Physics II

Physics II

LUMEN LEARNING

The original version of the materials are provided by OpenStax College, and may be accessed at http://openstaxcollege.org/ and downloaded for free at http://cnx.org/content/col11406/latest/

Cover Image: Part of the Large Hadron Collider at CERN, on the border of Switzerland and France. The LHC is a particle accelerator, designed to study fundamental particles. (credit: Image Editor, Flickr)

Contents

Part VII. Electromagnetic Induction, AC Circuits, and Electrical Technologies, Part II

Part VIII. Electromagnetic Waves

Part IX. Geometric Optics

Part XII. Special Relativity

Part XIII. Introduction to Quantum Physics

Part XIV. Atomic Physics

Part XV. Radioactivity and Nuclear Physics

How to Succeed in Physics Guide

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

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

Instructor Resources

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

PART I
ELECTRIC CHARGE AND ELECTRIC FIELD

1. Introduction to Electric Charge and Electric Field

Figure 1. Static electricity from this plastic slide causes the child's hair to stand on end. The sliding motion stripped electrons away from the child's body, leaving an excess of positive charges, which repel each other along each strand of hair. (credit: Ken Bosma/Wikimedia Commons)

The image of American politician and scientist Benjamin Franklin (1706–1790) flying a kite in a thunderstorm is familiar to every schoolchild. (See Figure 2.) In this experiment, Franklin demonstrated a connection between lightning and *static electricity*. Sparks were drawn from a key hung on a kite string during an electrical storm. These sparks were like those produced by static electricity, such as the spark that jumps from your finger to a metal doorknob after you walk across a wool carpet. What Franklin demonstrated in his dangerous experiment was a connection between phenomena on two different scales: one the grand power of an electrical storm, the other an effect of more human proportions. Connections like this one reveal the underlying unity of the laws of nature, an aspect we humans find particularly appealing.

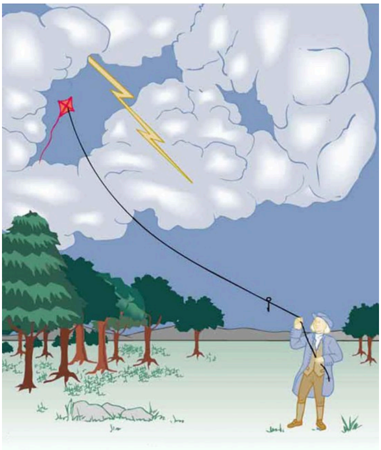


Figure 2. When Benjamin Franklin demonstrated that lightning was related to static electricity, he made a connection that is now part of the evidence that all directly experienced forces except the gravitational force are manifestations of the electromagnetic force.

Much has been written about Franklin. His experiments were only part of the life of a man who was a scientist, inventor, revolutionary, statesman, and writer. Franklin's experiments were not performed in isolation, nor were they the only ones to reveal connections.

For example, the Italian scientist Luigi Galvani (1737–1798) performed a series of experiments in which static electricity was used to stimulate contractions of leg muscles of dead frogs, an effect already known in humans subjected to static discharges. But Galvani also found that if he joined two metal wires (say copper and zinc) end to end and touched the other ends to muscles, he produced the same effect in frogs as static discharge. Alessandro

Volta (1745–1827), partly inspired by Galvani's work, experimented with various combinations of metals and developed the battery.

During the same era, other scientists made progress in discovering fundamental connections. The periodic table was developed as the systematic properties of the elements were discovered. This influenced the development and refinement of the concept of atoms as the basis of matter. Such submicroscopic descriptions of matter also help explain a great deal more.

Atomic and molecular interactions, such as the forces of friction, cohesion, and adhesion, are now known to be manifestations of the *electromagnetic force*. Static electricity is just one aspect of the electromagnetic force, which also includes moving electricity and magnetism.

All the macroscopic forces that we experience directly, such as the sensations of touch and the tension in a rope, are due to the electromagnetic force, one of the four fundamental forces in nature. The gravitational force, another fundamental force, is actually sensed through the electromagnetic interaction of molecules, such as between those in our feet and those on the top of a bathroom scale. (The other two fundamental forces, the strong nuclear force and the weak nuclear force, cannot be sensed on the human scale.)

This chapter begins the study of electromagnetic phenomena at a fundamental level. The next several chapters will cover static electricity, moving electricity, and magnetism—collectively known as electromagnetism. In this chapter, we begin with the study of electric phenomena due to charges that are at least temporarily stationary, called electrostatics, or static electricity.

2. Static Electricity and Charge: Conservation of Charge

Learning Objectives

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

- 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.

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 1). The very word *electric* derives from the Greek word for amber (*electron*).

Figure 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)

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 2 shows how these simple materials can be used to explore the nature of the force between charges.

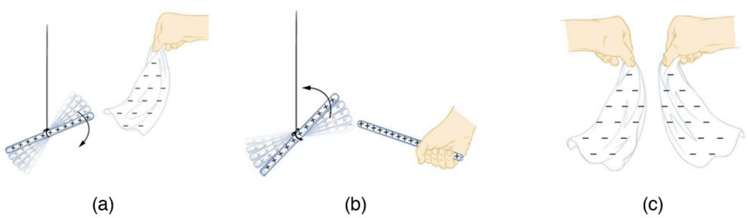


Figure 2. A glass rod becomes positively charged when rubbed with silk, while the silk becomes negatively charged. (a) The glass rod is attracted to the silk because their charges are opposite. (b) Two similarly charged glass rods repel. (c) Two similarly charged silk cloths repel.

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 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.

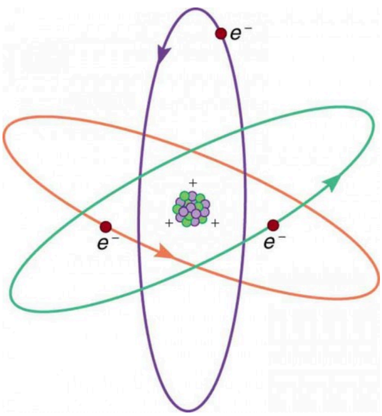


Figure 3. This simplified (and not to scale) view of an atom is called the planetary model of the atom. Negative electrons orbit a much heavier positive nucleus, as the planets orbit the much heavier sun. There the similarity ends, because forces in the atom are electromagnetic, whereas those in the 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 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 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.

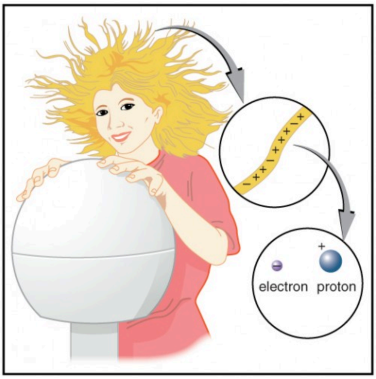


Figure 4. When this person touches a Van de Graaff generator, she receives an excess of positive charge, causing her hair to stand on end. The charges in one hair are shown. An artist's conception of an electron and a proton illustrate the particles carrying the negative and positive charges. We cannot 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 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.

