Infrared	Thermal agitations and atomic/molecular electron transitions	Thermal imaging, Heating	Absorbed by atmosphere	Greenhouse effect
Visible light	Thermal agitations and atomic/molecular electron transitions	All pervasive	Photosynthesis, Human vision	
Ultraviolet	Thermal agitations and atomic/molecular electron transitions	Sterilization, Cancer control	Vitamin D production	Ozone depletion, Cancer causing
X-rays	Inner atomic electron transitions and fast collisions	Medical, Security	Medical diagnosis, Cancer therapy	Cancer causing
Gamma rays	Nuclear decay	Nuclear medicine, Security	Medical diagnosis, Cancer therapy	Cancer causing, Radiation damage

CONNECTIONS: WAVES

There are many types of waves, such as water waves and even earthquakes. Among the many shared attributes of waves are propagation speed, frequency, and wavelength. These are always related by the expression $v_{\mathrm{W}} = f\lambda$. This module concentrates on EM waves, but other modules contain examples of all of these characteristics for sound waves and submicroscopic particles.

As noted before, an electromagnetic wave has a frequency and a wavelength associated with it and travels at the speed of light, or c. The relationship among these wave characteristics can be described by $v_{\mathrm{W}} = f\lambda$, where v_{W} is the propagation speed of the wave, f is the frequency, and λ is the wavelength. Here $v_{\mathrm{W}} = c$, so that for all electromagnetic waves,

$$c = f\lambda.$$

Thus, for all electromagnetic waves, the greater the frequency, the smaller the wavelength.

Figure 4.2.4.1 shows how the various types of electromagnetic waves are categorized according to their wavelengths and frequencies—that is, it shows the electromagnetic spectrum. Many of the characteristics of the various types of electromagnetic

waves are related to their frequencies and wavelengths, as we shall see.

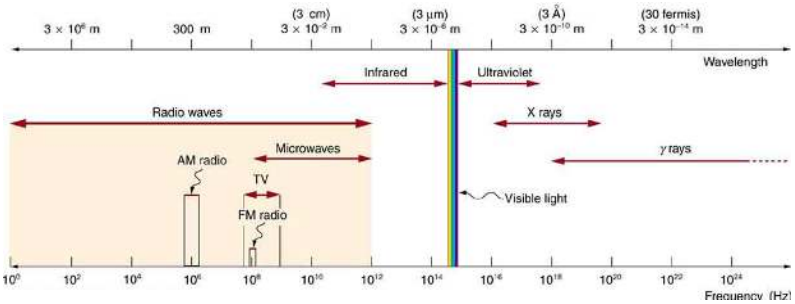


Figure 4.2.4.1: The electromagnetic spectrum, showing the major categories of electromagnetic waves. The range of frequencies and wavelengths is remarkable. The dividing line between some categories is distinct, whereas other categories overlap.

ELECTROMAGNETIC SPECTRUM: RULES OF THUMB

Three rules that apply to electromagnetic waves in general are as follows:

- High-frequency electromagnetic waves are more energetic and are more able to penetrate than low-frequency waves.
- High-frequency electromagnetic waves can carry more information per unit time than low-frequency waves.
- The shorter the wavelength of any electromagnetic wave probing a material, the smaller the detail it is possible to resolve.

Note that there are exceptions to these rules of thumb.

Transmission, Reflection, and Absorption

What happens when an electromagnetic wave impinges on a material? If the material is transparent to the particular frequency, then the wave can largely be transmitted. If the material is opaque to the frequency, then the wave can be totally reflected. The wave can also be absorbed by the material, indicating that there is some interaction between the wave and the material, such as the thermal agitation of molecules.

Of course it is possible to have partial transmission, reflection, and absorption. We normally associate these properties with visible light, but they do apply to all electromagnetic waves. What is not obvious is that something that is transparent to light may be opaque at other frequencies. For example, ordinary glass is transparent to visible light but largely opaque to ultraviolet radiation. Human skin is opaque to visible light—we cannot see through people—but transparent to X-rays.

Section Summary

- The relationship among the speed of propagation, wavelength, and frequency for any wave is given by $v_{\mathrm{W}} = f\lambda$, so that for electromagnetic waves,

$$c = f\lambda,$$

  where f is the frequency, λ is the wavelength, and c is the speed of light.
- The electromagnetic spectrum is separated into many categories and subcategories, based on the frequency and wavelength, source, and uses of the electromagnetic waves.

Glossary

electromagnetic spectrum
: the full range of wavelengths or frequencies of electromagnetic radiation

radio waves
: electromagnetic waves with wavelengths in the range from 1 mm to 100 km; they are produced by currents in wires and circuits and by astronomical phenomena

microwaves
: electromagnetic waves with wavelengths in the range from 1 mm to 1 m; they can be produced by currents in macroscopic circuits and devices

infrared radiation (IR)

a region of the electromagnetic spectrum with a frequency range that extends from just below the red region of the visible light spectrum up to the microwave region, or from 0.74 μm to 300 μm

ultraviolet radiation (UV)

electromagnetic radiation in the range extending upward in frequency from violet light and overlapping with the lowest X-ray frequencies, with wavelengths from 400 nm down to about 10 nm

visible light

the narrow segment of the electromagnetic spectrum to which the normal human eye responds

X-ray

invisible, penetrating form of very high frequency electromagnetic radiation, overlapping both the ultraviolet range and the γ-ray range

gamma ray

(γ ray); extremely high frequency electromagnetic radiation emitted by the nucleus of an atom, either from natural nuclear decay or induced nuclear processes in nuclear reactors and weapons. The lower end of the γ-ray frequency range overlaps the upper end of the X-ray range, but γ rays can have the highest frequency of any electromagnetic radiation

---

This page titled 4.2.4: The Electromagnetic Spectrum- an Overview is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- **11.4: The Electromagnetic Spectrum- an Overview** by OpenStax is licensed CC BY 4.0.

4.2.5: The Electromagnetic Spectrum- Application Notes

Learning Objectives

- List and explain the characteristics and applications of different parts of the electromagnetic spectrum.

In this module, we look at the properties of different types of electromagnetic waves. Again, Figure 4.2.5.1 shows the electromagnetic spectrum. The characteristics of the various types of electromagnetic waves you will read about below are related to their frequencies and wavelengths.

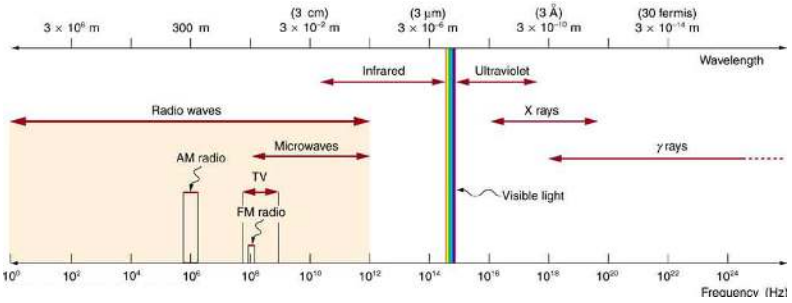


Figure 4.2.5.1: The electromagnetic spectrum, showing the major categories of electromagnetic waves. The range of frequencies and wavelengths is remarkable. The dividing line between some categories is distinct, whereas other categories overlap.

Radio and TV Waves

The broad category of **radio waves** is defined to contain any electromagnetic wave produced by currents in wires and circuits. Its name derives from their most common use as a carrier of audio information (i.e., radio). The name is applied to electromagnetic waves of similar frequencies regardless of source. Radio waves from outer space, for example, do not come from alien radio stations. They are created by many astronomical phenomena, and their study has revealed much about nature on the largest scales.

There are many uses for radio waves, and so the category is divided into many subcategories, including microwaves and those electromagnetic waves used for AM and FM radio, cellular telephones, and TV.

The lowest commonly encountered radio frequencies are produced by high-voltage AC power transmission lines at frequencies of 50 or 60 Hz. (See Figure 4.2.5.2.) These extremely long wavelength electromagnetic waves (about 6000 km!) are one means of energy loss in long-distance power transmission.



Figure 4.2.5.2: This high-voltage traction power line running to Eutingen Railway Substation in Germany radiates electromagnetic waves with very long wavelengths. (credit: Zonk43, Wikimedia Commons)

There is an ongoing controversy regarding potential health hazards associated with exposure to these electromagnetic fields (E-fields). Some people suspect that living near such transmission lines may cause a variety of illnesses, including cancer. But demographic data are either inconclusive or simply do not support the hazard theory. Recent reports that have looked at many European and American epidemiological studies have found no increase in risk for cancer due to exposure to E-fields.

Extremely low frequency (ELF) radio waves of about 1 kHz are used to communicate with submerged submarines. The ability of radio waves to penetrate salt water is related to their wavelength (much like ultrasound penetrating tissue)—the longer the wavelength, the farther they penetrate. Since salt water is a good conductor, radio waves are strongly absorbed by it, and very long wavelengths are needed to reach a submarine under the surface. (See Figure 4.2.5.3.)

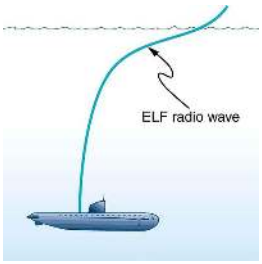


Figure 4.2.5.3: Very long wavelength radio waves are needed to reach this submarine, requiring extremely low frequency signals (ELF). Shorter wavelengths do not penetrate to any significant depth.

AM radio waves are used to carry commercial radio signals in the frequency range from 540 to 1600 kHz. The abbreviation AM stands for **amplitude modulation**, which is the method for placing information on these waves. (See Figure 4.2.5.4.) A **carrier wave** having the basic frequency of the radio station, say 1530 kHz, is varied or modulated in amplitude by an audio signal. The resulting wave has a constant frequency, but a varying amplitude.

A radio receiver tuned to have the same resonant frequency as the carrier wave can pick up the signal, while rejecting the many other frequencies impinging on its antenna. The receiver's circuitry is designed to respond to variations in amplitude of the carrier wave to replicate the original audio signal. That audio signal is amplified to drive a speaker or perhaps to be recorded.

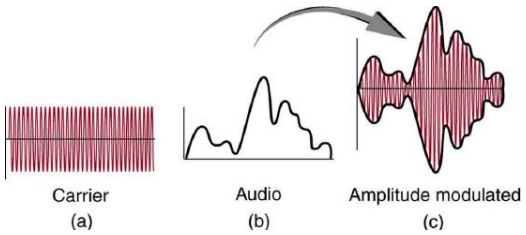


Figure 4.2.5.4: Amplitude modulation for AM radio. (a) A carrier wave at the station's basic frequency. (b) An audio signal at much lower audible frequencies. (c) The amplitude of the carrier is modulated by the audio signal without changing its basic frequency.

FM Radio Waves

FM radio waves are also used for commercial radio transmission, but in the frequency range of 88 to 108 MHz. FM stands for **frequency modulation**, another method of carrying information. (See Figure 4.2.5.5.) Here a carrier wave having the basic frequency of the radio station, perhaps 105.1 MHz, is modulated in frequency by the audio signal, producing a wave of constant amplitude but varying frequency.

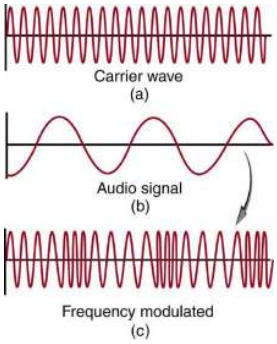


Figure 4.2.5.5: Frequency modulation for FM radio. (a) A carrier wave at the station's basic frequency. (b) An audio signal at much lower audible frequencies. (c) The frequency of the carrier is modulated by the audio signal without changing its amplitude.

Since audible frequencies range up to 20 kHz (or 0.020 MHz) at most, the frequency of the FM radio wave can vary from the carrier by as much as 0.020 MHz. Thus the carrier frequencies of two different radio stations cannot be closer than 0.020 MHz. An FM receiver is tuned to resonate at the carrier frequency and has circuitry that responds to variations in frequency, reproducing the audio information.

FM radio is inherently less subject to noise from stray radio sources than AM radio. The reason is that amplitudes of waves add. So an AM receiver would interpret noise added onto the amplitude of its carrier wave as part of the information. An FM receiver can be made to reject amplitudes other than that of the basic carrier wave and only look for variations in frequency. It is thus easier to reject noise from FM, since noise produces a variation in amplitude.

Television is also broadcast on electromagnetic waves. Since the waves must carry a great deal of visual as well as audio information, each channel requires a larger range of frequencies than simple radio transmission. TV channels utilize frequencies in the range of 54 to 88 MHz and 174 to 222 MHz. (The entire FM radio band lies between channels 88 MHz and 174 MHz.) These TV channels are called VHF (for **very high frequency**). Other channels called UHF (for **ultra high frequency**) utilize an even higher frequency range of 470 to 1000 MHz.

The TV video signal is AM, while the TV audio is FM. Note that these frequencies are those of free transmission with the user utilizing an old-fashioned roof antenna. Satellite dishes and cable transmission of TV occurs at significantly higher frequencies and is rapidly evolving with the use of the high-definition or HD format.

Example 4.2.5.1: Calculating Wavelengths of Radio Waves

Calculate the wavelengths of a 1530-kHz AM radio signal, a 105.1-MHz FM radio signal, and a 1.90-GHz cell phone signal.

Strategy

The relationship between wavelength and frequency is $c = f\lambda$, where $c = 3.00 \times 10^8\ \text{m/s}$ is the speed of light (the speed of light is only very slightly smaller in air than it is in a vacuum). We can rearrange this equation to find the wavelength for all three frequencies.

Solution

Rearranging gives

$$\lambda = \frac{c}{f}.$$

(a) For the $f = 1530\ \text{kHz}$ AM radio signal, then,

$$\begin{aligned}\lambda &= \frac{3.00 \times 10^8\ \text{m/s}}{1530 \times 10^3 \text{cycles/s}} \\ &= 196\ \text{m}.\end{aligned}$$

(b) For the $f = 105.1\ \text{MHz}$ FM radio signal,

$$\begin{aligned}\lambda &= \frac{3.00 \times 10^8\ \text{m/s}}{105.1 \times 10^6 \text{cycles/s}} \\ &= 2.85\ \text{m}.\end{aligned}$$

(c) And for the $f = 1.90\ \text{GHz}$ cell phone,

$$\begin{aligned}\lambda &= \frac{3.00 \times 10^8\ \text{m/s}}{1.90 \times 10^9 \text{cycles/s}} \\ &= 0.158\ \text{m}.\end{aligned}$$

Discussion

These wavelengths are consistent with the spectrum in Figure 4.2.5.1. The wavelengths are also related to other properties of these electromagnetic waves, as we shall see.

The wavelengths found in the preceding example are representative of AM, FM, and cell phones, and account for some of the differences in how they are broadcast and how well they travel. The most efficient length for a linear antenna, such as discussed in "Production of Electromagnetic Waves", is $\lambda/2$, half the wavelength of the electromagnetic wave. Thus a very large antenna is needed to efficiently broadcast typical AM radio with its carrier wavelengths on the order of hundreds of meters.

One benefit to these long AM wavelengths is that they can go over and around rather large obstacles (like buildings and hills), just as ocean waves can go around large rocks. FM and TV are best received when there is a line of sight between the broadcast antenna

and receiver, and they are often sent from very tall structures. FM, TV, and mobile phone antennas themselves are much smaller than those used for AM, but they are elevated to achieve an unobstructed line of sight. (See Figure 4.2.5.6.)

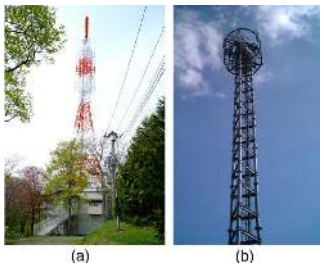


Figure 4.2.5.6: (a) A large tower is used to broadcast TV signals. The actual antennas are small structures on top of the tower—they are placed at great heights to have a clear line of sight over a large broadcast area. (credit: Ozizo, Wikimedia Commons) (b) The NTT Dokomo mobile phone tower at Tokorozawa City, Japan. (credit: tokoroten, Wikimedia Commons)

Radio Wave Interference

Astronomers and astrophysicists collect signals from outer space using electromagnetic waves. A common problem for astrophysicists is the "pollution" from electromagnetic radiation pervading our surroundings from communication systems in general. Even everyday gadgets like our car keys having the facility to lock car doors remotely and being able to turn TVs on and off using remotes involve radio-wave frequencies. In order to prevent interference between all these electromagnetic signals, strict regulations are drawn up for different organizations to utilize different radio frequency bands.

One reason why we are sometimes asked to switch off our mobile phones (operating in the range of 1.9 GHz) on airplanes and in hospitals is that important communications or medical equipment often uses similar radio frequencies and their operation can be affected by frequencies used in the communication devices.

For example, radio waves used in magnetic resonance imaging (MRI) have frequencies on the order of 100 MHz, although this varies significantly depending on the strength of the magnetic field used and the nuclear type being scanned. MRI is an important medical imaging and research tool, producing highly detailed two- and three-dimensional images. Radio waves are broadcast, absorbed, and reemitted in a resonance process that is sensitive to the density of nuclei (usually protons or hydrogen nuclei).

The wavelength of 100-MHz radio waves is 3 m, yet using the sensitivity of the resonant frequency to the magnetic field strength, details smaller than a millimeter can be imaged. This is a good example of an exception to a rule of thumb (in this case, the rubric that details much smaller than the probe's wavelength cannot be detected). The intensity of the radio waves used in MRI presents little or no hazard to human health.

Microwaves

Microwaves are the highest-frequency electromagnetic waves that can be produced by currents in macroscopic circuits and devices. Microwave frequencies range from about 10^9 Hz to the highest practical LC resonance at nearly 10^{12} Hz. Since they have high frequencies, their wavelengths are short compared with those of other radio waves—hence the name "microwave."

Microwaves can also be produced by atoms and molecules. They are, for example, a component of electromagnetic radiation generated by **thermal agitation**. The thermal motion of atoms and molecules in any object at a temperature above absolute zero causes them to emit and absorb radiation.

Since it is possible to carry more information per unit time on high frequencies, microwaves are quite suitable for communications. Most satellite-transmitted information is carried on microwaves, as are land-based long-distance transmissions. A clear line of sight between transmitter and receiver is needed because of the short wavelengths involved.

Radar is a common application of microwaves that was first developed in World War II. By detecting and timing microwave echoes, radar systems can determine the distance to objects as diverse as clouds and aircraft. A Doppler shift in the radar echo can be used to determine the speed of a car or the intensity of a rainstorm. Sophisticated radar systems are used to map the Earth and other planets, with a resolution limited by wavelength. (See Figure 4.2.5.7.) The shorter the wavelength of any probe, the smaller the detail it is possible to observe.

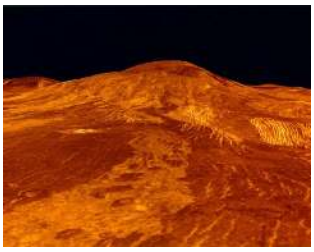


Figure 4.2.5.7: An image of Sif Mons with lava flows on Venus, based on Magellan synthetic aperture radar data combined with radar altimetry to produce a three-dimensional map of the surface. The Venusian atmosphere is opaque to visible light, but not to the microwaves that were used to create this image. (credit: NSSDC, NASA/JPL)

Heating with Microwaves

How does the ubiquitous microwave oven produce microwaves electronically, and why does food absorb them preferentially? Microwaves at a frequency of 2.45 GHz are produced by accelerating electrons. The microwaves are then used to induce an alternating electric field in the oven.

Water and some other constituents of food have a slightly negative charge at one end and a slightly positive charge at one end (called polar molecules). The range of microwave frequencies is specially selected so that the polar molecules, in trying to keep orienting themselves with the electric field, absorb these energies and increase their temperatures—called dielectric heating.

The energy thereby absorbed results in thermal agitation heating food and not the plate, which does not contain water. Hot spots in the food are related to constructive and destructive interference patterns. Rotating antennas and food turntables help spread out the hot spots.

Another use of microwaves for heating is within the human body. Microwaves will penetrate more than shorter wavelengths into tissue and so can accomplish “deep heating” (called microwave diathermy). This is used for treating muscular pains, spasms, tendonitis, and rheumatoid arthritis.

MAKING CONNECTIONS: TAKE-HOME EXPERIMENT—MICROWAVE OVENS

1. Look at the door of a microwave oven. Describe the structure of the door. Why is there a metal grid on the door? How does the size of the holes in the grid compare with the wavelengths of microwaves used in microwave ovens? What is this wavelength?
2. Place a glass of water (about 250 ml) in the microwave and heat it for 30 seconds. Measure the temperature gain (the ΔT). Assuming that the power output of the oven is 1000 W, calculate the efficiency of the heat-transfer process.
3. Remove the rotating turntable or moving plate and place a cup of water in several places along a line parallel with the opening. Heat for 30 seconds and measure the ΔT for each position. Do you see cases of destructive interference?

Microwaves generated by atoms and molecules far away in time and space can be received and detected by electronic circuits. Deep space acts like a blackbody with a 2.7 K temperature, radiating most of its energy in the microwave frequency range. In 1964, Penzias and Wilson detected this radiation and eventually recognized that it was the radiation of the Big Bang’s cooled remnants.

Infrared Radiation

The microwave and infrared regions of the electromagnetic spectrum overlap (see Figure 4.2.5.1). **Infrared radiation** is generally produced by thermal motion and the vibration and rotation of atoms and molecules. Electronic transitions in atoms and molecules can also produce infrared radiation.

The range of infrared frequencies extends up to the lower limit of visible light, just below red. In fact, infrared means “below red.” Frequencies at its upper limit are too high to be produced by accelerating electrons in circuits, but small systems, such as atoms and molecules, can vibrate fast enough to produce these waves.

Water molecules rotate and vibrate particularly well at infrared frequencies, emitting and absorbing them so efficiently that the emissivity for skin is $e = 0.97$ in the infrared. Night-vision scopes can detect the infrared emitted by various warm objects,

including humans, and convert it to visible light.

We can examine radiant heat transfer from a house by using a camera capable of detecting infrared radiation. Reconnaissance satellites can detect buildings, vehicles, and even individual humans by their infrared emissions, whose power radiation is proportional to the fourth power of the absolute temperature. More mundanely, we use infrared lamps, some of which are called quartz heaters, to preferentially warm us because we absorb infrared better than our surroundings.

The Sun radiates like a nearly perfect blackbody (that is, it has $e = 1$), with a 6000 K surface temperature. About half of the solar energy arriving at the Earth is in the infrared region, with most of the rest in the visible part of the spectrum, and a relatively small amount in the ultraviolet. On average, 50 percent of the incident solar energy is absorbed by the Earth.

The relatively constant temperature of the Earth is a result of the energy balance between the incoming solar radiation and the energy radiated from the Earth. Most of the infrared radiation emitted from the Earth is absorbed by CO_2 and H_2O in the atmosphere and then radiated back to Earth or into outer space. This radiation back to Earth is known as the greenhouse effect, and it maintains the surface temperature of the Earth about $40^{\circ}C$ higher than it would be if there is no absorption. Some scientists think that the increased concentration of CO_2 and other greenhouse gases in the atmosphere, resulting from increases in fossil fuel burning, has increased global average temperatures.

Visible Light

Visible light is the narrow segment of the electromagnetic spectrum to which the normal human eye responds. Visible light is produced by vibrations and rotations of atoms and molecules, as well as by electronic transitions within atoms and molecules. The receivers or detectors of light largely utilize electronic transitions. We say the atoms and molecules are excited when they absorb and relax when they emit through electronic transitions.

Figure 4.2.5.8 shows this part of the spectrum, together with the colors associated with particular pure wavelengths. We usually refer to visible light as having wavelengths of between 400 nm and 750 nm. (The retina of the eye actually responds to the lowest ultraviolet frequencies, but these do not normally reach the retina because they are absorbed by the cornea and lens of the eye.)

Red light has the lowest frequencies and longest wavelengths, while violet has the highest frequencies and shortest wavelengths. Blackbody radiation from the Sun peaks in the visible part of the spectrum but is more intense in the red than in the violet, making the Sun yellowish in appearance.

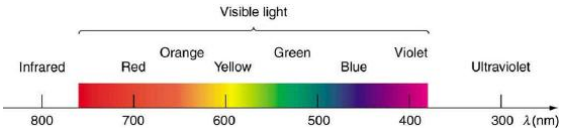


Figure 4.2.5.8: A small part of the electromagnetic spectrum that includes its visible components. The divisions between infrared, visible, and ultraviolet are not perfectly distinct, nor are those between the seven rainbow colors.

Living things—plants and animals—have evolved to utilize and respond to parts of the electromagnetic spectrum they are embedded in. Visible light is the most predominant and we enjoy the beauty of nature through visible light. Plants are more selective. Photosynthesis makes use of parts of the visible spectrum to make sugars.

Optics is the study of the behavior of visible light and other forms of electromagnetic waves. Optics falls into two distinct categories. When electromagnetic radiation, such as visible light, interacts with objects that are large compared with its wavelength, its motion can be represented by straight lines like rays. Ray optics is the study of such situations and includes lenses and mirrors.

When electromagnetic radiation interacts with objects about the same size as the wavelength or smaller, its wave nature becomes apparent. For example, observable detail is limited by the wavelength, and so visible light can never detect individual atoms, because they are so much smaller than its wavelength. Physical or wave optics is the study of such situations and includes all wave characteristics.

TAKE-HOME EXPERIMENT: COLORS THAT MATCH

When you light a match you see largely orange light; when you light a gas stove you see blue light. Why are the colors different? What other colors are present in these?

Ultraviolet Radiation

Ultraviolet means “above violet.” The electromagnetic frequencies of **ultraviolet radiation (UV)** extend upward from violet, the highest-frequency visible light. Ultraviolet is also produced by atomic and molecular motions and electronic transitions. The wavelengths of ultraviolet extend from 400 nm down to about 10 nm at its highest frequencies, which overlap with the lowest X-ray frequencies. It was recognized as early as 1801 by Johann Ritter that the solar spectrum had an invisible component beyond the violet range.

Solar UV radiation is broadly subdivided into three regions: UV-A (320–400 nm), UV-B (290–320 nm), and UV-C (220–290 nm), ranked from long to shorter wavelengths (from smaller to larger energies). Most UV-B and all UV-C is absorbed by ozone (O_3) molecules in the upper atmosphere. Consequently, 99% of the solar UV radiation reaching the Earth’s surface is UV-A.

Human Exposure to UV Radiation

It is largely exposure to UV-B that causes skin cancer. It is estimated that as many as 20% of adults will develop skin cancer over the course of their lifetime. Again, treatment is often successful if caught early. Despite very little UV-B reaching the Earth’s surface, there are substantial increases in skin-cancer rates in countries such as Australia, indicating how important it is that UV-B and UV-C continue to be absorbed by the upper atmosphere.

All UV radiation can damage collagen fibers, resulting in an acceleration of the aging process of skin and the formation of wrinkles. Because there is so little UV-B and UV-C reaching the Earth’s surface, sunburn is caused by large exposures, and skin cancer from repeated exposure. Some studies indicate a link between overexposure to the Sun when young and melanoma later in life.

The tanning response is a defense mechanism in which the body produces pigments to absorb future exposures in inert skin layers above living cells. Basically UV-B radiation excites DNA molecules, distorting the DNA helix, leading to mutations and the possible formation of cancerous cells.

Repeated exposure to UV-B may also lead to the formation of cataracts in the eyes—a cause of blindness among people living in the equatorial belt where medical treatment is limited. Cataracts, clouding in the eye’s lens and a loss of vision, are age related; 60% of those between the ages of 65 and 74 will develop cataracts. However, treatment is easy and successful, as one replaces the lens of the eye with a plastic lens. Prevention is important. Eye protection from UV is more effective with plastic sunglasses than those made of glass.

A major acute effect of extreme UV exposure is the suppression of the immune system, both locally and throughout the body.

Low-intensity ultraviolet is used to sterilize haircutting implements, implying that the energy associated with ultraviolet is deposited in a manner different from lower-frequency electromagnetic waves. (Actually this is true for all electromagnetic waves with frequencies greater than visible light.)

Flash photography is generally not allowed of precious artworks and colored prints because the UV radiation from the flash can cause photo-degradation in the artworks. Often artworks will have an extra-thick layer of glass in front of them, which is especially designed to absorb UV radiation.

UV Light and the Ozone Layer

If all of the Sun’s ultraviolet radiation reached the Earth’s surface, there would be extremely grave effects on the biosphere from the severe cell damage it causes. However, the layer of ozone (O_3) in our upper atmosphere (10 to 50 km above the Earth) protects life by absorbing most of the dangerous UV radiation.

Unfortunately, today we are observing a depletion in ozone concentrations in the upper atmosphere. This depletion has led to the formation of an “ozone hole” in the upper atmosphere. The hole is more centered over the southern hemisphere, and changes with the seasons, being largest in the spring. This depletion is attributed to the breakdown of ozone molecules by refrigerant gases called chlorofluorocarbons (CFCs).

The UV radiation helps dissociate the CFC’s, releasing highly reactive chlorine (Cl) atoms, which catalyze the destruction of the ozone layer. For example, the reaction of $CFCl_3$ with a photon of light $(h\upsilon)$ can be written as:

$$CFCl_3 + h\upsilon \rightarrow CFCl_2 + Cl.$$

The Cl atom then catalyzes the breakdown of ozone as follows:

$$Cl + O_3 \rightarrow ClO + O_2 \text{ and } ClO + O_3 \rightarrow Cl + 2O_2.$$

A single chlorine atom could destroy ozone molecules for up to two years before being transported down to the surface. The CFCs are relatively stable and will contribute to ozone depletion for years to come. CFCs are found in refrigerants, air conditioning systems, foams, and aerosols.

International concern over this problem led to the establishment of the “Montreal Protocol” agreement (1987) to phase out CFC production in most countries. However, developing-country participation is needed if worldwide production and elimination of CFCs is to be achieved. Probably the largest contributor to CFC emissions today is India. But the protocol seems to be working, as there are signs of an ozone recovery. (See Figure 4.2.5.9.)

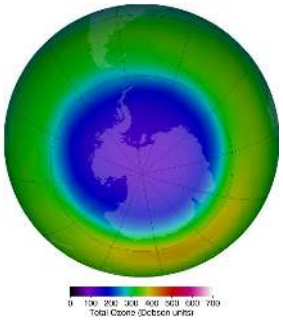


Figure 4.2.5.9: This map of ozone concentration over Antarctica in October 2011 shows severe depletion suspected to be caused by CFCs. Less dramatic but more general depletion has been observed over northern latitudes, suggesting the effect is global. With less ozone, more ultraviolet radiation from the Sun reaches the surface, causing more damage. (credit: NASA Ozone Watch)

Benefits of UV Light

Besides the adverse effects of ultraviolet radiation, there are also benefits of exposure in nature and uses in technology. Vitamin D production in the skin (epidermis) results from exposure to UVB radiation, generally from sunlight. A number of studies indicate lack of vitamin D can result in the development of a range of cancers (prostate, breast, colon), so a certain amount of UV exposure is helpful. Lack of vitamin D is also linked to osteoporosis. Exposures (with no sunscreen) of 10 minutes a day to arms, face, and legs might be sufficient to provide the accepted dietary level. However, in the winter time north of about 37° latitude, most UVB gets blocked by the atmosphere.

UV radiation is used in the treatment of infantile jaundice and in some skin conditions. It is also used in sterilizing workspaces and tools, and killing germs in a wide range of applications. It is also used as an analytical tool to identify substances.

When exposed to ultraviolet, some substances, such as minerals, glow in characteristic visible wavelengths, a process called fluorescence. So-called black lights emit ultraviolet to cause posters and clothing to fluoresce in the visible. Ultraviolet is also used in special microscopes to detect details smaller than those observable with longer-wavelength visible-light microscopes.

X-Rays

In the 1850s, scientists (such as Faraday) began experimenting with high-voltage electrical discharges in tubes filled with rarefied gases. It was later found that these discharges created an invisible, penetrating form of very high frequency electromagnetic radiation. This radiation was called an **X-ray**, because its identity and nature were unknown.

THINGS GREAT AND SMALL: A SUBMICROSCOPIC VIEW OF X-RAY PRODUCTION

X-rays can be created in a high-voltage discharge. They are emitted in the material struck by electrons in the discharge current. There are two mechanisms by which the electrons create X-rays.

The first method is illustrated in Figure 4.2.5.10. An electron is accelerated in an evacuated tube by a high positive voltage. The electron strikes a metal plate (e.g., copper) and produces X-rays. Since this is a high-voltage discharge, the electron gains sufficient energy to ionize the atom.

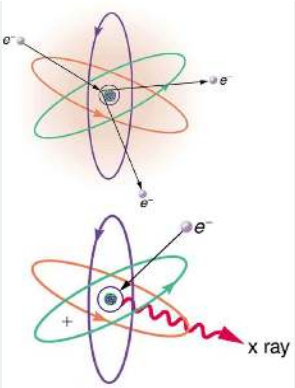


Figure 4.2.5.10: Artist's conception of an electron ionizing an atom followed by the recapture of an electron and emission of an X-ray. An energetic electron strikes an atom and knocks an electron out of one of the orbits closest to the nucleus. Later, the atom captures another electron, and the energy released by its fall into a low orbit generates a high-energy EM wave called an X-ray.

In the case shown, an inner-shell electron (one in an orbit relatively close to and tightly bound to the nucleus) is ejected. A short time later, another electron is captured and falls into the orbit in a single great plunge. The energy released by this fall is given to an EM wave known as an X-ray. Since the orbits of the atom are unique to the type of atom, the energy of the X-ray is characteristic of the atom, hence the name characteristic X-ray.

The second method by which an energetic electron creates an X-ray when it strikes a material is illustrated in Figure 4.2.5.11. The electron interacts with charges in the material as it penetrates. These collisions transfer kinetic energy from the electron to the electrons and atoms in the material.

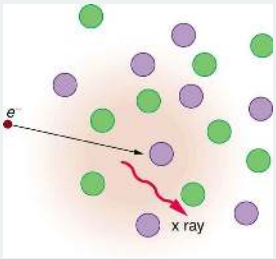


Figure 4.2.5.11: Artist's conception of an electron being slowed by collisions in a material and emitting X-ray radiation. This energetic electron makes numerous collisions with electrons and atoms in a material it penetrates. An accelerated charge radiates EM waves, a second method by which X-rays are created.

A loss of kinetic energy implies an acceleration, in this case decreasing the electron's velocity. Whenever a charge is accelerated, it radiates EM waves. Given the high energy of the electron, these EM waves can have high energy. We call them X-rays. Since the process is random, a broad spectrum of X-ray energy is emitted that is more characteristic of the electron energy than the type of material the electron encounters. Such EM radiation is called "bremsstrahlung" (German for "braking radiation").

As described above, there are two methods by which X-rays are created—both are submicroscopic processes and can be caused by high-voltage discharges. While the low-frequency end of the X-ray range overlaps with the ultraviolet, X-rays extend to much higher frequencies (and energies).

X-rays have adverse effects on living cells similar to those of ultraviolet radiation, and they have the additional liability of being more penetrating, affecting more than the surface layers of cells. Cancer and genetic defects can be induced by exposure to X-rays.

Because of their effect on rapidly dividing cells, X-rays can also be used to treat and even cure cancer.

The widest use of X-rays is for imaging objects that are opaque to visible light, such as the human body or aircraft parts. In humans, the risk of cell damage is weighed carefully against the benefit of the diagnostic information obtained. However, questions have risen in recent years as to accidental overexposure of some people during CT scans—a mistake at least in part due to poor monitoring of radiation dose.

The ability of X-rays to penetrate matter depends on density, and so an X-ray image can reveal very detailed density information. Figure 4.2.5.12 shows an example of the simplest type of X-ray image, an X-ray shadow on film. The amount of information in a simple X-ray image is impressive, but more sophisticated techniques, such as CT scans, can reveal three-dimensional information with details smaller than a millimeter.



Figure 4.2.5.12: This shadow X-ray image shows many interesting features, such as artificial heart valves, a pacemaker, and the wires used to close the sternum. (credit: P. P. Urone)

The use of X-ray technology in medicine is called radiology—an established and relatively cheap tool in comparison to more sophisticated technologies. Consequently, X-rays are widely available and used extensively in medical diagnostics. During World War I, mobile X-ray units, advocated by Madame Marie Curie, were used to diagnose soldiers.

Because they can have wavelengths less than 0.01 nm, X-rays can be scattered (a process called X-ray diffraction) to detect the shape of molecules and the structure of crystals. X-ray diffraction was crucial to Crick, Watson, and Wilkins in the determination of the shape of the double-helix DNA molecule.

X-rays are also used as a precise tool for trace-metal analysis in X-ray induced fluorescence, in which the energy of the X-ray emissions are related to the specific types of elements and amounts of materials present.

Gamma Rays

Soon after nuclear radioactivity was first detected in 1896, it was found that at least three distinct types of radiation were being emitted. The most penetrating nuclear radiation was called a **gamma ray (γ ray)** (again a name given because its identity and character were unknown), and it was later found to be an extremely high frequency electromagnetic wave.

In fact, γ rays are any electromagnetic radiation emitted by a nucleus. This can be from natural nuclear decay or induced nuclear processes in nuclear reactors and weapons. The lower end of the γ-ray frequency range overlaps the upper end of the X-ray range, but γ rays can have the highest frequency of any electromagnetic radiation.