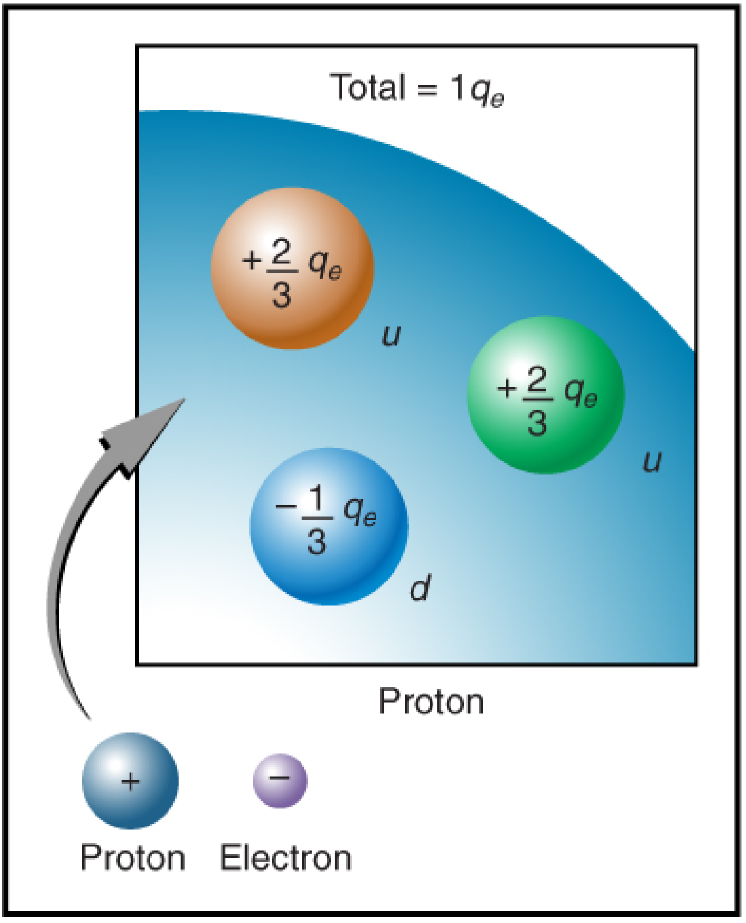


Figure 5. Artist's conception of fractional quark charges inside a proton. A group of three quark charges add up to the single positive charge on the proton: [latex]-\frac{1}{3}q_e+\frac{2}{3}q_e+\frac{2}{3}q_e=+1q_e\\[/latex].

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 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.

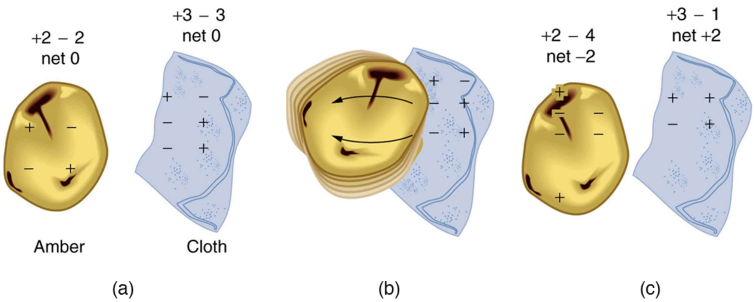


Figure 6. When materials are rubbed together, charges can be separated, particularly if one material has a greater affinity for electrons than another. (a) Both the amber and cloth are originally neutral, with equal positive and negative charges. Only a tiny fraction of the charges are involved, and 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 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.

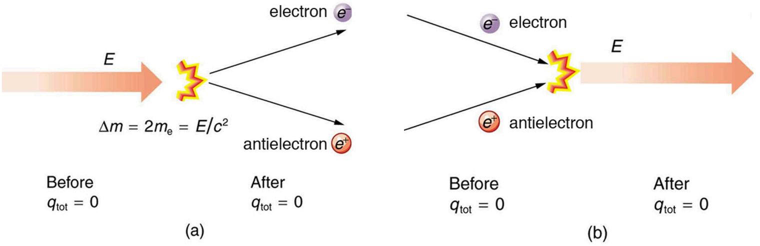


Figure 7. (a) When enough energy is present, it can be converted into matter. Here the matter created is an electron–antielectron pair. (me is the electron's mass.) The total charge before and after this event is zero. (b) When matter and antimatter collide, they annihilate each other; the total charge is conserved at zero before and after the annihilation.

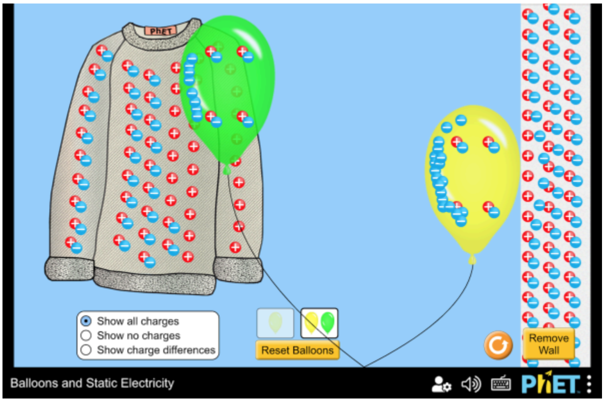
Making Connections: Conservation Laws

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.

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.

PhET Explorations: Balloons and Static Electricity

Why does a balloon stick to your sweater? Rub a balloon on a sweater, then let go of the balloon and it flies over and sticks to the sweater. View the charges in the sweater, balloons, and the wall.



Click to run the simulation.

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 $|q_e| = 1.60 \times 10^{-19}$ 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.

Conceptual Questions

1. There are very large numbers of charged particles in most objects. Why, then, don't most objects exhibit static electricity?
2. Why do most objects tend to contain nearly equal numbers of positive and negative charges?

Problems & Exercises

1. Common static electricity involves charges ranging from nanocoulombs to microcoulombs. (a) How many electrons are needed to form a charge of −2.00 nC (b) How many electrons must be removed from a neutral object to leave a net charge of 0.500 μC?
2. If 1.80×10^{20} electrons move through a pocket calculator during a full day's operation, how many coulombs of charge moved through it?
3. To start a car engine, the car battery moves 3.75×10^{21} electrons through the starter motor. How many coulombs of charge were moved?
4. A certain lightning bolt moves 40.0 C of charge. How many fundamental units of charge $|q_e|$ is this?

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

Selected Solutions to Problems & Exercises

1. (a) 1.25×10^{10}; (b) 3.13×10^{12}

3. −600 C

3. Conductors and Insulators

Learning Objectives

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

- Define conductor and insulator, explain the difference, and give examples of each.
- Describe three methods for charging an object.
- Explain what happens to an electric force as you move farther from the source.
- Define polarization.

Figure 1. This power adapter uses metal wires and connectors to conduct electricity from the wall socket to a laptop computer. The conducting wires allow electrons to move freely through the cables, which are shielded by rubber and plastic. These materials act as insulators that don't allow electric charge to escape outward. (credit: Evan-Amos, Wikimedia Commons)

Some substances, such as metals and salty water, allow charges to move through them with relative ease. Some of the electrons in metals and similar conductors are not bound to individual atoms or sites in the material. These *free electrons* can move through the material much as air moves through loose sand. Any substance that has free electrons and allows charge to move relatively freely through it is called a *conductor*. The moving electrons may collide with fixed atoms and molecules, losing some energy, but they can move in a conductor. Superconductors allow the movement of charge without any loss of energy. Salty water and other similar conducting materials contain free ions that can move through them. An ion is an atom or molecule having a positive or negative (nonzero) total charge. In other words, the total number of electrons is not equal to the total number of protons.

Other substances, such as glass, do not allow charges to move through them. These are called *insulators*. Electrons and ions in insulators are bound in the structure and cannot move easily—as much as 10^{23} times more slowly than in conductors. Pure water and dry table salt are insulators, for example, whereas molten salt and salty water are conductors.

Charging by Contact

Figure 2 shows an electroscope being charged by touching it with a positively charged glass rod. Because the glass rod is an insulator, it must actually touch the electroscope to transfer charge to or from it. (Note that the extra positive charges reside on the surface of the glass rod as a result of rubbing it with silk before starting the experiment.) Since only electrons move in metals, we see that they are attracted to the top of the electroscope. There, some are transferred to the positive rod by touch, leaving the electroscope with a net positive charge.

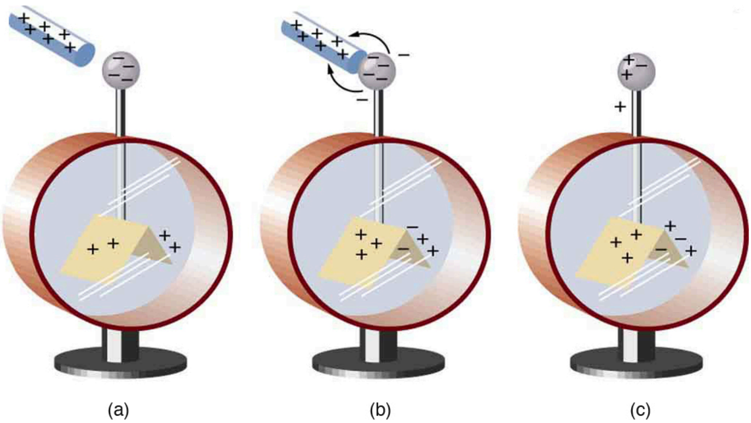


Figure 2. An electroscope is a favorite instrument in physics demonstrations and student laboratories. It is typically made with gold foil leaves hung from a (conducting) metal stem and is insulated from the room air in a glass-walled container. (a) A positively charged glass rod is brought near the tip of the electroscope, attracting electrons to the top and leaving a net positive charge on the leaves. Like charges in the light flexible gold leaves repel, separating them. (b) When the rod is touched against the ball, electrons are attracted and transferred, reducing the net charge on the glass rod but leaving the electroscope positively charged. (c) The excess charges are evenly distributed in the stem and leaves of the electroscope once the glass rod is removed.

Electrostatic repulsion in the leaves of the charged electroscope separates them. The electrostatic force has a horizontal component that results in the leaves moving apart as well as a vertical component that is balanced by the gravitational force. Similarly, the electroscope can be negatively charged by contact with a negatively charged object.

Charging by Induction

It is not necessary to transfer excess charge directly to an object in order to charge it. Figure 3 shows a method of *induction* wherein a charge is created in a nearby object, without direct contact. Here we see two neutral metal spheres in contact with one another but insulated from the rest of the world. A positively charged rod is brought near one of them, attracting negative charge to that side, leaving the other sphere positively charged.

This is an example of induced *polarization* of neutral objects. Polarization is the separation of charges in an object that remains neutral. If the spheres are now separated (before the rod is pulled away), each sphere will have a net charge. Note that the object closest to the charged rod receives an opposite charge when charged by induction. Note also that no charge is removed from the charged rod, so that this process can be repeated without depleting the supply of excess charge.

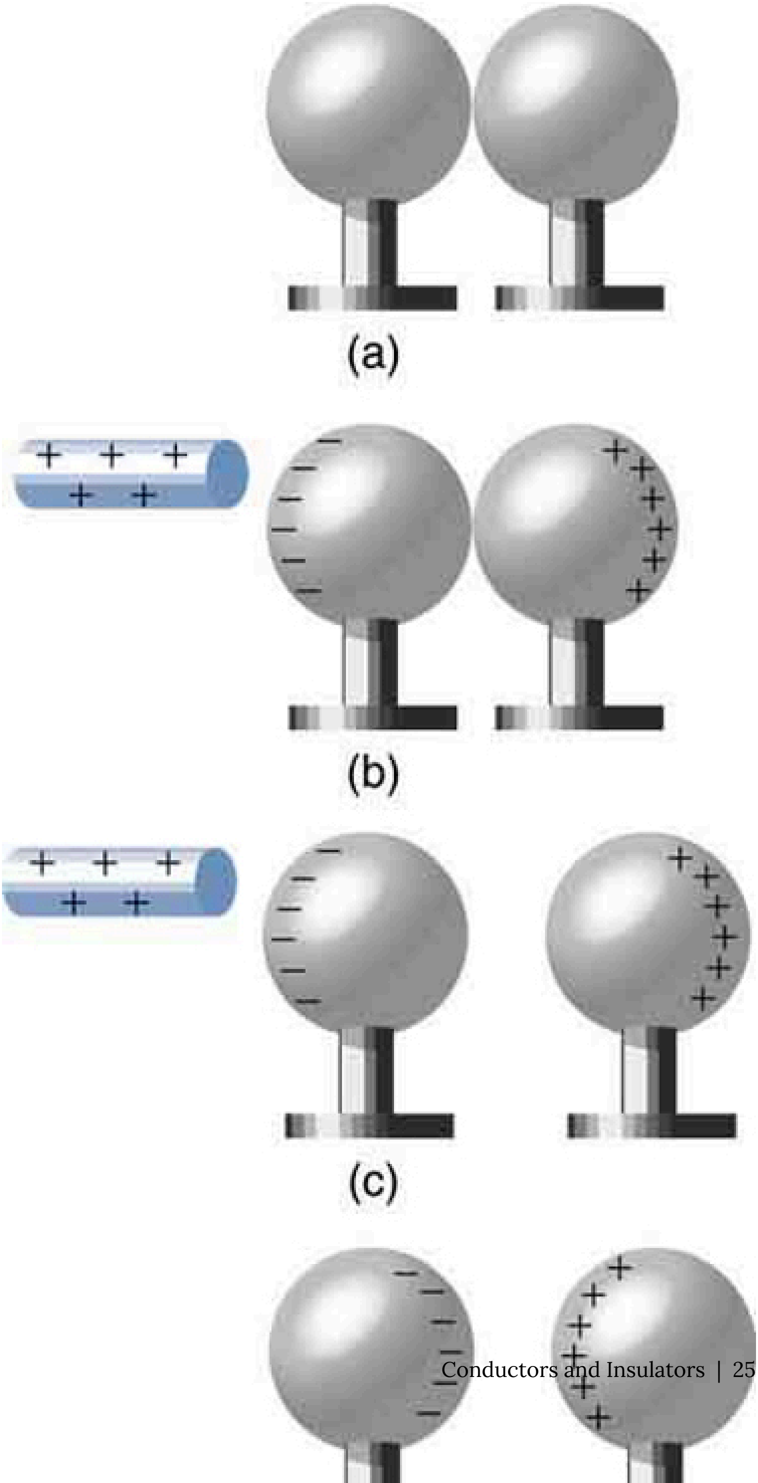

(a)
(b)
(c)

Figure 3. Charging by induction. (a) Two uncharged or neutral metal spheres are in contact with each other but insulated from the rest of the world. (b) A positively charged glass rod is brought near the sphere on the left, attracting negative charge and leaving the other sphere positively charged. (c) The spheres are separated before the rod is removed, thus separating negative and positive charge. (d) The spheres retain net charges after the inducing rod is removed—without ever having been touched by a charged object.

Another method of charging by induction is shown in Figure 4. The neutral metal sphere is polarized when a charged rod is brought near it. The sphere is then grounded, meaning that a conducting wire is run from the sphere to the ground. Since the earth is large and most ground is a good conductor, it can supply or accept excess charge easily. In this case, electrons are attracted to the sphere through a wire called the ground wire, because it supplies a conducting path to the ground. The ground connection is broken before the charged rod is removed, leaving the sphere with an excess charge opposite to that of the rod. Again, an opposite charge is achieved when charging by induction and the charged rod loses none of its excess charge.