Gamma rays have characteristics identical to X-rays of the same frequency—they differ only in source. At higher frequencies, γ rays are more penetrating and more damaging to living tissue. They have many of the same uses as X-rays, including cancer therapy. Gamma radiation from radioactive materials is used in nuclear medicine.

Figure 4.2.5.13 shows a medical image based on γ rays. Food spoilage can be greatly inhibited by exposing it to large doses of γ radiation, thereby obliterating responsible microorganisms. Damage to food cells through irradiation occurs as well, and the long-term hazards of consuming radiation-preserved food are unknown and controversial for some groups. Both X-ray and γ-ray technologies are also used in scanning luggage at airports.



Figure 4.2.5.13: This is an image of the γ rays emitted by nuclei in a compound that is concentrated in the bones and eliminated through the kidneys. Bone cancer is evidenced by nonuniform concentration in similar structures. For example, some ribs are darker than others. (credit: P. P. Urone)

Detecting Electromagnetic Waves from Space

A final note on star gazing. The entire electromagnetic spectrum is used by researchers for investigating stars, space, and time. As noted earlier, Penzias and Wilson detected microwaves to identify the background radiation originating from the Big Bang. Radio telescopes such as the Arecibo Radio Telescope in Puerto Rico and Parkes Observatory in Australia were designed to detect radio waves.

Infrared telescopes need to have their detectors cooled by liquid nitrogen to be able to gather useful signals. Since infrared radiation is predominantly from thermal agitation, if the detectors were not cooled, the vibrations of the molecules in the antenna would be stronger than the signal being collected.

The most famous of these infrared sensitive telescopes is the James Clerk Maxwell Telescope in Hawaii. The earliest telescopes, developed in the seventeenth century, were optical telescopes, collecting visible light. Telescopes in the ultraviolet, X-ray, and γ-ray regions are placed outside the atmosphere on satellites orbiting the Earth.

The Hubble Space Telescope (launched in 1990) gathers ultraviolet radiation as well as visible light. In the X-ray region, there is the Chandra X-ray Observatory (launched in 1999), and in the γ-ray region, there is the new Fermi Gamma-ray Space Telescope (launched in 2008—taking the place of the Compton Gamma Ray Observatory, 1991–2000.).

Section Summary

- The electromagnetic spectrum is separated into many categories and subcategories, based on the frequency and wavelength, source, and uses of the electromagnetic waves.
- Any electromagnetic wave produced by currents in wires is classified as a radio wave, the lowest frequency electromagnetic waves. Radio waves are divided into many types, depending on their applications, ranging up to microwaves at their highest frequencies.

- Infrared radiation lies below visible light in frequency and is produced by thermal motion and the vibration and rotation of atoms and molecules. Infrared's lower frequencies overlap with the highest-frequency microwaves.
- Visible light is largely produced by electronic transitions in atoms and molecules, and is defined as being detectable by the human eye. Its colors vary with frequency, from red at the lowest to violet at the highest.
- Ultraviolet radiation starts with frequencies just above violet in the visible range and is produced primarily by electronic transitions in atoms and molecules.
- X-rays are created in high-voltage discharges and by electron bombardment of metal targets. Their lowest frequencies overlap the ultraviolet range but extend to much higher values, overlapping at the high end with gamma rays.
- Gamma rays are nuclear in origin and are defined to include the highest-frequency electromagnetic radiation of any type.

Glossary

electromagnetic spectrum
the full range of wavelengths or frequencies of electromagnetic radiation

radio waves
electromagnetic waves with wavelengths in the range from 1 mm to 100 km; they are produced by currents in wires and circuits and by astronomical phenomena

microwaves
electromagnetic waves with wavelengths in the range from 1 mm to 1 m; they can be produced by currents in macroscopic circuits and devices

thermal agitation
the thermal motion of atoms and molecules in any object at a temperature above absolute zero, which causes them to emit and absorb radiation

radar
a common application of microwaves. Radar can determine the distance to objects as diverse as clouds and aircraft, as well as determine the speed of a car or the intensity of a rainstorm

infrared radiation (IR)
a region of the electromagnetic spectrum with a frequency range that extends from just below the red region of the visible light spectrum up to the microwave region, or from 0.74 μm to 300 μm

ultraviolet radiation (UV)
electromagnetic radiation in the range extending upward in frequency from violet light and overlapping with the lowest X-ray frequencies, with wavelengths from 400 nm down to about 10 nm

visible light
the narrow segment of the electromagnetic spectrum to which the normal human eye responds

amplitude modulation (AM)
a method for placing information on electromagnetic waves by modulating the amplitude of a carrier wave with an audio signal, resulting in a wave with constant frequency but varying amplitude

extremely low frequency (ELF)
electromagnetic radiation with wavelengths usually in the range of 0 to 300 Hz, but also about 1kHz

carrier wave
an electromagnetic wave that carries a signal by modulation of its amplitude or frequency

frequency modulation (FM)
a method of placing information on electromagnetic waves by modulating the frequency of a carrier wave with an audio signal, producing a wave of constant amplitude but varying frequency

TV

video and audio signals broadcast on electromagnetic waves

very high frequency (VHF)

TV channels utilizing frequencies in the two ranges of 54 to 88 MHz and 174 to 222 MHz

ultra-high frequency (UHF)

TV channels in an even higher frequency range than VHF, of 470 to 1000 MHz

X-ray

invisible, penetrating form of very high frequency electromagnetic radiation, overlapping both the ultraviolet range and the γ-ray range

gamma ray

(γ ray); extremely high frequency electromagnetic radiation emitted by the nucleus of an atom, either from natural nuclear decay or induced nuclear processes in nuclear reactors and weapons. The lower end of the γ-ray frequency range overlaps the upper end of the X-ray range, but γ rays can have the highest frequency of any electromagnetic radiation

---

This page titled 4.2.5: The Electromagnetic Spectrum- Application Notes is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- **11.5: The Electromagnetic Spectrum- Application Notes** by OpenStax is licensed CC BY 4.0.

4.2.6: Reflection

Learning Objectives

- Explain reflection of light from polished and rough surfaces.

Whenever we look into a mirror, or squint at sunlight glinting from a lake, we are seeing a reflection. When you look at this page, too, you are seeing light reflected from it. Large telescopes use reflection to form an image of stars and other astronomical objects.

The law of reflection is illustrated in Figure 4.2.6.1, which also shows how the angles are measured relative to the perpendicular to the surface at the point where the light ray strikes. We expect to see reflections from smooth surfaces, but Figure 4.2.6.2 illustrates how a rough surface reflects light. Since the light strikes different parts of the surface at different angles, it is reflected in many different directions, or diffused. Diffused light is what allows us to see a sheet of paper from any angle, as illustrated in Figure 4.2.6.3. Many objects, such as people, clothing, leaves, and walls, have rough surfaces and can be seen from all sides. A mirror, on the other hand, has a smooth surface (compared with the wavelength of light) and reflects light at specific angles, as illustrated in Figure 4.2.6.4. When the moon reflects from a lake, as shown in Figure 4.2.6.5, a combination of these effects takes place.

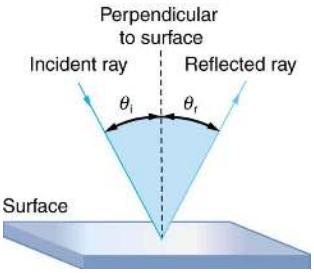


Figure 4.2.6.1: The law of reflection states that the angle of reflection equals the angle of incidence— $\theta_r = \theta_i$. The angles are measured relative to the perpendicular to the surface at the point where the ray strikes the surface.

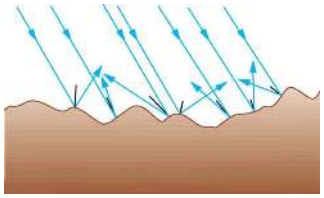


Figure 4.2.6.2: Light is diffused when it reflects from a rough surface. Here many parallel rays are incident, but they are reflected at many different angles since the surface is rough.

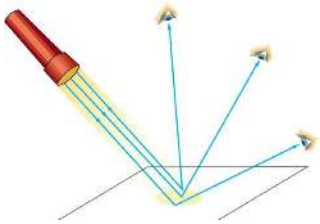


Figure 4.2.6.3: When a sheet of paper is illuminated with many parallel incident rays, it can be seen at many different angles, because its surface is rough and diffuses the light.

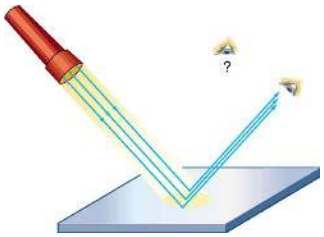


Figure 4.2.6.4: A mirror illuminated by many parallel rays reflects them in only one direction, since its surface is very smooth. Only the observer at a particular angle will see the reflected light.



Figure 4.2.6.5: Moonlight is spread out when it is reflected by the lake, since the surface is shiny but uneven. (credit: Diego Torres Silvestre, Flickr)

The law of reflection is very simple: The angle of reflection equals the angle of incidence.

THE LAW OF REFLECTION

The angle of reflection equals the angle of incidence.

When we see ourselves in a mirror, it appears that our image is actually behind the mirror. This is illustrated in Figure 4.2.6.6. We see the light coming from a direction determined by the law of reflection. The angles are such that our image is exactly the same distance behind the mirror as we stand away from the mirror. If the mirror is on the wall of a room, the images in it are all behind the mirror, which can make the room seem bigger. Although these mirror images make objects appear to be where they cannot be (like behind a solid wall), the images are not figments of our imagination. Mirror images can be photographed and videotaped by instruments and look just as they do with our eyes (optical instruments themselves). The precise manner in which images are formed by mirrors and lenses will be treated in later sections of this chapter.

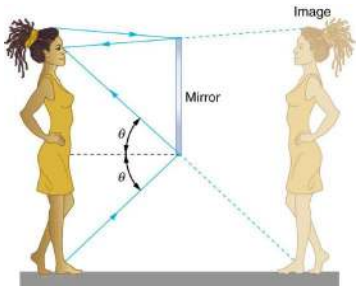


Figure 4.2.6.6: Our image in a mirror is behind the mirror. The two rays shown are those that strike the mirror at just the correct angles to be reflected into the eyes of the person. The image appears to be in the direction the rays are coming from when they enter the eyes.

TAKE-HOME EXPERIMENT: LAW OF REFLECTION

Take a piece of paper and shine a flashlight at an angle at the paper, as shown in Figure 4.2.6.3. Now shine the flashlight at a mirror at an angle. Do your observations confirm the predictions in Figure 4.2.6.3 and Figure 4.2.6.4? Shine the flashlight on various surfaces and determine whether the reflected light is diffuse or not. You can choose a shiny metallic lid of a pot or your skin. Using the mirror and flashlight, can you confirm the law of reflection? You will need to draw lines on a piece of paper showing the incident and reflected rays. (This part works even better if you use a laser pencil.)

Section Summary

- The angle of reflection equals the angle of incidence.
- A mirror has a smooth surface and reflects light at specific angles.
- Light is diffused when it reflects from a rough surface.
- Mirror images can be photographed and videotaped by instruments.

Glossary

mirror

smooth surface that reflects light at specific angles, forming an image of the person or object in front of it

law of reflection

angle of reflection equals the angle of incidence

This page titled 4.2.6: Reflection is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- **11.6: Reflection** by OpenStax is licensed CC BY 4.0.

4.2.7: Refraction

Learning Objectives

- Determine the index of refraction, given the speed of light in a medium.

It is easy to notice some odd things when looking into a fish tank. For example, you may see the same fish appearing to be in two different places. (See Figure 4.2.7.1.) This is because light coming from the fish to us changes direction when it leaves the tank, and in this case, it can travel two different paths to get to our eyes. The changing of a light ray's direction (loosely called bending) when it passes through variations in matter is called **refraction**. Refraction is responsible for a tremendous range of optical phenomena, from the action of lenses to voice transmission through optical fibers.

Definition: REFRACTION

The changing of a light ray's direction (loosely called bending) when it passes through variations in matter is called refraction.

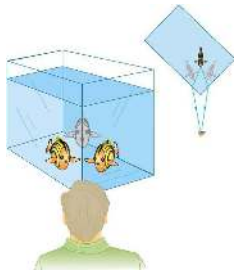


Figure 4.2.7.1: Looking at the fish tank as shown, we can see the same fish in two different locations, because light changes directions when it passes from water to air. In this case, the light can reach the observer by two different paths, and so the fish seems to be in two different places. This bending of light is called refraction and is responsible for many optical phenomena.

Why does light change direction when passing from one material (medium) to another? It is because light changes speed when going from one material to another. So before we study the law of refraction, it is useful to discuss the speed of light and how it varies in different media.

The Speed of Light

Early attempts to measure the speed of light, such as those made by Galileo, determined that light moved extremely fast, perhaps instantaneously. The first real evidence that light traveled at a finite speed came from the Danish astronomer Ole Roemer in the late 17th century. Roemer had noted that the average orbital period of one of Jupiter's moons, as measured from Earth, varied depending on whether Earth was moving toward or away from Jupiter. He correctly concluded that the apparent change in period was due to the change in distance between Earth and Jupiter and the time it took light to travel this distance. From his 1676 data, a value of the speed of light was calculated to be $2.26 \times 10^8\ \mathrm{m/s}$ (only 25% different from today's accepted value). In more recent times, physicists have measured the speed of light in numerous ways and with increasing accuracy. One particularly direct method, used in 1887 by the American physicist Albert Michelson (1852–1931), is illustrated in Figure 4.2.7.2. Light reflected from a rotating set of mirrors was reflected from a stationary mirror 35 km away and returned to the rotating mirrors. The time for the light to travel can be determined by how fast the mirrors must rotate for the light to be returned to the observer's eye.

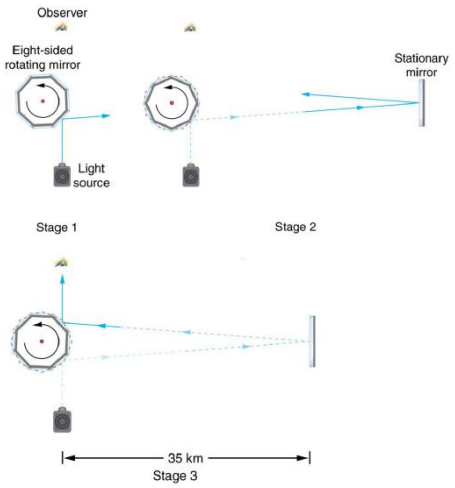


Figure 4.2.7.2: A schematic of early apparatus used by Michelson and others to determine the speed of light. As the mirrors rotate, the reflected ray is only briefly directed at the stationary mirror. The returning ray will be reflected into the observer's eye only if the next mirror has rotated into the correct position just as the ray returns. By measuring the correct rotation rate, the time for the round trip can be measured and the speed of light calculated. Michelson's calculated value of the speed of light was only 0.04% different from the value used today.

The speed of light is now known to great precision. In fact, the speed of light in a vacuum c is so important that it is accepted as one of the basic physical quantities and has the fixed value

$$c = 2.9972458 \times 10^8 \text{ m/s} \approx 3.00 \times 10^8 \text{ m/s},$$

where the approximate value of 3.00×10^8 m/s is used whenever three-digit accuracy is sufficient. The speed of light through matter is less than it is in a vacuum, because light interacts with atoms in a material. The speed of light depends strongly on the type of material, since its interaction with different atoms, crystal lattices, and other substructures varies. We define the **index of refraction** n of a material to be

$$n = \frac{c}{v},$$

where v is the observed speed of light in the material. Since the speed of light is always less than c in matter and equals c only in a vacuum, the index of refraction is always greater than or equal to one.

VALUE OF THE SPEED OF LIGHT

$$c = 2.9972458 \times 10^8 \text{ m/s} \approx 3.00 \times 10^8 \text{ m/s}$$

Definition: INDEX OF REFRACTION

$$n = \frac{c}{v}$$

That is, $n \geq 1$. Table 4.2.7.1 gives the indices of refraction for some representative substances. The values are listed for a particular wavelength of light, because they vary slightly with wavelength. (This can have important effects, such as colors produced by a prism.) Note that for gases, n is close to 1.0. This seems reasonable, since atoms in gases are widely separated and light travels at c in the vacuum between atoms. It is common to take $n = 1$ for gases unless great precision is needed. Although the speed of light v in a medium varies considerably from its value c in a vacuum, it is still a large speed.

Table 4.2.7.1: Index of Refraction in Various Media

Medium	*n*
Gases at 0ºC, 1 atm	
Air	1.000293
Carbon dioxide	1.00045
Hydrogen	1.000139
Oxygen	1.000271
Liquids at 20ºC	
Benzene	1.501
Carbon disulfide	1.628
Carbon tetrachloride	1.461
Ethanol	1.361
Glycerine	1.473
Water, fresh	1.333
Solids at 20ºC	
Diamond	2.419
Fluorite	1.434
Glass, crown	1.52
Glass, flint	1.66
Ice at 20ºC	1.309
Polystyrene	1.49
Plexiglas	1.51
Quartz, crystalline	1.544
Quartz, fused	1.458
Sodium chloride	1.544
Zircon	1.923

Example 4.2.7.1: Speed of Light in Matter

Calculate the speed of light in zircon, a material used in jewelry to imitate diamond.

Strategy

The speed of light in a material, v, can be calculated from the index of refraction n of the material using the equation $n = c/v$.

Solution

The equation for index of refraction states that $n = c/v$. Rearranging this to determine v gives

$$c$$

$$v = \frac{c}{n}.$$

The index of refraction for zircon is given as 1.923 in Table 4.2.7.1, and c is given in the equation for speed of light. Entering these values in the last expression gives

$$v = \frac{3.00 \times 10^8 \text{ m/s}}{1.923}$$
$$= 1.56 \times 10^8 \text{ m/s}.$$

Discussion

This speed is slightly larger than half the speed of light in a vacuum and is still high compared with speeds we normally experience. The only substance listed in Table 4.2.7.1 that has a greater index of refraction than zircon is diamond. We shall see later that the large index of refraction for zircon makes it sparkle more than glass, but less than diamond.

Law of Refraction

Figure 4.2.7.3 shows how a ray of light changes direction when it passes from one medium to another. As before, the angles are measured relative to a perpendicular to the surface at the point where the light ray crosses it. (Some of the incident light will be reflected from the surface, but for now we will concentrate on the light that is transmitted.) The change in direction of the light ray depends on how the speed of light changes. The change in the speed of light is related to the indices of refraction of the media involved. In the situations shown in Figure 4.2.7.3, medium 2 has a greater index of refraction than medium 1. This means that the speed of light is less in medium 2 than in medium 1. Note that as shown in Figure 4.2.7.3(a), the direction of the ray moves closer to the perpendicular when it slows down. Conversely, as shown in Figure 4.2.7.3(b), the direction of the ray moves away from the perpendicular when it speeds up. The path is exactly reversible. In both cases, you can imagine what happens by thinking about pushing a lawn mower from a footpath onto grass, and vice versa. Going from the footpath to grass, the front wheels are slowed and pulled to the side as shown. This is the same change in direction as for light when it goes from a fast medium to a slow one. When going from the grass to the footpath, the front wheels can move faster and the mower changes direction as shown. This, too, is the same change in direction as for light going from slow to fast.

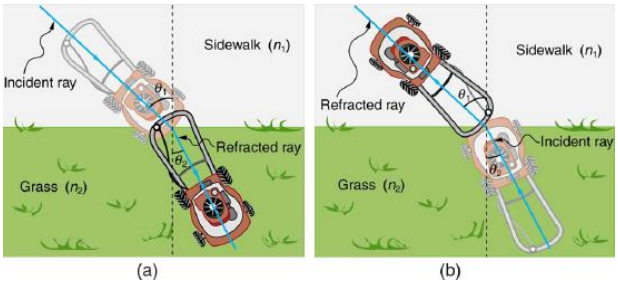


Figure 4.2.7.3: The change in direction of a light ray depends on how the speed of light changes when it crosses from one medium to another. The speed of light is greater in medium 1 than in medium 2 in the situations shown here. (a) A ray of light moves closer to the perpendicular when it slows down. This is analogous to what happens when a lawn mower goes from a footpath to grass. (b) A ray of light moves away from the perpendicular when it speeds up. This is analogous to what happens when a lawn mower goes from grass to footpath. The paths are exactly reversible.

The amount that a light ray changes its direction depends both on the incident angle and the amount that the speed changes. For a ray at a given incident angle, a large change in speed causes a large change in direction, and thus a large change in angle.

TAKE-HOME EXPERIMENT: A BROKEN PENCIL

A classic observation of refraction occurs when a pencil is placed in a glass half filled with water. Do this and observe the shape of the pencil when you look at the pencil sideways, that is, through air, glass, water. Explain your observations. Draw ray diagrams for the situation.

Section Summary

- The changing of a light ray's direction when it passes through variations in matter is called refraction.
- The speed of light in vacuum $c = 2.9972458 \times 10^8 \text{ m/s} \approx 3.00 \times 10^8 \text{ m/s}$
- Index of refraction $n = \frac{c}{v}$, where v is the speed of light in the material, c is the speed of light in vacuum, and n is the index of refraction.

Glossary

refraction

changing of a light ray’s direction when it passes through variations in matter

index of refraction

for a material, the ratio of the speed of light in vacuum to that in the material

---

This page titled 4.2.7: Refraction is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- **11.7: Refraction** by OpenStax is licensed CC BY 4.0.

4.2.8: Dispersion- The Rainbow and Prisms

Learning Objectives

- Explain the phenomenon of dispersion and discuss its advantages and disadvantages.

Everyone enjoys the spectacle of a rainbow glimmering against a dark stormy sky. How does sunlight falling on clear drops of rain get broken into the rainbow of colors we see? The same process causes white light to be broken into colors by a clear glass prism or a diamond. (See Figure 4.2.8.1.)

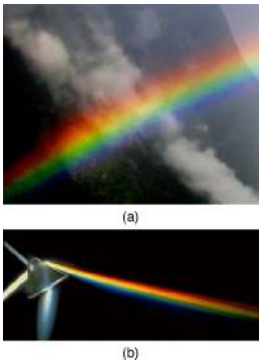


Figure 4.2.8.1: The colors of the rainbow (a) and those produced by a prism (b) are identical. (credit: Alfredo55, Wikimedia Commons; NASA)

We see about six colors in a rainbow—red, orange, yellow, green, blue, and violet; sometimes indigo is listed, too. Those colors are associated with different wavelengths of light, as shown in Figure 4.2.8.2. When our eye receives pure-wavelength light, we tend to see only one of the six colors, depending on wavelength. The thousands of other hues we can sense in other situations are our eye's response to various mixtures of wavelengths. White light, in particular, is a fairly uniform mixture of all visible wavelengths. Sunlight, considered to be white, actually appears to be a bit yellow because of its mixture of wavelengths, but it does contain all visible wavelengths. The sequence of colors in rainbows is the same sequence as the colors plotted versus wavelength in Figure 4.2.8.2. What this implies is that white light is spread out according to wavelength in a rainbow. **Dispersion** is defined as the spreading of white light into its full spectrum of wavelengths. More technically, dispersion occurs whenever there is a process that changes the direction of light in a manner that depends on wavelength. Dispersion, as a general phenomenon, can occur for any type of wave and always involves wavelength-dependent processes.

Definition: DISPERSION

Dispersion is defined to be the spreading of white light into its full spectrum of wavelengths.

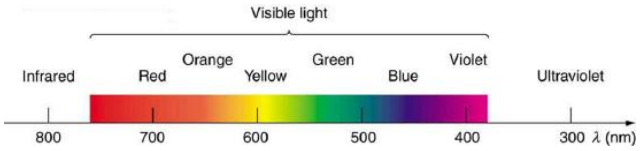


Figure 4.2.8.2: Even though rainbows are associated with seven colors, the rainbow is a continuous distribution of colors according to wavelengths.

Refraction is responsible for dispersion in rainbows and many other situations. The angle of refraction depends on the index of refraction. We know that the index of refraction n depends on the medium. But for a given medium, n also depends on wavelength. (See Table 4.2.8.1. Note that, for a given medium, n increases as wavelength decreases and is greatest for violet light. Thus violet light is bent more than red light, as shown for a prism in Figure 4.2.8.3(b), and the light is dispersed into the same sequence of wavelengths as seen in Figure 4.2.8.1 and Figure 4.2.8.2.

MAKING CONNECTIONS: DISPERSION

Any type of wave can exhibit dispersion. Sound waves, all types of electromagnetic waves, and water waves can be dispersed according to wavelength. Dispersion occurs whenever the speed of propagation depends on wavelength, thus separating and spreading out various wavelengths. Dispersion may require special circumstances and can result in spectacular displays such as in the production of a rainbow. This is also true for sound, since all frequencies ordinarily travel at the same speed. If you listen to sound through a long tube, such as a vacuum cleaner hose, you can easily hear it is dispersed by interaction with the tube. Dispersion, in fact, can reveal a great deal about what the wave has encountered that disperses its wavelengths. The dispersion of electromagnetic radiation from outer space, for example, has revealed much about what exists between the stars—the so-called empty space.

Table 4.2.8.1: Index of Refraction *n* in Selected Media at Various Wavelengths

Medium	Red (660 nm)	Orange (610 nm)	Yellow (580 nm)	Green (550 nm)	Blue (470 nm)	Violet (410 nm)
Water	1.331	1.332	1.333	1.335	1.338	1.342
Diamond	2.410	2.415	2.417	2.426	2.444	2.458
Glass, crown	1.512	1.514	1.518	1.519	1.524	1.530
Glass, flint	1.662	1.665	1.667	1.674	1.684	1.698
Polystyrene	1.488	1.490	1.492	1.493	1.499	1.506
Quartz, fused	1.455	1.456	1.458	1.459	1.462	1.468

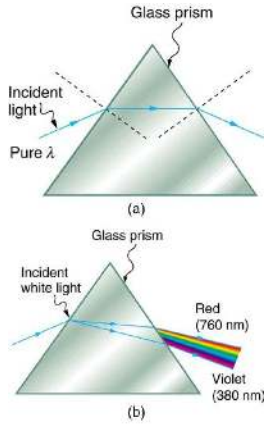


Figure 4.2.8.3: (a) A pure wavelength of light falls onto a prism and is refracted at both surfaces. (b) White light is dispersed by the prism (shown exaggerated). Since the index of refraction varies with wavelength, the angles of refraction vary with wavelength. A sequence of red to violet is produced, because the index of refraction increases steadily with decreasing wavelength.

Rainbows are produced by a combination of refraction and reflection. You may have noticed that you see a rainbow only when you look away from the sun. Light enters a drop of water and is reflected from the back of the drop, as shown in Figure 4.2.8.4. The light is refracted both as it enters and as it leaves the drop. Since the index of refraction of water varies with wavelength, the light is dispersed, and a rainbow is observed, as shown in Figure 4.2.8.5(a). (There is no dispersion caused by reflection at the back surface, since the law of reflection does not depend on wavelength.) The actual rainbow of colors seen by an observer depends on the myriad of rays being refracted and reflected toward the observer's eyes from numerous drops of water. The effect is most spectacular when the background is dark, as in stormy weather, but can also be observed in waterfalls and lawn sprinklers. The arc of a rainbow comes from the need to be looking at a specific angle relative to the direction of the sun, as illustrated in Figure 4.2.8.5(b). (If there are two reflections of light within the water drop, another "secondary" rainbow is produced. This rare event produces an arc that lies above the primary rainbow arc—see Figure 4.2.8.5(c).)

Definition: RAINBOWS

Rainbows are produced by a combination of refraction and reflection.

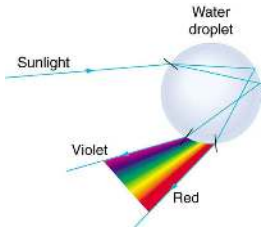


Figure 4.2.8.4: Part of the light falling on this water drop enters and is reflected from the back of the drop. This light is refracted and dispersed both as it enters and as it leaves the drop.

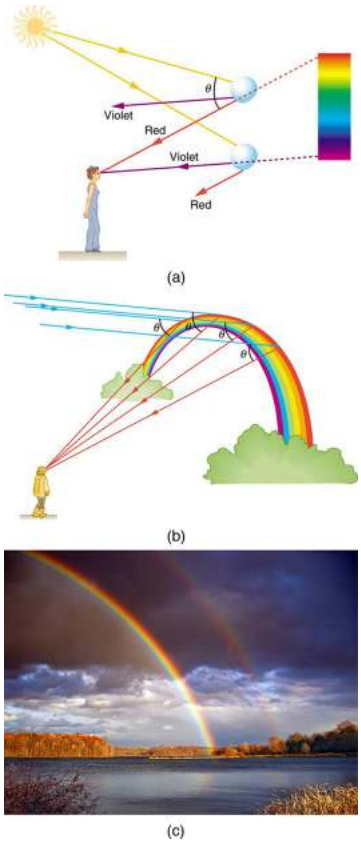


Figure 4.2.8.5: (a) Different colors emerge in different directions, and so you must look at different locations to see the various colors of a rainbow. (b) The arc of a rainbow results from the fact that a line between the observer and any point on the arc must make the correct angle with the parallel rays of sunlight to receive the refracted rays. (c) Double rainbow. (credit: Nicholas, Wikimedia Commons)

Dispersion may produce beautiful rainbows, but it can cause problems in optical systems. White light used to transmit messages in a fiber is dispersed, spreading out in time and eventually overlapping with other messages. Since a laser produces a nearly pure wavelength, its light experiences little dispersion, an advantage over white light for transmission of information. In contrast, dispersion of electromagnetic waves coming to us from outer space can be used to determine the amount of matter they pass through. As with many phenomena, dispersion can be useful or a nuisance, depending on the situation and our human goals.

Section Summary

- The spreading of white light into its full spectrum of wavelengths is called dispersion.
- Rainbows are produced by a combination of refraction and reflection and involve the dispersion of sunlight into a continuous distribution of colors.
- Dispersion produces beautiful rainbows but also causes problems in certain optical systems.

Glossary

dispersion

spreading of white light into its full spectrum of wavelengths

rainbow

dispersion of sunlight into a continuous distribution of colors according to wavelength, produced by the refraction and reflection of sunlight by water droplets in the sky

---

This page titled 4.2.8: Dispersion- The Rainbow and Prisms is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- **11.8: Dispersion- The Rainbow and Prisms** by OpenStax is licensed CC BY 4.0.

4.2.9: Image Formation by Mirrors

Learning Objectives

- Illustrate image formation in a flat mirror.
- Explain with ray diagrams the formation of an image using spherical mirrors.
- Determine focal length and magnification given radius of curvature, distance of object and image.

We only have to look as far as the nearest bathroom to find an example of an image formed by a mirror. Images in flat mirrors are the same size as the object and are located behind the mirror. Like lenses, mirrors can form a variety of images. For example, dental mirrors may produce a magnified image, just as makeup mirrors do. Security mirrors in shops, on the other hand, form images that are smaller than the object. We will use the law of reflection to understand how mirrors form images, and we will find that mirror images are analogous to those formed by lenses.

Figure 4.2.9.1 helps illustrate how a flat mirror forms an image. Two rays are shown emerging from the same point, striking the mirror, and being reflected into the observer's eye. The rays can diverge slightly, and both still get into the eye. If the rays are extrapolated backward, they seem to originate from a common point behind the mirror, locating the image. (The paths of the reflected rays into the eye are the same as if they had come directly from that point behind the mirror.) Using the law of reflection —the angle of reflection equals the angle of incidence—we can see that the image and object are the same distance from the mirror. This is a virtual image, since it cannot be projected—the rays only appear to originate from a common point behind the mirror. Obviously, if you walk behind the mirror, you cannot see the image, since the rays do not go there. But in front of the mirror, the rays behave exactly as if they had come from behind the mirror, so that is where the image is situated.

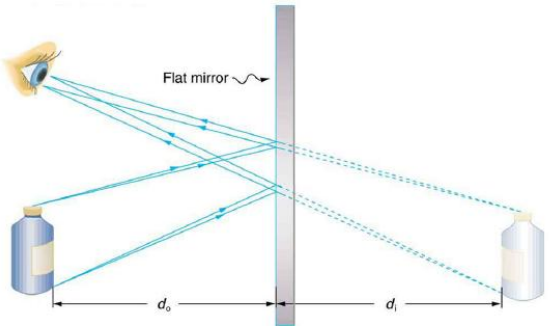


Figure 4.2.9.1: Two sets of rays from common points on an object are reflected by a flat mirror into the eye of an observer. The reflected rays seem to originate from behind the mirror, locating the virtual image.

Now let us consider the focal length of a mirror—for example, the concave spherical mirrors in Figure 4.2.9.2. Rays of light that strike the surface follow the law of reflection. For a mirror that is large compared with its radius of curvature, as in Figure 4.2.9.1(a), we see that the reflected rays do not cross at the same point, and the mirror does not have a well-defined focal point. If the mirror had the shape of a parabola, the rays would all cross at a single point, and the mirror would have a well-defined focal point. But parabolic mirrors are much more expensive to make than spherical mirrors. The solution is to use a mirror that is small compared with its radius of curvature, as shown in Figure 4.2.9.2(b). (This is the mirror equivalent of the thin lens approximation.) To a very good approximation, this mirror has a well-defined focal point at F that is the focal distance f from the center of the mirror. The focal length f of a concave mirror is positive, since it is a converging mirror.

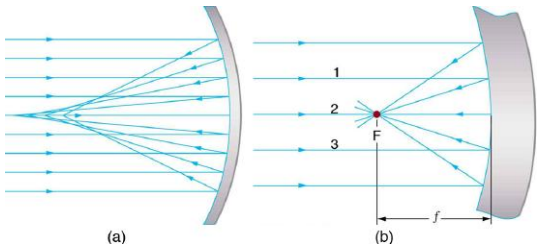


Figure 4.2.9.2: (a) Parallel rays reflected from a large spherical mirror do not all cross at a common point. (b) If a spherical mirror is small compared with its radius of curvature, parallel rays are focused to a common point. The distance of the focal point from the center of the mirror is its focal length f. Since this mirror is converging, it has a positive focal length.

Just as for lenses, the shorter the focal length, the more powerful the mirror; thus, $P = 1/f$ for a mirror, too. A more strongly curved mirror has a shorter focal length and a greater power. Using the law of reflection and some simple trigonometry, it can be shown that the focal length is half the radius of curvature, or

$$f = \frac{R}{2},$$

where R is the radius of curvature of a spherical mirror. The smaller the radius of curvature, the smaller the focal length and, thus, the more powerful the mirror.

The convex mirror shown in Figure 4.2.9.3 also has a focal point. Parallel rays of light reflected from the mirror seem to originate from the point F at the focal distance f behind the mirror. The focal length and power of a convex mirror are negative, since it is a diverging mirror.

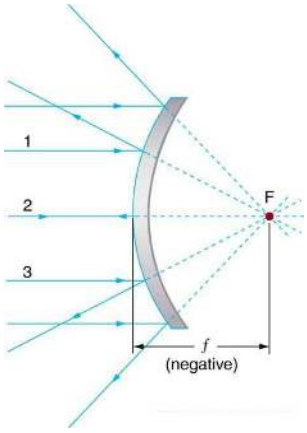


Figure 4.2.9.3: Parallel rays of light reflected from a convex spherical mirror (small in size compared with its radius of curvature) seem to originate from a well-defined focal point at the focal distance f behind the mirror. Convex mirrors diverge light rays and, thus, have a negative focal length.

Ray tracing is as useful for mirrors as for lenses. The rules for ray tracing for mirrors are based on the illustrations just discussed:

1. **Principal Ray 1**: a ray approaching a concave converging mirror parallel to its axis is reflected through the focal point F of the mirror on the same side. (See rays 1 and 3 in Figure 4.2.9.2(b).); a ray approaching a convex diverging mirror parallel to its axis is reflected so that it seems to come from the focal point F behind the mirror. (See rays 1 and 3 in Figure 4.2.9.3.)
2. **Principal Ray 2**: any ray striking the center of a mirror is followed by applying the law of reflection; it makes the same angle with the axis when leaving as when approaching. (See ray 2 in Figure 4.2.9.4.)
3. **Principal Ray 3**: a ray approaching a concave converging mirror through its focal point is reflected parallel to its axis. (The reverse of rays 1 and 3 in Figure 4.2.9.2.); a ray approaching a convex diverging mirror by heading toward its focal point on the opposite side is reflected parallel to the axis. (The reverse of rays 1 and 3 in Figure 4.2.9.3.) *The third principal ray is optional and may be used to verify the accuracy of image location.*

We will use ray tracing to illustrate how images are formed by mirrors, and we can use ray tracing quantitatively to obtain numerical information. But since we assume each mirror is small compared with its radius of curvature, we can use the thin lens equations for mirrors just as we did for lenses.

Consider the situation shown in Figure 4.2.9.4, concave spherical mirror reflection, in which an object is placed farther from a concave (converging) mirror than its focal length. That is, f is positive and $d_o > f$, so that we may expect an image similar to the case 1 real image formed by a converging lens. Ray tracing in Figure 4.2.9.4 shows that the rays from a common point on the object all cross at a point on the same side of the mirror as the object. Thus a real image can be projected onto a screen placed at this location. The image distance is positive, and the image is inverted, so its magnification is negative. This is a *case 1 image for mirrors*. It differs from the case 1 image for lenses only in that the image is on the same side of the mirror as the object. It is otherwise identical.