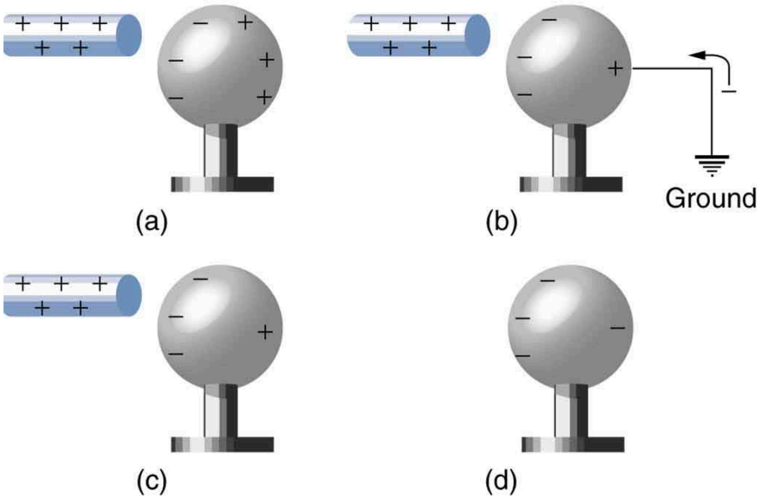


Figure 4. Charging by induction, using a ground connection. (a) A positively charged rod is brought near a neutral metal sphere, polarizing it. (b) The sphere is grounded, allowing electrons to be attracted from the earth's ample supply. (c) The ground connection is broken. (d) The positive rod is removed, leaving the sphere with an induced negative charge.

Neutral objects can be attracted to any charged object. The pieces of straw attracted to polished amber are neutral, for example. If you run a plastic comb through your hair, the charged comb can pick up neutral pieces of paper. Figure 5 shows how the polarization of atoms and molecules in neutral objects results in their attraction to a charged object.

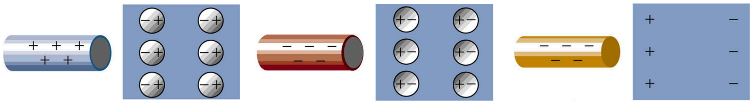


Figure 5. Both positive and negative objects attract a neutral object by polarizing its molecules. (a) A positive object brought near a neutral insulator polarizes its molecules. There is a slight shift in the distribution of the electrons orbiting the molecule, with unlike charges being brought nearer and like charges moved away. Since the electrostatic force decreases with distance, there is a net attraction. (b) A negative object produces the opposite polarization, but again attracts the neutral object. (c) The same effect occurs for a conductor; since the unlike charges are closer, there is a net attraction.

When a charged rod is brought near a neutral substance, an insulator in this case, the distribution of charge in atoms and molecules is shifted slightly. Opposite charge is attracted nearer the external charged rod, while like charge is repelled. Since the electrostatic force decreases with distance, the repulsion of like charges is weaker than the attraction of unlike charges, and so there is a net attraction. Thus a positively charged glass rod attracts neutral pieces of paper, as will a negatively charged rubber rod. Some molecules, like water, are polar molecules. Polar molecules have a natural or inherent separation of charge, although they are neutral overall. Polar molecules are particularly affected by other charged objects and show greater polarization effects than molecules with naturally uniform charge distributions.

Check Your Understanding

Can you explain the attraction of water to the charged rod in Figure 6?

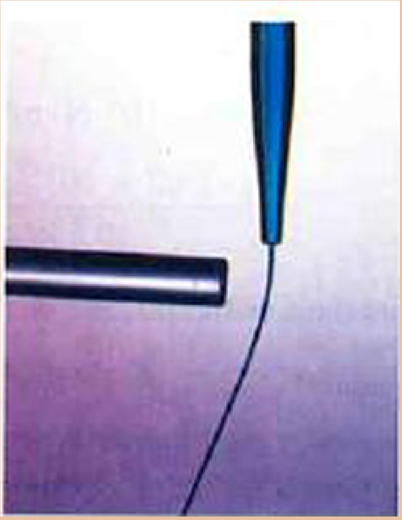


Figure 6.

Solution

Water molecules are polarized, giving them slightly positive and slightly negative sides. This makes water even more susceptible to a charged rod's attraction. As the water flows downward, due to the force of gravity, the charged conductor exerts a net attraction to the opposite charges in the stream of water, pulling it closer.

PhET Explorations: John Travoltage

Make sparks fly with John Travoltage. Wiggle Johnnie's foot and he picks up charges from the carpet. Bring his hand close to the door knob and get rid of the excess charge.

Click to run the simulation.

Section Summary

- Polarization is the separation of positive and negative charges in a neutral object.
- A conductor is a substance that allows charge to flow freely

through its atomic structure.

- An insulator holds charge within its atomic structure.
- Objects with like charges repel each other, while those with unlike charges attract each other.
- A conducting object is said to be grounded if it is connected to the Earth through a conductor. Grounding allows transfer of charge to and from the earth's large reservoir.
- Objects can be charged by contact with another charged object and obtain the same sign charge.
- If an object is temporarily grounded, it can be charged by induction, and obtains the opposite sign charge.
- Polarized objects have their positive and negative charges concentrated in different areas, giving them a non-symmetrical charge.
- Polar molecules have an inherent separation of charge.

Conceptual Questions

1. An eccentric inventor attempts to levitate by first placing a large negative charge on himself and then putting a large positive charge on the ceiling of his workshop. Instead, while attempting to place a large negative charge on himself, his clothes fly off. Explain.
2. If you have charged an electroscope by contact with a positively charged object, describe how you could use it to determine the charge of other objects. Specifically, what would the leaves of the electroscope do if other charged objects were brought near its knob?
3. When a glass rod is rubbed with silk, it becomes positive and the silk becomes negative—yet both

attract dust. Does the dust have a third type of charge that is attracted to both positive and negative? Explain.

4. Why does a car always attract dust right after it is polished? (Note that car wax and car tires are insulators.)
5. Describe how a positively charged object can be used to give another object a negative charge. What is the name of this process?
6. What is grounding? What effect does it have on a charged conductor? On a charged insulator?

Problems & Exercises

1. Suppose a speck of dust in an electrostatic precipitator has 1.0000×10^{12} protons in it and has a net charge of −5.00 nC (a very large charge for a small speck). How many electrons does it have?
2. An amoeba has 1.00×10^{16} protons and a net charge of 0.300 pC. (a) How many fewer electrons are there than protons? (b) If you paired them up, what fraction of the protons would have no electrons?
3. A 50.0 g ball of copper has a net charge of 2.00 μC. What fraction of the copper's electrons has been removed? (Each copper atom has 29 protons, and copper has an atomic mass of 63.5.)
4. What net charge would you place on a 100 g piece of sulfur if you put an extra electron on 1 in 10^{12} of its

atoms? (Sulfur has an atomic mass of 32.1.)

5. How many coulombs of positive charge are there in 4.00 kg of plutonium, given its atomic mass is 244 and that each plutonium atom has 94 protons?

Glossary

free electron: an electron that is free to move away from its atomic orbit

conductor: a material that allows electrons to move separately from their atomic orbits

insulator: a material that holds electrons securely within their atomic orbits

grounded: when a conductor is connected to the Earth, allowing charge to freely flow to and from Earth's unlimited reservoir

induction: the process by which an electrically charged object brought near a neutral object creates a charge in that object

polarization: slight shifting of positive and negative charges to opposite sides of an atom or molecule

electrostatic repulsion: the phenomenon of two objects with like charges repelling each other

Selected Solutions to Problems & Exercises

1. 1.03×10^{12}

3. 9.09×10^{-13}

5. 1.48×10^8 C

4. Coulomb's Law

Learning Objectives

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

- State Coulomb's law in terms of how the electrostatic force changes with the distance between two objects.
- Calculate the electrostatic force between two charged point forces, such as electrons or protons.
- Compare the electrostatic force to the gravitational attraction for a proton and an electron; for a human and the Earth.

Figure 1. This NASA image of Arp 87 shows the result of a strong gravitational attraction between two galaxies. In contrast, at the subatomic level, the electrostatic attraction between two objects, such as an electron and a proton, is far greater than their mutual attraction due to gravity. (credit: NASA/HST)

Through the work of scientists in the late 18th century, the main features of the *electrostatic force*—the existence of two types of charge, the observation that like charges repel, unlike charges attract, and the decrease of force with distance—were eventually refined, and expressed as a mathematical formula. The mathematical formula for the electrostatic force is called *Coulomb's law* after the French physicist Charles Coulomb (1736–1806), who performed experiments and first proposed a formula to calculate it.

Coulomb's Law

$$F = k\frac{\mid q_1 q_2 \mid}{r^2}$$

Coulomb's law calculates the magnitude of the force F between two point charges, q_1 and q_2, separated by a distance r. In SI units, the constant k is equal to

$$k = 8.988 \times 10^9 \frac{\mathrm{N \cdot m^2}}{\mathrm{C^2}} \approx 8.99 \times 10^9 \frac{\mathrm{N \cdot m^2}}{\mathrm{C^2}}$$

The electrostatic force is a vector quantity and is expressed in units of newtons. The force is understood to be along the line joining the two charges. (See Figure 2).

Although the formula for Coulomb's law is simple, it was no mean task to prove it. The experiments Coulomb did, with the primitive equipment then available, were difficult. Modern experiments have verified Coulomb's law to great precision. For example, it has been shown that the force is inversely proportional to distance between two objects squared $\left(F \propto \frac{1}{r^2}\right)$ to an accuracy of 1 part in 10^{16}. No exceptions have ever been found, even at the small distances within the atom.

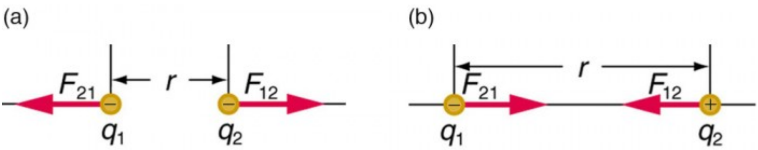


Figure 2. The magnitude of the electrostatic force F between point charges q_1 and q_2 separated by a distance r is given by Coulomb's law. Note that Newton's third law (every force exerted creates an equal and opposite force) applies as usual—the force on q_1 is equal in magnitude and opposite in direction to the force it exerts on q_2. (a) Like charges. (b) Unlike charges.

Example 1. How Strong is the Coulomb Force Relative to the Gravitational Force?

Compare the electrostatic force between an electron and proton separated by 0.530×10^{-10} m with the gravitational force between them. This distance is their average separation in a hydrogen atom.

Strategy

To compare the two forces, we first compute the electrostatic force using Coulomb's law, $F = k\frac{|q_1 q_2|}{r^2}$. We then calculate the gravitational force using Newton's universal law of gravitation. Finally, we take a ratio to see how the forces compare in magnitude.

Solution

Entering the given and known information about the charges and separation of the electron and proton into the expression of Coulomb's law yields

$$\begin{aligned} F &= k\frac{|q_1 q_2|}{r^2} \\ &= 8.99 \times 10^9 \frac{\mathrm{N{\cdot}m^2}}{\mathrm{C^2}} \times \frac{(1.60{\times}10^{-19}\ \mathrm{C})(1.60{\times}10^{-19}\ \mathrm{C})}{(0.530{\times}10^{-10}\ \mathrm{m})^2} \end{aligned}$$

Thus the Coulomb force is $F = 8.19 \times 10^{-8}$ N.

The charges are opposite in sign, so this is an attractive force. This is a very large force for an electron—it would cause an acceleration of $8.99 \times 10^{22}\ \mathrm{m/s^2}$ (verification is left as an end-of-section problem).The gravitational force is given by Newton's law of gravitation as:

$$F_G = G\frac{mM}{r^2},$$

where $G = 6.67 \times 10^{-11}\ \mathrm{N \cdot m^2/kg^2}$. Here m and M represent the electron and proton masses, which can be found in the appendices. Entering values for the knowns yields

$$F_G = \left(6.67 \times 10^{-11} \mathrm{N \cdot m^2/kg^2}\right) \times \frac{(9.11 \times 10^{-31}\ \mathrm{kg})\,(1.67 \times 10^{-27}\ \mathrm{kg})}{(0.530 \times 10^{-10}\ \mathrm{m})^2} = 3.61 \times 10^{-47}\ \mathrm{N}$$

This is also an attractive force, although it is traditionally shown as positive since gravitational force is always attractive. The ratio of the magnitude of the electrostatic force to gravitational force in this case is, thus,

$$\frac{F}{F_G} = 2.27 \times 10^{39}.$$

Discussion

This is a remarkably large ratio! Note that this will be the ratio of electrostatic force to gravitational force for an electron and a proton at any distance (taking the ratio before entering numerical values shows that the distance cancels). This ratio gives some indication of just how much larger the Coulomb force is than the gravitational force between two of the most common particles in nature.

As the example implies, gravitational force is completely negligible on a small scale, where the interactions of individual charged particles are important. On a large scale, such as between the Earth and a person, the reverse is true. Most objects are nearly electrically neutral, and so attractive and repulsive *Coulomb forces* nearly cancel. Gravitational force on a large scale dominates interactions between large objects because it is always attractive, while Coulomb forces tend to cancel.

Section Summary

- Frenchman Charles Coulomb was the first to publish the mathematical equation that describes the electrostatic force between two objects.
- Coulomb's law gives the magnitude of the force between point

charges. It is $F = k\frac{|\ q_1 q_2\ |}{r^2}$, where q_1 and q_2 are two point charges separated by a distance r, and $k \approx 8.99 \times 10^9 \frac{\mathrm{N \cdot m^2}}{\mathrm{C^2}}$

- This Coulomb force is extremely basic, since most charges are due to point-like particles. It is responsible for all electrostatic effects and underlies most macroscopic forces.
- The Coulomb force is extraordinarily strong compared with the gravitational force, another basic force—but unlike gravitational force it can cancel, since it can be either attractive or repulsive.
- The electrostatic force between two subatomic particles is far greater than the gravitational force between the same two particles.

Conceptual Questions

Use Figure 3 as a reference in the following questions. Figure 3 shows a schematic representation of the outer electron cloud of a neutral water molecule. The electrons spend more time near the oxygen than the hydrogens, giving a permanent charge separation as shown. Water is thus a *polar molecule*. It is more easily affected by electrostatic forces than molecules with uniform charge distributions.

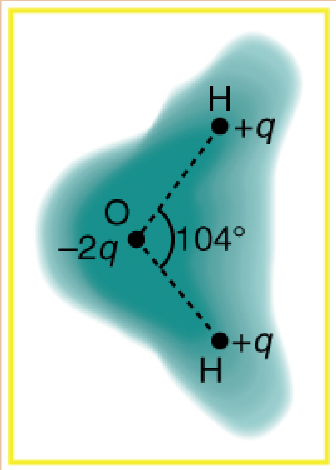


Figure 3. Schematic representation of the outer electron cloud of a neutral water molecule.

1. Figure 3 shows the charge distribution in a water molecule, which is called a polar molecule because it has an inherent separation of charge. Given water's polar character, explain what effect humidity has on removing excess charge from objects.
2. Using Figure 3, explain, in terms of Coulomb's law, why a polar molecule (such as in Figure 3) is attracted by both positive and negative charges.
3. Given the polar character of water molecules, explain how ions in the air form nucleation centers for rain droplets.