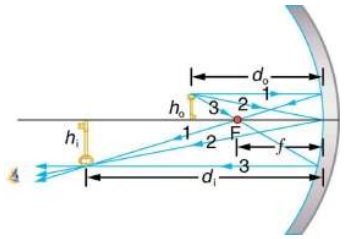


Figure 4.2.9.4: A case 1 image for a mirror. An object is farther from the converging mirror than its focal length. Rays from a common point on the object are traced using the rules in the text. Ray 1 approaches parallel to the axis, ray 2 strikes the center of the mirror, and ray 3 goes through the focal point on the way toward the mirror. All three rays cross at the same point after being reflected, locating the inverted real image. Although three rays are shown, only two of the three are needed to locate the image and determine its height.

Example 4.2.9.1: A Concave Reflector

Electric room heaters use a concave mirror to reflect infrared (IR) radiation from hot coils. Note that IR follows the same law of reflection as visible light. Given that the mirror has a radius of curvature of 50.0 cm and produces an image of the coils 3.00 m away from the mirror, where are the coils?

Strategy and Concept

We are given that the concave mirror projects a real image of the coils at an image distance $d_\mathrm{i} = 3.00\ \mathrm{m}$. The coils are the object, and we are asked to find their location—that is, to find the object distance d_o. We are also given the radius of curvature of the mirror, so that its focal length is $f = R/2 = 25.0\ \mathrm{cm}$ (positive since the mirror is concave or converging). Assuming the mirror is small compared with its radius of curvature, we can use the thin lens equations, to solve this problem.

Solution

Since d_i and f are known, thin lens equation can be used to find d_o:

$$\frac{1}{d_\mathrm{o}} + \frac{1}{d_\mathrm{i}} = \frac{1}{f}.$$

Rearranging to isolate d_o gives

$$\frac{1}{d_\mathrm{o}} = \frac{1}{f} - \frac{1}{d_\mathrm{i}}.$$

Entering known quantities gives a value for $1/d_\mathrm{o}$:

$$\frac{1}{d_\mathrm{o}} = \frac{1}{0.250\ \mathrm{m}} - \frac{1}{3.00\ \mathrm{m}} = \frac{3.667}{\mathrm{m}}.$$

This must be inverted to find d_o:

$$d_\mathrm{o} = \frac{1\ \mathrm{m}}{3.667} = 27.3\ \mathrm{cm}.$$

Discussion

Note that the object (the filament) is farther from the mirror than the mirror's focal length. This is a case 1 image ($d_\mathrm{o} > f$ and f positive), consistent with the fact that a real image is formed. You will get the most concentrated thermal energy directly in front of the mirror and 3.00 m away from it. Generally, this is not desirable, since it could cause burns. Usually, you want the rays to emerge parallel, and this is accomplished by having the filament at the focal point of the mirror.

Note that the filament here is not much farther from the mirror than its focal length and that the image produced is considerably farther away. This is exactly analogous to a slide projector. Placing a slide only slightly farther away from the projector lens than its focal length produces an image significantly farther away. As the object gets closer to the focal distance, the image gets

farther away. In fact, as the object distance approaches the focal length, the image distance approaches infinity and the rays are sent out parallel to one another.

Example 4.2.9.2: Solar Electric Generating System

One of the solar technologies used today for generating electricity is a device (called a parabolic trough or concentrating collector) that concentrates the sunlight onto a blackened pipe that contains a fluid. This heated fluid is pumped to a heat exchanger, where its heat energy is transferred to another system that is used to generate steam—and so generate electricity through a conventional steam cycle. Figure 4.2.9.5 shows such a working system in southern California. Concave mirrors are used to concentrate the sunlight onto the pipe. The mirror has the approximate shape of a section of a cylinder. For the problem, assume that the mirror is exactly one-quarter of a full cylinder.

a. If we wish to place the fluid-carrying pipe 40.0 cm from the concave mirror at the mirror's focal point, what will be the radius of curvature of the mirror?
b. Per meter of pipe, what will be the amount of sunlight concentrated onto the pipe, assuming the insolation (incident solar radiation) is $0.900\ \text{kW/m}^2$?
c. If the fluid-carrying pipe has a 2.00-cm diameter, what will be the temperature increase of the fluid per meter of pipe over a period of one minute? Assume all the solar radiation incident on the reflector is absorbed by the pipe, and that the fluid is mineral oil.

Strategy

To solve an *Integrated Concept Problem* we must first identify the physical principles involved. Part (a) is related to the current topic. Part (b) involves a little math, primarily geometry. Part (c) requires an understanding of heat and density.

Solution to (a)

To a good approximation for a concave or semi-spherical surface, the point where the parallel rays from the sun converge will be at the focal point, so $R = 2f = 80.0\ \text{cm}$.

Solution to (b)

The insolation is $900\ \text{W/m}^2$. We must find the cross-sectional area A of the concave mirror, since the power delivered is $900\ \text{W/m}^2 \times \text{A}$. The mirror in this case is a quarter-section of a cylinder, so the area for a length L of the mirror is $\text{A} = \frac{1}{4}(2\pi R)\text{L}$. The area for a length of 1.00 m is then

$$\text{A} = \frac{\pi}{2}R(1.00\ \text{m}) = \frac{(3.14)}{2}(0.800\ \text{m})(1.00\ \text{m}) = 1.26\ \text{m}^2.$$

The insolation on the 1.00-m length of pipe is then

$$\left(9.00 \times 10^2 \frac{\text{W}}{\text{m}^2}\right)\left(1.26\ \text{m}^2\right) = 1130\ \text{W}.$$

Solution to (c)

The increase in temperature is given by $Q = mc\Delta T$. The mass m of the mineral oil in the one-meter section of pipe is

$$\begin{aligned} m = \rho\text{V} &= \rho\pi\left(\frac{d}{2}\right)^2(1.00\ \text{m}) \\ &= \left(8.00 \times 10^2\ \text{kg/m}^3\right)(3.14)(0.0100\ \text{m})^2(1.00\ \text{m}) \\ &= 0.251\ \text{kg}. \end{aligned}$$

Therefore, the increase in temperature in one minute is

$$\begin{aligned} \Delta T &= Q/mc \\ &= \frac{(1130\ \text{W})(60.0\ \text{s})}{(0.251\ \text{kg})\,(1670\ \text{J}\cdot\text{kg}/^\circ\text{C})} \\ &= 162^\circ\text{C}. \end{aligned}$$

Discussion for (c)

An array of such pipes in the California desert can provide a thermal output of 250 MW on a sunny day, with fluids reaching temperatures as high as 400°C. We are considering only one meter of pipe here, and ignoring heat losses along the pipe.



Figure 4.2.9.5: Parabolic trough collectors are used to generate electricity in southern California. (credit: kjkolb, Wikimedia Commons)

What happens if an object is closer to a concave mirror than its focal length? This is analogous to a case 2 image for lenses ($d_o < f$ and f positive), which is a magnifier. In fact, this is how makeup mirrors act as magnifiers. Figure 4.2.9.6(a) uses ray tracing to locate the image of an object placed close to a concave mirror. Rays from a common point on the object are reflected in such a manner that they appear to be coming from behind the mirror, meaning that the image is virtual and cannot be projected. As with a magnifying glass, the image is upright and larger than the object. This is a *case 2 image for mirrors* and is exactly analogous to that for lenses.

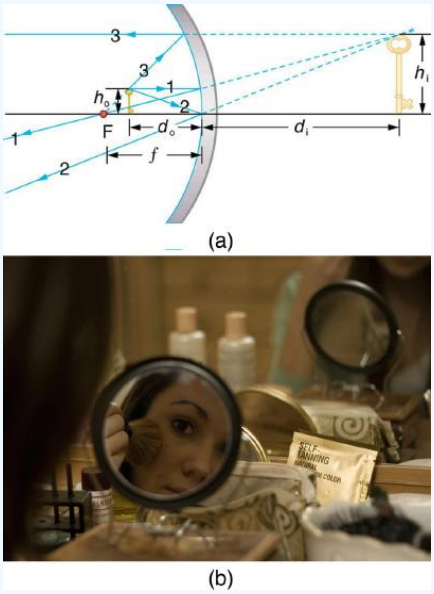
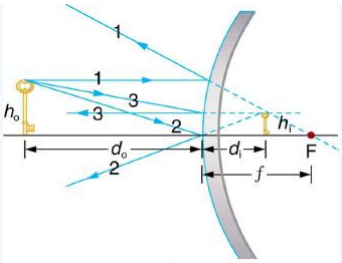


Figure 4.2.9.6: (a) Case 2 images for mirrors are formed when a converging mirror has an object closer to it than its focal length. Ray 1 approaches parallel to the axis, ray 2 strikes the center of the mirror, and ray 3 approaches the mirror as if it came from the focal point. (b) A magnifying mirror showing the reflection. (credit: Mike Melrose, Flickr)

All three rays appear to originate from the same point after being reflected, locating the upright virtual image behind the mirror and showing it to be larger than the object. (b) Makeup mirrors are perhaps the most common use of a concave mirror to produce a larger, upright image.

A convex mirror is a diverging mirror (f is negative) and forms only one type of image. It is a *case 3* image—one that is upright and smaller than the object, just as for diverging lenses. Figure 4.2.9.7(a) uses ray tracing to illustrate the location and size of the case 3 image for mirrors. Since the image is behind the mirror, it cannot be projected and is thus a virtual image. It is also seen to be smaller than the object.



(a)



(b)

Figure 4.2.9.7: Case 3 images for mirrors are formed by any convex mirror. Ray 1 approaches parallel to the axis, ray 2 strikes the center of the mirror, and ray 3 approaches toward the focal point. All three rays appear to originate from the same point after being reflected, locating the upright virtual image behind the mirror and showing it to be smaller than the object. (b) Security mirrors are convex, producing a smaller, upright image. Because the image is smaller, a larger area is imaged compared to what would be observed for a flat mirror (and hence security is improved). (credit: Laura D'Alessandro, Flickr)

Example 4.2.9.3: Image in a Convex Mirror

A keratometer is a device used to measure the curvature of the cornea, particularly for fitting contact lenses. Light is reflected from the cornea, which acts like a convex mirror, and the keratometer measures the magnification of the image. The smaller the magnification, the smaller the radius of curvature of the cornea. If the light source is 12.0 cm from the cornea and the image's magnification is 0.0320, what is the cornea's radius of curvature?

Strategy

If we can find the focal length of the convex mirror formed by the cornea, we can find its radius of curvature (the radius of curvature is twice the focal length of a spherical mirror). We are given that the object distance is $d_o = 12.0\ \text{cm}$ and that $m = 0.0320$. We first solve for the image distance d_i, and then for f.

Solution

$m = -d_i/d_o$. Solving this expression for d_i gives

$$d_i = -md_o.$$

Entering known values yields

$$d_i = -(0.0320)(12.0\ \text{cm}) = -0.384\ \text{cm}.$$

$$\frac{1}{f} = \frac{1}{d_o} + \frac{1}{d_i}$$

Substituting known values,

$$\frac{1}{f} = \frac{1}{12.0\ \text{cm}} + \frac{1}{-0.384\ \text{cm}} = \frac{-2.52}{\text{cm}}.$$

This must be inverted to find f:

$$f = \frac{\text{cm}}{-2.52} = -0.400\ \text{cm}.$$

The radius of curvature is twice the focal length, so that

$$R = 2|f| = 0.800\ \text{cm}.$$

Discussion

Although the focal length f of a convex mirror is defined to be negative, we take the absolute value to give us a positive value for R. The radius of curvature found here is reasonable for a cornea. The distance from cornea to retina in an adult eye is about 2.0 cm. In practice, many corneas are not spherical, complicating the job of fitting contact lenses. Note that the image distance here is negative, consistent with the fact that the image is behind the mirror, where it cannot be projected. In this section's Problems and Exercises, you will show that for a fixed object distance, the smaller the radius of curvature, the smaller the magnification.

The three types of images formed by mirrors (cases 1, 2, and 3) are exactly analogous to those formed by lenses, as summarized in the table at the end of "Image Formation by Lenses". It is easiest to concentrate on only three types of images—then remember that concave mirrors act like convex lenses, whereas convex mirrors act like concave lenses.

TAKE-HOME EXPERIMENT: CONCAVE MIRRORS CLOSE TO HOME

Find a flashlight and identify the curved mirror used in it. Find another flashlight and shine the first flashlight onto the second one, which is turned off. Estimate the focal length of the mirror. You might try shining a flashlight on the curved mirror behind the headlight of a car, keeping the headlight switched off, and determine its focal length.

Section Summary

- The characteristics of an image formed by a flat mirror are: (a) The image and object are the same distance from the mirror, (b) The image is a virtual image, and (c) The image is situated behind the mirror.
- Image length is half the radius of curvature.

$$f = \frac{R}{2}$$

- A convex mirror is a diverging mirror and forms only one type of image, namely a virtual image.

Glossary

converging mirror
a concave mirror in which light rays that strike it parallel to its axis converge at one or more points along the axis

diverging mirror
a convex mirror in which light rays that strike it parallel to its axis bend away (diverge) from its axis

law of reflection
angle of reflection equals the angle of incidence

This page titled 4.2.9: Image Formation by Mirrors is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- **11.10: Image Formation by Mirrors** by OpenStax is licensed CC BY 4.0.

4.2.10: Polarization

Learning Objectives

- Discuss the meaning of polarization.
- Discuss the property of optical activity of certain materials.

Polaroid sunglasses are familiar to most of us. They have a special ability to cut the glare of light reflected from water or glass (see Figure 4.2.10.1). Polaroids have this ability because of a wave characteristic of light called polarization. What is polarization? How is it produced? What are some of its uses? The answers to these questions are related to the wave character of light.



Figure 4.2.10.1: These two photographs of a river show the effect of a polarizing filter in reducing glare in light reflected from the surface of water. Part (b) of this figure was taken with a polarizing filter and part (a) was not. As a result, the reflection of clouds and sky observed in part (a) is not observed in part (b). Polarizing sunglasses are particularly useful on snow and water. (credit: Amithshs, Wikimedia Commons)

Light is one type of electromagnetic (EM) wave. As noted earlier, EM waves are *transverse waves* consisting of varying electric and magnetic fields that oscillate perpendicular to the direction of propagation (see Figure 4.2.10.2). There are specific directions for the oscillations of the electric and magnetic fields. **Polarization** is the attribute that a wave's oscillations have a definite direction relative to the direction of propagation of the wave. (This is not the same type of polarization as that discussed for the separation of charges.) Waves having such a direction are said to be **polarized**. For an EM wave, we define the **direction of polarization** to be the direction parallel to the electric field. Thus we can think of the electric field arrows as showing the direction of polarization, as in Figure 4.2.10.2.

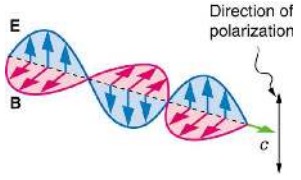


Figure 4.2.10.2: An EM wave, such as light, is a transverse wave. The electric and magnetic fields are perpendicular to the direction of propagation.

To examine this further, consider the transverse waves in the ropes shown in Figure 4.2.10.3. The oscillations in one rope are in a vertical plane and are said to be **vertically polarized**. Those in the other rope are in a horizontal plane and are **horizontally polarized**. If a vertical slit is placed on the first rope, the waves pass through. However, a vertical slit blocks the horizontally polarized waves. For EM waves, the direction of the electric field is analogous to the disturbances on the ropes.

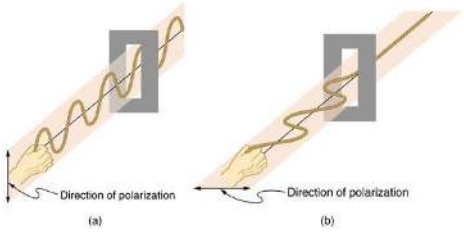


Figure 4.2.10.3: The transverse oscillations in one rope are in a vertical plane, and those in the other rope are in a horizontal plane. The first is said to be vertically polarized, and the other is said to be horizontally polarized. Vertical slits pass vertically polarized waves and block horizontally polarized waves.

The Sun and many other light sources produce waves that are randomly polarized (see Figure 4.2.10.4). Such light is said to be **unpolarized** because it is composed of many waves with all possible directions of polarization. Polaroid materials, invented by the founder of Polaroid Corporation, Edwin Land, act as a *polarizing* slit for light, allowing only polarization in one direction to

pass through. Polarizing filters are composed of long molecules aligned in one direction. Thinking of the molecules as many slits, analogous to those for the oscillating ropes, we can understand why only light with a specific polarization can get through. The **axis of a polarizing filter** is the direction along which the filter passes the electric field of an EM wave (see Figure 4.2.10.5).

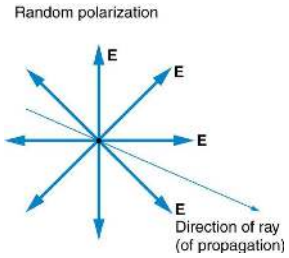


Figure 4.2.10.4: The slender arrow represents a ray of unpolarized light. The bold arrows represent the direction of polarization of the individual waves composing the ray. Since the light is unpolarized, the arrows point in all directions.

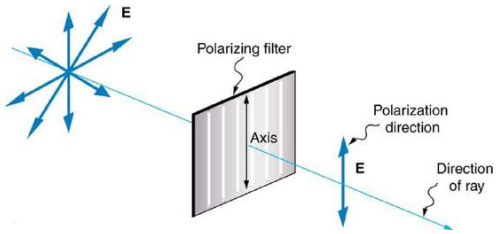


Figure 4.2.10.5: A polarizing filter has a polarization axis that acts as a slit passing through electric fields parallel to its direction. The direction of polarization of an EM wave is defined to be the direction of its electric field.

Figure 4.2.10.6 shows the effect of two polarizing filters on originally unpolarized light. The first filter polarizes the light along its axis. When the axes of the first and second filters are aligned (parallel), then all of the polarized light passed by the first filter is also passed by the second. If the second polarizing filter is rotated, only the component of the light parallel to the second filter's axis is passed. When the axes are perpendicular, no light is passed by the second.

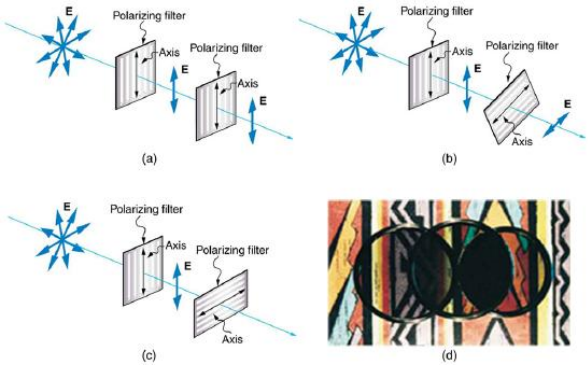


Figure 4.2.10.6: The effect of rotating two polarizing filters, where the first polarizes the light. (a) All of the polarized light is passed by the second polarizing filter, because its axis is parallel to the first. (b) As the second is rotated, only part of the light is passed. (c) When the second is perpendicular to the first, no light is passed. (d) In this photograph, a polarizing filter is placed above two others. Its axis is perpendicular to the filter on the right (dark area) and parallel to the filter on the left (lighter area). (credit: P.P. Urone)

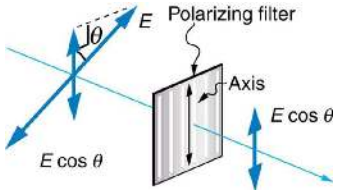


Figure 4.2.10.7: A polarizing filter transmits only the component of the wave parallel to its axis, $E\cos\theta$, reducing the intensity of any light not polarized parallel to its axis.

Polarization by Reflection

By now you can probably guess that Polaroid sunglasses cut the glare in reflected light because that light is polarized. You can check this for yourself by holding Polaroid sunglasses in front of you and rotating them while looking at light reflected from water or glass. As you rotate the sunglasses, you will notice the light gets bright and dim, but not completely black. This implies the reflected light is partially polarized and cannot be completely blocked by a polarizing filter.

Figure 4.2.10.8 illustrates what happens when unpolarized light is reflected from a surface. Vertically polarized light is preferentially refracted at the surface, so that *the reflected light is left more horizontally polarized*. The reasons for this phenomenon are beyond the scope of this text, but a convenient mnemonic for remembering this is to imagine the polarization direction to be like an arrow. Vertical polarization would be like an arrow perpendicular to the surface and would be more likely to stick and not be reflected. Horizontal polarization is like an arrow bouncing on its side and would be more likely to be reflected. Sunglasses with vertical axes would then block more reflected light than unpolarized light from other sources.

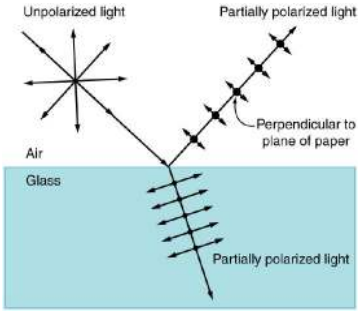


Figure 4.2.10.8: Polarization by reflection. Unpolarized light has equal amounts of vertical and horizontal polarization. After interaction with a surface, the vertical components are preferentially absorbed or refracted, leaving the reflected light more horizontally polarized. This is akin to arrows striking on their sides bouncing off, whereas arrows striking on their tips go into the surface.

THINGS GREAT AND SMALL: ATOMIC EXPLANATION OF POLARIZING FILTERS

Polarizing filters have a polarization axis that acts as a slit. This slit passes electromagnetic waves (often visible light) that have an electric field parallel to the axis. This is accomplished with long molecules aligned perpendicular to the axis as shown in Figure 4.2.10.9.

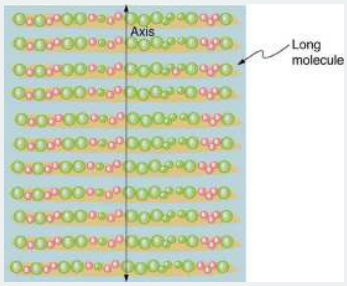


Figure 4.2.10.9: Long molecules are aligned perpendicular to the axis of a polarizing filter. The component of the electric field in an EM wave perpendicular to these molecules passes through the filter, while the component parallel to the molecules is absorbed.

Figure 4.2.10.10 illustrates how the component of the electric field parallel to the long molecules is absorbed. An electromagnetic wave is composed of oscillating electric and magnetic fields. The electric field is strong compared with the magnetic field and is more effective in exerting force on charges in the molecules. The most affected charged particles are the electrons in the molecules, since electron masses are small. If the electron is forced to oscillate, it can absorb energy from the EM wave. This reduces the fields in the wave and, hence, reduces its intensity. In long molecules, electrons can more easily oscillate parallel to the molecule than in the perpendicular direction. The electrons are bound to the molecule and are more restricted in their movement perpendicular to the molecule. Thus, the electrons can absorb EM waves that have a component of their electric field parallel to the molecule. The electrons are much less responsive to electric fields perpendicular to the molecule and will allow those fields to pass. Thus the axis of the polarizing filter is perpendicular to the length of the molecule.

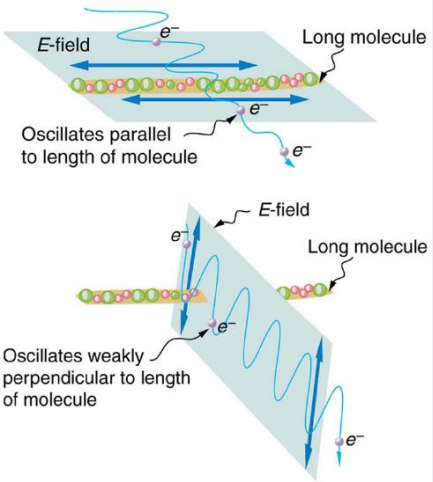


Figure 4.2.10.10: Artist's conception of an electron in a long molecule oscillating parallel to the molecule. The oscillation of the electron absorbs energy and reduces the intensity of the component of the EM wave that is parallel to the molecule.

Polarization by Scattering

If you hold your Polaroid sunglasses in front of you and rotate them while looking at blue sky, you will see the sky get bright and dim. This is a clear indication that light scattered by air is partially polarized. Figure 4.2.10.11 helps illustrate how this happens. Since light is a transverse EM wave, it vibrates the electrons of air molecules perpendicular to the direction it is traveling. The electrons then radiate like small antennae. Since they are oscillating perpendicular to the direction of the light ray, they produce EM radiation that is polarized perpendicular to the direction of the ray. When viewing the light along a line perpendicular to the original ray, as in Figure 4.2.10.11, there can be no polarization in the scattered light parallel to the original ray, because that would require the original ray to be a longitudinal wave. Along other directions, a component of the other polarization can be projected along the line of sight, and the scattered light will only be partially polarized. Furthermore, multiple scattering can bring light to your eyes from other directions and can contain different polarizations.

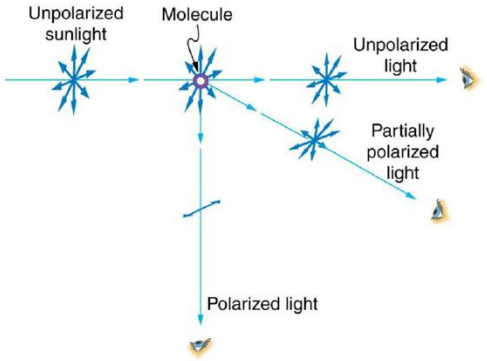


Figure 4.2.10.11: Polarization by scattering. Unpolarized light scattering from air molecules shakes their electrons perpendicular to the direction of the original ray. The scattered light therefore has a polarization perpendicular to the original direction and none parallel to the original direction.

Photographs of the sky can be darkened by polarizing filters, a trick used by many photographers to make clouds brighter by contrast. Scattering from other particles, such as smoke or dust, can also polarize light. Detecting polarization in scattered EM waves can be a useful analytical tool in determining the scattering source.

There is a range of optical effects used in sunglasses. Besides being Polaroid, other sunglasses have colored pigments embedded in them, while others use non-reflective or even reflective coatings. A recent development is photochromic lenses, which darken in

the sunlight and become clear indoors. Photochromic lenses are embedded with organic microcrystalline molecules that change their properties when exposed to UV in sunlight, but become clear in artificial lighting with no UV.

TAKE-HOME EXPERIMENT: POLARIZATION

Find Polaroid sunglasses and rotate one while holding the other still and look at different surfaces and objects. Explain your observations. What is the difference in angle from when you see a maximum intensity to when you see a minimum intensity? Find a reflective glass surface and do the same. At what angle does the glass need to be oriented to give minimum glare?

Liquid Crystals and Other Polarization Effects in Materials

While you are undoubtedly aware of liquid crystal displays (LCDs) found in watches, calculators, computer screens, cellphones, flat screen televisions, and other myriad places, you may not be aware that they are based on polarization. Liquid crystals are so named because their molecules can be aligned even though they are in a liquid. Liquid crystals have the property that they can rotate the polarization of light passing through them by 90°. Furthermore, this property can be turned off by the application of a voltage, as illustrated in Figure 4.2.10.12. It is possible to manipulate this characteristic quickly and in small well-defined regions to create the contrast patterns we see in so many LCD devices.

In flat screen LCD televisions, there is a large light at the back of the TV. The light travels to the front screen through millions of tiny units called pixels (picture elements). One of these is shown in Figure 4.2.10.1 (a) and (b). Each unit has three cells, with red, blue, or green filters, each controlled independently. When the voltage across a liquid crystal is switched off, the liquid crystal passes the light through the particular filter. One can vary the picture contrast by varying the strength of the voltage applied to the liquid crystal.

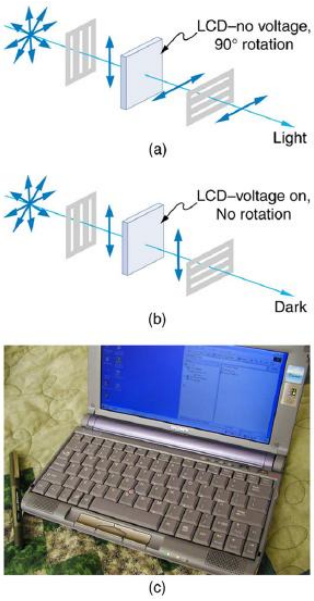


Figure 4.2.10.12: (a) Polarized light is rotated 90° by a liquid crystal and then passed by a polarizing filter that has its axis perpendicular to the original polarization direction. (b) When a voltage is applied to the liquid crystal, the polarized light is not rotated and is blocked by the filter, making the region dark in comparison with its surroundings. (c) LCDs can be made color specific, small, and fast enough to use in laptop computers and TVs. (credit: Jon Sullivan)

Many crystals and solutions rotate the plane of polarization of light passing through them. Such substances are said to be **optically active**. Examples include sugar water, insulin, and collagen (see Figure 4.2.10.13). In addition to depending on the type of substance, the amount and direction of rotation depends on a number of factors. Among these is the concentration of the substance, the distance the light travels through it, and the wavelength of light. Optical activity is due to the asymmetric shape of molecules in the substance, such as being helical. Measurements of the rotation of polarized light passing through substances can thus be used to measure concentrations, a standard technique for sugars. It can also give information on the shapes of molecules, such as proteins, and factors that affect their shapes, such as temperature and pH.

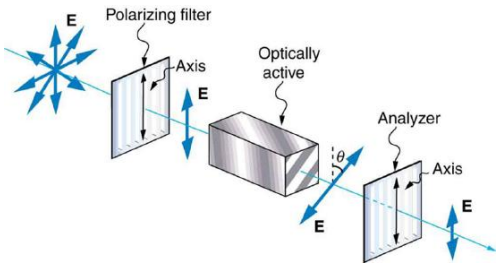


Figure 4.2.10.13: Optical activity is the ability of some substances to rotate the plane of polarization of light passing through them. The rotation is detected with a polarizing filter or analyzer.

Glass and plastic become optically active when stressed; the greater the stress, the greater the effect. Optical stress analysis on complicated shapes can be performed by making plastic models of them and observing them through crossed filters, as seen in Figure 4.2.10.14. It is apparent that the effect depends on wavelength as well as stress. The wavelength dependence is sometimes also used for artistic purposes.

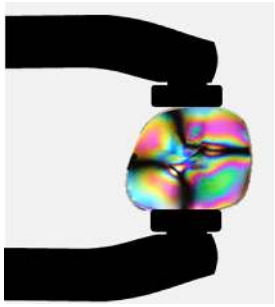


Figure 4.2.10.14: Optical stress analysis of a plastic lens placed between crossed polarizers. (credit: Infopro, Wikimedia Commons)

Another interesting phenomenon associated with polarized light is the ability of some crystals to split an unpolarized beam of light into two. Such crystals are said to be **birefringent** (see Figure 4.2.10.15). Each of the separated rays has a specific polarization. One behaves normally and is called the ordinary ray, whereas the other does not obey Snell's law and is called the extraordinary ray. Birefringent crystals can be used to produce polarized beams from unpolarized light. Some birefringent materials preferentially absorb one of the polarizations. These materials are called dichroic and can produce polarization by this preferential absorption. This is fundamentally how polarizing filters and other polarizers work. The interested reader is invited to further pursue the numerous properties of materials related to polarization.

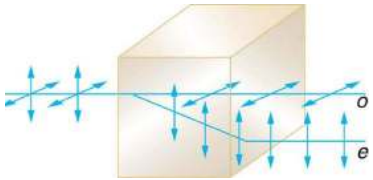


Figure 4.2.10.15: Birefringent materials, such as the common mineral calcite, split unpolarized beams of light into two. The ordinary ray behaves as expected, but the extraordinary ray does not obey Snell's law.

Section Summary

- Polarization is the attribute that wave oscillations have a definite direction relative to the direction of propagation of the wave.
- EM waves are transverse waves that may be polarized.
- The direction of polarization is defined to be the direction parallel to the electric field of the EM wave.
- Unpolarized light is composed of many rays having random polarization directions.
- Light can be polarized by passing it through a polarizing filter or other polarizing material. The intensity of polarized light after passing through a polarizing filter depends on the direction of polarization and the axis of the filter.
- Polarization is also produced by reflection.
- Polarization can also be produced by scattering.
- There are a number of types of optically active substances that rotate the direction of polarization of light passing through them.

Glossary

axis of a polarizing filter
the direction along which the filter passes the electric field of an EM wave

birefringent
crystals that split an unpolarized beam of light into two beams

direction of polarization
the direction parallel to the electric field for EM waves

horizontally polarized
the oscillations are in a horizontal plane

optically active
substances that rotate the plane of polarization of light passing through them

polarization
the attribute that wave oscillations have a definite direction relative to the direction of propagation of the wave

polarized
waves having the electric and magnetic field oscillations in a definite direction

unpolarized
waves that are randomly polarized

vertically polarized
the oscillations are in a vertical plane

This page titled 4.2.10: Polarization is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- **11.11: Polarization** by OpenStax is licensed CC BY 4.0.

4.3: Image Formation by Lenses

Learning Objectives

- List the rules for ray tracking for thin lenses.
- Illustrate the formation of images using the technique of ray tracking.
- Determine power of a lens given the focal length.

Lenses are found in a huge array of optical instruments, ranging from a simple magnifying glass to the eye to a camera's zoom lens. In this section, we will use the law of refraction to explore the properties of lenses and how they form images.

The word *lens* derives from the Latin word for a lentil bean, the shape of which is similar to the convex lens in Figure 4.3.1. The convex lens shown has been shaped so that all light rays that enter it parallel to its axis cross one another at a single point on the opposite side of the lens. (The axis is defined to be a line normal to the lens at its center, as shown in Figure 4.3.1.) Such a lens is called a **converging (or convex) lens** for the converging effect it has on light rays. An expanded view of the path of one ray through the lens is shown, to illustrate how the ray changes direction both as it enters and as it leaves the lens. Since the index of refraction of the lens is greater than that of air, the ray moves towards the perpendicular as it enters and away from the perpendicular as it leaves. (This is in accordance with the law of refraction.) Due to the lens's shape, light is thus bent toward the axis at both surfaces. The point at which the rays cross is defined to be the **focal point** F of the lens. The distance from the center of the lens to its focal point is defined to be the **focal length** f of the lens. Figure 4.3.2 shows how a converging lens, such as that in a magnifying glass, can converge the nearly parallel light rays from the sun to a small spot.

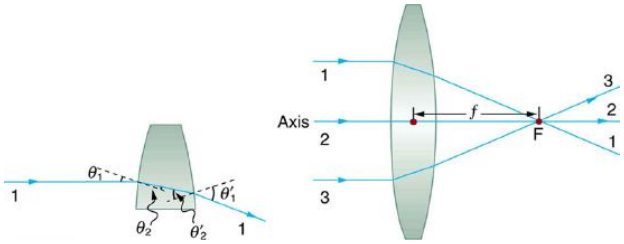


Figure 4.3.1: Rays of light entering a converging lens parallel to its axis converge at its focal point F. (Ray 2 lies on the axis of the lens.) The distance from the center of the lens to the focal point is the lens's focal length f. An expanded view of the path taken by ray 1 shows the perpendiculars and the angles of incidence and refraction at both surfaces.

Definition: CONVERGING OR CONVEX LENS

The lens in which light rays that enter it parallel to its axis cross one another at a single point on the opposite side with a converging effect is called converging lens.

Definition: FOCAL POINT F

The point at which the light rays cross is called the focal point F of the lens.

Definition: FOCAL LENGTH f

The distance from the center of the lens to its focal point is called focal length f.



Figure 4.3.2: Sunlight focused by a converging magnifying glass can burn paper. Light rays from the sun are nearly parallel and cross at the focal point of the lens. The more powerful the lens, the closer to the lens the rays will cross.

The greater effect a lens has on light rays, the more powerful it is said to be. For example, a powerful converging lens will focus parallel light rays closer to itself and will have a smaller focal length than a weak lens. The light will also focus into a smaller and more intense spot for a more powerful lens. The **power** P of a lens is defined to be the inverse of its focal length. In equation form, this is

$$P = \frac{1}{f}.$$

Definition: POWER P

The **power** P of a lens is defined to be the inverse of its focal length. In equation form, this is

$$P = \frac{1}{f}.$$

where f is the focal length of the lens, which must be given in meters (and not cm or mm). The power of a lens P has the unit diopters (D), provided that the focal length is given in meters. That is, $1\ \text{D} = 1/\text{m}$, or $1\ \text{m}^{-1}$. (Note that this power (optical power, actually) is not the same as power in watts. It is a concept related to the effect of optical devices on light.) Optometrists prescribe common spectacles and contact lenses in units of diopters.

Example 4.3.1: What is the Power of a Common Magnifying Glass?

Suppose you take a magnifying glass out on a sunny day and you find that it concentrates sunlight to a small spot 8.00 cm away from the lens. What are the focal length and power of the lens?

Strategy

The situation here is the same as those shown in Figure 4.3.1 and Figure 4.3.2. The Sun is so far away that the Sun's rays are nearly parallel when they reach Earth. The magnifying glass is a convex (or converging) lens, focusing the nearly parallel rays of sunlight. Thus the focal length of the lens is the distance from the lens to the spot, and its power is the inverse of this distance (in m).