Problems & Exercises

1. What is the repulsive force between two pith balls that are 8.00 cm apart and have equal charges of −30.0 nC?
2. (a) How strong is the attractive force between a glass rod with a 0.700 μC charge and a silk cloth with a −0.600 μC charge, which are 12.0 cm apart, using the approximation that they act like point charges? (b) Discuss how the answer to this problem might be affected if the charges are distributed over some area and do not act like point charges.
3. Two point charges exert a 5.00 N force on each other. What will the force become if the distance between them is increased by a factor of three?
4. Two point charges are brought closer together, increasing the force between them by a factor of 25. By what factor was their separation decreased?
5. How far apart must two point charges of 75.0 nC (typical of static electricity) be to have a force of 1.00 N between them?
6. If two equal charges each of 1 C each are separated in air by a distance of 1 km, what is the magnitude of the force acting between them? You will see that even at a distance as large as 1 km, the repulsive force is substantial because 1 C is a very significant amount of charge.
7. A test charge of +2 μC is placed halfway between a charge of +6 μC and another of +4 μC separated by 10 cm. (a) What is the magnitude of the force on the test charge? (b) What is the direction of this force (away

from or toward the +6 μC charge)?

8. Bare free charges do not remain stationary when close together. To illustrate this, calculate the acceleration of two isolated protons separated by 2.00 nm (a typical distance between gas atoms). Explicitly show how you follow the steps in the Problem-Solving Strategy for electrostatics.
9. (a) By what factor must you change the distance between two point charges to change the force between them by a factor of 10? (b) Explain how the distance can either increase or decrease by this factor and still cause a factor of 10 change in the force.
10. Suppose you have a total charge q_{tot} that you can split in any manner. Once split, the separation distance is fixed. How do you split the charge to achieve the greatest force?
11. (a) Common transparent tape becomes charged when pulled from a dispenser. If one piece is placed above another, the repulsive force can be great enough to support the top piece's weight. Assuming equal point charges (only an approximation), calculate the magnitude of the charge if electrostatic force is great enough to support the weight of a 10.0 mg piece of tape held 1.00 cm above another. (b) Discuss whether the magnitude of this charge is consistent with what is typical of static electricity.
12. (a) Find the ratio of the electrostatic to gravitational force between two electrons. (b) What is this ratio for two protons? (c) Why is the ratio different for electrons and protons?
13. At what distance is the electrostatic force between

two protons equal to the weight of one proton?

14. A certain five cent coin contains 5.00 g of nickel. What fraction of the nickel atoms' electrons, removed and placed 1.00 m above it, would support the weight of this coin? The atomic mass of nickel is 58.7, and each nickel atom contains 28 electrons and 28 protons.
15. (a) Two point charges totaling 8.00 μC exert a repulsive force of 0.150 N on one another when separated by 0.500 m. What is the charge on each? (b) What is the charge on each if the force is attractive?
16. Point charges of 5.00 μC and −3.00 μC are placed 0.250 m apart. (a) Where can a third charge be placed so that the net force on it is zero? (b) What if both charges are positive?
17. Two point charges q_1 and q_2 are 3.00 m apart, and their total charge is 20 μC. (a) If the force of repulsion between them is 0.075N, what are magnitudes of the two charges? (b) If one charge attracts the other with a force of 0.525N, what are the magnitudes of the two charges? Note that you may need to solve a quadratic equation to reach your answer.

Glossary

Coulomb's law: the mathematical equation calculating the electrostatic force vector between two charged particles

Coulomb force: another term for the electrostatic force

electrostatic force: the amount and direction of attraction or repulsion between two charged bodies

Selected Solutions to Problems & Exercises

2. (a) 0.263 N; (b) If the charges are distributed over some area, there will be a concentration of charge along the side closest to the oppositely charged object. This effect will increase the net force.

4. The separation decreased by a factor of 5.

8.
$$\begin{aligned} F &= k\frac{|q_1 q_2|}{r^2} = ma \Rightarrow a = \frac{kq^2}{mr^2} \\ &= \frac{\left(9.00\times10^{9}\ \mathrm{N\cdot m^2/C^2}\right)\left(1.60\times10^{-19}\ \mathrm{m}\right)^2}{\left(1.67\times10^{-27}\ \mathrm{kg}\right)\left(2.00\times10^{-9}\ \mathrm{m}\right)^2} \\ &= 3.45\times10^{16}\ \mathrm{m/s^2} \end{aligned}$$

9. (a) 3.2; (b) If the distance increases by 3.2, then the force will decrease by a factor of 10 ; if the distance decreases by 3.2, then the force will increase by a factor of 10. Either way, the force changes by a factor of 10.

11. (a) 1.04×10^{-9} C; (b) This charge is approximately 1 nC, which is consistent with the magnitude of charge typical for static electricity

14. 1.02×10^{-11}

16. (a) 0.859 m beyond negative charge on line connecting two charges; (b) 0.109 m from lesser charge on line connecting two charges

5. Electric Field: Concept of a Field Revisited

Learning Objectives

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

- Describe a force field and calculate the strength of an electric field due to a point charge.
- Calculate the force exerted on a test charge by an electric field.
- Explain the relationship between electrical force (**F**) on a test charge and electrical field strength (**E**).

Contact forces, such as between a baseball and a bat, are explained on the small scale by the interaction of the charges in atoms and molecules in close proximity. They interact through forces that include the *Coulomb force*. Action at a distance is a force between objects that are not close enough for their atoms to "touch." That is, they are separated by more than a few atomic diameters.

For example, a charged rubber comb attracts neutral bits of paper from a distance via the Coulomb force. It is very useful to think of an object being surrounded in space by a *force field*. The force field carries the force to another object (called a test object) some distance away.

Concept of a Field

A field is a way of conceptualizing and mapping the force that surrounds any object and acts on another object at a distance without apparent physical connection. For example, the gravitational field surrounding the earth (and all other masses) represents the gravitational force that would be experienced if another mass were placed at a given point within the field.

In the same way, the Coulomb force field surrounding any charge extends throughout space. Using Coulomb's law, $F = k\frac{\mid q_1 q_2 \mid}{r^2}$, its magnitude is given by the equation $F = k\frac{\mid qQ \mid}{r^2}$, for a *point charge* (a particle having a charge Q) acting on a *test charge q* at a distance *r* (see Figure 1). Both the magnitude and direction of the Coulomb force field depend on Q and the test charge *q*.

To simplify things, we would prefer to have a field that depends only on Q and not on the test charge q. The electric field is defined in such a manner that it represents only the charge creating it and is unique at every point in space. Specifically, the electric field E is defined to be the ratio of the Coulomb force to the test charge:

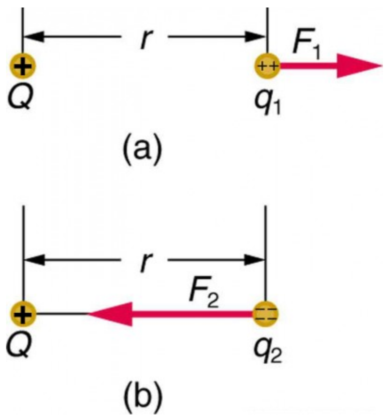


Figure 1. The Coulomb force field due to a positive charge Q is shown acting on two different charges. Both charges are the same distance from Q. (a) Since q_1 *is positive, the force* F_1 *acting on it is repulsive. (b) The charge* q_2 *is negative and greater in magnitude than* q_1*, and so the force* F_2 *acting on it is attractive and stronger than* F_1*. The Coulomb force field is thus not unique at any point in space, because it depends on the test charges* q_1 *and* q_2 *as well as the charge Q.*

$$\mathbf{E} = \frac{\mathbf{E}}{q},$$

where **F** is the electrostatic force (or Coulomb force) exerted on a positive test charge q. It is understood that E is in the same direction as **F**. It is also assumed that q is so small that it does not alter the charge distribution creating the electric field. The units of electric field are newtons per coulomb (N/C). If the electric field is known, then the electrostatic force on any charge q is simply obtained by multiplying charge times electric field, or **F** = q**E**. Consider the electric field due to a point charge Q. According to Coulomb's law, the force it exerts on a test charge q is $F = k\frac{|\,qQ\,|}{r^2}$. Thus the magnitude of the electric field, E, for a point charge is

$$E = \left|\frac{F}{q}\right| = k\left|\frac{qQ}{qr^2}\right| = k\frac{|Q|}{r^2}.$$

Since the test charge cancels, we see that

$$E = k\frac{|Q|}{r^2}.$$

The electric field is thus seen to depend only on the charge Q and the distance *r*; it is completely independent of the test charge *q*.