Solution

The focal length of the lens is the distance from the center of the lens to the spot, given to be 8.00 cm. Thus,

$$f = 8.00\ \text{cm}.$$

To find the power of the lens, we must first convert the focal length to meters; then, we substitute this value into the equation for power. This gives

$$P = \frac{1}{f} = \frac{1}{0.0800\ \text{m}} = 12.5\ \text{D}.$$

Discussion

This is a relatively powerful lens. The power of a lens in diopters should not be confused with the familiar concept of power in watts. It is an unfortunate fact that the word "power" is used for two completely different concepts. If you examine a prescription for eyeglasses, you will note lens powers given in diopters. If you examine the label on a motor, you will note energy consumption rate given as a power in watts.

Figure 4.3.3 shows a concave lens and the effect it has on rays of light that enter it parallel to its axis (the path taken by ray 2 in the figure is the axis of the lens). The concave lens is a **diverging lens**, because it causes the light rays to bend away (diverge) from its axis. In this case, the lens has been shaped so that all light rays entering it parallel to its axis appear to originate from the same point, F, defined to be the focal point of a diverging lens. The distance from the center of the lens to the focal point is again called the focal length f of the lens. Note that the focal length and power of a diverging lens are defined to be negative. For example, if the distance to F in Figure 4.3.3 is 5.00 cm, then the focal length is $f = -5.00\ \text{cm}$ and the power of the lens is $P = -20\ \text{D}$. An expanded view of the path of one ray through the lens is shown in the figure to illustrate how the shape of the lens, together with the law of refraction, causes the ray to follow its particular path and be diverged.

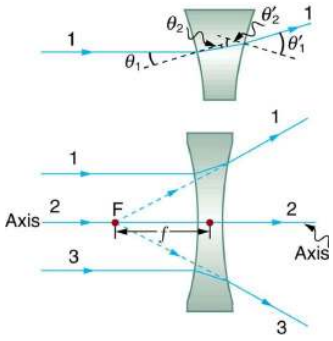


Figure 4.3.3: Rays of light entering a diverging lens parallel to its axis are diverged, and all appear to originate at its focal point F. The dashed lines are not rays—they indicate the directions from which the rays appear to come. The focal length f of a diverging lens is negative. An expanded view of the path taken by ray 1 shows the perpendiculars and the angles of incidence and refraction at both surfaces.

Definition: DIVERGING LENS

A lens that causes the light rays to bend away from its axis is called a diverging lens.

As noted in the initial discussion of the law of refraction, the paths of light rays are exactly reversible. This means that the direction of the arrows could be reversed for all of the rays in Figure 4.3.1 and Figure 4.3.3. For example, if a point light source is placed at the focal point of a convex lens, as shown in Figure 4.3.4, parallel light rays emerge from the other side.

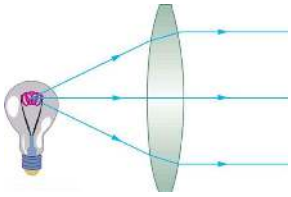


Figure 4.3.4: A small light source, like a light bulb filament, placed at the focal point of a convex lens, results in parallel rays of light emerging from the other side. The paths are exactly the reverse of those shown in Figure 4.3.1. This technique is used in lighthouses and sometimes in traffic lights to produce a directional beam of light from a source that emits light in all directions.

Ray Tracing and Thin Lenses

Ray tracing is the technique of determining or following (tracing) the paths that light rays take. For rays passing through matter, the law of refraction is used to trace the paths. Here we use ray tracing to help us understand the action of lenses in situations ranging from forming images on film to magnifying small print to correcting nearsightedness. While ray tracing for complicated lenses, such as those found in sophisticated cameras, may require computer techniques, there is a set of simple rules for tracing rays through thin lenses. A **thin lens** is defined to be one whose thickness allows rays to refract, as illustrated in Figure 4.3.1, but does not allow properties such as dispersion and aberrations. An ideal thin lens has two refracting surfaces but the lens is thin enough to assume that light rays bend only once. A thin symmetrical lens has two focal points, one on either side and both at the same distance from the lens. (See Figure 4.3.5.) Another important characteristic of a thin lens is that light rays through its center are deflected by a negligible amount, as seen in Figure 4.3.6.

Definition: THIN LENS

A thin lens is defined to be one whose thickness allows rays to refract but does not allow properties such as dispersion and aberrations.

TAKE-HOME EXPERIMENT: A VISIT TO THE OPTICIAN

Look through your eyeglasses (or those of a friend) backward and forward and comment on whether they act like thin lenses.

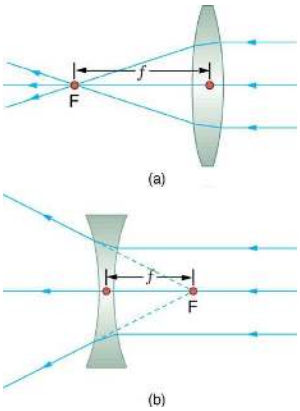


Figure 4.3.5: Thin lenses have the same focal length on either side. (a) Parallel light rays entering a converging lens from the right cross at its focal point on the left. (b) Parallel light rays entering a diverging lens from the right seem to come from the focal point on the right.

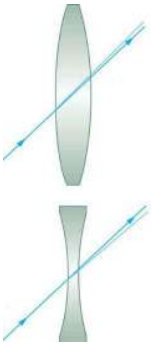


Figure 4.3.6: The light ray through the center of a thin lens is deflected by a negligible amount and is assumed to emerge parallel to its original path (shown as a shaded line).

Using paper, pencil, and a straight edge, ray tracing can accurately describe the operation of a lens. The rules for ray tracing for thin lenses are based on the illustrations already discussed:

- **Principal Ray 1**: a ray entering a converging lens parallel to its axis passes through the focal point F of the lens on the other side (see rays 1 and 3 in Figure 4.3.1); a ray entering a diverging lens parallel to its axis seems to come from the focal point F (see rays 1 and 3 in Figure 4.3.3)
- **Principal Ray 2**: a ray passing through the center of either a converging or a diverging lens does not change direction (see Figure 4.3.6, and see ray 2 in Figure 4.3.1 and Figure 4.3.3)
- **Principal Ray 3**: a ray entering a converging lens through its focal point exits parallel to its axis (the reverse of rays 1 and 3 in Figure 4.3.1); a ray that enters a diverging lens by heading toward the focal point on the opposite side exits parallel to the axis (the reverse of rays 1 and 3 in Figure 4.3.3). *The third principal ray is optional and may be used to verify the accuracy of image location.*

Image Formation by Thin Lenses

In some circumstances, a lens forms an obvious image, such as when a movie projector casts an image onto a screen. In other cases, the image is less obvious. Where, for example, is the image formed by eyeglasses? We will use ray tracing for thin lenses to illustrate how they form images, and we will develop equations to describe the image formation quantitatively.

Consider an object some distance away from a converging lens, as shown in Figure 4.3.7. To find the location and size of the image formed, we trace the paths of principal rays originating from one point on the object, in this case the top of the person's head. The figure shows the three principal rays from the top of the object as described above. (Note that there are many light rays leaving this point going in many directions, but we concentrate on the three principal rays which can be traced by simple rules.) The first ray is one that enters the lens parallel to its axis and passes through the focal point on the other side (Principal Ray 1). The second ray passes through the center of the lens without changing direction (Principal Ray 2). The third ray passes through the nearer focal point on its way into the lens and leaves the lens parallel to its axis (Principal Ray 3). If correctly drawn, all three rays cross at the same point on the other side of the lens. The image of the top of the person's head is located at this point. *All* rays, including those that are not principal rays, that come from the same point on the top of the person's head are refracted in such a way as to cross at the point shown; the principal rays are what we use to *find* this point. Rays from another point on the object, such as her belt buckle, will also cross at another common point, forming a complete image, as shown. Although three rays are traced in Figure 4.3.7, only two are necessary to locate the image. Before applying ray tracing to other situations, let us consider the example shown in Figure 4.3.7 in more detail.

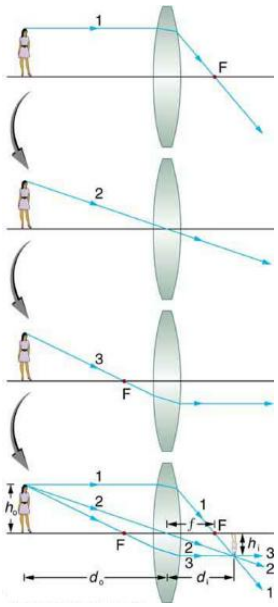


Figure 4.3.7: Ray tracing is used to locate the image formed by a lens. Rays originating from the same point on the object are traced—the three chosen rays each follow one of the rules for ray tracing, so that their paths are easy to determine. The image is located at the point where the rays cross. In this case, a real image—one that can be projected on a screen—is formed.

The image formed in Figure 4.3.7 is a **real image**, meaning that it can be projected. That is, light rays from one point on the object actually cross at the location of the image and can be projected onto a screen, a piece of film, or the retina of an eye, for example. Figure 4.3.8 shows how such an image would be projected onto film by a camera lens. This figure also shows how a real image is projected onto the retina by the lens of an eye. Note that the image is there whether it is projected onto a screen or not.

Definition: REAL IMAGE

The image in which light rays from one point on the object actually cross at the location of the image and can be projected onto a screen, a piece of film, or the retina of an eye is called a real image.

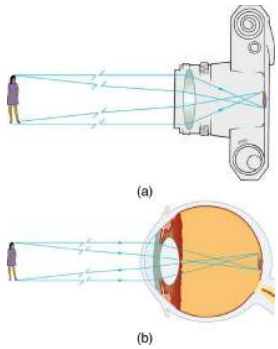


Figure 4.3.8: Real images can be projected. (a) A real image of the person is projected onto film. (b) The converging nature of the multiple surfaces that make up the eye result in the projection of a real image on the retina.

Several important distances appear in Figure 4.3.7. We define d_o to be the object distance, the distance of an object from the center of a lens. **Image distance** d_i is defined to be the distance of the image from the center of a lens. The height of the object and height of the image are given the symbols h_o and h_i, respectively. Images that appear upright relative to the object have heights that are positive and those that are inverted have negative heights. Using the rules of ray tracing and making a scale drawing with paper and pencil, like that in Figure 4.3.8, we can accurately describe the location and size of an image. But the real benefit of ray tracing is in visualizing how images are formed in a variety of situations. To obtain numerical information, we use a pair of equations that can be derived from a geometric analysis of ray tracing for thin lenses. The **thin lens equations** are

$$\frac{1}{d_o} + \frac{1}{d_i} = \frac{1}{f}$$

and

$$\frac{h_i}{h_o} = -\frac{d_i}{d_o} = m.$$

We define the ratio of image height to object height (h_i/h_o) to be the **magnification** m. (The minus sign in the equation above will be discussed shortly.) The thin lens equations are broadly applicable to all situations involving thin lenses (and "thin" mirrors, as we will see later). We will explore many features of image formation in the following worked examples.

Definition: IMAGE DISTANCE

The distance of the image from the center of the lens is called image distance.

THIN LENS EQUATIONS AND MAGNIFICATION

$$\frac{1}{d_o} + \frac{1}{d_i} = \frac{1}{f}$$

$$\frac{h_i}{h_o} = -\frac{d_i}{d_o} = m$$

Example 4.3.2: Finding the Image of a Light Bulb Filament by Ray Tracing and by the Thin Lens Equations

A clear glass light bulb is placed 0.750 m from a convex lens having a 0.500 m focal length, as shown in Figure 4.3.9. Use ray tracing to get an approximate location for the image. Then use the thin lens equations to calculate (a) the location of the image and (b) its magnification. Verify that ray tracing and the thin lens equations produce consistent results.

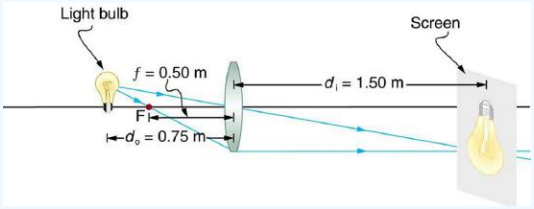


Figure 4.3.9: A light bulb placed 0.750 m from a lens having a 0.500 m focal length produces a real image on a poster board as discussed in the example above. Ray tracing predicts the image location and size.

Strategy and Concept

Since the object is placed farther away from a converging lens than the focal length of the lens, this situation is analogous to those illustrated in Figure 4.3.7 and Figure 4.3.8. Ray tracing to scale should produce similar results for di. Numerical solutions for d_i and m can be obtained using the thin lens equations, noting that $d_\mathrm{o} = 0.750\ \mathrm{m}$ and $f = 0.500\ \mathrm{m}$.

Solutions (Ray tracing)

The ray tracing to scale in Figure 4.3.9 shows two rays from a point on the bulb's filament crossing about 1.50 m on the far side of the lens. Thus the image distance d_i is about 1.50 m. Similarly, the image height based on ray tracing is greater than the object height by about a factor of 2, and the image is inverted. Thus m is about –2. The minus sign indicates that the image is inverted.

The thin lens equations can be used to find d_i from the given information:

$$\frac{1}{d_\mathrm{o}} + \frac{1}{d_\mathrm{i}} = \frac{1}{f}.$$

Rearranging to isolate d_i gives

$$\frac{1}{d_\mathrm{i}} = \frac{1}{f} - \frac{1}{d_\mathrm{o}}.$$

Entering known quantities gives a value for $1/d_\mathrm{i}$:

$$\frac{1}{d_\mathrm{i}} = \frac{1}{0.500\ \mathrm{m}} - \frac{1}{0.750\ \mathrm{m}} = \frac{0.667}{\mathrm{m}}.$$

This must be inverted to find d_i:

$$d_\mathrm{i} = \frac{\mathrm{m}}{0.667} = 1.50\ \mathrm{m}.$$

Note that another way to find d_i is to rearrange the equation:

$$\frac{1}{d_\mathrm{i}} = \frac{1}{f} - \frac{1}{d_\mathrm{o}}.$$

This yields the equation for the image distance as:

$$d_\mathrm{i} = \frac{f d_\mathrm{o}}{d_\mathrm{o} - f}.$$

Note that there is no inverting here.

The thin lens equations can be used to find the magnification m, since both d_i and d_o are known. Entering their values gives

$$m = -\frac{d_\mathrm{i}}{d_\mathrm{o}} = -\frac{1.50\ \mathrm{m}}{0.750\ \mathrm{m}} = -2.00.$$

Discussion

Note that the minus sign causes the magnification to be negative when the image is inverted. Ray tracing and the use of the thin lens equations produce consistent results. The thin lens equations give the most precise results, being limited only by the accuracy of the given information. Ray tracing is limited by the accuracy with which you can draw, but it is highly useful both conceptually and visually.

Real images, such as the one considered in the previous example, are formed by converging lenses whenever an object is farther from the lens than its focal length. This is true for movie projectors, cameras, and the eye. We shall refer to these as *case 1* images. A case 1 image is formed when $d_0 > f$ and f is positive, as in Figure 4.3.10(a). (A summary of the three cases or types of image formation appears at the end of this section.)

A different type of image is formed when an object, such as a person's face, is held close to a convex lens. The image is upright and larger than the object, as seen in Figure 4.3.10(b), and so the lens is called a magnifier. If you slowly pull the magnifier away from the face, you will see that the magnification steadily increases until the image begins to blur. Pulling the magnifier even farther away produces an inverted image as seen in Figure 4.3.10(a). The distance at which the image blurs, and beyond which it inverts, is the focal length of the lens. To use a convex lens as a magnifier, the object must be closer to the converging lens than its focal length. This is called a *case 2* image. A case 2 image is formed when $d_o < f$ and f is positive.

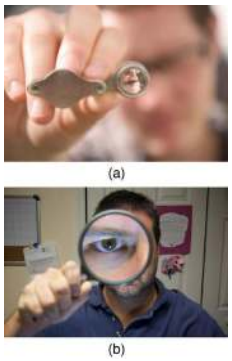


Figure 4.3.10: (a) When a converging lens is held farther away from the face than the lens's focal length, an inverted image is formed. This is a case 1 image. Note that the image is in focus but the face is not, because the image is much closer to the camera taking this photograph than the face. (credit: DaMongMan, Flickr) (b) A magnified image of a face is produced by placing it closer to the converging lens than its focal length. This is a case 2 image. (credit: Casey Fleser, Flickr)

Figure 4.3.11 uses ray tracing to show how an image is formed when an object is held closer to a converging lens than its focal length. Rays coming from a common point on the object continue to diverge after passing through the lens, but all appear to originate from a point at the location of the image. The image is on the same side of the lens as the object and is farther away from the lens than the object. This image, like all case 2 images, cannot be projected and, hence, is called a **virtual image**. Light rays only appear to originate at a virtual image; they do not actually pass through that location in space. A screen placed at the location of a virtual image will receive only diffuse light from the object, not focused rays from the lens. Additionally, a screen placed on the opposite side of the lens will receive rays that are still diverging, and so no image will be projected on it. We can see the magnified image with our eyes, because the lens of the eye converges the rays into a real image projected on our retina. Finally, we note that a virtual image is upright and larger than the object, meaning that the magnification is positive and greater than 1.

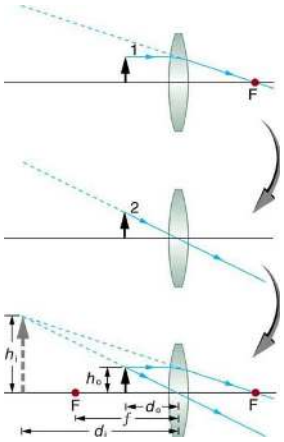


Figure 4.3.11: Ray tracing predicts the image location and size for an object held closer to a converging lens than its focal length. Ray 1 enters parallel to the axis and exits through the focal point on the opposite side, while ray 2 passes through the center of the lens without changing path. The two rays continue to diverge on the other side of the lens, but both appear to come from a common point, locating the upright, magnified, virtual image. This is a case 2 image.

Definition: VIRTUAL IMAGE

An image that is on the same side of the lens as the object and cannot be projected on a screen is called a virtual image.

Example 4.3.3: Image Produced by a Magnifying Glass

Suppose the book page in Figure 4.3.11(a) is held 7.50 cm from a convex lens of focal length 10.0 cm, such as a typical magnifying glass might have. What magnification is produced?

Strategy and Concept

We are given that $d_o = 7.50$ cm and $f = 10.0$ cm, so we have a situation where the object is placed closer to the lens than its focal length. We therefore expect to get a case 2 virtual image with a positive magnification that is greater than 1. Ray tracing produces an image like that shown in Figure 4.3.11, but we will use the thin lens equations to get numerical solutions in this example.

Solution

To find the magnification m, we try to use magnification equation, $m = -d_i/d_o$. We do not have a value for d_i, so that we must first find the location of the image using lens equation. (The procedure is the same as followed in the preceding example, where d_o and f were known.) Rearranging the magnification equation to isolate d_i gives

$$\frac{1}{d_i} = \frac{1}{f} - \frac{1}{d_o}.$$

Entering known values, we obtain a value for $1/d_i$:

$$\frac{1}{d_i} = \frac{1}{10.0 \text{ cm}} - \frac{1}{7.50 \text{ cm}} = \frac{-0.0333}{\text{cm}}.$$

This must be inverted to find d_i:

$$d_i = -\frac{\text{cm}}{0.0333} = -30.0 \text{ cm}.$$

Now the thin lens equation can be used to find the magnification m, since both d_i and d_o are known. Entering their values gives

$$m = -\frac{d_i}{d_o} = -\frac{-30.0 \text{ cm}}{7.5 \text{ cm}} = 4.00.$$

Discussion

A number of results in this example are true of all case 2 images, as well as being consistent with Figure 4.3.11. Magnification is indeed positive (as predicted), meaning the image is upright. The magnification is also greater than 1, meaning that the image is larger than the object—in this case, by a factor of 3. Note that the image distance is negative. This means the image is

on the same side of the lens as the object. Thus the image cannot be projected and is virtual. (Negative values of d_i occur for virtual images.) The image is farther from the lens than the object, since the image distance is greater in magnitude than the object distance. The location of the image is not obvious when you look through a magnifier. In fact, since the image is bigger than the object, you may think the image is closer than the object. But the image is farther away, a fact that is useful in correcting farsightedness, as we shall see in a later section.

A third type of image is formed by a diverging or concave lens. Try looking through eyeglasses meant to correct nearsightedness. (See Figure 4.3.12.) You will see an image that is upright but smaller than the object. This means that the magnification is positive but less than 1. The ray diagram in Figure 4.3.13 shows that the image is on the same side of the lens as the object and, hence, cannot be projected—it is a virtual image. Note that the image is closer to the lens than the object. This is a *case 3* image, formed for any object by a negative focal length or diverging lens.



Figure 4.3.12: A car viewed through a concave or diverging lens looks upright. This is a case 3 image. (credit: Daniel Oines, Flickr)

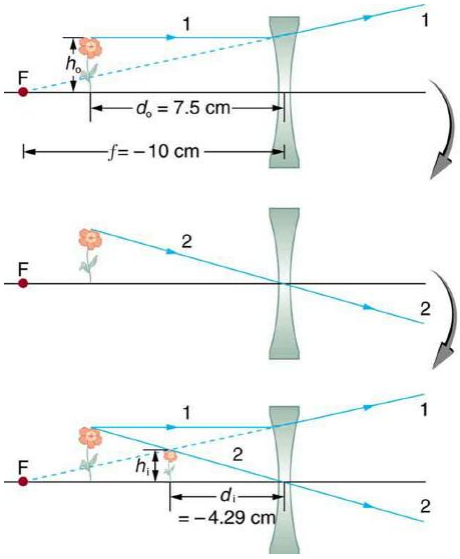


Figure 4.3.13: Ray tracing predicts the image location and size for a concave or diverging lens. Ray 1 enters parallel to the axis and is bent so that it appears to originate from the focal point. Ray 2 passes through the center of the lens without changing path. The two rays appear to come from a common point, locating the upright image. This is a case 3 image, which is closer to the lens than the object and smaller in height.

Example 4.3.4: Image Produced by a Concave Lens

Suppose an object such as a book page is held 7.50 cm from a concave lens of focal length –10.0 cm. Such a lens could be used in eyeglasses to correct pronounced nearsightedness. What magnification is produced?

Strategy and Concept

This example is identical to the preceding one, except that the focal length is negative for a concave or diverging lens. The method of solution is thus the same, but the results are different in important ways.

Solution

To find the magnification m, we must first find the image distance d_i using thin lens equation

$$\frac{1}{d_i} = \frac{1}{f} - \frac{1}{d_o},$$

or its alternative rearrangement

$$\frac{fd_o}{}$$

$$d_i = \frac{}{d_o - f}.$$

We are given that $f = -10.0$ cm and $d_o = 7.50$ cm. Entering these yields a value for $1/d_i$:

$$\frac{1}{d_i} = \frac{1}{-10.0 \text{ cm}} - \frac{1}{7.50 \text{ cm}} = \frac{-0.2333}{\text{cm}}.$$

This must be inverted to find d_i:

$$d_i = -\frac{\text{cm}}{0.2333} = -4.29 \text{ cm}.$$

Or

$$d_i = \frac{(7.5)(-10)}{(7.5 - (-10))} = -75/17.5 = -4.29 \text{ cm}.$$

Now the magnification equation can be used to find the magnification m, since both d_i and d_o are known. Entering their values gives

$$m = -\frac{d_i}{d_o} = -\frac{-4.29 \text{ cm}}{7.50 \text{ cm}} = 0.571.$$

Discussion

A number of results in this example are true of all case 3 images, as well as being consistent with Figure 4.3.13. Magnification is positive (as predicted), meaning the image is upright. The magnification is also less than 1, meaning the image is smaller than the object—in this case, a little over half its size. The image distance is negative, meaning the image is on the same side of the lens as the object. (The image is virtual.) The image is closer to the lens than the object, since the image distance is smaller in magnitude than the object distance. The location of the image is not obvious when you look through a concave lens. In fact, since the image is smaller than the object, you may think it is farther away. But the image is closer than the object, a fact that is useful in correcting nearsightedness, as we shall see in a later section.

Table 4.3.1 summarizes the three types of images formed by single thin lenses. These are referred to as case 1, 2, and 3 images. Convex (converging) lenses can form either real or virtual images (cases 1 and 2, respectively), whereas concave (diverging) lenses can form only virtual images (always case 3). Real images are always inverted, but they can be either larger or smaller than the object. For example, a slide projector forms an image larger than the slide, whereas a camera makes an image smaller than the object being photographed. Virtual images are always upright and cannot be projected. Virtual images are larger than the object only in case 2, where a convex lens is used. The virtual image produced by a concave lens is always smaller than the object—a case 3 image. We can see and photograph virtual images only by using an additional lens to form a real image.

Table 4.3.1: Three Types of Images Formed By Thin Lenses

Type	Formed when	Image type	d_i	m
Case 1	f positive, $d_o > f$	real	positive	negative
Case 2	f positive, $d_o < f$	virtual	negative	positive, $m > 1$
Case 3	f negative	virtual	negative	positive, $m < 1$

TAKE-HOME EXPERIMENT: CONCENTRATING SUNLIGHT

Find several lenses and determine whether they are converging or diverging. In general those that are thicker near the edges are diverging and those that are thicker near the center are converging. On a bright sunny day take the converging lenses outside and try focusing the sunlight onto a piece of paper. Determine the focal lengths of the lenses. Be careful because the paper may start to burn, depending on the type of lens you have selected.

Section Summary

- Light rays entering a converging lens parallel to its axis cross one another at a single point on the opposite side.
- For a converging lens, the focal point is the point at which converging light rays cross; for a diverging lens, the focal point is the point from which diverging light rays appear to originate.

- The distance from the center of the lens to its focal point is called the focal length f.
- Power P of a lens is defined to be the inverse of its focal length, $P = \frac{1}{f}$.
- A lens that causes the light rays to bend away from its axis is called a diverging lens.
- Ray tracing is the technique of graphically determining the paths that light rays take.
- The image in which light rays from one point on the object actually cross at the location of the image and can be projected onto a screen, a piece of film, or the retina of an eye is called a real image.
- Thin lens equations are $\frac{1}{d_o} + \frac{1}{d_i} = \frac{1}{f}$ and $\frac{h_i}{h_o} = -\frac{d_i}{d_o} = m$ (magnification).
- The distance of the image from the center of the lens is called image distance.
- An image that is on the same side of the lens as the object and cannot be projected on a screen is called a virtual image.

Glossary

converging lens
a convex lens in which light rays that enter it parallel to its axis converge at a single point on the opposite side

diverging lens
a concave lens in which light rays that enter it parallel to its axis bend away (diverge) from its axis

focal point
for a converging lens or mirror, the point at which converging light rays cross; for a diverging lens or mirror, the point from which diverging light rays appear to originate

focal length
distance from the center of a lens or curved mirror to its focal point

magnification
ratio of image height to object height

power
inverse of focal length

real image
image that can be projected

virtual image
image that cannot be projected

This page titled 4.3: Image Formation by Lenses is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- **11.9: Image Formation by Lenses** by OpenStax is licensed CC BY 4.0.

CHAPTER OVERVIEW

5: Energy



Under Construction(Public Domain ; Public Domain Pictures)

5.1: Prelude to Thermodynamics

5.1.1: The First Law of Thermodynamics

5.1.2: The First Law of Thermodynamics and Some Simple Processes

5.1.3: Gravitational Potential Energy

5.2: Introduction to the Second Law of Thermodynamics - Heat Engines and their Efficiency

5.2.1: Carnot’s Perfect Heat Engine- The Second Law of Thermodynamics Restated

5.2.2: Applications of Thermodynamics- Heat Pumps and Refrigerators

5.2.3: Entropy and the Second Law of Thermodynamics- Disorder and the Unavailability of Energy

5.3: Simple Machines

5: Energy is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

5.1: Prelude to Thermodynamics

Heat transfer is energy in transit, and it can be used to do work. It can also be converted to any other form of energy. A car engine, for example, burns fuel for heat transfer into a gas. Work is done by the gas as it exerts a force through a distance, converting its energy into a variety of other forms—into the car's kinetic or gravitational potential energy; into electrical energy to run the spark plugs, radio, and lights; and back into stored energy in the car's battery. But most of the heat transfer produced from burning fuel in the engine does not do work on the gas. Rather, the energy is released into the environment, implying that the engine is quite inefficient.



Figure 5.1.1. A steam engine uses heat transfer to do work. Tourists regularly ride this narrow-gauge steam engine train near the San Juan Skyway in Durango, Colorado, part of the National Scenic Byways Program. (credit: Dennis Adams)

It is often said that modern gasoline engines cannot be made to be significantly more efficient. We hear the same about heat transfer to electrical energy in large power stations, whether they are coal, oil, natural gas, or nuclear powered. Why is that the case? Is the inefficiency caused by design problems that could be solved with better engineering and superior materials? Is it part of some money-making conspiracy by those who sell energy? Actually, the truth is more interesting, and reveals much about the nature of heat transfer.

Basic physical laws govern how heat transfer for doing work takes place and place insurmountable limits onto its efficiency. This chapter will explore these laws as well as many applications and concepts associated with them. These topics are part of *thermodynamics*—the study of heat transfer and its relationship to doing work.

This page titled 5.1: Prelude to Thermodynamics is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **15.0: Prelude to Thermodynamics** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/college-physics.

5.1.1: The First Law of Thermodynamics

Learning Objectives

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

- Define the first law of thermodynamics.
- Describe how conservation of energy relates to the first law of thermodynamics.
- Identify instances of the first law of thermodynamics working in everyday situations, including biological metabolism.
- Calculate changes in the internal energy of a system, after accounting for heat transfer and work done.

If we are interested in how heat transfer is converted into doing work, then the conservation of energy principle is important. The **first law of thermodynamics** applies the conservation of energy principle to systems where heat transfer and doing work are the methods of transferring energy into and out of the system.



Figure 5.1.1.1: This boiling tea kettle represents energy in motion. The water in the kettle is turning to water vapor because heat is being transferred from the stove to the kettle. As the entire system gets hotter, work is done—from the evaporation of the water to the whistling of the kettle. (credit: Gina Hamilton)

The first law of thermodynamics states that the change in internal energy of a system equals the net heat transfer *into* the system minus the net work done *by* the system. In equation form, the first law of thermodynamics is

$$\Delta U = Q - W. \tag{5.1.1.1}$$

Here ΔU is the *change in internal energy* U of the system. Q is the *net heat transferred into the system*—that is, Q is the sum of all heat transfer into and out of the system. W is the *net work done by the system*—that is, W is the sum of all work done on or by the system. We use the following sign conventions: if Q is positive, then there is a net heat transfer into the system; if W is positive, then there is net work done by the system. So positive Q adds energy to the system and positive W takes energy from the system. Thus $\Delta U = Q - W$. Note also that if more heat transfer into the system occurs than work done, the difference is stored as internal energy. Heat engines are a good example of this—heat transfer into them takes place so that they can do work (Figure 5.1.1.2). We will now examine Q, W and ΔU further.

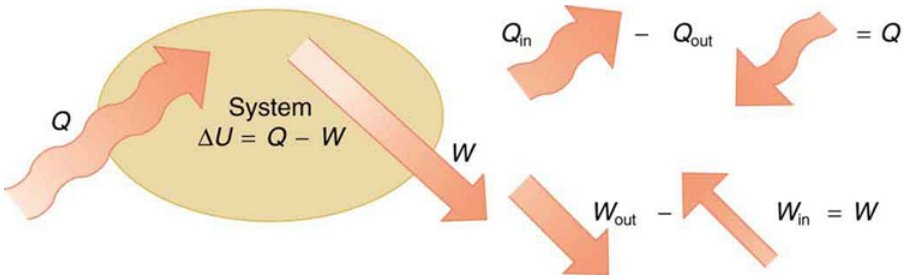


Figure 5.1.1.2: The first law of thermodynamics is the conservation-of-energy principle stated for a system where heat and work are the methods of transferring energy for a system in thermal equilibrium. Q represents the net heat transfer—it is the sum of all heat transfers into and out of the system. Q is positive for net heat transfer into the system. W is the total work done on and by the system. W is positive when more work is done by the system than on it. The change in the internal energy of the system, ΔU, is related to heat and work by the first law of thermodynamics (Equation 5.1.1.1).

LAW OF THERMODYNAMICS AND LAW OF CONSERVATION OF ENERGY

The first law of thermodynamics is actually the law of conservation of energy stated in a form most useful in thermodynamics. The first law gives the relationship between heat transfer, work done, and the change in internal energy of a system.

Heat Q and Work W

Heat transfer Q and doing work W are the two everyday means of bringing energy into or taking energy out of a system. The processes are quite different. Heat transfer, a less organized process, is driven by temperature differences. Work, a quite organized process, involves a macroscopic force exerted through a distance. Nevertheless, heat and work can produce identical results.For example, both can cause a temperature increase. Heat transfer into a system, such as when the Sun warms the air in a bicycle tire, can increase its temperature, and so can work done on the system, as when the bicyclist pumps air into the tire. Once the temperature increase has occurred, it is impossible to tell whether it was caused by heat transfer or by doing work. This uncertainty is an important point. Heat transfer and work are both energy in transit—neither is stored as such in a system. However, both can change the internal energy U of a system. Internal energy is a form of energy completely different from either heat or work.

Internal Energy U

We can think about the internal energy of a system in two different but consistent ways. The first is the atomic and molecular view, which examines the system on the atomic and molecular scale. The **internal energy** U of a system is the sum of the kinetic and potential energies of its atoms and molecules. Recall that kinetic plus potential energy is called mechanical energy. Thus internal energy is the sum of atomic and molecular mechanical energy. Because it is impossible to keep track of all individual atoms and molecules, we must deal with averages and distributions. A second way to view the internal energy of a system is in terms of its macroscopic characteristics, which are very similar to atomic and molecular average values.

Macroscopically, we define the change in internal energy ΔU to be that given by the first law of thermodynamics (Equation 5.1.1.1):

$$\Delta U = Q - W$$

Many detailed experiments have verified that $\Delta U = Q - W$, where ΔU is the change in total kinetic and potential energy of all atoms and molecules in a system. It has also been determined experimentally that the internal energy U of a system depends only on the state of the system and *not how it reached that state*. More specifically, U is found to be a function of a few macroscopic quantities (pressure, volume, and temperature, for example), independent of past history such as whether there has been heat transfer or work done. This independence means that if we know the state of a system, we can calculate changes in its internal energy U from a few macroscopic variables.

MACROSCOPIC vs. MICROSCOPIC

In thermodynamics, we often use the macroscopic picture when making calculations of how a system behaves, while the atomic and molecular picture gives underlying explanations in terms of averages and distributions. We shall see this again in later sections of this chapter. For example, in the topic of entropy, calculations will be made using the atomic and molecular view.

To get a better idea of how to think about the internal energy of a system, let us examine a system going from State 1 to State 2. The system has internal energy U_1 in State 1, and it has internal energy U_2 in State 2, no matter how it got to either state. So the change in internal energy

$$\Delta U = U_2 - U_1 \tag{5.1.1.2}$$

is independent of what caused the change. In other words, δU *is independent of path*. By path, we mean the method of getting from the starting point to the ending point. Why is this independence important? Both Q and W *depend on path*, but ΔU does not (Equation 5.1.1.1). This path independence means that internal energy U is easier to consider than either heat transfer or work done.

Example 5.1.1.1: Calculating Change in Internal Energy - The Same Change in U is Produced by Two Different Processes

a. Suppose there is heat transfer of 40.00 J to a system, while the system does 10.00 J of work. Later, there is heat transfer of 25.00 J out of the system while 4.00 J of work is done on the system. What is the net change in internal energy of the system?

b. What is the change in internal energy of a system when a total of 150.00 J of heat transfer occurs out of (from) the system and 159.00 J of work is done on the system (Figure 5.1.1.3)?

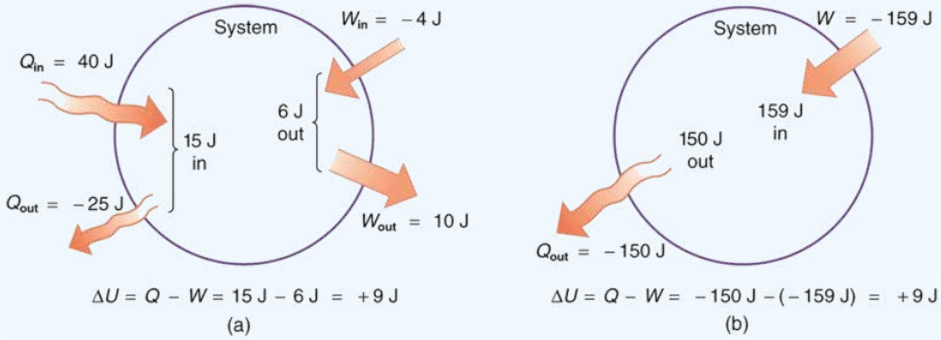


Figure 5.1.1.3: Two different processes produce the same change in a system. (a) A total of 15.00 J of heat transfer occurs into the system, while work takes out a total of 6.00 J. The change in internal energy is $\delta U = Q - W = 9.00\ J$. (b) Heat transfer removes 150.00 J from the system while work puts 159.00 J into it, producing an increase of 9.00 J in internal energy. If the system starts out in the same state in (a) and (b), it will end up in the same final state in either case—its final state is related to internal energy, not how that energy was acquired.

Strategy

In part (a), we must first find the net heat transfer and net work done from the given information. Then the first law of thermodynamics (Equation 5.1.1.1).

can be used to find the change in internal energy. In part (b), the net heat transfer and work done are given, so the equation can be used directly.