Example 1. Calculating the Electric Field of a Point Charge

Calculate the strength and direction of the electric field E due to a point charge of 2.00 nC (nano-Coulombs) at a distance of 5.00 mm from the charge.

Strategy

We can find the electric field created by a point charge by using the equation $E = \frac{kQ}{r^2}$.

Solution

Here Q = 2.00 × 10^{-9} C and *r* = 5.00 × 10^{-3} m. Entering those values into the above equation gives

$$\begin{aligned} E &= k\frac{Q}{r^2} \\ &= \left(8.99\times 10^9\,\frac{\text{N}\cdot\text{m}^2}{\text{C}^2}\right)\times\frac{(2.00\times 10^{-9}\text{ C})}{(5.00\times 10^{-3}\text{ m})^2} \\ &= 7.19\times 10^5\text{ N/C} \end{aligned}$$

Discussion

This *electric field strength* is the same at any point 5.00 mm away from the charge Q that creates the field. It is positive, meaning that it has a direction pointing away from the charge Q.

Example 2. Calculating the Force Exerted on a Point Charge by an Electric Field

What force does the electric field found in the previous example exert on a point charge of −0.250 μC?

Strategy

Since we know the electric field strength and the charge in the field, the force on that charge can be calculated using the definition of electric field $\mathbf{E} = \frac{\mathbf{F}}{q}$ rearranged to **F** = q**E**.

Solution

The magnitude of the force on a charge q = −0.250 μC exerted by a field of strength E = 7.20 × 10^5 N/C is thus,

$$\begin{aligned} F &= -qE \\ &= \left(0.250 \times 10^{-6}\ \text{C}\right)\left(7.20 \times 10^{5}\ \text{N/C}\right) \\ &= 0.180\ \text{N} \end{aligned}$$

Because q is negative, the force is directed opposite to the direction of the field.

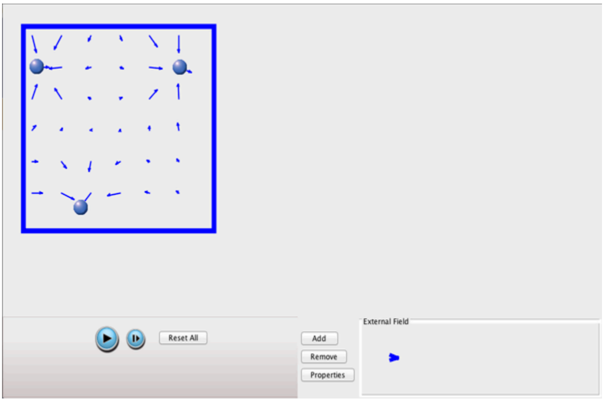
Discussion

The force is attractive, as expected for unlike charges. (The field was created by a positive charge and here acts on a negative charge.) The charges in this example are typical of common static electricity, and the modest attractive force obtained is similar to forces experienced in static cling and similar situations.

PhET Explorations: Electric Field of

Dreams

Play ball! Add charges to the Field of Dreams and see how they react to the electric field. Turn on a background electric field and adjust the direction and magnitude.



Click to download the simulation. Run using Java.

Section Summary

- The electrostatic force field surrounding a charged object extends out into space in all directions.
- The electrostatic force exerted by a point charge on a test

charge at a distance r depends on the charge of both charges, as well as the distance between the two.

- The electric field **E** is defined to be $\mathbf{E} = \frac{\mathbf{F}}{q}$, where **F** is the Coulomb or electrostatic force exerted on a small positive test charge q. **E** has units of N/C.
- The magnitude of the electric field **E** created by a point charge Q is $\mathrm{E} = k\frac{|Q|}{r^2}$, where r is the distance from Q. The electric field **E** is a vector and fields due to multiple charges add like vectors.

Conceptual Questions

1. Why must the test charge q in the definition of the electric field be vanishingly small?
2. Are the direction and magnitude of the Coulomb force unique at a given point in space? What about the electric field?

Problems & Exercises

1. What is the magnitude and direction of an electric field that exerts a 2.00×10^{-5} N upward force on a $-1.75\ \mu\text{C}$ charge?
2. What is the magnitude and direction of the force

exerted on a 3.50 μC charge by a 250 N/C electric field that points due east?

3. Calculate the magnitude of the electric field 2.00 m from a point charge of 5.00 mC (such as found on the terminal of a Van de Graaff).
4. (a) What magnitude point charge creates a 10,000 N/C electric field at a distance of 0.250 m? (b) How large is the field at 10.0 m?
5. Calculate the initial (from rest) acceleration of a proton in a 5.00×10^{6} N/C electric field (such as created by a research Van de Graaff). Explicitly show how you follow the steps in the Problem-Solving Strategy for electrostatics.
6. (a) Find the direction and magnitude of an electric field that exerts a 4.80×10^{-17} N westward force on an electron. (b) What magnitude and direction force does this field exert on a proton?

Glossary

field: a map of the amount and direction of a force acting on other objects, extending out into space

point charge: A charged particle, designated Q, generating an electric field

test charge: A particle (designated *q*) with either a positive or negative charge set down within an electric field generated by a point charge

Selected Solutions to Problems & Exercises

2. 8.75×10^{-4} N

4. (a) 6.94×10^{-8} C; (b) 6.25 N/C

6. (a) 300 N/C (east); (b) 4.80×10^{-17} N (east)

6. Video: Electric Fields

Watch the following Physics Concept Trailer to learn more about the electric fields created by lightning.

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://library.achievingthedream.org/austinccphysics2/?p=24#oembed-1

7. Electric Field Lines: Multiple Charges

Learning Objectives

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

- Calculate the total force (magnitude and direction) exerted on a test charge from more than one charge
- Describe an electric field diagram of a positive point charge; of a negative point charge with twice the magnitude of positive charge
- Draw the electric field lines between two points of the same charge; between two points of opposite charge.

Drawings using lines to represent *electric fields* around charged objects are very useful in visualizing field strength and direction. Since the electric field has both magnitude and direction, it is a vector. Like all *vectors*, the electric field can be represented by an arrow that has length proportional to its magnitude and that points in the correct direction. (We have used arrows extensively to represent force vectors, for example.)

Figure 1 shows two pictorial representations of the same electric field created by a positive point charge Q. Figure 1b shows the standard representation using continuous lines. Figure 1b shows numerous individual arrows with each arrow representing the force on a test charge *q*. Field lines are essentially a map of infinitesimal force vectors.

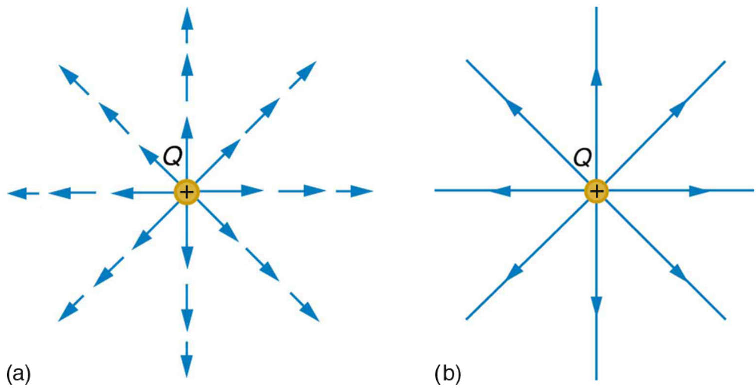


Figure 1. Two equivalent representations of the electric field due to a positive charge Q. (a) Arrows representing the electric field's magnitude and direction. (b) In the standard representation, the arrows are replaced by continuous field lines having the same direction at any point as the electric field. The closeness of the lines is directly related to the strength of the electric field. A test charge placed anywhere will feel a force in the direction of the field line; this force will have a strength proportional to the density of the lines (being greater near the charge, for example).