Solution for (a)

The net heat transfer is the heat transfer into the system minus the heat transfer out of the system, or

$$\begin{aligned} Q &= 40.00\ J - 25.00\ J \\ &= 15.00\ J \end{aligned}$$

Similarly, the total work is the work done by the system minus the work done on the system, or

$$\begin{aligned} W &= 10.00\ J - 4.00\ J \\ &= 6.00\ J. \end{aligned}$$

Discussion on (a)

No matter whether you look at the overall process or break it into steps, the change in internal energy is the same.

Solution for (b)

Here the net heat transfer and total work are given directly to be $Q = -150.00\ J$ and $W = -159.00\ J$, so that

$$\Delta U = Q - W = -150.00 - (-159.00)$$
$$= 9.00\ J.$$

Discussion on (b)

A very different process in part (b) produces the same 9.00-J change in internal energy as in part (a). Note that the change in the system in both parts is related to ΔU and not to the individual Qs or Ws involved. The system ends up in the *same* state in both (a) and (b). Parts (a) and (b) present two different paths for the system to follow between the same starting and ending points, and the change in internal energy for each is the same—it is independent of path.

Human Metabolism and the First Law of Thermodynamics

Human metabolism is the conversion of food into heat transfer, work, and stored fat. Metabolism is an interesting example of the first law of thermodynamics in action. We now take another look at these topics via the first law of thermodynamics. Considering the body as the system of interest, we can use the first law to examine heat transfer, doing work, and internal energy in activities ranging from sleep to heavy exercise. What are some of the major characteristics of heat transfer, doing work, and energy in the body? For one, body temperature is normally kept constant by heat transfer to the surroundings. This means Q is negative. Another fact is that the body usually does work on the outside world. This means W is positive. In such situations, then, the body loses internal energy, since $\Delta U = Q - W$ is negative.

Now consider the effects of eating. Eating increases the internal energy of the body by adding chemical potential energy (this is an unromantic view of a good steak). The body *metabolizes* all the food we consume. Basically, metabolism is an oxidation process in which the chemical potential energy of food is released. This implies that food input is in the form of work. Food energy is reported in a special unit, known as the Calorie. This energy is measured by burning food in a calorimeter, which is how the units are determined.

In chemistry and biochemistry, one calorie (spelled with a *lowercase* c) is defined as the energy (or heat transfer) required to raise the temperature of one gram of pure water by one degree Celsius. Nutritionists and weight-watchers tend to use the *dietary* calorie, which is frequently called a Calorie (spelled with a *capital* C). One food Calorie is the energy needed to raise the temperature of one *kilogram* of water by one degree Celsius. This means that one dietary Calorie is equal to one kilocalorie for the chemist, and one must be careful to avoid confusion between the two.

Again, consider the internal energy the body has lost. There are three places this internal energy can go—to heat transfer, to doing work, and to stored fat (a tiny fraction also goes to cell repair and growth). Heat transfer and doing work take internal energy out of the body, and food puts it back. If you eat just the right amount of food, then your average internal energy remains constant. Whatever you lose to heat transfer and doing work is replaced by food, so that, in the long run, $\Delta U = 0$. If you overeat repeatedly, then ΔU is always positive, and your body stores this extra internal energy as fat. The reverse is true if you eat too little. If ΔU is negative for a few days, then the body metabolizes its own fat to maintain body temperature and do work that takes energy from the body. This process is how dieting produces weight loss.

Life is not always this simple, as any dieter knows. The body stores fat or metabolizes it only if energy intake changes for a period of several days. Once you have been on a major diet, the next one is less successful because your body alters the way it responds to low energy intake. Your basal metabolic rate (BMR) is the rate at which food is converted into heat transfer and work done while the body is at complete rest. The body adjusts its basal metabolic rate to partially compensate for over-eating or under-eating. The body will decrease the metabolic rate rather than eliminate its own fat to replace lost food intake. You will chill more easily and feel less energetic as a result of the lower metabolic rate, and you will not lose weight as fast as before. Exercise helps to lose weight, because it produces both heat transfer from your body and work, and raises your metabolic rate even when you are at rest. Weight loss is also aided by the quite low efficiency of the body in converting internal energy to work, so that the loss of internal energy resulting from doing work is much greater than the work done.It should be noted, however, that living systems are not in thermal equilibrium.

The body provides us with an excellent indication that many thermodynamic processes are *irreversible*. An irreversible process can go in one direction but not the reverse, under a given set of conditions. For example, although body fat can be converted to do work and produce heat transfer, work done on the body and heat transfer into it cannot be converted to body fat. Otherwise, we could skip lunch by sunning ourselves or by walking down stairs. Another example of an irreversible thermodynamic process is photosynthesis. This process is the intake of one form of energy—light—by plants and its conversion to chemical potential energy. Both applications of the first law of thermodynamics are illustrated in Figure 5.1.1.4. One great advantage of conservation laws

such as the first law of thermodynamics is that they accurately describe the beginning and ending points of complex processes, such as metabolism and photosynthesis, without regard to the complications in between.

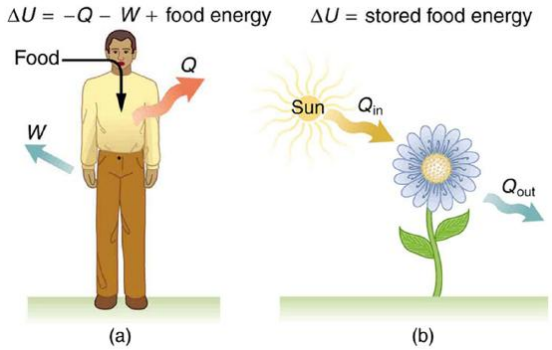


Figure 5.1.1.4: (a) The first law of thermodynamics applied to metabolism. Heat transferred out of the body Q and work done by the body W remove internal energy, while food intake replaces it. (Food intake may be considered as work done on the body.) (b) Plants convert part of the radiant heat transfer in sunlight to stored chemical energy, a process called photosynthesis.

Summary

The table presents a summary of terms relevant to the first law of thermodynamics.

Term	Definition
U	Internal energy—the sum of the kinetic and potential energies of a system's atoms and molecules. Can be divided into many subcategories, such as thermal and chemical energy. Depends only on the state of a system (such as its P, V and T, not on how the energy entered the system. Change in internal energy is path independent.
Q	Heat—energy transferred because of a temperature difference. Characterized by random molecular motion. Highly dependent on path. Q entering a system is positive.
W	Work—energy transferred by a force moving through a distance. An organized, orderly process. Path dependent. W done by a system (either against an external force or to increase the volume of the system) is positive.

- The first law of thermodynamics is given as $\Delta U = Q - W$, where ΔU is the change in internal energy of a system, Q is the net heat transfer (the sum of all heat transfer into and out of the system), and W is the net work done (the sum of all work done on or by the system).
- Both Q and W are energy in transit; only ΔU represents an independent quantity capable of being stored.
- The internal energy U of a system depends only on the state of the system and not how it reached that state.
- Metabolism of living organisms, and photosynthesis of plants, are specialized types of heat transfer, doing work, and internal energy of systems.

Glossary

first law of thermodynamics

states that the change in internal energy of a system equals the net heat transfer *into* the system minus the net work done *by* the system

internal energy

the sum of the kinetic and potential energies of a system's atoms and molecules

human metabolism

conversion of food into heat transfer, work, and stored fat

This page titled 5.1.1: The First Law of Thermodynamics is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **15.1: The First Law of Thermodynamics** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/college-physics.

5.1.2: The First Law of Thermodynamics and Some Simple Processes

Learning Objectives

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

- Define the first law of thermodynamics.
- Describe how conservation of energy relates to the first law of thermodynamics.
- Identify instances of the first law of thermodynamics working in everyday situations, including biological metabolism.
- Calculate changes in the internal energy of a system, after accounting for heat transfer and work done.

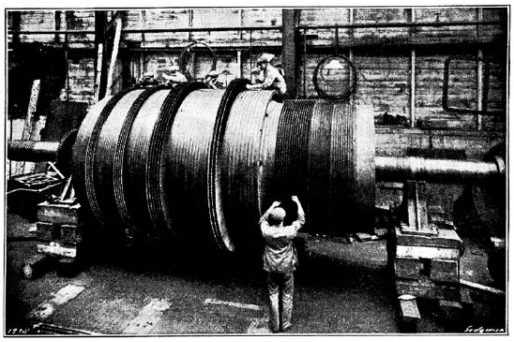


Figure 5.1.2.1: Beginning with the Industrial Revolution, humans have harnessed power through the use of the first law of thermodynamics, before we even understood it completely. This photo, of a steam engine at the Turbinia Works, dates from 1911, a mere 61 years after the first explicit statement of the first law of thermodynamics by Rudolph Clausius. (credit: public domain; author unknown)

One of the most important things we can do with heat transfer is to use it to do work for us. Such a device is called a **heat engine**. Car engines and steam turbines that generate electricity are examples of heat engines. Figure 5.1.2.2 shows schematically how the first law of thermodynamics applies to the typical heat engine.

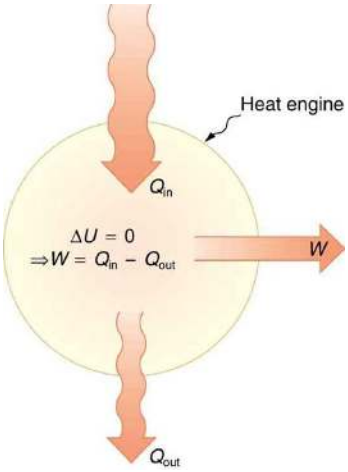


Figure 5.1.2.2: Schematic representation of a heat engine, governed, of course, by the first law of thermodynamics. It is impossible to devise a system where $Q_{out} = 0$, that is, in which no heat transfer occurs to the environment.

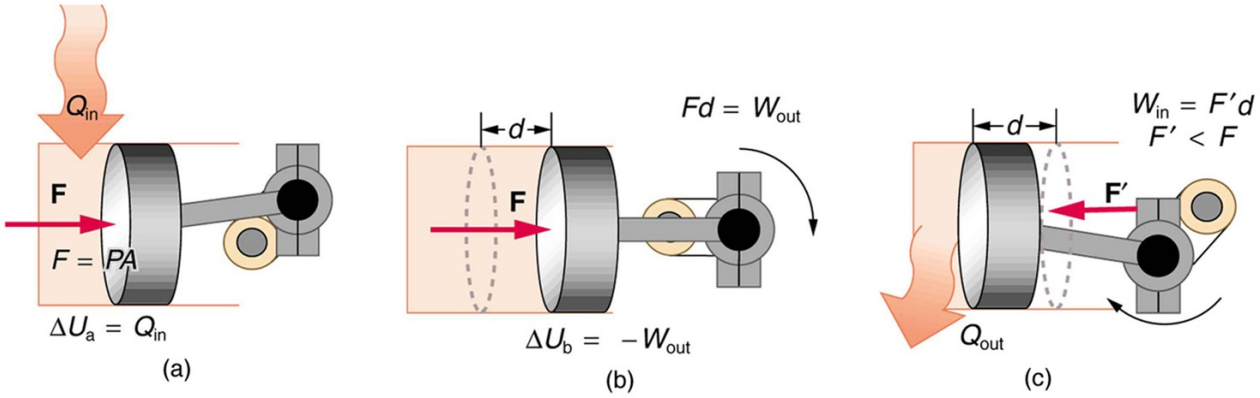


Figure 5.1.2.3***:*** *(a) Heat transfer to the gas in a cylinder increases the internal energy of the gas, creating higher pressure and temperature. (b) The force exerted on the movable cylinder does work as the gas expands. Gas pressure and temperature decrease when it expands, indicating that the gas's internal energy has been decreased by doing work. (c) Heat transfer to the environment further reduces pressure in the gas so that the piston can be more easily returned to its starting position.*

The illustrations above show one of the ways in which heat transfer does work. Fuel combustion produces heat transfer to a gas in a cylinder, increasing the pressure of the gas and thereby the force it exerts on a movable piston. The gas does work on the outside world, as this force moves the piston through some distance. Heat transfer to the gas cylinder results in work being done. To repeat this process, the piston needs to be returned to its starting point. Heat transfer now occurs from the gas to the surroundings so that its pressure decreases, and a force is exerted by the surroundings to push the piston back through some distance. Variations of this process are employed daily in hundreds of millions of heat engines. We will examine heat engines in detail in the next section. In this section, we consider some of the simpler underlying processes on which heat engines are based.

PV Diagrams and their Relationship to Work Done on or by a Gas

A process by which a gas does work on a piston at constant pressure is called an isobaric process. Since the pressure is constant, the force exerted is constant and the work done is given as

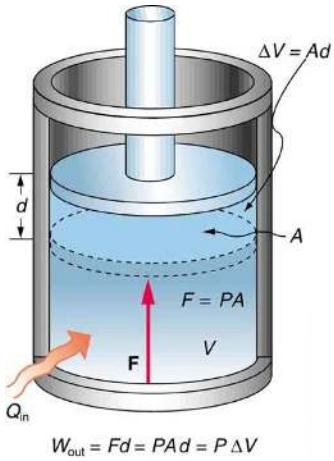


Figure 5.1.2.4: An isobaric expansion of a gas requires heat transfer to keep the pressure constant. Since pressure is constant, the work done is $P\Delta V$.

$$W = Fd \tag{5.1.2.1}$$

See the symbols as shown in Figure 5.1.2.4. Now $F = PA$, and so

$$W = PAd. \tag{5.1.2.2}$$

Because the volume of a cylinder is its cross-sectional area A times its length d, we see that $Ad = \Delta V$, the change in volume; thus,

$$W = P\Delta V\,(isobaric\,process). \tag{5.1.2.3}$$

Note that if ΔV is positive, then W is positive, meaning that work is done *by* the gas on the outside world.

(Note that the pressure involved in this work that we've called P is the pressure of the gas *inside* the tank. If we call the pressure outside the tank P_{ext}, an expanding gas would be working *against* the external pressure; the work done would therefore be $W = -P_{ext}\delta V$ (isobaric process). Many texts use this definition of work, and not the definition based on internal pressure, as the basis of the First Law of Thermodynamics. This definition reverses the sign conventions for work, and results in a statement of the first law that becomes $\Delta U = Q + W$.)

It is not surprising that $W = P\Delta V$, since we have already noted in our treatment of fluids that pressure is a type of potential energy per unit volume and that pressure in fact has units of energy divided by volume. We also noted in our discussion of the ideal gas law that PV

has units of energy. In this case, some of the energy associated with pressure becomes work.

Figure 5.1.2.5 shows a graph of pressure versus volume (that is, a PV diagram for an isobaric process. You can see in the figure that the work done is the area under the graph. This property of PV diagrams is very useful and broadly applicable: *the work done on or by a system in going from one state to another equals the area under the curve on a PV diagram.*

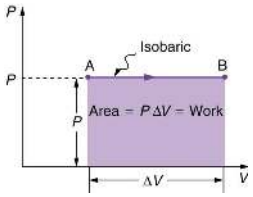


Figure \(\PageIndex{4\).. The area under the curve equals the work done by the gas, since $W = P\Delta V$.

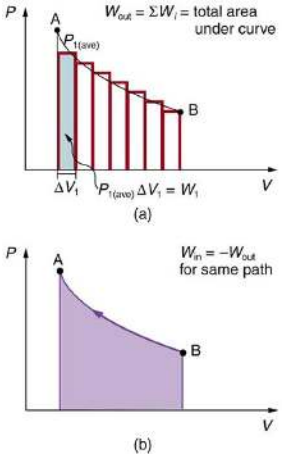


Figure 5.1.2.6: (a) A PV diagram in which pressure varies as well as volume. The work done for each interval is its average pressure times the change in volume, or the area under the curve over that interval. Thus the total area under the curve equals the total work done. (b) Work must be done on the system to follow the reverse path. This is interpreted as a negative area under the curve.

We can see where this leads by considering Figure 5.1.2.6(a), which shows a more general process in which both pressure and volume change. The area under the curve is closely approximated by dividing it into strips, each having an average constant pressure $P_{i(ave)}$. The work done is $W_i = P_{i(ave)}\Delta V_i$ for each strip, and the total work done is the sum of the W_i. Thus the total work done is the total area under the curve. If the path is reversed, as in Figure 5.1.2.6(b), then work is done on the system. The area under the curve in that case is negative, because ΔV is negative.

PV diagrams clearly illustrate that *the work done depends on the path taken and not just the endpoints*. This path dependence is seen in Figure 5.1.2.7(a), where more work is done in going from A to C by the path via point B than by the path via point D. The vertical paths, where volume is constant, are called isochoric processes. Since volume is constant, $\Delta V = 0$, and no work is done in an isochoric process. Now, if the system follows the cyclical path ABCDA, as in Figure 5.1.2.7(b), then the total work done is the area inside the loop. The negative area below path CD subtracts, leaving only the area inside the rectangle. In fact, the work done in any cyclical process (one that returns to its starting point) is the area inside the loop it forms on a PV diagram, as Figure 5.1.2.7(c) illustrates for a general cyclical process. Note that the loop must be traversed in the clockwise direction for work to be positive—that is, for there to be a net work output.

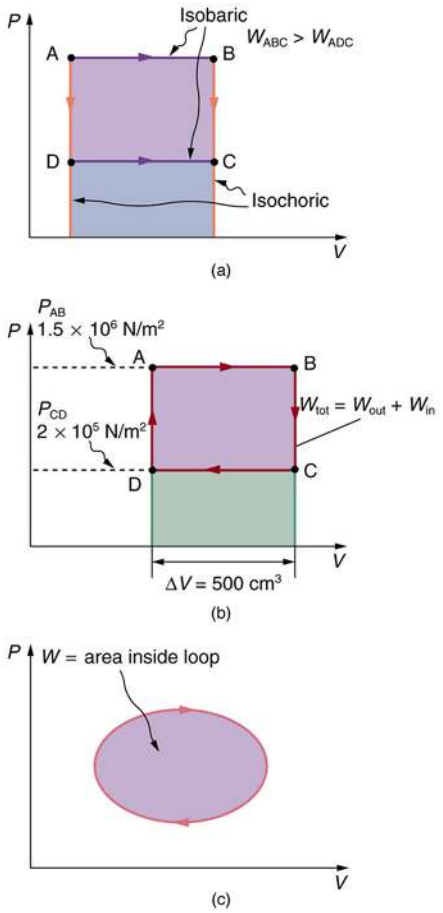


Figure 5.1.2.7: (a) The work done in going from A to C depends on path. The work is greater for the path ABC than for the path ADC, because the former is at higher pressure. In both cases, the work done is the area under the path. This area is greater for path ABC. (b) The total work done in the cyclical process ABCDA is the area inside the loop, since the negative area below CD subtracts out, leaving just the area inside the rectangle. (The values given for the pressures and the change in volume are intended for use in the example below.) (c) The area inside any closed loop is the work done in the cyclical process. If the loop is traversed in a clockwise direction, W is positive—it is work done on the outside environment. If the loop is traveled in a counter-clockwise direction, W is negative—it is work that is done to the system.

Example 5.1.2.1: Total Work Done in a Cyclical Process Equals the Area Inside the Closed Loop on a *PV* Diagram

Calculate the total work done in the cyclical process ABCDA shown in Figure 5.1.2.7(b) by the following two methods to verify that work equals the area inside the closed loop on the PV

diagram. (Take the data in the figure to be precise to three significant figures.) (a) Calculate the work done along each segment of the path and add these values to get the total work. (b) Calculate the area inside the rectangle ABCDA.

Strategy

To find the work along any path on a PV diagram, you use the fact that work is pressure times change in volume, or $W = P\Delta V$.

So in part (a), this value is calculated for each leg of the path around the closed loop.

Solution for (a)

The work along path AB is

$$W_{AB} = P_{AB}\Delta Y_{AB} \tag{5.1.2.4}$$

$$= (1.50 \times 10^6\, N/m^2)(5.00 \times 10^{-4}\, m^3) = 750\, J. \tag{5.1.2.5}$$

Since the path BC is isochoric, $\Delta V_{BC} = 0$, and so $W_{BC} = 0$. The work along path CD is negative, since ΔV_{CD} is negative (the volume decreases). The work is

$$W_{CD} = P_{CD}\Delta V_{CD} \tag{5.1.2.6}$$

$$= (2.00 \times 10^5\, N/m^2)(-5.00 \times 10^{-4}\, m^3) = -100\, J. \tag{5.1.2.7}$$

Again, since the path DA is isochoric, $\delta V_{DA} = 0$, and so $W_{DA} = 0$. Now the total work is

$$W = W_{AB} + W_{BC} + W_{CD} + W_{DA} \tag{5.1.2.8}$$

$$= 750\, J + 0 + (-100\, J) + 0 = 650\, J. \tag{5.1.2.9}$$

Solution for (b)

The area inside the rectangle is its height times its width, or

$$area = (P_{AB} - P_{CD})\Delta V \tag{5.1.2.10}$$

$$= [(1.50 \times 10^6\, N/m^2) - (2.00 \times 10^5\, N/m^2)](5.00 \times 1 -^{-4}\, m^3) \tag{5.1.2.11}$$

$$= 650\, J. \tag{5.1.2.12}$$

Thus,

$$area = 650\, J = W. \tag{5.1.2.13}$$

Discussion

The result, as anticipated, is that the area inside the closed loop equals the work done. The area is often easier to calculate than is the work done along each path. It is also convenient to visualize the area inside different curves on PV diagrams in order to see which processes might produce the most work. Recall that work can be done to the system, or by the system, depending on the sign of W. A positive W is work that is done by the system on the outside environment; a negative W is work that is done by the system on the outside environment; a negative W represents work done by the environment on the system.

Figure 5.1.2.8(a) shows two other important processes on a PV diagram. For comparison, both are shown starting from the same point A. The upper curve ending at point B is an isothermal process—that is, one in which temperature is kept constant. If the gas behaves like an ideal gas, as is often the case, and if no phase change occurs, then $PV = nRT$. Since T is constant, PV is a constant for an isothermal process. We ordinarily expect the temperature of a gas to decrease as it expands, and so we correctly suspect that heat transfer must occur from the surroundings to the gas to keep the temperature constant during an isothermal expansion. To show this more rigorously for the special case of a monatomic ideal gas, we note that the average kinetic energy of an atom in such a gas is given by

$$\frac{1}{2}m\bar{v}^2 = \frac{3}{2}kT. \tag{5.1.2.14}$$

The kinetic energy of the atoms in a monatomic ideal gas is its only form of internal energy, and so its total internal energy U is

$$U = N\frac{1}{2}m\bar{v}^2 = \frac{3}{2}NkT,\ (monatomic\, ideal\, gas), \tag{5.1.2.15}$$

where N is the number of atoms in the gas. This relationship means that the internal energy of an ideal monatomic gas is constant during an isothermal process—that is, $\Delta U = 0$. If the internal energy does not change, then the net heat transfer into the gas must equal the net work done by the gas. That is, because $\Delta U = Q - W = 0$ here, $Q = W$. We must have just enough heat transfer to replace the work done. An isothermal process is inherently slow, because heat transfer occurs continuously to keep the gas temperature constant at all times and must be allowed to spread through the gas so that there are no hot or cold regions.

Also shown in Figure 5.1.2.8(a) is a curve AC for an **adiabatic** process, defined to be one in which there is no heat transfer—that is, $Q = 0$. Processes that are nearly adiabatic can be achieved either by using very effective insulation or by performing the process so fast that there is little time for heat transfer. Temperature must decrease during an adiabatic expansion process, since work is done at the expense of internal energy:

$$U = \frac{3}{2}NkT. \tag{5.1.2.16}$$

(You might have noted that a gas released into atmospheric pressure from a pressurized cylinder is substantially colder than the gas in the cylinder.) In fact, because $Q = 0,\ \Delta U = -W$ for an adiabatic process. Lower temperature results in lower pressure

along the way, so that curve AC is lower than curve AB, and less work is done. If the path ABCA could be followed by cooling the gas from B to C at constant volume (isochorically), Figure 5.1.2.8(b), there would be a net work output.

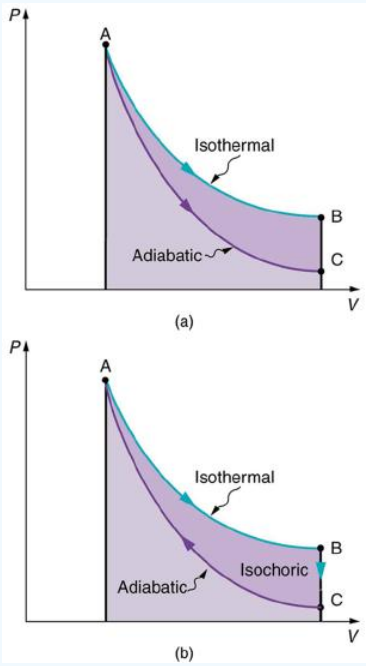


Figure 5.1.2.8: (a) The upper curve is an isothermal process ($\Delta T = 0$), whereas the lower curve is an adiabatic process ($Q = 0$). Both start from the same point A, but the isothermal process does more work than the adiabatic because heat transfer into the gas takes place to keep its temperature constant. This keeps the pressure higher all along the isothermal path than along the adiabatic path, producing more work. The adiabatic path thus ends up with a lower pressure and temperature at point C, even though the final volume is the same as for the isothermal process. (b) The cycle ABCA produces a net work output.

Reversible Processes

Both isothermal and adiabatic processes such as shown in Figure 5.1.2.8 are reversible in principle. A reversible process is one in which both the system and its environment can return to exactly the states they were in by following the reverse path. The reverse isothermal and adiabatic paths are BA and CA, respectively. Real macroscopic processes are never exactly reversible. In the previous examples, our system is a gas (like that in Figure 5.1.2.4), and its environment is the piston, cylinder, and the rest of the universe. If there are any energy-dissipating mechanisms, such as friction or turbulence, then heat transfer to the environment occurs for either direction of the piston. So, for example, if the path BA is followed and there is friction, then the gas will be returned to its original state but the environment will not—it will have been heated in both directions. Reversibility requires the direction of heat transfer to reverse for the reverse path. Since dissipative mechanisms cannot be completely eliminated, real processes cannot be reversible.

There must be reasons that real macroscopic processes cannot be reversible. We can imagine them going in reverse. For example, heat transfer occurs spontaneously from hot to cold and never spontaneously the reverse. Yet it would not violate the first law of thermodynamics for this to happen. In fact, all spontaneous processes, such as bubbles bursting, never go in reverse. There is a second thermodynamic law that forbids them from going in reverse. When we study this law, we will learn something about nature and also find that such a law limits the efficiency of heat engines. We will find that heat engines with the greatest possible theoretical efficiency would have to use reversible processes, and even they cannot convert all heat transfer into doing work. The table summarizes the simpler thermodynamic processes and their definitions.

Isobaric	Constant pressure	$W = P\Delta V$
Isochoric	Constant volume	$W = 0$
Isothermal	Constant temperature	$Q = W$
Adiabatic	No heat transfer	$Q = 0$

PHET EXPLORATIONS: STATES OF MATTER

Watch different types of molecules form a solid, liquid, or gas in teh States of Matter simuator. Add or remove heat and watch the phase change. Change the temperature or volume of a container and see a pressure-temperature diagram respond in real time. Relate the interaction potential to the forces between molecules.

Summary

- One of the important implications of the first law of thermodynamics is that machines can be harnessed to do work that humans previously did by hand or by external energy supplies such as running water or the heat of the Sun. A machine that uses heat transfer to do work is known as a heat engine.
- There are several simple processes, used by heat engines, that flow from the first law of thermodynamics. Among them are the isobaric, isochoric, isothermal and adiabatic processes.
- These processes differ from one another based on how they affect pressure, volume, temperature, and heat transfer.
- If the work done is performed on the outside environment, work (W) will be a positive value. If the work done is done to the heat engine system, work (W) will be a negative value.
- Some thermodynamic processes, including isothermal and adiabatic processes, are reversible in theory; that is, both the thermodynamic system and the environment can be returned to their initial states. However, because of loss of energy owing to the second law of thermodynamics, complete reversibility does not work in practice.

Glossary

heat engine

a machine that uses heat transfer to do work

isobaric process

constant-pressure process in which a gas does work

isochoric process

a constant-volume process

isothermal process

a constant-temperature process

adiabatic process

a process in which no heat transfer takes place

reversible process

a process in which both the heat engine system and the external environment theoretically can be returned to their original states

---

This page titled 5.1.2: The First Law of Thermodynamics and Some Simple Processes is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **15.2: The First Law of Thermodynamics and Some Simple Processes** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/college-physics.

5.1.3: Gravitational Potential Energy

Learning Objectives

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

- Explain gravitational potential energy in terms of work done against gravity.
- Show that the gravitational potential energy of an object of mass *m* at height *h* on Earth is given by $PE_g = mgh$
- Show how knowledge of the potential energy as a function of position can be used to simplify calculations and explain physical phenomena.

Work Done Against Gravity

Climbing stairs and lifting objects is work in both the scientific and everyday sense—it is work done against the gravitational force. When there is work, there is a transformation of energy. The work done against the gravitational force goes into an important form of stored energy that we will explore in this section.

Let us calculate the work done in lifting an object of mass m through a height h such as in Figure. If the object is lifted straight up at constant speed, then the force needed to lift it is equal to its weight mg. The work done on the mass is then $W = Fd = mgh$. We define this to be the **gravitational potential energy** (PE_g) put into (or gained by) the object-Earth system. This energy is associated with the state of separation between two objects that attract each other by the gravitational force. For convenience, we refer to this as the PE_g gained by the object, recognizing that this is energy stored in the gravitational field of Earth. Why do we use the word "system"? Potential energy is a property of a system rather than of a single object—due to its physical position. An object's gravitational potential is due to its position relative to the surroundings within the Earth-object system. The force applied to the object is an external force, from outside the system. When it does positive work it increases the gravitational potential energy of the system. Because gravitational potential energy depends on relative position, we need a reference level at which to set the potential energy equal to 0. We usually choose this point to be Earth's surface, but this point is arbitrary; what is important is the *difference* in gravitational potential energy, because this difference is what relates to the work done. The difference in gravitational potential energy of an object (in the Earth-object system) between two rungs of a ladder will be the same for the first two rungs as for the last two rungs.

Converting Between Potential Energy and Kinetic Energy

Gravitational potential energy may be converted to other forms of energy, such as kinetic energy. If we release the mass, gravitational force will do an amount of work equal to mgh on it, thereby increasing its kinetic energy by that same amount (by the work-energy theorem). We will find it more useful to consider just the conversion of PE_g to KE without explicitly considering the intermediate step of work. (See Example 5.1.3.2.) This shortcut makes it is easier to solve problems using energy (if possible) rather than explicitly using forces.

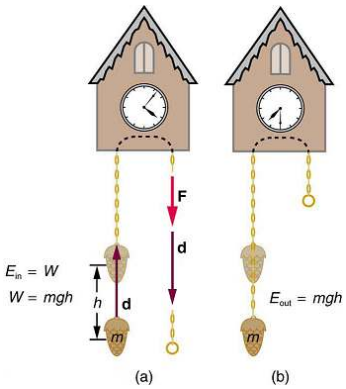


Figure 5.1.3.1: (a) The work done to lift the weight is stored in the mass-Earth system as gravitational potential energy. (b) As the weight moves downward, this gravitational potential energy is transferred to the cuckoo clock.

More precisely, we define the *change* in gravitational potential energy ΔPE_g to be

$$\Delta PE_g = mgh, \tag{5.1.3.1}$$

where, for simplicity, we denote the change in height by h rather than the usual Δh. Note that h is positive when the final height is greater than the initial height, and vice versa. For example, if a 0.500-kg mass hung from a cuckoo clock is raised 1.00 m, then its change in gravitational potential energy is

$$mgh = (0.500\,kg)(9.80\,m/s^2)(1.00\,m) \tag{5.1.3.2}$$

$$= 4.90\,kg \cdot m^2/s^2 = 4.90\,J. \tag{5.1.3.3}$$

Note that the units of gravitational potential energy turn out to be joules, the same as for work and other forms of energy. As the clock runs, the mass is lowered. We can think of the mass as gradually giving up its 4.90 J of gravitational potential energy, *without directly considering the force of gravity that does the work.*

Using Potential Energy to Simplify Calculations

The equation $\Delta PE_g = mgh$ applies for any path that has a change in height of h, not just when the mass is lifted straight up. (See Figure.) It is much easier to calculate mgh (a simple multiplication) than it is to calculate the work done along a complicated path. The idea of gravitational potential energy has the double advantage that it is very broadly applicable and it makes calculations easier. From now on, we will consider that any change in vertical position h of a mass m is accompanied by a change in gravitational potential energy mgh, and we will avoid the equivalent but more difficult task of calculating work done by or against the gravitational force.

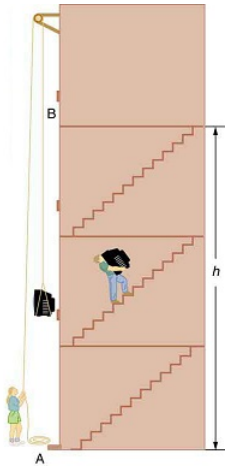


Figure 5.1.3.2: The change in gravitational potential energy (ΔPE_g) between points A and B is independent of the path $\Delta PE_g = mgh$ for any path between the two points. Gravity is one of a small class of forces where the work done by or against the force depends only on the starting and ending points, not on the path between them.

Example 5.1.3.1: The Force to Stop Falling

A 60.0-kg person jumps onto the floor from a height of 3.00 m. If he lands stiffly (with his knee joints compressing by 0.500 cm), calculate the force on the knee joints.

Strategy

This person's energy is brought to zero in this situation by the work done on him by the floor as he stops. The initial PE_g is transformed into KE as he falls. The work done by the floor reduces this kinetic energy to zero.

Solution

The work done on the person by the floor as he stops is given by

$$W = Fd\cos\theta = -Fd \tag{5.1.3.4}$$

with a minus sign because the displacement while stopping and the force from floor are in opposite directions ($\cos\theta = \cos 180^o = -1$.) The floor removes energy from the system, so it does negative work.

The kinetic energy the person has upon reaching the floor is the amount of potential energy lost by falling through height h:

$$KE = -\Delta PE_g = -mgh \tag{5.1.3.5}$$

The distance d that the person's knees bend is much smaller than the height h of the fall, so the additional change in gravitational potential energy during the knee bend is ignored. The work W done by the floor on the person stops the person and brings the person's kinetic energy to zero:

$$W = -KE = mgh \tag{5.1.3.6}$$

Combining this equation with the expression for W gives

$$-Fd = mgh. \tag{5.1.3.7}$$

Recalling that h is negative because the person fell *down*, the force on the knee joints is given by

$$F = -\frac{mgh}{d} = -\frac{(60.0\,kg)(9.80\,m/s^2)(-3.00\,m)}{5.00 \times 10^{-3}\,m} = 3.53 \times 10^5\,N. \tag{5.1.3.8}$$

Discussion

Such a large force (500 times more than the person's weight) over the short impact time is enough to break bones. A much better way to cushion the shock is by bending the legs or rolling on the ground, increasing the time over which the force acts. A bending motion of 0.5 m this way yields a force 100 times smaller than in the example. A kangaroo's hopping shows this method in action. The kangaroo is the only large animal to use hopping for locomotion, but the shock in hopping is cushioned by the bending of its hind legs in each jump.(See Figure.)



Figure 5.1.3.3: The work done by the ground upon the kangaroo reduces its kinetic energy to zero as it lands. However, by applying the force of the ground on the hind legs over a longer distance, the impact on the bones is reduced. (credit: Chris Samuel, Flickr)

Example 5.1.3.2: Finding the Speed of a Roller Coaster from its Height

(a) What is the final speed of the roller coaster shown in Figure, if it starts from rest at the top of the 20.0 m hill and work done by frictional forces is negligible? (b) What is its final speed (again assuming negligible friction) if its initial speed is 5.00 m/s?

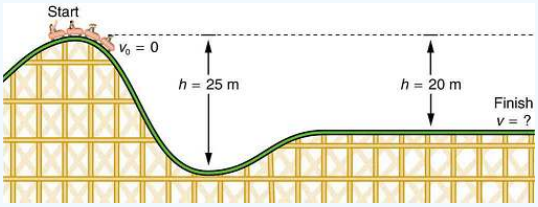


Figure 5.1.3.4: The speed of a roller coaster increases as gravity pulls it downhill and is greatest at its lowest point. Viewed in terms of energy, the roller-coaster-Earth system's gravitational potential energy is converted to kinetic energy. If work done by friction is negligible, all ΔPE_g is converted to KE.

Strategy

The roller coaster loses potential energy as it goes downhill. We neglect friction, so that the remaining force exerted by the track is the normal force, which is perpendicular to the direction of motion and does no work. The net work on the roller coaster is then done by gravity alone. The *loss* of gravitational potential energy from moving *downward* through a distance h

equals the *gain* in kinetic energy. This can be written in equation form as $-\Delta PE = \Delta KE$. Using the equations for PE_g and KE we can solve for the final speed v, which is the desired quantity.

Solution for (a)

Here the initial kinetic energy is zero, so that $\Delta KE = \frac{1}{2}mv^2$. The equation for change in potential energy states that $\Delta PE_g = mgh$. Since h is negative in this case, we will rewrite this as $\Delta PE_g = -mg|h|$ to show the minus sign clearly. Thus,

$$-\Delta PE_g = \Delta KE \tag{5.1.3.9}$$

becomes

$$mg|h| = \frac{1}{2}mv^2 \tag{5.1.3.10}$$

Solving for v we find that mass cancels and that

$$v = \sqrt{2g|h|}. \tag{5.1.3.11}$$

Substituting known values,

$$v = \sqrt{2(9.80\,m/s^2)(20.0\,m)} \tag{5.1.3.12}$$

$$= 19.8\,m/s \tag{5.1.3.13}$$

Solution for (b)

Again $-\Delta PE_g = \Delta KE$. In this case there is initial kinetic energy, so $\Delta KE = \frac{1}{2}mv^2 - \frac{1}{2}mv_0^2$. Thus,

$$mgh = \frac{1}{2}mv^2 - \frac{1}{2}m_0^2. \tag{5.1.3.14}$$

Rearranging gives

$$\frac{1}{2}mv^2 = mg|h| + \frac{1}{2}mv_0^2. \tag{5.1.3.15}$$

This means that the final kinetic energy is the sum of the initial kinetic energy and the gravitational potential energy. Mass again cancels, and

$$v = \sqrt{2g|h| + v_0^2}. \tag{5.1.3.16}$$

This equation is very similar to the kinematics equation $v = \sqrt{v_0^2 + 2ad}$, but it is more general—the kinematics equation is valid only for constant acceleration, whereas our equation above is valid for any path regardless of whether the object moves with a constant acceleration. Now, substituting known values gives

$$v = \sqrt{2(9.80\,m/s^2)(20.0\,m) + (5.00)^2} \tag{5.1.3.17}$$

$$= 20.4\,m/s. \tag{5.1.3.18}$$

Discussion and Implications

First, note that mass cancels. This is quite consistent with observations made in Falling Objects that all objects fall at the same rate if friction is negligible. Second, only the speed of the roller coaster is considered; there is no information about its direction at any point. This reveals another general truth. When friction is negligible, the speed of a falling body depends only on its initial speed and height, and not on its mass or the path taken. For example, the roller coaster will have the same final speed whether it falls 20.0 m straight down or takes a more complicated path like the one in the figure. Third, and perhaps unexpectedly, the final speed in part (b) is greater than in part (a), but by far less than 5.00 m/s. Finally, note that speed can be found at *any* height along the way by simply using the appropriate value of h at the point of interest.

We have seen that work done by or against the gravitational force depends only on the starting and ending points, and not on the path between, allowing us to define the simplifying concept of gravitational potential energy. We can do the same thing for a few other forces, and we will see that this leads to a formal definition of the law of conservation of energy.

Making Connections: Take-Home Investigation—Converting Potential to Kinetic Energy

One can study the conversion of gravitational potential energy into kinetic energy in this experiment. On a smooth, level surface, use a ruler of the kind that has a groove running along its length and a book to make an incline (see Figure). Place a marble at the 10-cm position on the ruler and let it roll down the ruler. When it hits the level surface, measure the time it takes to roll one meter. Now place the marble at the 20-cm and the 30-cm positions and again measure the times it takes to roll 1 m on the level surface. Find the velocity of the marble on the level surface for all three positions. Plot velocity squared versus the distance traveled by the marble. What is the shape of each plot? If the shape is a straight line, the plot shows that the marble's kinetic energy at the bottom is proportional to its potential energy at the release point.

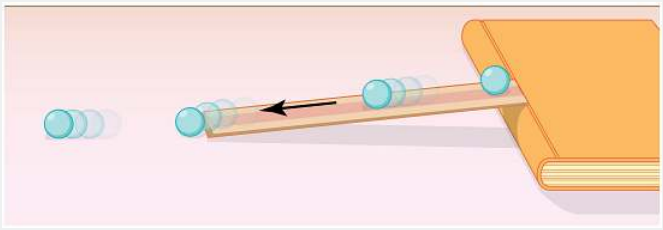


Figure 5.1.3.5: A marble rolls down a ruler, and its speed on the level surface is measured.

Summary

- Work done against gravity in lifting an object becomes potential energy of the object-Earth system.
- The change in gravitational potential energy ΔPE_g, is $\Delta PE_g = mgh$, with h being the increase in height and g the acceleration due to gravity.
- The gravitational potential energy of an object near Earth's surface is due to its position in the mass-Earth system. Only differences in gravitational potential energy, ΔPE_g, have physical significance.
- As an object descends without friction, its gravitational potential energy changes into kinetic energy corresponding to increasing speed, so that $\Delta KE = -\Delta PE_g$.

Glossary

gravitational potential energy
the energy an object has due to its position in a gravitational field

This page titled 5.1.3: Gravitational Potential Energy is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **7.3: Gravitational Potential Energy** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/college-physics.

5.2: Introduction to the Second Law of Thermodynamics - Heat Engines and their Efficiency

Learning Objectives

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

- State the expressions of the second law of thermodynamics.
- Calculate the efficiency and carbon dioxide emission of a coal-fired electricity plant, using second law characteristics.
- Describe and define the Otto cycle.



Figure 5.2.1: These ice floes melt during the Arctic summer. Some of them refreeze in the winter, but the second law of thermodynamics predicts that it would be extremely unlikely for the water molecules contained in these particular floes to reform the distinctive alligator-like shape they formed when the picture was taken in the summer of 2009. (credit: Patrick Kelley, U.S. Coast Guard, U.S. Geological Survey)

The second law of thermodynamics deals with the direction taken by spontaneous processes. Many processes occur spontaneously in one direction only—that is, they are irreversible, under a given set of conditions. Although irreversibility is seen in day-to-day life—a broken glass does not resume its original state, for instance—complete irreversibility is a statistical statement that cannot be seen during the lifetime of the universe. More precisely, an **irreversible process** is one that depends on path. If the process can go in only one direction, then the reverse path differs fundamentally and the process cannot be reversible. For example, as noted in the previous section, heat involves the transfer of energy from higher to lower temperature. A cold object in contact with a hot one never gets colder, transferring heat to the hot object and making it hotter. Furthermore, mechanical energy, such as kinetic energy, can be completely converted to thermal energy by friction, but the reverse is impossible. A hot stationary object never spontaneously cools off and starts moving. Yet another example is the expansion of a puff of gas introduced into one corner of a vacuum chamber. The gas expands to fill the chamber, but it never regroups in the corner. The random motion of the gas molecules could take them all back to the corner, but this is never observed to happen. (See Figure 5.2.2)

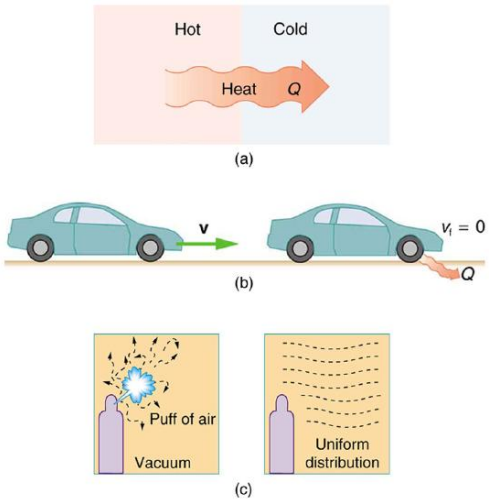


Figure 5.2.2: Examples of one-way processes in nature. (a) Heat transfer occurs spontaneously from hot to cold and not from cold to hot. (b) The brakes of this car convert its kinetic energy to heat transfer to the environment. The reverse process is impossible. (c) The burst of gas let into this vacuum chamber quickly expands to uniformly fill every part of the chamber. The random motions of the gas molecules will never return them to the corner.

The fact that certain processes never occur suggests that there is a law forbidding them to occur. The first law of thermodynamics would allow them to occur—none of those processes violate conservation of energy. The law that forbids these processes is called the second law of thermodynamics. We shall see that the second law can be stated in many ways that may seem different, but which in fact are equivalent. Like all natural laws, the second law of thermodynamics gives insights into nature, and its several statements imply that it is broadly applicable, fundamentally affecting many apparently disparate processes.

The already familiar direction of heat transfer from hot to cold is the basis of our first version of the **second law of thermodynamics**.

The Second Law of Thermodynamics (first expression)

Heat transfer occurs spontaneously from higher- to lower-temperature bodies but never spontaneously in the reverse direction.

Another way of stating this: It is impossible for any process to have as its sole result heat transfer from a cooler to a hotter object.

Heat Engines

Now let us consider a device that uses heat transfer to do work. As noted in the previous section, such a device is called a heat engine, and one is shown schematically in 5.2.3(b). Gasoline and diesel engines, jet engines, and steam turbines are all heat engines that do work by using part of the heat transfer from some source. Heat transfer from the hot object (or hot reservoir) is denoted as Q_f, while heat transfer into the cold object (or cold reservoir) is Q_c, and the work done by the engine is W. The temperatures of the hot and cold reservoirs are T_h and T_c, respectively.

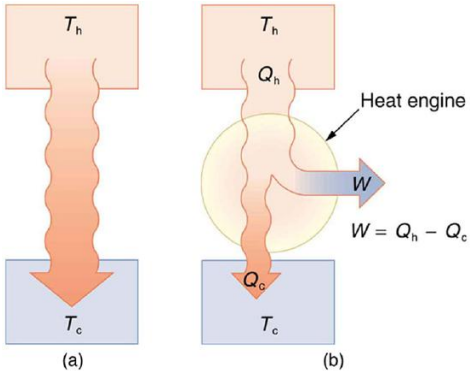


Figure 5.2.3: (a) Heat transfer occurs spontaneously from a hot object to a cold one, consistent with the second law of thermodynamics. (b) A heat engine, represented here by a circle, uses part of the heat transfer to do work. The hot and cold objects are called the hot and cold reservoirs Q_h is the heat transfer out of the hot reservoir, W is the work output, and Q_c is the heat transfer into the cold reservoir.

Because the hot reservoir is heated externally, which is energy intensive, it is important that the work is done as efficiently as possible. In fact, we would like W to equal Q_h and for there to be no heat transfer to the environment $(Q_c = 0)$. Unfortunately, this is impossible. The second law of thermodynamics also states, with regard to using heat transfer to do work (the second expression of the second law):

The Second Law of Thermodynamics (second expression)

It is impossible in any system for heat transfer from a reservoir to completely convert to work in a cyclical process in which the system returns to its initial state.

A cyclical process brings a system, such as the gas in a cylinder, back to its original state at the end of every cycle. Most heat engines, such as reciprocating piston engines and rotating turbines, use cyclical processes. The second law, just stated in its second form, clearly states that such engines cannot have perfect conversion of heat transfer into work done. Before going into the underlying reasons for the limits on converting heat transfer into work, we need to explore the relationships among W, Q_h, and Q_c and to define the efficiency of a cyclical heat engine. As noted, a cyclical process brings the system back to its original condition at the end of every cycle. Such a system's internal energy U is the same at the beginning and end of every cycle—that is, $\Delta U = 0$. The first law of thermodynamics states that

$$\Delta U = Q - W, \tag{5.2.1}$$

where Q is the *net* heat transfer during the cycle $(Q = Q_h - Q_c)$ and W is the net work done by the system. Since $\Delta U = 0$ for a complete cycle, we have

$$0 = Q - W, \tag{5.2.2}$$

so that

$$W = Q. \tag{5.2.3}$$

Thus the net work done by the system equals the net heat transfer into the system, or

$$W = Q_h - Q_c \, (cyclical \, process), \tag{5.2.4}$$

just as shown schematically in Figure 5.2.3(b). The problem is that in all processes, there is some heat transfer Q_c to the environment—and usually a very significant amount at that.

In the conversion of energy to work, we are always faced with the problem of getting less out than we put in. We define *conversion efficiency* Eff to be the ratio of useful work output to the energy input (or, in other words, the ratio of what we get to what we spend). In that spirit, we define the efficiency of a heat engine to be its net work output W divided by heat transfer to the engine Q_h, that is,

$$Eff = \frac{W}{Q_h}. \tag{5.2.5}$$

Since $W = Q_h - Q_c$ in a cyclical process, we can also express this as

$$Eff = \frac{Q_h - Q_c}{Q_h} = 1 - \frac{Q_c}{Q_h} \, (cyclical \, process), \tag{5.2.6}$$

making it clear that an efficiency of 1, or 100%, is possible only if there is no heat transfer to the environment $(Q_c = 0)$. Note that all Qs are positive. The direction of heat transfer is indicated by a plus or minus sign. For example, $Q - c$ is out of the system and so is preceded by a minus sign.

Example 5.2.1: Daily Work Done by a Coal-Fired Power Station, Its Efficiency and Carbon Dioxide Emissions

A coal-fired power station is a huge heat engine. It uses heat transfer from burning coal to do work to turn turbines, which are used to generate electricity. In a single day, a large coal power station has $2.50 \times 10^{14} \, J$ of heat transfer from coal and $1.48 \times 10^{1}4 \, J$ of heat transfer into the environment. (a) What is the work done by the power station? (b) What is the efficiency of the power station? (c) In the combustion process, the following chemical reaction occurs: $C + O_2 \rightarrow CO_2$ This implies that every 12 kg of coal puts 12 kg + 16 kg + 16 kg = 44 kg of carbon dioxide into the atmosphere. Assuming that 1 kg of coal can provide $2.5 \times 10^6 \, J$ of heat transfer upon combustion, how much CO_2 is emitted per day by this power plant?

Strategy for (a)

We can use $W = Q_h - Q_c$ to find the work output W,

assuming a cyclical process is used in the power station. In this process, water is boiled under pressure to form high-temperature steam, which is used to run steam turbine-generators, and then condensed back to water to start the cycle again.

Solution for (a)

Work output is given by:

$$W = Q_h - Q_c.$$

Substituting the given values:

$$\begin{aligned} W &= 2.50 \times 10^{1}4 \, J - 1.48 \times 10^{14} \, J \\ &= 1.02 \times 10^{14} \, J. \end{aligned}$$

Strategy for (b)

The efficiency can be calculated with $Eff = \frac{W}{Q_h}$ since Q_H is given and work W was found in the first part of this example.

Solution for (b)

Efficiency is given by $Eff = \frac{W}{Q_h}$. The work W was just found to be 1.01×10^{14}, and Q_h is given, so the efficiency is

$$Eff = \frac{1.01 \times 10^{14}\, J}{2.50 \times 10^{14}\, J} = 0.408,\ or\ 40.8\% \tag{5.2.7}$$

Strategy for (c)

The daily consumption of coal is calculated using the information that each day there is

$$\frac{2.50 \times 10^{14}\, J}{2.50 \times 10^{6}\, J/kg} = 1.0 \times 10^{8}\, kg.$$

Assuming that the coal is pure and that all the coal goes toward producing carbon dioxide, the carbon dioxide produced per day is

$$1.0 \times 10^{8}\, kg\, coal \times \frac{44\, kg\, CO_2}{12\, kg\, coal} = 3.7 \times 10^{8}\, kg\, CO_2.$$

This is 370,000 metric tons of $CO2$ produced every day.

Discussion

If all the work output is converted to electricity in a period of one day, the average power output is 1180 MW (this is left to you as an end-of-chapter problem). This value is about the size of a large-scale conventional power plant. The efficiency found is acceptably close to the value of 42% given for coal power stations. It means that fully 59.2% of the energy is heat transfer to the environment, which usually results in warming lakes, rivers, or the ocean near the power station, and is implicated in a warming planet generally. While the laws of thermodynamics limit the efficiency of such plants—including plants fired by nuclear fuel, oil, and natural gas—the heat transfer to the environment could be, and sometimes is, used for heating homes or for industrial processes. The generally low cost of energy has not made it economical to make better use of the waste heat transfer from most heat engines. Coal-fired power plants produce the greatest amount of CO_2 per unit energy output (compared to natural gas or oil), making coal the least efficient fossil fuel.

With the information given in Example 5.2.1, we can find characteristics such as the efficiency of a heat engine without any knowledge of how the heat engine operates, but looking further into the mechanism of the engine will give us greater insight. Figure 5.2.4 illustrates the operation of the common four-stroke gasoline engine. The four steps shown complete this heat engine's cycle, bringing the gasoline-air mixture back to its original condition.

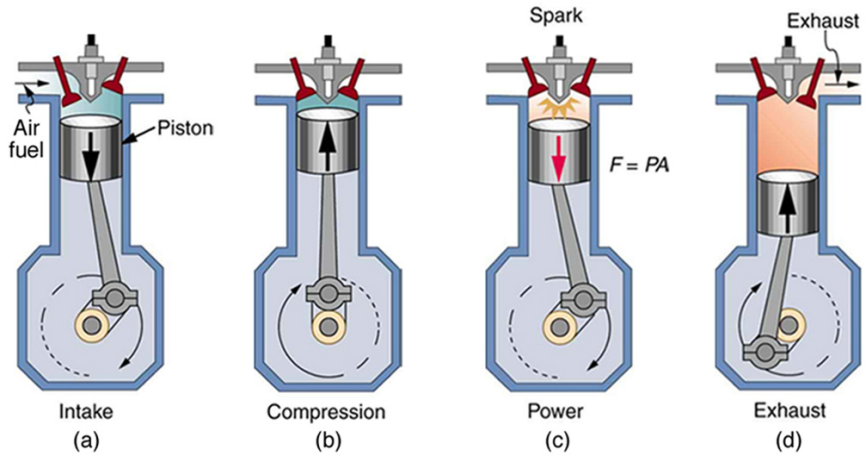


Figure 5.2.4: In the four-stroke internal combustion gasoline engine, heat transfer into work takes place in the cyclical process shown here. The piston is connected to a rotating crankshaft, which both takes work out of and does work on the gas in the cylinder. (a) Air is mixed with fuel during the intake stroke. (b) During the compression stroke, the air-fuel mixture is rapidly compressed in a nearly adiabatic process, as the piston rises with the valves closed. Work is done on the gas. (c) The power stroke has two distinct parts. First, the air-fuel mixture is ignited, converting chemical potential energy into thermal energy almost instantaneously, which leads to a great increase in pressure. Then the piston descends, and the gas does work by exerting a force through a distance in a nearly adiabatic process. (d) The exhaust stroke expels the hot gas to prepare the engine for another cycle, starting again with the intake stroke.

The Otto cycle shown in Figure 5.2.5a is used in four-stroke internal combustion engines, although in fact the true Otto cycle paths do not correspond exactly to the strokes of the engine.

The adiabatic process AB corresponds to the nearly adiabatic compression stroke of the gasoline engine. In both cases, work is done on the system (the gas mixture in the cylinder), increasing its temperature and pressure. Along path BC of the Otto cycle, heat transfer Q_h into the gas occurs at constant volume, causing a further increase in pressure and temperature. This process corresponds to burning fuel in an internal combustion engine, and takes place so rapidly that the volume is nearly constant. Path CD in the Otto cycle is an adiabatic expansion that does work on the outside world, just as the power stroke of an internal combustion engine does in its nearly adiabatic expansion. The work done by the system along path CD is greater than the work done on the system along path AB, because the pressure is greater, and so there is a net work output. Along path DA in the Otto cycle, heat transfer Q_c

from the gas at constant volume reduces its temperature and pressure, returning it to its original state. In an internal combustion engine, this process corresponds to the exhaust of hot gases and the intake of an air-gasoline mixture at a considerably lower temperature. In both cases, heat transfer into the environment occurs along this final path.

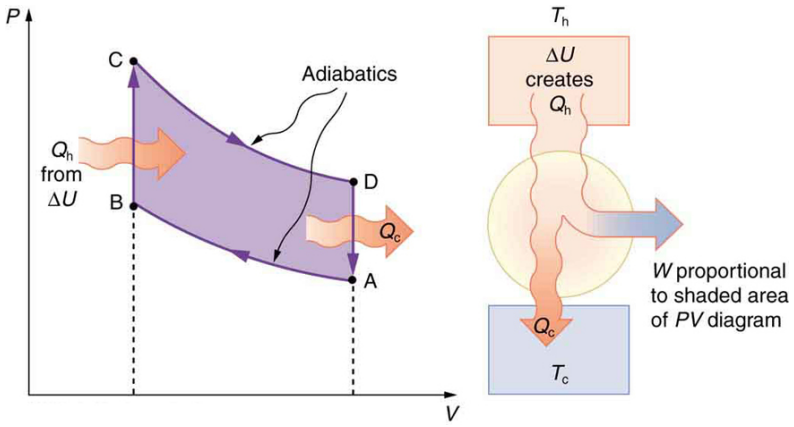


Figure 5.2.5: PV diagram for a simplified Otto cycle, analogous to that employed in an internal combustion engine. Point A corresponds to the start of the compression stroke of an internal combustion engine. Paths AB and CD are adiabatic and correspond to the compression and power strokes of an internal combustion engine, respectively. Paths BC and DA are isochoric and accomplish similar results to the ignition and exhaust-intake portions, respectively, of the internal combustion engine's cycle. Work is done on the gas along path AB, but more work is done by the gas along path CD, so that there is a net work output.

The net work done by a cyclical process is the area inside the closed path on a PV diagram, such as that inside path ABCDA in Figure 5.2.5. Note that in every imaginable cyclical process, it is absolutely necessary for heat transfer from the system to occur in order to get a net work output. In the Otto cycle, heat transfer occurs along path DA. If no heat transfer occurs, then the return path is the same, and the net work output is zero. The lower the temperature on the path AB, the less work has to be done to compress the gas. The area inside the closed path is then greater, and so the engine does more work and is thus more efficient. Similarly, the higher the temperature along path CD, the more work output there is (Figure 5.2.6. So efficiency is related to the temperatures of the hot and cold reservoirs. In the next section, we shall see what the absolute limit to the efficiency of a heat engine is, and how it is related to temperature.

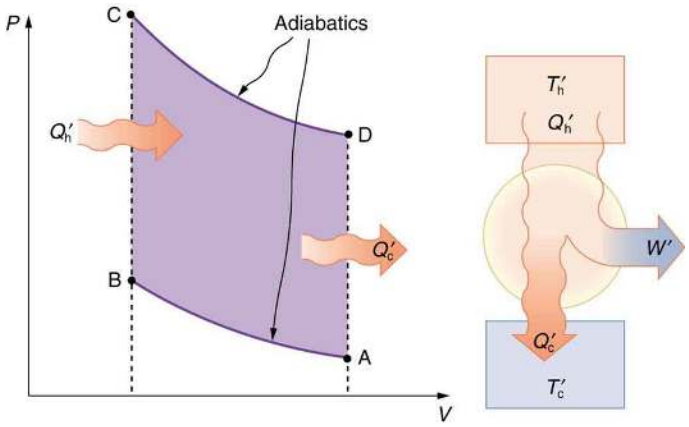


Figure 5.2.5: This Otto cycle produces a greater work output than the one in Figure 5, because the starting temperature of path CD is higher because the starting temperature of path CD is higher and the starting temperature of path AB is lower. The area inside the loop is greater, corresponding to greater net work output.

Summary

- The two expressions of the second law of thermodynamics are: (i) Heat transfer occurs spontaneously from higher- to lower-temperature bodies but never spontaneously in the reverse direction; and (ii) It is impossible in any system for heat transfer from a reservoir to completely convert to work in a cyclical process in which the system returns to its initial state.
- Irreversible processes depend on path and do not return to their original state. Cyclical processes are processes that return to their original state at the end of every cycle.
- In a cyclical process, such as a heat engine, the net work done by the system equals the net heat transfer into the system, or $W = Q_h - Q_c$, where Q_h is the heat transfer from the hot object (hot reservoir), and Q_c is the heat transfer into the cold object (cold reservoir).
- Efficiency can be expressed as $Eff = \frac{W}{Q_h}$, the ratio of work output divided by the amount of energy input.
- The four-stroke gasoline engine is often explained in terms of the Otto cycle, which is a repeating sequence of processes that convert heat into work.

Glossary

irreversible process
: any process that depends on path direction

second law of thermodynamics
: heat transfer flows from a hotter to a cooler object, never the reverse, and some heat energy in any process is lost to available work in a cyclical process

cyclical process
: a process in which the path returns to its original state at the end of every cycle

Otto cycle
: a thermodynamic cycle, consisting of a pair of adiabatic processes and a pair of isochoric processes, that converts heat into work, e.g., the four-stroke engine cycle of intake, compression, ignition, and exhaust

This page titled 5.2: Introduction to the Second Law of Thermodynamics - Heat Engines and their Efficiency is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **15.3: Introduction to the Second Law of Thermodynamics - Heat Engines and their Efficiency** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/college-physics.

5.2.1: Carnot's Perfect Heat Engine- The Second Law of Thermodynamics Restated

Learning Objectives

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

- Identify a Carnot cycle.
- Calculate maximum theoretical efficiency of a nuclear reactor.
- Explain how dissipative processes affect the ideal Carnot engine.

We know from the second law of thermodynamics that a heat engine cannot be 100% efficient, since there must always be some heat transfer Q_c to the environment, which is often called waste heat. How efficient, then, can a heat engine be? This question was answered at a theoretical level in 1824 by a young French engineer, Sadi Carnot (1796–1832), in his study of the then-emerging heat engine technology crucial to the Industrial Revolution. He devised a theoretical cycle, now called the **Carnot cycle**, which is the most efficient cyclical process possible. The second law of thermodynamics can be restated in terms of the Carnot cycle, and so what Carnot actually discovered was this fundamental law. Any heat engine employing the Carnot cycle is called a **Carnot engine**.



Figure 5.2.1.1: This novelty toy, known as the drinking bird, is an example of Carnot's engine. It contains methylene chloride (mixed with a dye) in the abdomen, which boils at a very low temperature—about 100^oF. To operate, one gets the bird's head wet. As the water evaporates, fluid moves up into the head, causing the bird to become top-heavy and dip forward back into the water. This cools down the methylene chloride in the head, and it moves back into the abdomen, causing the bird to become bottom heavy and tip up. Except for a very small input of energy—the original head-wetting—the bird becomes a perpetual motion machine of sorts. (credit: Arabesk.nl, Wikimedia Commons)

What is crucial to the Carnot cycle—and, in fact, defines it—is that only reversible processes are used. Irreversible processes involve dissipative factors, such as friction and turbulence. This increases heat transfer Q_c to the environment and reduces the efficiency of the engine. Obviously, then, reversible processes are superior.

Carnot Engine

Stated in terms of reversible processes, the **second law of thermodynamics** has a third form:

A Carnot engine operating between two given temperatures has the greatest possible efficiency of any heat engine operating between these two temperatures. Furthermore, all engines employing only reversible processes have this same maximum efficiency when operating between the same given temperatures.

Figure 5.2.1.2 shows the PV diagram for a Carnot cycle. The cycle comprises two isothermal and two adiabatic processes. Recall that both isothermal and adiabatic processes are, in principle, reversible.

Carnot also determined the efficiency of a perfect heat engine—that is, a Carnot engine. It is always true that the efficiency of a cyclical heat engine is given by:

$$Eff = \frac{Q_h - Q_c}{Q_h} = 1 - \frac{Q_c}{Q_h}. \tag{5.2.1.1}$$

What Carnot found was that for a perfect heat engine, the ratio Q_c/Q_h equals the ratio of the absolute temperatures of the heat reservoirs. That is, $Q_c/Q_h = T_c/T_h$ for a Carnot engine, so that the maximum or **Carnot efficiency** Eff_c is given by

$$Eff_c = 1 - \frac{T_c}{T_h}, \tag{5.2.1.2}$$

where T_h and T_c are in kelvins (or any other absolute temperature scale). No real heat engine can do as well as the Carnot efficiency —an actual efficiency of about 0.7 of this maximum is usually the best that can be accomplished. But the ideal Carnot engine, like the drinking bird above, while a fascinating novelty, has zero power. This makes it unrealistic for any applications.

Carnot's interesting result implies that 100% efficiency would be possible only if $T_c = 0$ - that is, only if the cold reservoir were at absolute zero, a practical and theoretical impossibility. But the physical implication is this—the only way to have all heat transfer go into doing work is to remove *all* thermal energy, and this requires a cold reservoir at absolute zero.

It is also apparent that the greatest efficiencies are obtained when the ratio T_c/T_h is as small as possible. Just as discussed for the Otto cycle in the previous section, this means that efficiency is greatest for the highest possible temperature of the hot reservoir and lowest possible temperature of the cold reservoir. (This setup increases the area inside the closed loop on the PV diagram; also, it seems reasonable that the greater the temperature difference, the easier it is to divert the heat transfer to work.) The actual reservoir temperatures of a heat engine are usually related to the type of heat source and the temperature of the environment into which heat transfer occurs. Consider the following example.

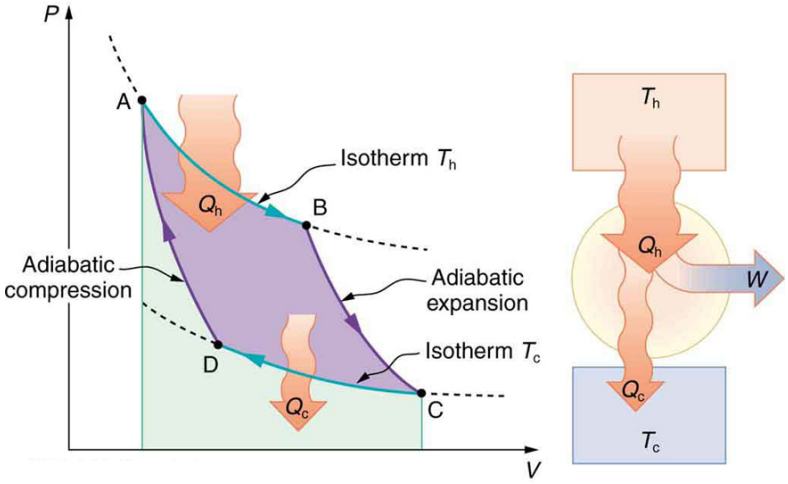


Figure 5.2.1.2: PV diagram for a Carnot cycle, employing only reversible isothermal and adiabatic processes. Heat transfer Q_h occurs into the working substance during the isothermal path AB, which takes place at constant temperature T_h. Heat transfer Q_c occurs out of the working substance during the isothermal path CD, which takes place at constant temperature T_c. The net work output W equals the area inside the path ABCDA. Also shown is a schematic of a Carnot engine operating between hot and cold reservoirs at temperatures T_h and T_c. Any heat engine using reversible processes and operating between these two temperatures will have the same maximum efficiency as the Carnot engine.

Example 5.2.1.1: Maximum Theoretical Efficiency for a Nuclear Reactor

A nuclear power reactor has pressurized water at 300^oC. (Higher temperatures are theoretically possible but practically not, due to limitations with materials used in the reactor.) Heat transfer from this water is a complex process (see Figure 5.2.1.3). Steam, produced in the steam generator, is used to drive the turbine-generators. Eventually the steam is condensed to water at 27^oC and then heated again to start the cycle over. Calculate the maximum theoretical efficiency for a heat engine operating between these two temperatures.

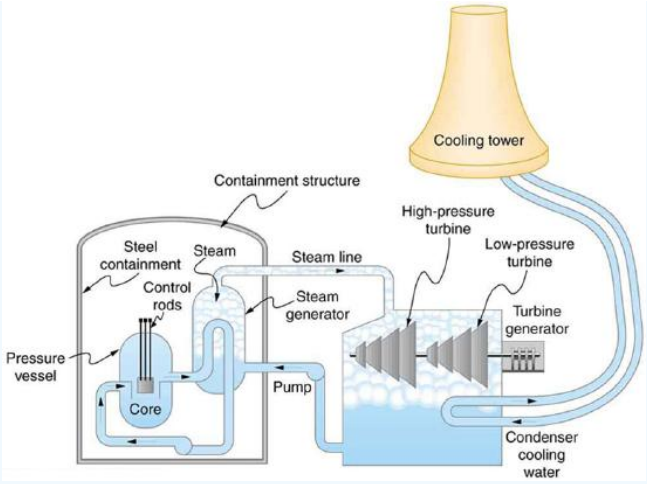


Figure 5.2.1.3. Schematic diagram of a pressurized water nuclear reactor and the steam turbines that convert work into electrical energy. Heat exchange is used to generate steam, in part to avoid contamination of the generators with radioactivity. Two turbines are used because this is less expensive than operating a single generator that produces the same amount of electrical energy. The steam is condensed to liquid before being returned to the heat exchanger, to keep exit steam pressure low and aid the flow of steam through the turbines (equivalent to using a lower-temperature cold reservoir). The considerable energy associated with condensation must be dissipated into the local environment; in this example, a cooling tower is used so there is no direct heat transfer to an aquatic environment. (Note that the water going to the cooling tower does not come into contact with the steam flowing over the turbines.)

Strategy

Since temperatures are given for the hot and cold reservoirs of this heat engine, $Eff_c = 1 - \frac{T_c}{T_h}$ can be used to calculate the Carnot (maximum theoretical) efficiency. Those temperatures must first be converted to kelvins.

Solution

The hot and cold reservoir temperatures are given as 300^oC and 27^oC, respectively. In kelvins, then, $T_h = 573\,K$ and $T_c = 300\,K$, so that the maximum efficiency is

$$Eff_c = 1 - \frac{T_c}{T_h}. \tag{5.2.1.3}$$

Thus,

$$Eff_c = 1 - \frac{300\,K}{573\,K} \tag{5.2.1.4}$$

$$= 0.476,\ or\ 47.6\%. \tag{5.2.1.5}$$

Discussion

A typical nuclear power station's actual efficiency is about 35%, a little better than 0.7 times the maximum possible value, a tribute to superior engineering. Electrical power stations fired by coal, oil, and natural gas have greater actual efficiencies (about 42%), because their boilers can reach higher temperatures and pressures. The cold reservoir temperature in any of these power stations is limited by the local environment. Figure 5.2.1.4 shows (a) the exterior of a nuclear power station and (b) the exterior of a coal-fired power station. Both have cooling towers into which water from the condenser enters the tower near the top and is sprayed downward, cooled by evaporation.

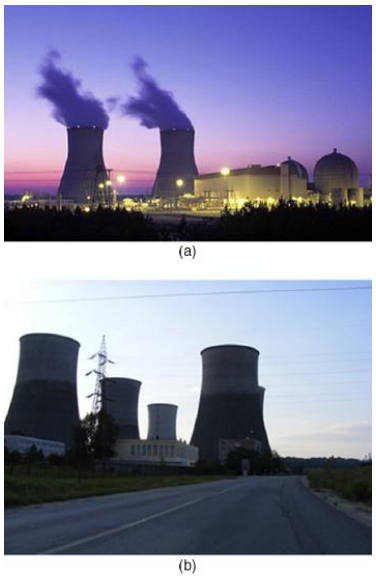


Figure 5.2.1.4: (a) A nuclear power station (credit: BlatantWorld.com) and (b) a coal-fired power station. Both have cooling towers in which water evaporates into the environment, representing Q_c. The nuclear reactor, which supplies Q_h, is housed inside the dome-shaped containment buildings. (credit: Robert & Mihaela Vicol, publicphoto.org)

Since all real processes are irreversible, the actual efficiency of a heat engine can never be as great as that of a Carnot engine, as illustrated in Figure 5.2.1.5(a). Even with the best heat engine possible, there are always dissipative processes in peripheral equipment, such as electrical transformers or car transmissions. These further reduce the overall efficiency by converting some of the engine's work output back into heat transfer, as shown in Figure 5.2.1.5(b).

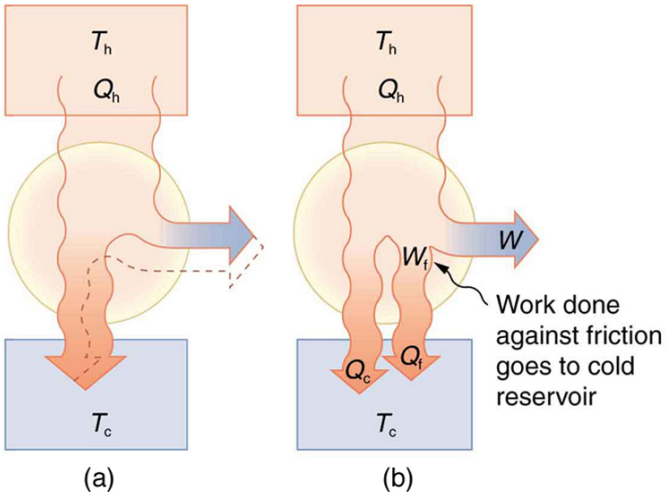


Figure 5.2.1.5: Real heat engines are less efficient than Carnot engines. (a) Real engines use irreversible processes, reducing the heat transfer to work. Solid lines represent the actual process; the dashed lines are what a Carnot engine would do between the same two reservoirs. (b) Friction and other dissipative processes in the output mechanisms of a heat engine convert some of its work output into heat transfer to the environment.

Summary

- The Carnot cycle is a theoretical cycle that is the most efficient cyclical process possible. Any engine using the Carnot cycle, which uses only reversible processes (adiabatic and isothermal), is known as a Carnot engine.
- Any engine that uses the Carnot cycle enjoys the maximum theoretical efficiency.
- While Carnot engines are ideal engines, in reality, no engine achieves Carnot's theoretical maximum efficiency, since dissipative processes, such as friction, play a role. Carnot cycles without heat loss may be possible at absolute zero, but this has never been seen in nature.

Glossary

Carnot cycle

a cyclical process that uses only reversible processes, the adiabatic and isothermal processes

Carnot engine

a heat engine that uses a Carnot cycle

Carnot efficiency

the maximum theoretical efficiency for a heat engine

This page titled 5.2.1: Carnot's Perfect Heat Engine- The Second Law of Thermodynamics Restated is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **15.4: Carnot's Perfect Heat Engine- The Second Law of Thermodynamics Restated** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/college-physics.

5.2.2: Applications of Thermodynamics- Heat Pumps and Refrigerators

Learning Objectives

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

- Describe the use of heat engines in heat pumps and refrigerators.
- Demonstrate how a heat pump works to warm an interior space.
- Explain the differences between heat pumps and refrigerators.
- Calculate a heat pump's coefficient of performance.



Figure 5.2.2.1: Almost every home contains a refrigerator. Most people don't realize they are also sharing their homes with a heat pump. (credit: Id1337x, Wikimedia Commons)

Heat pumps, air conditioners, and refrigerators utilize heat transfer from cold to hot. They are heat engines run backward. We say backward, rather than reverse, because except for Carnot engines, all heat engines, though they can be run backward, cannot truly be reversed. Heat transfer occurs from a cold reservoir Q_c and into a hot one. This requires work input W, which is also converted to heat transfer. Thus the heat transfer to the hot reservoir is $Q_h = Q_c + W$. (Note that Q_h, Q_c and W are positive, with their directions indicated on schematics rather than by sign.) A heat pump's mission is for heat transfer $Q - h$ to occur into a warm environment, such as a home in the winter. The mission of air conditioners and refrigerators is for heat transfer Q_c to occur from a cool environment, such as chilling a room or keeping food at lower temperatures than the environment. (Actually, a heat pump can be used both to heat and cool a space. It is essentially an air conditioner and a heating unit all in one. In this section we will concentrate on its heating mode.)

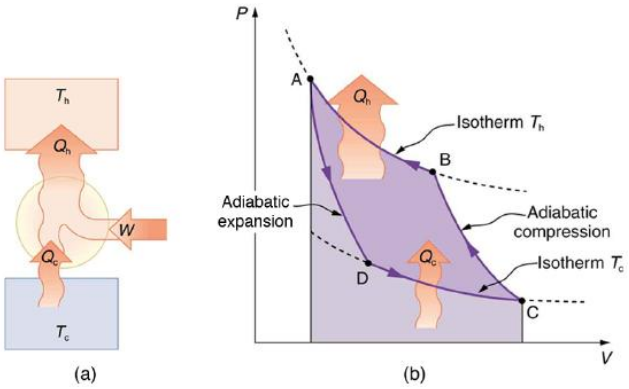


Figure 5.2.2.3 but reversed, following path ADCBA. The area inside the loop is negative, meaning there is a net work input. There is heat transfer Q_c into the system from a cold reservoir along path DC, and heat transfer Q_h out of the system into a hot reservoir along path BA.

Heat Pumps

The great advantage of using a heat pump to keep your home warm, rather than just burning fuel, is that a heat pump supplies $Q_h = Q_c + W$. Heat transfer is from the outside air, even at a temperature below freezing, to the indoor space. You only pay for W, and you get an additional heat transfer of Q_c from the outside at no cost; in many cases, at least twice as much energy is transferred to the heated space as is used to run the heat pump. When you burn fuel to keep warm, you pay for all of it. The disadvantage is that the work input (required by the second law of thermodynamics) is sometimes more expensive than simply burning fuel, especially if the work is done by electrical energy.

The basic components of a heat pump in its heating mode are shown in Figure 5.2.2.3. A working fluid such as a non-CFC refrigerant is used. In the outdoor coils (the evaporator), heat transfer Q_c occurs to the working fluid from the cold outdoor air, turning it into a gas.

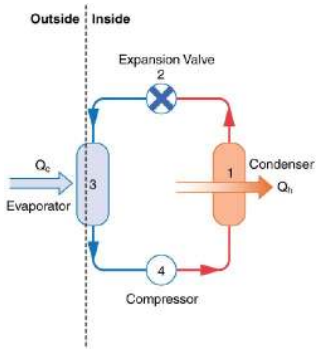


Figure 5.2.2.3: A simple heat pump has four basic components: (1) condenser, (2) expansion valve, (3) evaporator, and (4) compressor. In the heating mode, heat transfer Q_c occurs to the working fluid in the evaporator (3) from the colder outdoor air, turning it into a gas. The electrically driven compressor (4) increases the temperature and pressure of the gas and forces it into the condenser coils (1) inside the heated space. Because the temperature of the gas is higher than the temperature in the room, heat transfer from the gas to the room occurs as the gas condenses to a liquid. The working fluid is then cooled as it flows back through an expansion valve (2) to the outdoor evaporator coils.

The electrically driven compressor (work input W) raises the temperature and pressure of the gas and forces it into the condenser coils that are inside the heated space. Because the temperature of the gas is higher than the temperature inside the room, heat transfer to the room occurs and the gas condenses to a liquid. The liquid then flows back through a pressure-reducing valve to the outdoor evaporator coils, being cooled through expansion. (In a cooling cycle, the evaporator and condenser coils exchange roles and the flow direction of the fluid is reversed.)

The quality of a heat pump is judged by how much heat transfer Q_h occurs into the warm space compared with how much work input W is required. In the spirit of taking the ratio of what you get to what you spend, we define a heat pump's coefficient of performance (COP_{hp}) to be

$$COP_{hp} = \frac{Q_h}{W}. \tag{5.2.2.1}$$

Since the efficiency of a heat engine is $Eff = W/Q_h$, wee see that $COP_{hp} = 1/Eff$, an important and interesting fact. First, since the efficiency of any heat engine is less than 1, it means that COP_{hp} is always greater than 1—that is, a heat pump always has more heat transfer Q_h than work put into it. Second, it means that heat pumps work best when temperature differences are small. The efficiency of a perfect, or Carnot, engine is $Eff_c = 1 - (T_c?T_h)$ thus, the smaller the temperature difference, the smaller the efficiency and the greater the COP_{hp} (because $COP_{hp} = 1/Eff$). In other words, heat pumps do not work as well in very cold climates as they do in more moderate climates.

Friction and other irreversible processes reduce heat engine efficiency, but they do *not* benefit the operation of a heat pump—instead, they reduce the work input by converting part of it to heat transfer back into the cold reservoir before it gets into the heat pump.

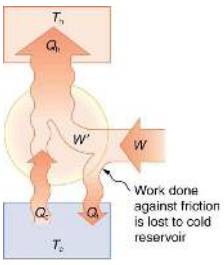


Figure 5.2.2.4: When a real heat engine is run backward, some of the intended work input W goes into heat transfer before it gets into the heat engine, thereby reducing its coefficient of performance COP_{hp}. In this figure, W' represents the portion of W that goes into the heat pump, while the remainder of W is lost in the form of frictional heat Q_f to the cold reservoir. If all of W had gone into the heat pump, then Q_h would have been greater. The best heat pump uses adiabatic and isothermal processes, since, in theory, there would be no dissipative processes to reduce the heat transfer to the hot reservoir.

Example 5.2.2.1: The Best COP_{hp} of a Heat Pump for Home Use

A heat pump used to warm a home must employ a cycle that produces a working fluid at temperatures greater than typical indoor temperature so that heat transfer to the inside can take place. Similarly, it must produce a working fluid at temperatures that are colder than the outdoor temperature so that heat transfer occurs from outside. Its hot and cold reservoir temperatures therefore cannot be too close, placing a limit on its $COPhp$. (See Figure 5.2.2.5.) What is the best coefficient of performance possible for such a heat pump, if it has a hot reservoir temperature of $45.0^{o}C$ and a cold reservoir temperature of $-15.0^{o}C$?

Strategy

A Carnot engine reversed will give the best possible performance as a heat pump. As noted above, $COP_{hp} = 1/Eff$, so that we need to first calculate the Carnot efficiency to solve this problem.

Solution

Carnot efficiency in terms of absolute temperature is given by**:**

$$Eff_c = 1 - \frac{T_c}{T_h}. \tag{5.2.2.2}$$

The temperatures in kelvins are $T_h = 318\,K$ and $T_c = 258\,K$, so that

$$Eff_c = 1 - \frac{258}{318}K = 0.1887. \tag{5.2.2.3}$$

Thus, from the discussion above,

$$COP_{hp} = \frac{1}{0.1887} = 5.30, \tag{5.2.2.4}$$

or

$$COP_{hp} = \frac{Q_h}{W} = 5.30, \tag{5.2.2.5}$$

so that

$$Q_h = 5.30\,W. \tag{5.2.2.6}$$

Discussion

This result means that the heat transfer by the heat pump is 5.30 times as much as the work put into it. It would cost 5.30 times as much for the same heat transfer by an electric room heater as it does for that produced by this heat pump. This is not a violation of conservation of energy. Cold ambient air provides 4.3 J per 1 J of work from the electrical outlet.

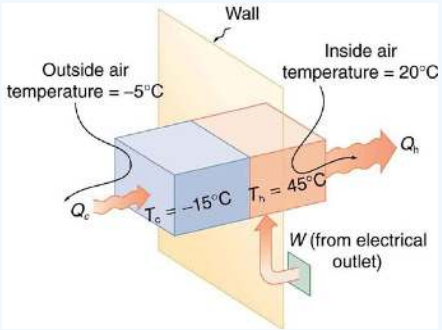


Figure 5.2.2.5: Heat transfer from the outside to the inside, along with work done to run the pump, takes place in the heat pump of the example above. Note that the cold temperature produced by the heat pump is lower than the outside temperature, so that heat transfer into the working fluid occurs. The pump's compressor produces a temperature greater than the indoor temperature in order for heat transfer into the house to occur.

Real heat pumps do not perform quite as well as the ideal one in the previous example; their values of COP_{hp} range from about 2 to 4. This range means that the heat transfer Q_h from the heat pumps is 2 to 4 times as great as the work W put into them. Their economical feasibility is still limited, however, since W is usually supplied by electrical energy that costs more per joule than heat transfer by burning fuels like natural gas. Furthermore, the initial cost of a heat pump is greater than that of many furnaces, so that

a heat pump must last longer for its cost to be recovered. Heat pumps are most likely to be economically superior where winter temperatures are mild, electricity is relatively cheap, and other fuels are relatively expensive. Also, since they can cool as well as heat a space, they have advantages where cooling in summer months is also desired. Thus some of the best locations for heat pumps are in warm summer climates with cool winters. Figure 5.2.2.6 shows a heat pump, called a "*reverse cycle*" or "*split-system cooler*" in some countries.



Figure 5.2.2.6: In hot weather, heat transfer occurs from air inside the room to air outside, cooling the room. In cool weather, heat transfer occurs from air outside to air inside, warming the room. This switching is achieved by reversing the direction of flow of the working fluid.

Air Conditioners and Refrigerators

Air conditioners and refrigerators are designed to cool something down in a warm environment. As with heat pumps, work input is required for heat transfer from cold to hot, and this is expensive. The quality of air conditioners and refrigerators is judged by how much heat transfer Q_c occurs from a cold environment compared with how much work input W is required. What is considered the benefit in a heat pump is considered waste heat in a refrigerator. We thus define the **coefficient of performance** COP_{ref} of an air conditioner or refrigerator to be

$$COP_{ref} = \frac{Q_c}{W}. \tag{5.2.2.7}$$

Noting again that $Q_h = Q_c + W$, we can see that an air conditioner will have a lower coefficient of performance than a heat pump, because $COP_{hp} = Q_h/W$ and Q_h is greater than Q_c. In this module's Problems and Exercises, you will show that

$$COP_{ref} = COP_{hp} - 1 \tag{5.2.2.8}$$

for a heat engine used as either an air conditioner or a heat pump operating between the same two temperatures. Real air conditioners and refrigerators typically do remarkably well, having values of COP_{ref} ranging from 2 to 6. These numbers are better than the COP_{hp} values for the heat pumps mentioned above, because the temperature differences are smaller, but they are less than those for Carnot engines operating between the same two temperatures.

A type of COP rating system called the "energy efficiency rating" (EER) has been developed. This rating is an example where non-SI units are still used and relevant to consumers. To make it easier for the consumer, Australia, Canada, New Zealand, and the U.S. use an Energy Star Rating out of 5 stars—the more stars, the more energy efficient the appliance. EERs are expressed in mixed units of British thermal units (Btu) per hour of heating or cooling divided by the power input in watts. Room air conditioners are readily available with EERs ranging from 6 to 12. Although not the same as the COP just described, these EERs are good for comparison purposes—the greater the EER, the cheaper an air conditioner is to operate (but the higher its purchase price is likely to be).

The EER of an air conditioner or refrigerator can be expressed as

$$EER = \frac{Q_c/t_1}{W/t_2}, \tag{5.2.2.9}$$

where Q_c is the amount of heat transfer from a cold environment in British thermal units, t_1 is time in hours, W is the work input in joules, and t_2 is time in seconds.

PROBLEM SOLVING STRATEGIES FOR THERMODYNAMICS

1. *Examine the situation to determine whether heat, work, or internal energy are involved.* Look for any system where the primary methods of transferring energy are heat and work. Heat engines, heat pumps, refrigerators, and air conditioners are examples of such systems.
2. *Identify the system of interest and draw a labeled diagram of the system showing energy flow.*
3. *Identify exactly what needs to be determined in the problem (identify the unknowns).* A written list is useful. Maximum efficiency means a Carnot engine is involved. Efficiency is not the same as the coefficient of performance.

4. *Make a list of what is given or can be inferred from the problem as stated (identify the knowns).* Be sure to distinguish heat transfer into a system from heat transfer out of the system, as well as work input from work output. In many situations, it is useful to determine the type of process, such as isothermal or adiabatic.
5. *Solve the appropriate equation for the quantity to be determined (the unknown).*
6. *Substitute the known quantities along with their units into the appropriate equation and obtain numerical solutions complete with units.*
7. *Check the answer to see if it is reasonable: Does it make sense?* For example, efficiency is always less than 1, whereas coefficients of performance are greater than 1.

Summary

- An artifact of the second law of thermodynamics is the ability to heat an interior space using a heat pump. Heat pumps compress cold ambient air and, in so doing, heat it to room temperature without violation of conservation principles.
- To calculate the heat pump's coefficient of performance, use the equation $COP_{hp} = \frac{Q_h}{W}$.
- A refrigerator is a heat pump; it takes warm ambient air and expands it to chill it.

Glossary

heat pump

a machine that generates heat transfer from cold to hot

coefficient of performance

for a heat pump, it is the ratio of heat transfer at the output (the hot reservoir) to the work supplied; for a refrigerator or air conditioner, it is the ratio of heat transfer from the cold reservoir to the work supplied

This page titled 5.2.2: Applications of Thermodynamics- Heat Pumps and Refrigerators is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **15.5: Applications of Thermodynamics- Heat Pumps and Refrigerators** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/college-physics.

5.2.3: Entropy and the Second Law of Thermodynamics- Disorder and the Unavailability of Energy

Learning Objectives

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

- Define entropy.
- Calculate the increase of entropy in a system with reversible and irreversible processes.
- Explain the expected fate of the universe in entropic terms.
- Calculate the increasing disorder of a system.



Figure 5.2.3.1: The ice in this drink is slowly melting. Eventually the liquid will reach thermal equilibrium, as predicted by the second law of thermodynamics. (credit: Jon Sullivan, PDPhoto.org)

There is yet another way of expressing the second law of thermodynamics. This version relates to a concept called entropy. By examining it, we shall see that the directions associated with the second law—heat transfer from hot to cold, for example—are related to the tendency in nature for systems to become disordered and for less energy to be available for use as work. The entropy of a system can in fact be shown to be a measure of its disorder and of the unavailability of energy to do work.

MAKING CONNECTIONS: ENTROPY, ENERGY, AND WORK

Recall that the simple definition of energy is the ability to do work. Entropy is a measure of how much energy is not available to do work. Although all forms of energy are interconvertible, and all can be used to do work, it is not always possible, even in principle, to convert the entire available energy into work. That unavailable energy is of interest in thermodynamics, because the field of thermodynamics arose from efforts to convert heat to work.

We can see how entropy is defined by recalling our discussion of the Carnot engine. We noted that for a Carnot cycle, and hence for any reversible processes, $Q_c/Q_h = T_c/T_h$. Rearranging terms yields

$$\frac{Q_c}{T_c} = \frac{Q_h}{T_h} \tag{5.2.3.1}$$

for any reversible process. Q_c and Q_h are absolute values of the heat transfer at temperatures T_c and T_h, respectively. This ratio of Q/T is defined to be the **change in entropy** ΔS for a reversible process,

$$\Delta S = \left(\frac{Q}{T}\right)_{rev}, \tag{5.2.3.2}$$

where Q is the heat transfer, which is positive for heat transfer into and negative for heat transfer out of, and T is the absolute temperature at which the reversible process takes place. The SI unit for entropy is joules per kelvin (J/K). If temperature changes during the process, then it is usually a good approximation (for small changes in temperature) to take T to be the average temperature, avoiding the need to use integral calculus to find ΔS.

The definition of ΔS is strictly valid only for reversible processes, such as used in a Carnot engine. However, we can find ΔS precisely even for real, irreversible processes. The reason is that the entropy S of a system, like internal energy U depends only on the state of the system and not how it reached that condition. Entropy is a property of state. Thus the change in entropy ΔS of a system between state 1 and state 2 is the same no matter how the change occurs. We just need to find or imagine a reversible

process that takes us from state 1 to state 2 and calculate ΔS for that process. That will be the change in entropy for any process going from state 1 to state 2. (See Figure 5.2.3.2.)

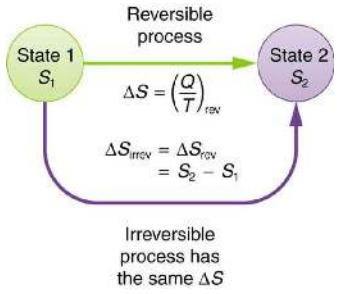


Figure 5.2.3.2: When a system goes from state 1 to state 2, its entropy changes by the same amount ΔS whether a hypothetical reversible path is followed or a real irreversible path is taken.

Now let us take a look at the change in entropy of a Carnot engine and its heat reservoirs for one full cycle. The hot reservoir has a loss of entropy $\Delta S_h = -Q_h/T_h$, because heat transfer occurs out of it (remember that when heat transfers out, then Q has a negative sign). The cold reservoir has a gain of entropy $\Delta S_c = Q_c/T_c$, because heat transfer occurs into it. (We assume the reservoirs are sufficiently large that their temperatures are constant.) So the total change in entropy is

$$\Delta S_{tot} = \Delta S_h + \Delta S_c. \tag{5.2.3.3}$$

Thus, since we know that $Q_h/T_h = Q_c/T_c$ for a Carnot engine,

$$\Delta S_{tot} = -\frac{Q_h}{T_h} + \frac{Q_c}{T_c} = 0. \tag{5.2.3.4}$$

This result, which has general validity, means that *the total change in entropy for a system in any reversible process is zero.*

The entropy of various parts of the system may change, but the total change is zero. Furthermore, the system does not affect the entropy of its surroundings, since heat transfer between them does not occur. Thus the reversible process changes neither the total entropy of the system nor the entropy of its surroundings. Sometimes this is stated as follows: *Reversible processes do not affect the total entropy of the universe.* Real processes are not reversible, though, and they do change total entropy. We can, however, use hypothetical reversible processes to determine the value of entropy in real, irreversible processes. The following example illustrates this point.

Example 5.2.3.1: Entropy Increases in an Irreversible (Real) Process

Spontaneous heat transfer from hot to cold is an irreversible process. Calculate the total change in entropy if 4000 J of heat transfer occurs from a hot reservoir at $T_h = 600\ K\ (327^oC)$ to a cold reservoir at $T_c = 250\ K\ (-23^oC)$, assuming there is no temperature change in either reservoir. (See Figure 5.2.3.3.)

Strategy

How can we calculate the change in entropy for an irreversible process when $\Delta S_{tot} = \Delta S_h + \Delta S_c$ is valid only for reversible processes? Remember that the total change in entropy of the hot and cold reservoirs will be the same whether a reversible or irreversible process is involved in heat transfer from hot to cold. So we can calculate the change in entropy of the hot reservoir for a hypothetical reversible process in which 4000 J of heat transfer occurs from it; then we do the same for a hypothetical reversible process in which 4000 J of heat transfer occurs to the cold reservoir. This produces the same changes in the hot and cold reservoirs that would occur if the heat transfer were allowed to occur irreversibly between them, and so it also produces the same changes in entropy.

Solution

We now calculate the two changes in entropy using $\Delta S_{tot} = \Delta S_h + \Delta S_c$. First, for the heat transfer from the hot reservoir,

$$\Delta S_h = \frac{-Q_h}{T_h} = \frac{-4000\ J}{600\ K} = -6.67\ J/K. \tag{5.2.3.5}$$

And for the cold reservoir,

$$\Delta S_c = \frac{Q_c}{T_c} = \frac{4000\ J}{250\ K} = 16.0\ J/K. \tag{5.2.3.6}$$

Thus the total is

$$\Delta S_{tot} = \Delta S_h + \Delta S_c \tag{5.2.3.7}$$

$$= (-6.67 + 16.0)\, J/K \tag{5.2.3.8}$$

$$= 9.33\, J/K. \tag{5.2.3.9}$$

Discussion

There is an *increase* in entropy for the system of two heat reservoirs undergoing this irreversible heat transfer. We will see that this means there is a loss of ability to do work with this transferred energy. Entropy has increased, and energy has become unavailable to do work.

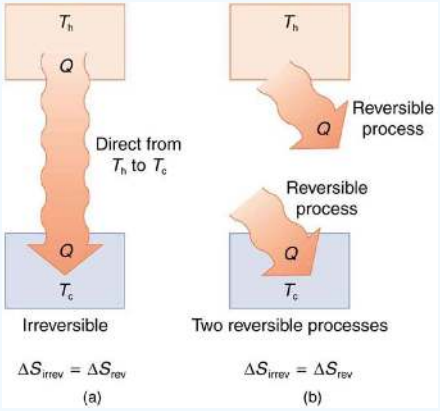


Figure 5.2.3.3: (a) Heat transfer from a hot object to a cold one is an irreversible process that produces an overall increase in entropy. (b) The same final state and, thus, the same change in entropy is achieved for the objects if reversible heat transfer processes occur between the two objects whose temperatures are the same as the temperatures of the corresponding objects in the irreversible process.

It is reasonable that entropy increases for heat transfer from hot to cold. Since the change in entropy is Q/T there is a larger change at lower temperatures. The decrease in entropy of the hot object is therefore less than the increase in entropy of the cold object, producing an overall increase, just as in the previous example. This result is very general:

There is an increase in entropy for any system undergoing an irreversible process.

With respect to entropy, there are only two possibilities: entropy is constant for a reversible process, and it increases for an irreversible process. There is a fourth version of **the second law of thermodynamics stated in terms of entropy:**

The total entropy of a system either increases or remains constant in any process; it never decreases.

For example, heat transfer cannot occur spontaneously from cold to hot, because entropy would decrease.

Entropy is very different from energy. Entropy is *not* conserved but increases in all real processes. Reversible processes (such as in Carnot engines) are the processes in which the most heat transfer to work takes place and are also the ones that keep entropy constant. Thus we are led to make a connection between entropy and the availability of energy to do work.

Entropy and the Unavailability of Energy to Do Work

What does a change in entropy mean, and why should we be interested in it? One reason is that entropy is directly related to the fact that not all heat transfer can be converted into work. The next example gives some indication of how an increase in entropy results in less heat transfer into work.

Example 5.2.3.2: Less Work is Produced by a Given Heat Transfer When Entropy Change is Greater

(a) Calculate the work output of a Carnot engine operating between temperatures of 600 K and 100 K for 4000 J of heat transfer to the engine. (b) Now suppose that the 4000 J of heat transfer occurs first from the 600 K reservoir to a 250 K reservoir (without doing any work, and this produces the increase in entropy calculated above) before transferring into a Carnot engine operating between 250 K and 100 K. What work output is produced? (See Figure 5.2.3.4.)

Strategy

In both parts, we must first calculate the Carnot efficiency and then the work output.

Solution (a)

The Carnot efficiency is given by

$$Eff_c = 1 - \frac{T_c}{T_h}. \tag{5.2.3.10}$$

Substituting the given temperatures yields

$$Eff_c = 1 - \frac{100\,K}{600\,K} = 0.833. \tag{5.2.3.11}$$

Now the work output can be calculated using the definition of efficiency for any heat engine as given by

$$Eff = \frac{W}{Q_h}. \tag{5.2.3.12}$$

Solving for W and substituting known terms gives

$$W = Eff_c Q_h \tag{5.2.3.13}$$

$$= (0.833)(4000\,J) = 3333\,J. \tag{5.2.3.14}$$

Solution (b)

Similarly,

$$Eff_c = 1 - \frac{T_c}{T_c'} = 1 - \frac{100\,K}{250\,K} = 0.600, \tag{5.2.3.15}$$

so that

$$W = Eff_c Q_h \tag{5.2.3.16}$$

$$= (0.600)(4000\,J) = 2400\,J \tag{5.2.3.17}$$

Discussion

There is 933 J less work from the same heat transfer in the second process. This result is important. The same heat transfer into two perfect engines produces different work outputs, because the entropy change differs in the two cases. In the second case, entropy is greater and less work is produced. Entropy is associated with the *un*availability of energy to do work.

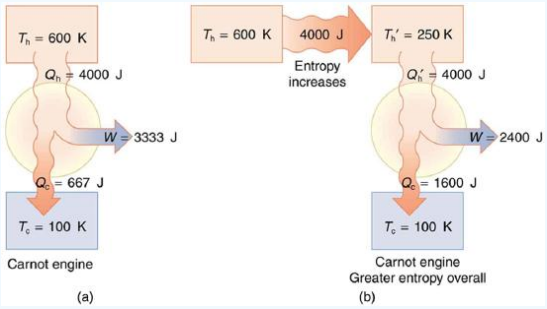


Figure 5.2.3.4: (a) A Carnot engine working at between 600 K and 100 K has 4000 J of heat transfer and performs 3333 J of work. (b) The 4000 J of heat transfer occurs first irreversibly to a 250 K reservoir and then goes into a Carnot engine. The increase in entropy caused by the heat transfer to a colder reservoir results in a smaller work output of 2400 J. There is a permanent loss of 933 J of energy for the purpose of doing work.

When entropy increases, a certain amount of energy becomes *permanently* unavailable to do work. The energy is not lost, but its character is changed, so that some of it can never be converted to doing work—that is, to an organized force acting through a distance. For instance, in the previous example, 933 J less work was done after an increase in entropy of 9.33 J/K occurred in the 4000 J heat transfer from the 600 K reservoir to the 250 K reservoir. It can be shown that the amount of energy that becomes unavailable for work is

$$W_{unavail} = \Delta S \cdot T_0, \tag{5.2.3.18}$$

where T_0 is the lowest temperature utilized. In the previous example,

$$W_{unavail} = (9.33\,J/K)(100\,K) = 933\,J \tag{5.2.3.19}$$

Heat Death of the Universe: An Overdose of Entropy

In the early, energetic universe, all matter and energy were easily interchangeable and identical in nature. Gravity played a vital role in the young universe. Although it may have *seemed* disorderly, and therefore, superficially entropic, in fact, there was enormous potential energy available to do work—all the future energy in the universe.

As the universe matured, temperature differences arose, which created more opportunity for work. Stars are hotter than planets, for example, which are warmer than icy asteroids, which are warmer still than the vacuum of the space between them.

Most of these are cooling down from their usually violent births, at which time they were provided with energy of their own—nuclear energy in the case of stars, volcanic energy on Earth and other planets, and so on. Without additional energy input, however, their days are numbered.

As entropy increases, less and less energy in the universe is available to do work. On Earth, we still have great stores of energy such as fossil and nuclear fuels; large-scale temperature differences, which can provide wind energy; geothermal energies due to differences in temperature in Earth's layers; and tidal energies owing to our abundance of liquid water. As these are used, a certain fraction of the energy they contain can never be converted into doing work. Eventually, all fuels will be exhausted, all temperatures will equalize, and it will be impossible for heat engines to function, or for work to be done.

Entropy increases in a closed system, such as the universe. But in parts of the universe, for instance, in the Solar system, it is not a locally closed system. Energy flows from the Sun to the planets, replenishing Earth's stores of energy. The Sun will continue to supply us with energy for about another five billion years. We will enjoy direct solar energy, as well as side effects of solar energy, such as wind power and biomass energy from photosynthetic plants. The energy from the Sun will keep our water at the liquid state, and the Moon's gravitational pull will continue to provide tidal energy. But Earth's geothermal energy will slowly run down and won't be replenished.

But in terms of the universe, and the very long-term, very large-scale picture, the entropy of the universe is increasing, and so the availability of energy to do work is constantly decreasing. Eventually, when all stars have died, all forms of potential energy have been utilized, and all temperatures have equalized (depending on the mass of the universe, either at a very high temperature following a universal contraction, or a very low one, just before all activity ceases) there will be no possibility of doing work.

Either way, the universe is destined for thermodynamic equilibrium—maximum entropy. This is often called the *heat death of the universe*, and will mean the end of all activity. However, whether the universe contracts and heats up, or continues to expand and cools down, the end is not near. Calculations of black holes suggest that entropy can easily continue for at least 10^{100} years.

Order to Disorder

Entropy is related not only to the unavailability of energy to do work—it is also a measure of disorder. This notion was initially postulated by Ludwig Boltzmann in the 1800s. For example, melting a block of ice means taking a highly structured and orderly system of water molecules and converting it into a disorderly liquid in which molecules have no fixed positions. (See Figure 5.2.3.5.) There is a large increase in entropy in the process, as seen in the following example.

Example 5.2.3.3: Entropy Associated with Disorder

Find the increase in entropy of 1.00 kg of ice originally at 0^oC, that is melted to form water at 0^oC.

Strategy

As before, the change in entropy can be calculated from the definition of ΔS once we find the energy Q needed to melt the ice.

Solution

The change in entropy is defined as:

$$\Delta S = \frac{Q}{T}. \tag{5.2.3.20}$$

Here Q is the heat transfer necessary to melt 1.00 kg of ice and is given by

$$Q = mL_f, \tag{5.2.3.21}$$

where m is the mass and L_f is the latent heat of fusion. $L_f = 334\,kJ/kg$ for water, so that

$$Q = 1.00\,kg)(334\,kJ/kg) = 3.34 \times 10^5\,J. \tag{5.2.3.22}$$

Now the change in entropy is positive, since heat transfer occurs into the ice to cause the phase change; thus,

$$\Delta S = \frac{Q}{T} = \frac{3.34 \times 10^5\,J}{T}. \tag{5.2.3.23}$$

T is the melting temperature of ice. That is $T = 0^oC = 273\,K$. So the change in entropy is

$$\Delta S = \frac{3.34 \times 10^5\,J}{273\,K} \tag{5.2.3.24}$$

$$= 1.22 \times 10^3\,J/K. \tag{5.2.3.25}$$

Discussion

This is a significant increase in entropy accompanying an increase in disorder.

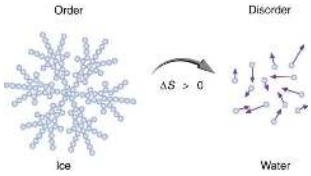


Figure 5.2.3.5: When ice melts, it becomes more disordered and less structured. The systematic arrangement of molecules in a crystal structure is replaced by a more random and less orderly movement of molecules without fixed locations or orientations. Its entropy increases because heat transfer occurs into it. Entropy is a measure of disorder.

In another easily imagined example, suppose we mix equal masses of water originally at two different temperatures, say 20.0^oC and 40.0^oC. The result is water at an intermediate temperature of 30.0^oC. Three outcomes have resulted: entropy has increased, some energy has become unavailable to do work, and the system has become less orderly. Let us think about each of these results.

First, entropy has increased for the same reason that it did in the example above. Mixing the two bodies of water has the same effect as heat transfer from the hot one and the same heat transfer into the cold one. The mixing decreases the entropy of the hot water but increases the entropy of the cold water by a greater amount, producing an overall increase in entropy.

Second, once the two masses of water are mixed, there is only one temperature—you cannot run a heat engine with them. The energy that could have been used to run a heat engine is now unavailable to do work.

Third, the mixture is less orderly, or to use another term, less structured. Rather than having two masses at different temperatures and with different distributions of molecular speeds, we now have a single mass with a uniform temperature.

These three results—entropy, unavailability of energy, and disorder—are not only related but are in fact essentially equivalent.

Life, Evolution, and the Second Law of Thermodynamics

Some people misunderstand the second law of thermodynamics, stated in terms of entropy, to say that the process of the evolution of life violates this law. Over time, complex organisms evolved from much simpler ancestors, representing a large decrease in entropy of the Earth's biosphere. It is a fact that living organisms have evolved to be highly structured, and much lower in entropy than the substances from which they grow. But it is *always* possible for the entropy of one part of the universe to decrease, provided the total change in entropy of the universe increases. In equation form, we can write this as

$$\Delta S_{tot} = \Delta S_{syst} + \Delta S_{envir} > 0. \tag{5.2.3.26}$$

Thus ΔS_{yst} can be negative as long as ΔS_{envir} is positive and greater in magnitude.

How is it possible for a system to decrease its entropy? Energy transfer is necessary. If I pick up marbles that are scattered about the room and put them into a cup, my work has decreased the entropy of that system. If I gather iron ore from the ground and convert it into steel and build a bridge, my work has decreased the entropy of that system. Energy coming from the Sun can decrease the entropy of local systems on Earth—that is, ΔS_{syst} is negative. But the overall entropy of the rest of the universe increases by a greater amount—that is, ΔS_{envir} is positive and greater in magnitude. Thus, $\Delta S_{tot} = \Delta S_{syst} + \Delta S_{envir} > 0$, and the second law of thermodynamics is *not* violated.

Every time a plant stores some solar energy in the form of chemical potential energy, or an updraft of warm air lifts a soaring bird, the Earth can be viewed as a heat engine operating between a hot reservoir supplied by the Sun and a cold reservoir supplied by dark outer space—a heat engine of high complexity, causing local decreases in entropy as it uses part of the heat transfer from the Sun into deep space. There is a large total increase in entropy resulting from this massive heat transfer. A small part of this heat transfer is stored in structured systems on Earth, producing much smaller local decreases in entropy. (See Figure 5.2.3.6.)

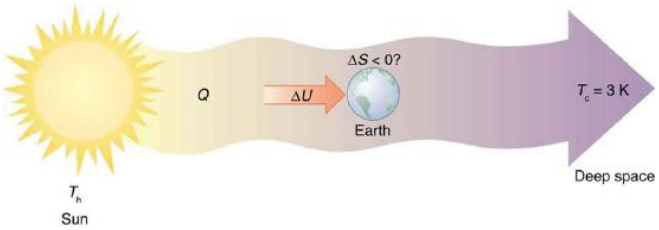


Figure 5.2.3.6: Earth's entropy may decrease in the process of intercepting a small part of the heat transfer from the Sun into deep space. Entropy for the entire process increases greatly while Earth becomes more structured with living systems and stored energy in various forms.

PHET EXPLORATIONS: REVERSIBLE REACTIONS

Watch a reaction proceed over time. How does total energy affect a reaction rate? Vary temperature, barrier height, and potential energies. Record concentrations and time in order to extract rate coefficients. Do temperature dependent studies to extract Arrhenius parameters. This simulation is best used with teacher guidance because it presents an analogy of chemical reactions.

PhET Interactive Simulation

5.2.3.7: *Reversible Reaction*

Summary

- Entropy is the loss of energy available to do work.
- Another form of the second law of thermodynamics states that the total entropy of a system either increases or remains constant; it never decreases.
- Entropy is zero in a reversible process; it increases in an irreversible process.
- The ultimate fate of the universe is likely to be thermodynamic equilibrium, where the universal temperature is constant and no energy is available to do work.
- Entropy is also associated with the tendency toward disorder in a closed system.

Glossary

entropy
a measurement of a system's disorder and its inability to do work in a system

change in entropy
the ratio of heat transfer to temperature Q/T

second law of thermodynamics stated in terms of entropy
the total entropy of a system either increases or remains constant; it never decrease

This page titled 5.2.3: Entropy and the Second Law of Thermodynamics- Disorder and the Unavailability of Energy is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **15.6: Entropy and the Second Law of Thermodynamics- Disorder and the Unavailability of Energy** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/college-physics.

5.3: Simple Machines

Learning Objectives

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

- Describe different simple machines.
- Calculate the mechanical advantage.

Simple machines are devices that can be used to multiply or augment a force that we apply – often at the expense of a distance through which we apply the force. The word for "machine" comes from the Greek word meaning "to help make things easier." Levers, gears, pulleys, wedges, and screws are some examples of machines. Energy is still conserved for these devices because a machine cannot do more work than the energy put into it. However, machines can reduce the input force that is needed to perform the job. The ratio of output to input force magnitudes for any simple machine is called its mechanical advantage (MA).

$$MA = \frac{F_o}{F_i} \tag{5.3.1}$$

One of the simplest machines is the lever, which is a rigid bar pivoted at a fixed place called the fulcrum. Torques are involved in levers, since there is rotation about a pivot point. Distances from the physical pivot of the lever are crucial, and we can obtain a useful expression for the MA in terms of these distances.

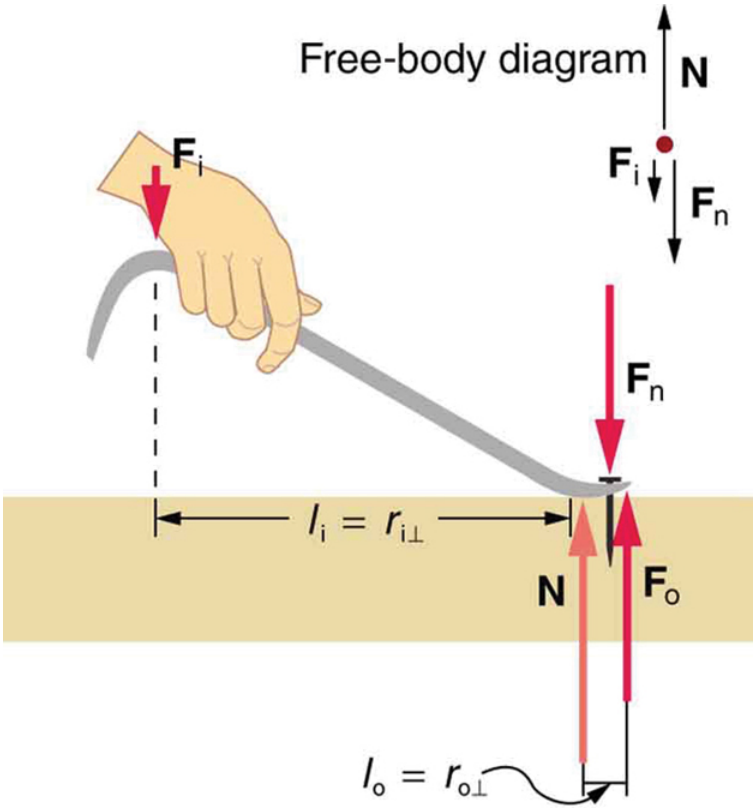


Figure 5.3.1: A nail puller is a lever with a large mechanical advantage. The external forces on the nail puller are represented by solid arrows. The force that the nail puller applies to the nail (F_o) is not a force on the nail puller. The reaction force the nail exerts back on the puller (F_n) is an external force and is equal and opposite to F_o. The perpendicular lever arms of the input and output forces are l_i and l_o.

Figure shows a lever type that is used as a nail puller. Crowbars, seesaws, and other such levers are all analogous to this one F_i is the input force and F_o is the output force. There are three vertical forces acting on the nail puller (the system of interest) – these are F_i, F_o, and N. F_n is the reaction force back on the system, equal and opposite to F_o. (note that F_o is not a force on the system.) N is the normal force upon the lever, and its torque is zero since it is exerted at the pivot. The torques due to F_i and F_n must be equal to each other if the nail is not moving, to satisfy the second condition for equilibrium $(net\ \tau = 0)$. (In order for the nail to actually move, the torque due to F_i must be ever-so-slightly greater than torque due to F_n.) Hence,

$$l_i F_i = l_o F_o \tag{5.3.2}$$

where l_i and l_o are the distances from where the input and output forces are applied to the pivot, as shown in the figure. Rearranging the last equation gives

$$\frac{F_o}{F_i} = \frac{l_i}{l_o}. \tag{5.3.3}$$

What interests us most here is that the magnitude of the force exerted by the nail puller,F_o, is much greater than the magnitude of the input force applied to the puller at the other end, F_i. For the nail puller,

$$MA = \frac{F_o}{F_i} = \frac{l_i}{l_o}. \tag{5.3.4}$$

This equation is true for levers in general. For the nail puller, the MA is certainly greater than one. The longer the handle on the nail puller, the greater the force you can exert with it.

Two other types of levers that differ slightly from the nail puller are a wheelbarrow and a shovel, shown in Figure. All these lever types are similar in that only three forces are involved – the input force, the output force, and the force on the pivot – and thus their MAs are given by $MA = \frac{F_o}{F_i}$ and $MA = \frac{d_1}{d_2}$, with distances being measured relative to the physical pivot. The wheelbarrow and shovel differ from the nail puller because both the input and output forces are on the same side of the pivot.

In the case of the wheelbarrow, the output force or load is between the pivot (the wheel's axle) and the input or applied force. In the case of the shovel, the input force is between the pivot (at the end of the handle) and the load, but the input lever arm is shorter than the output lever arm. In this case, the MA is less than one.

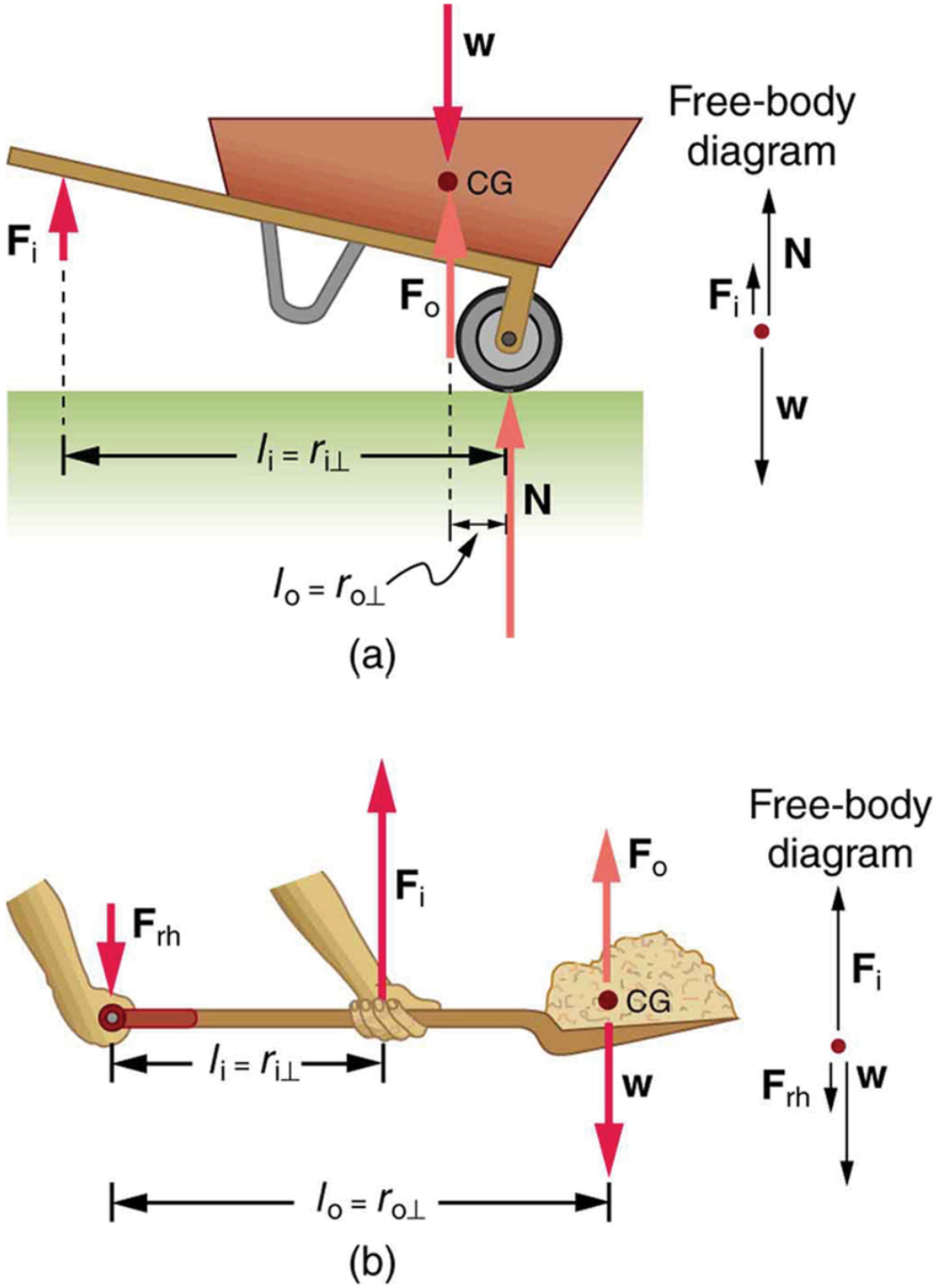


Figure 5.3.2: (a) In the case of the wheelbarrow, the output force or load is between the pivot and the input force. The pivot is the wheel's axle. Here, the output force is greater than the input force. Thus, a wheelbarrow enables you to lift much heavier loads than you could with your body alone. (b) In the case of the shovel, the input force is between the pivot and the load, but the input lever arm is shorter than the output lever arm. The pivot is at the handle held by the right hand. Here, the output force (supporting the shovel's load) is less than the input force (from the hand nearest the load), because the input is exerted closer to the pivot than is the output.

Example 5.3.1: What is the Advantage for the Wheelbarrow?

In the wheelbarrow of Figure, the load has a perpendicular lever arm of 7.50 cm, while the hands have a perpendicular lever arm of 1.02 m. (a) What upward force must you exert to support the wheelbarrow and its load if their combined mass is 45.0 kg? (b) What force does the wheelbarrow exert on the ground?

Strategy

Here, we use the concept of mechanical advantage.

Solution

(a) In this case, $\frac{F_o}{F_i} = \frac{l_i}{l_o}$ becomes

$$F_i = F_o \frac{l_o}{l_i}, \tag{5.3.5}$$

Adding values into this equation yields

$$F_i = (45.0\,kg)(9.80\,m/s^2)\frac{0.075\ m}{1.02\,m} = 32.4\,N. \tag{5.3.6}$$

The free-body diagram (see Figure) gives the following normal force:

$$F_i = +N = W. \tag{5.3.7}$$

Therefore,

$$N = (45.0\,kg)(9.80\,m/s^2) - 32.4\,N = 409\,N. \tag{5.3.8}$$

N is the normal force acting on the wheel; by Newton's third law, the force the wheel exerts on the ground is $409\,N$.

Discussion

An even longer handle would reduce the force needed to lift the load. The MA here is $MA = 1.01/0.0750 = 13.6$

Another very simple machine is the inclined plane. Pushing a cart up a plane is easier than lifting the same cart straight up to the top using a ladder, because the applied force is less. However, the work done in both cases (assuming the work done by friction is negligible) is the same. Inclined lanes or ramps were probably used during the construction of the Egyptian pyramids to move large blocks of stone to the top.

A crank is a lever that can be rotated 360^o about its pivot, as shown in Figure. Such a machine may not look like a lever, but the physics of its actions remain the same. The MA for a crank is simply the ratio of the radii r_i/r_o. Wheels and gears have this simple expression for their MAs too. The MA can be greater than 1, as it is for the crank, or less than 1, as it is for the simplified car axle driving the wheels, as shown. If the axle's radius is $2.0\,cm$ and the wheel's radius is $24.0\,cm$, then $MA = 2.0/24.0 = 0.083$ and the axle would have to exert a force of $12,000\,N$ on the wheel to enable it to exert a force of $1000\,N$ on the ground.

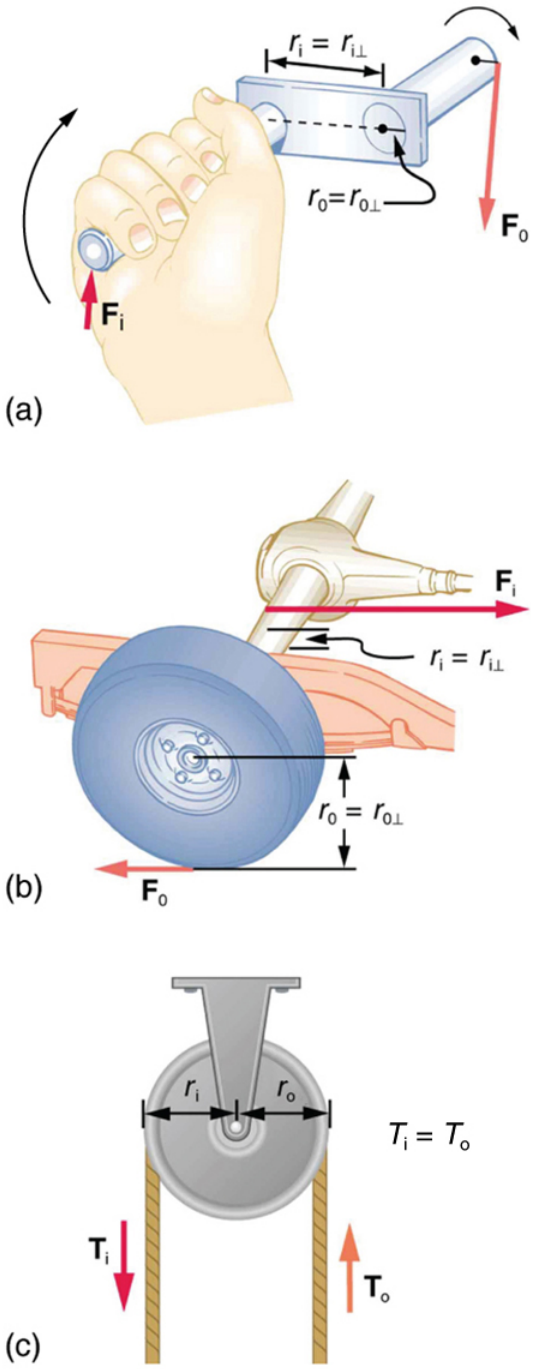


Figure 5.3.3: (a) A crank is a type of lever that can be rotated 360^o about its pivot. Cranks are usually designed to have a large MA. (b) A simplified automobile axle drives a wheel, which has a much larger diameter than the axle. The MA is less than 1. (c) An ordinary pulley is used to lift a heavy load. The pulley changes the direction of the force T exerted by the cord without changing its magnitude. Hence, this machine has an MA of 1.

An ordinary pulley has an MA of 1; it only changes the direction of the force and not its magnitude. Combinations of pulleys, such as those illustrated in Figure, are used to multiply force. If the pulleys are friction-free, then the force output is approximately an integral multiple of the tension in the cable. The number of cables pulling directly upward on the system of interest, as illustrated in the figures given below, is approximately the MA of the pulley system. Since each attachment applies an external force in approximately the same direction as the others, they add, producing a total force that is nearly an integral multiple of the input force T.

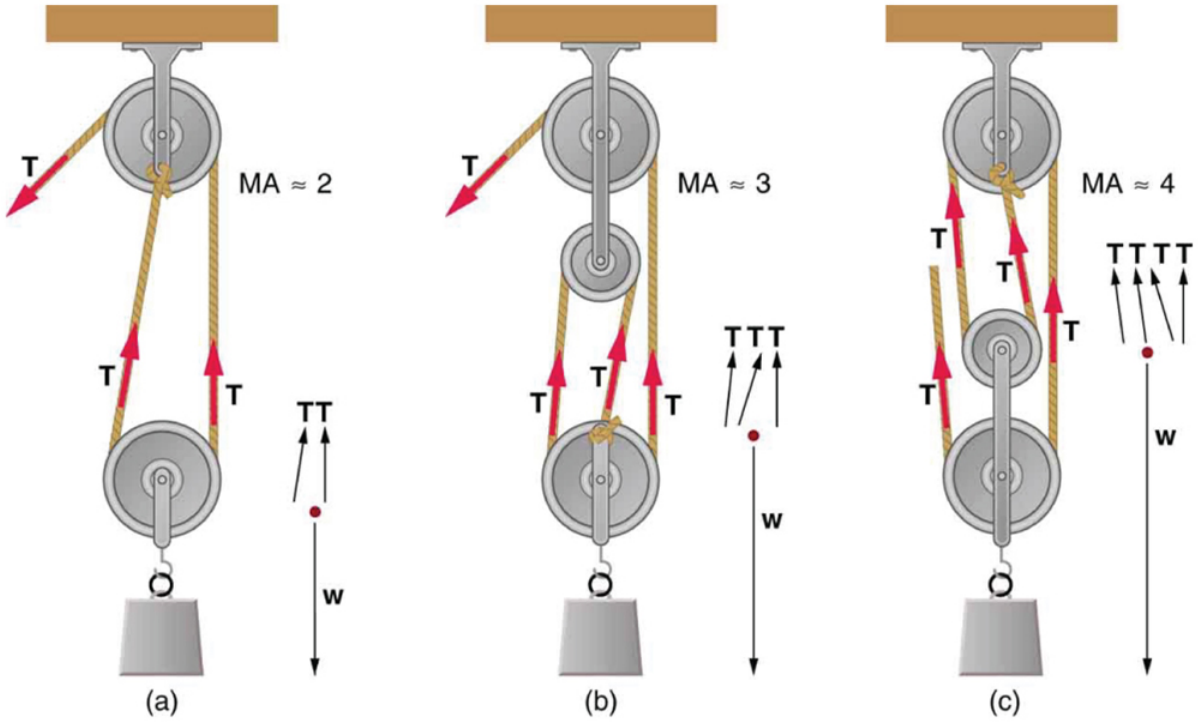


Figure 9.6.4. (a) The combination of pulleys is used to multiply force. The force is an integral multiple of tension if the pulleys are frictionless. This pulley system has two cables attached to its load, thus applying a force of approximately $2T$. This machine has $MA \approx 2$. (b) Three pulleys are used to lift a load in such a way that the mechanical advantage is about 3. Effectively, there are three cables attached to the load. (c) This pulley system applies a force of $4T$, so that it has $MA \approx 4$. Effectively, four cables are pulling on the system of interest.

Summary

- Simple machines are devices that can be used to multiply or augment a force that we apply – often at the expense of a distance through which we have to apply the force.
- The ratio of output to input forces for any simple machine is called its mechanical advantage
- A few simple machines are the lever, nail puller, wheelbarrow, crank, etc.

Glossary

mechanical advantage

the ratio of output to input forces for any simple machine

This page titled 5.3: Simple Machines is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **9.5: Simple Machines** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/college-physics.

Index

A

accuracy
1.1.1: Accuracy, Precision, and Significant Figures
adiabatic process
5.1.2: The First Law of Thermodynamics and Some Simple Processes
anion
1.2.5: Ions - Losing and Gaining Electrons
atomic mass unit
1.2.3: The Properties of Protons, Neutrons, and Electrons

B

background radiation
1.2.10: Natural Radioactivity and Half-Life

C

capacitance
3.3: Capacitors and Dielectrics
capacitor
3.3: Capacitors and Dielectrics
Carnot cycle
5.2.1: Carnot's Perfect Heat Engine- The Second Law of Thermodynamics Restated
Carnot efficiency
5.2.1: Carnot's Perfect Heat Engine- The Second Law of Thermodynamics Restated
Carnot engine
5.2.1: Carnot's Perfect Heat Engine- The Second Law of Thermodynamics Restated
cation
1.2.5: Ions - Losing and Gaining Electrons
change in entropy
5.2.3: Entropy and the Second Law of Thermodynamics- Disorder and the Unavailability of Energy
coefficient of performance
5.2.2: Applications of Thermodynamics- Heat Pumps and Refrigerators
compound
1.2.2: Classifying Matter According to Its Composition
Curie temperature
3.5: Ferromagnets and Electromagnets
current
3.2: Resistors in Series and Parallel
cyclical process
5.2: Introduction to the Second Law of Thermodynamics - Heat Engines and their Efficiency

D

dielectric
3.3: Capacitors and Dielectrics
dielectric strength
3.3: Capacitors and Dielectrics
domains
3.5: Ferromagnets and Electromagnets

E

electromagnet
3.5: Ferromagnets and Electromagnets
Electromagnetism
3.5: Ferromagnets and Electromagnets
electron
1.2.3: The Properties of Protons, Neutrons, and Electrons
element
1.2.2: Classifying Matter According to Its Composition
entropy
5.2.3: Entropy and the Second Law of Thermodynamics- Disorder and the Unavailability of Energy

F

farad (Units)
3.3: Capacitors and Dielectrics
ferromagnetic
3.5: Ferromagnets and Electromagnets
first law of thermodynamics
5.1.1: The First Law of Thermodynamics

G

Geiger counter
1.2.9: Detecting Radioactivity
gravitational potential energy
5.1.3: Gravitational Potential Energy

H

heat engine
5.1.2: The First Law of Thermodynamics and Some Simple Processes
heat pump
5.2.2: Applications of Thermodynamics- Heat Pumps and Refrigerators
human metabolism
5.1.1: The First Law of Thermodynamics

I

internal energy
5.1.1: The First Law of Thermodynamics
ions
1.2.5: Ions - Losing and Gaining Electrons
irreversible process
5.2: Introduction to the Second Law of Thermodynamics - Heat Engines and their Efficiency
isobaric process
5.1.2: The First Law of Thermodynamics and Some Simple Processes
isochoric process
5.1.2: The First Law of Thermodynamics and Some Simple Processes
isothermal process
5.1.2: The First Law of Thermodynamics and Some Simple Processes
isotopes
1.2.6: Isotopes - When the Number of Neutrons Varies

J

Joule's law
3.2: Resistors in Series and Parallel

M

machines
5.3: Simple Machines
magnetic monopoles
3.5: Ferromagnets and Electromagnets
magnetic pole
3.4: Magnets
magnetized
3.5: Ferromagnets and Electromagnets
matter
1.2.1: What is Matter?
mechanical advantage
5.3: Simple Machines
method of adding percents
1.1.1: Accuracy, Precision, and Significant Figures
mixture
1.2.2: Classifying Matter According to Its Composition
monopoles
3.5: Ferromagnets and Electromagnets

N

neutron
1.2.3: The Properties of Protons, Neutrons, and Electrons
north magnetic pole
3.4: Magnets

O

Ohm's law
3.2: Resistors in Series and Parallel
Otto cycle
5.2: Introduction to the Second Law of Thermodynamics - Heat Engines and their Efficiency

P

parallel
3.2: Resistors in Series and Parallel
Parallel Plate Capacitor
3.3: Capacitors and Dielectrics
Pascal's Principle
1.3.6: Pascal's Principle
percent uncertainty
1.1.1: Accuracy, Precision, and Significant Figures
polar molecule
3.3: Capacitors and Dielectrics
precision
1.1.1: Accuracy, Precision, and Significant Figures
proton
1.2.3: The Properties of Protons, Neutrons, and Electrons

R

resistance
3.2: Resistors in Series and Parallel
resistor
3.2: Resistors in Series and Parallel
Resistors in Parallel
3.2: Resistors in Series and Parallel
resistors in series
3.2: Resistors in Series and Parallel
reversible process
5.1.2: The First Law of Thermodynamics and Some Simple Processes

S

second law of thermodynamics
5.2: Introduction to the Second Law of Thermodynamics - Heat Engines and their Efficiency

second law of thermodynamics stated in terms of entropy
5.2.3: Entropy and the Second Law of Thermodynamics- Disorder and the Unavailability of Energy
series
3.2: Resistors in Series and Parallel
significant figures
1.1.1: Accuracy, Precision, and Significant Figures
south magnetic pole
3.4: Magnets
Stability of Isotopes
1.2.6: Isotopes - When the Number of Neutrons Varies
V
voltage drop
3.2: Resistors in Series and Parallel

Glossary

Sample Word 1 | Sample Definition 1

Detailed Licensing

Many of the pages in this book are re-mixes of open content from other books in the LibreTexts Library. Sometimes sections from several sources are used on the same page and there may be no unique open license. Where there is a conflict, the licensing may appear as *undeclared*. On any page, by toggling AutoAttibution on the Tools menu, you can display the detailed attribution on a paragraph level. Material from different books will be color coded and the source displayed by mousing over the paragraph.

Content in the Front and Back Matter, directories and similar can be uncopywriteable, being simple listings and these are also shown as *undeclared*.

Overview

Title: MLC - Physical Science

Webpages: 89

Applicable Restrictions: Noncommercial

All licenses found:

- CC BY 4.0: 67.4% (60 pages)
- Undeclared: 21.3% (19 pages)
- CK-12 License: 10.1% (9 pages)
- CC BY-NC-SA 3.0: 1.1% (1 page)

By Page

- MLC - Physical Science - *Undeclared*
  - Front Matter - *Undeclared*
    - TitlePage - *Undeclared*
    - InfoPage - *Undeclared*
    - Table of Contents - *Undeclared*
    - Licensing - *Undeclared*
  - 1: Matter and Measurement - *Undeclared*
    - 1.1: Measurement and Significant Figures - *Undeclared*
      - 1.1.1: Accuracy, Precision, and Significant Figures - *CC BY 4.0*
    - 1.2: Matter, Radioactive Decay, and Experimentation - *Undeclared*
      - 1.2.1: What is Matter? - *CK-12 License*
      - 1.2.2: Classifying Matter According to Its Composition - *CK-12 License*
      - 1.2.3: The Properties of Protons, Neutrons, and Electrons - *CK-12 License*
      - 1.2.4: Elements- Defined by Their Number of Protons - *CK-12 License*
      - 1.2.5: Ions - Losing and Gaining Electrons - *CC BY-NC-SA 3.0*
      - 1.2.6: Isotopes - When the Number of Neutrons Varies - *CK-12 License*
      - 1.2.7: The Discovery of Radioactivity - *CK-12 License*
      - 1.2.8: Types of Radioactivity- Alpha, Beta, and Gamma Decay - *CK-12 License*
      - 1.2.9: Detecting Radioactivity - *CK-12 License*
      - 1.2.10: Natural Radioactivity and Half-Life - *CK-12 License*
    - 1.3: Fluids - *Undeclared*
      - 1.3.1: What Is a Fluid? - *CC BY 4.0*
      - 1.3.2: Pressure - *CC BY 4.0*
      - 1.3.3: Pressure Due to the Weight of Fluid - *CC BY 4.0*
      - 1.3.4: Archimedes' Principle - *CC BY 4.0*
      - 1.3.5: Bernoulli's Equation - *CC BY 4.0*
      - 1.3.6: Pascal's Principle - *CC BY 4.0*
  - 2: Motion - *Undeclared*
    - 2.1: Introduction to One-Dimensional Kinematics - *CC BY 4.0*
      - 2.1.1: Displacement - *CC BY 4.0*
      - 2.1.2: Time, Velocity, and Speed - *CC BY 4.0*
      - 2.1.3: Acceleration - *CC BY 4.0*
    - 2.2: Motion Equations for Constant Acceleration in One Dimension - *CC BY 4.0*
    - 2.3: Falling Objects - *CC BY 4.0*
      - 2.3.1: Projectile Motion - *CC BY 4.0*
      - 2.3.E: Kinematics (Exercise) - *CC BY 4.0*
    - 2.4: Introduction to Dynamics- Newton's Laws of Motion - *CC BY 4.0*
      - 2.4.1: Development of Force Concept - *CC BY 4.0*
      - 2.4.2: Newton's First Law of Motion- Inertia - *CC BY 4.0*

- 2.4.3: Newton's Second Law of Motion- Force and Acceleration - *CC BY 4.0*
- 2.4.4: Newton's Third Law of Motion- Symmetry in Forces - *CC BY 4.0*
- 2.4.5: Friction - *CC BY 4.0*
- 2.4.E: Dynamics (Exercise) - *CC BY 4.0*

- 3: Electricity and Magnetism - *Undeclared*
  - 3.1: Static Electricity and Charge- Conservation of Charge - *CC BY 4.0*
    - 3.1.1: Ohm's Law- Resistance and Simple Circuits - *CC BY 4.0*
  - 3.2: Resistors in Series and Parallel - *CC BY 4.0*
  - 3.3: Capacitors and Dielectrics - *CC BY 4.0*
  - 3.4: Magnets - *CC BY 4.0*
  - 3.5: Ferromagnets and Electromagnets - *CC BY 4.0*
- 4: Waves - *Undeclared*
  - 4.1: Oscillations and Waves - *CC BY 4.0*
    - 4.1.1: Introduction to Oscillatory Motion and Waves - *CC BY 4.0*
    - 4.1.2: Period and Frequency in Oscillations - *CC BY 4.0*
    - 4.1.3: Simple Harmonic Motion- A Special Periodic Motion - *CC BY 4.0*
    - 4.1.4: Forced Oscillations and Resonance - *CC BY 4.0*
    - 4.1.5: Waves - *CC BY 4.0*
    - 4.1.6: Wave Interference- Standing Waves and Beats - *CC BY 4.0*
    - 4.1.7: Sound - *CC BY 4.0*
    - 4.1.8: Speed of Sound, Frequency, and Wavelength - *CC BY 4.0*
    - 4.1.9: Doppler Effect and Sonic Booms - *CC BY 4.0*
    - 4.1.E: Oscillations and Waves (Exercise) - *CC BY 4.0*
  - 4.2: Light - *CC BY 4.0*
    - 4.2.1: Introduction to Light - *CC BY 4.0*
    - 4.2.2: Maxwell's Equations- Electromagnetic Waves Predicted and Observed - *CC BY 4.0*
    - 4.2.3: Production and Properties of Electromagnetic Waves - *CC BY 4.0*
    - 4.2.4: The Electromagnetic Spectrum- an Overview - *CC BY 4.0*
    - 4.2.5: The Electromagnetic Spectrum- Application Notes - *CC BY 4.0*
    - 4.2.6: Reflection - *CC BY 4.0*
    - 4.2.7: Refraction - *CC BY 4.0*
    - 4.2.8: Dispersion- The Rainbow and Prisms - *CC BY 4.0*
    - 4.2.9: Image Formation by Mirrors - *CC BY 4.0*
    - 4.2.10: Polarization - *CC BY 4.0*
  - 4.3: Image Formation by Lenses - *CC BY 4.0*
- 5: Energy - *Undeclared*
  - 5.1: Prelude to Thermodynamics - *CC BY 4.0*
    - 5.1.1: The First Law of Thermodynamics - *CC BY 4.0*
    - 5.1.2: The First Law of Thermodynamics and Some Simple Processes - *CC BY 4.0*
    - 5.1.3: Gravitational Potential Energy - *CC BY 4.0*
  - 5.2: Introduction to the Second Law of Thermodynamics - Heat Engines and their Efficiency - *CC BY 4.0*
    - 5.2.1: Carnot's Perfect Heat Engine- The Second Law of Thermodynamics Restated - *CC BY 4.0*
    - 5.2.2: Applications of Thermodynamics- Heat Pumps and Refrigerators - *CC BY 4.0*
    - 5.2.3: Entropy and the Second Law of Thermodynamics- Disorder and the Unavailability of Energy - *CC BY 4.0*
  - 5.3: Simple Machines - *CC BY 4.0*
- Back Matter - *Undeclared*
  - Index - *Undeclared*
  - Glossary - *Undeclared*
  - Detailed Licensing - *Undeclared*
  - Detailed Licensing - *Undeclared*

Detailed Licensing

Overview

Title: MLC - Physical Science

Webpages: 89

Applicable Restrictions: Noncommercial

All licenses found:

- CC BY 4.0: 67.4% (60 pages)
- Undeclared: 21.3% (19 pages)
- CK-12 License: 10.1% (9 pages)
- CC BY-NC-SA 3.0: 1.1% (1 page)

By Page

- MLC - Physical Science - *Undeclared*
  - Front Matter - *Undeclared*
    - TitlePage - *Undeclared*
    - InfoPage - *Undeclared*
    - Table of Contents - *Undeclared*
    - Licensing - *Undeclared*
  - 1: Matter and Measurement - *Undeclared*
    - 1.1: Measurement and Significant Figures - *Undeclared*
      - 1.1.1: Accuracy, Precision, and Significant Figures - *CC BY 4.0*
    - 1.2: Matter, Radioactive Decay, and Experimentation - *Undeclared*
      - 1.2.1: What is Matter? - *CK-12 License*
      - 1.2.2: Classifying Matter According to Its Composition - *CK-12 License*
      - 1.2.3: The Properties of Protons, Neutrons, and Electrons - *CK-12 License*
      - 1.2.4: Elements- Defined by Their Number of Protons - *CK-12 License*
      - 1.2.5: Ions - Losing and Gaining Electrons - *CC BY-NC-SA 3.0*
      - 1.2.6: Isotopes - When the Number of Neutrons Varies - *CK-12 License*
      - 1.2.7: The Discovery of Radioactivity - *CK-12 License*
      - 1.2.8: Types of Radioactivity- Alpha, Beta, and Gamma Decay - *CK-12 License*
      - 1.2.9: Detecting Radioactivity - *CK-12 License*
      - 1.2.10: Natural Radioactivity and Half-Life - *CK-12 License*
    - 1.3: Fluids - *Undeclared*
      - 1.3.1: What Is a Fluid? - *CC BY 4.0*
      - 1.3.2: Pressure - *CC BY 4.0*
      - 1.3.3: Pressure Due to the Weight of Fluid - *CC BY 4.0*
      - 1.3.4: Archimedes' Principle - *CC BY 4.0*
      - 1.3.5: Bernoulli's Equation - *CC BY 4.0*
      - 1.3.6: Pascal's Principle - *CC BY 4.0*
  - 2: Motion - *Undeclared*
    - 2.1: Introduction to One-Dimensional Kinematics - *CC BY 4.0*
      - 2.1.1: Displacement - *CC BY 4.0*
      - 2.1.2: Time, Velocity, and Speed - *CC BY 4.0*
      - 2.1.3: Acceleration - *CC BY 4.0*
    - 2.2: Motion Equations for Constant Acceleration in One Dimension - *CC BY 4.0*
    - 2.3: Falling Objects - *CC BY 4.0*
      - 2.3.1: Projectile Motion - *CC BY 4.0*
      - 2.3.E: Kinematics (Exercise) - *CC BY 4.0*
    - 2.4: Introduction to Dynamics- Newton's Laws of Motion - *CC BY 4.0*
      - 2.4.1: Development of Force Concept - *CC BY 4.0*
      - 2.4.2: Newton's First Law of Motion- Inertia - *CC BY 4.0*
      - 2.4.3: Newton's Second Law of Motion- Force and Acceleration - *CC BY 4.0*
      - 2.4.4: Newton's Third Law of Motion- Symmetry in Forces - *CC BY 4.0*
      - 2.4.5: Friction - *CC BY 4.0*
      - 2.4.E: Dynamics (Exercise) - *CC BY 4.0*
  - 3: Electricity and Magnetism - *Undeclared*
    - 3.1: Static Electricity and Charge- Conservation of Charge - *CC BY 4.0*
      - 3.1.1: Ohm's Law- Resistance and Simple Circuits - *CC BY 4.0*
    - 3.2: Resistors in Series and Parallel - *CC BY 4.0*
    - 3.3: Capacitors and Dielectrics - *CC BY 4.0*

- 3.4: Magnets - *CC BY 4.0*
- 3.5: Ferromagnets and Electromagnets - *CC BY 4.0*

4: Waves - *Undeclared*

- 4.1: Oscillations and Waves - *CC BY 4.0*
  - 4.1.1: Introduction to Oscillatory Motion and Waves - *CC BY 4.0*
  - 4.1.2: Period and Frequency in Oscillations - *CC BY 4.0*
  - 4.1.3: Simple Harmonic Motion- A Special Periodic Motion - *CC BY 4.0*
  - 4.1.4: Forced Oscillations and Resonance - *CC BY 4.0*
  - 4.1.5: Waves - *CC BY 4.0*
  - 4.1.6: Wave Interference- Standing Waves and Beats - *CC BY 4.0*
  - 4.1.7: Sound - *CC BY 4.0*
  - 4.1.8: Speed of Sound, Frequency, and Wavelength - *CC BY 4.0*
  - 4.1.9: Doppler Effect and Sonic Booms - *CC BY 4.0*
  - 4.1.E: Oscillations and Waves (Exercise) - *CC BY 4.0*
- 4.2: Light - *CC BY 4.0*
  - 4.2.1: Introduction to Light - *CC BY 4.0*
  - 4.2.2: Maxwell's Equations- Electromagnetic Waves Predicted and Observed - *CC BY 4.0*
  - 4.2.3: Production and Properties of Electromagnetic Waves - *CC BY 4.0*
  - 4.2.4: The Electromagnetic Spectrum- an Overview - *CC BY 4.0*
  - 4.2.5: The Electromagnetic Spectrum- Application Notes - *CC BY 4.0*
  - 4.2.6: Reflection - *CC BY 4.0*
  - 4.2.7: Refraction - *CC BY 4.0*
  - 4.2.8: Dispersion- The Rainbow and Prisms - *CC BY 4.0*
  - 4.2.9: Image Formation by Mirrors - *CC BY 4.0*
  - 4.2.10: Polarization - *CC BY 4.0*
- 4.3: Image Formation by Lenses - *CC BY 4.0*

5: Energy - *Undeclared*

- 5.1: Prelude to Thermodynamics - *CC BY 4.0*
  - 5.1.1: The First Law of Thermodynamics - *CC BY 4.0*
  - 5.1.2: The First Law of Thermodynamics and Some Simple Processes - *CC BY 4.0*
  - 5.1.3: Gravitational Potential Energy - *CC BY 4.0*
- 5.2: Introduction to the Second Law of Thermodynamics - Heat Engines and their Efficiency - *CC BY 4.0*
  - 5.2.1: Carnot's Perfect Heat Engine- The Second Law of Thermodynamics Restated - *CC BY 4.0*
  - 5.2.2: Applications of Thermodynamics- Heat Pumps and Refrigerators - *CC BY 4.0*
  - 5.2.3: Entropy and the Second Law of Thermodynamics- Disorder and the Unavailability of Energy - *CC BY 4.0*
- 5.3: Simple Machines - *CC BY 4.0*

Back Matter - *Undeclared*

- Index - *Undeclared*
- Glossary - *Undeclared*
- Detailed Licensing - *Undeclared*
- Detailed Licensing - *Undeclared*