Note that the electric field is defined for a positive test charge q, so that the field lines point away from a positive charge and toward a negative charge. (See Figure 2.) The electric field strength is exactly proportional to the number of field lines per unit area, since the magnitude of the electric field for a point charge is $E = k\frac{|Q|}{r^2}$ and area is proportional to r^2. This pictorial representation, in which field lines represent the direction and their closeness (that is, their areal density or the number of lines crossing a unit area) represents strength, is used for all fields: electrostatic, gravitational, magnetic, and others.

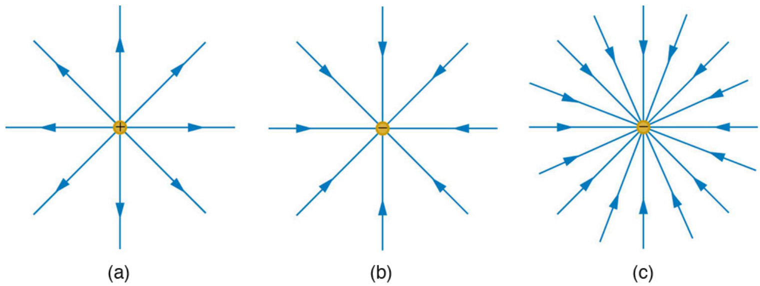


Figure 2. The electric field surrounding three different point charges. (a) A positive charge. (b) A negative charge of equal magnitude. (c) A larger negative charge.

In many situations, there are multiple charges. The total electric field created by multiple charges is the vector sum of the individual fields created by each charge. The following example shows how to add electric field vectors.

Example 1. Adding Electric Fields

Find the magnitude and direction of the total electric field due to the two point charges, q_1 and q_2, at the origin of the coordinate system as shown in Figure 3.

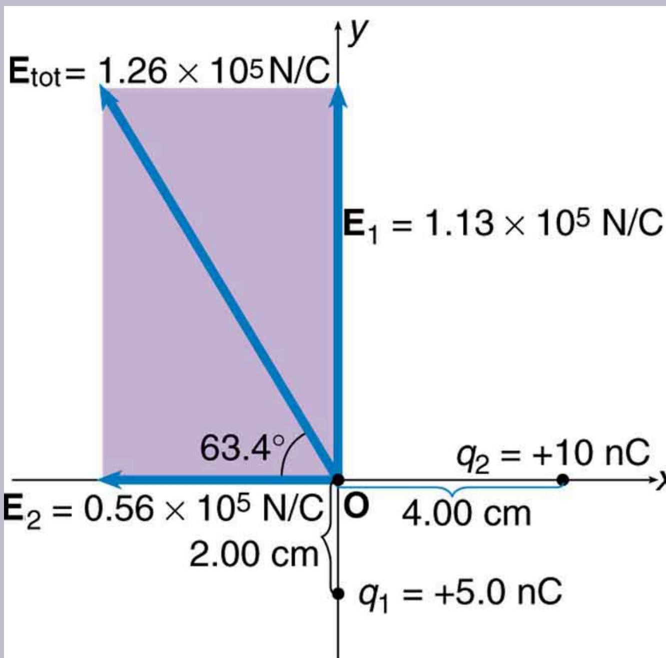


Figure 3. The electric fields $\mathbf{E}_1$ *and* $\mathbf{E}_2$ *at the origin O add to* $\mathbf{E}_{tot}$.

Strategy

Since the electric field is a vector (having magnitude and direction), we add electric fields with the same vector techniques used for other types of vectors. We first must find the electric field due to each charge at the point of interest, which is the origin of the coordinate system (O) in this instance. We pretend that there is a positive test charge, *q*, at point O, which allows us to determine the direction of the fields $\mathbf{E}_1$ and $\mathbf{E}_2$. Once those fields are found, the total field can be determined using *vector addition.*

Solution

The electric field strength at the origin due to q_1 is labeled E_1 and is calculated:

$$\begin{aligned} E_1 &= k\frac{q_1}{r_1^2} = \left(8.99 \times 10^9 \, \frac{\text{N}\cdot\text{m}^2}{\text{C}^2}\right) \frac{\left(5.00 \times 10^{-9}\ \text{C}\right)}{\left(2.00 \times 10^{-2}\ \text{m}\right)^2} \\ E_1 &= 1.124 \times 10^5\ \text{N/C} \end{aligned}$$

Similarly, E_2 is

$$\begin{aligned} E_2 &= k\frac{q_2}{r_2^2} = \left(8.99 \times 10^9 \, \frac{\text{N}\cdot\text{m}^2}{\text{C}^2}\right) \frac{\left(10.0 \times 10^{-9}\ \text{C}\right)}{\left(4.00 \times 10^{-2}\ \text{m}\right)^2} \\ E_2 &= 0.5619 \times 10^5\ \text{N/C} \end{aligned}$$

Four digits have been retained in this solution to illustrate that E_1 is exactly twice the magnitude of E_2. Now arrows are drawn to represent the magnitudes and directions of $\mathbf{E}_1$ and $\mathbf{E}_2$. (See Figure 3.) The direction of the electric field is that of the force on a positive charge so both arrows point directly away from the positive charges that create them. The arrow for $\mathbf{E}_1$ is exactly twice the length of that for $\mathbf{E}_2$. The arrows form a right triangle in this case and can be added using the Pythagorean theorem. The magnitude of the total field E_{tot} is

$$\begin{aligned} E_{\text{tot}} &= \left(E_1^2 + E_2^2\right)^{1/2} \\ E_{\text{tot}} &= \left\{\left(1.124 \times 10^5\ \text{N/C}\right)^2 + \left(0.5619 \times 10^5\ \text{N/C}\right)^2\right\}^{1/2} \\ E_{\text{tot}} &= 1.26 \times 10^5\ \text{N/C} \end{aligned}$$

The direction is

$$\begin{aligned}\theta &= \tan^{-1}\left(\frac{E_1}{E_2}\right)\\ &= \tan^{-1}\left(\frac{1.124\times10^5\ \text{N/C}}{0.5619\times10^5\ \text{N/C}}\right)\\ &= 63.4^\circ\end{aligned}$$

or 63.4° above the x-axis.

Discussion

In cases where the electric field vectors to be added are not perpendicular, vector components or graphical techniques can be used. The total electric field found in this example is the total electric field at only one point in space. To find the total electric field due to these two charges over an entire region, the same technique must be repeated for each point in the region. This impossibly lengthy task (there are an infinite number of points in space) can be avoided by calculating the total field at representative points and using some of the unifying features noted next.

Figure 4 shows how the electric field from two point charges can be drawn by finding the total field at representative points and drawing electric field lines consistent with those points. While the electric fields from multiple charges are more complex than those of single charges, some simple features are easily noticed.

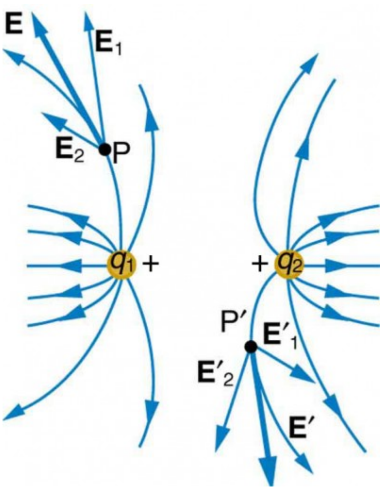


Figure 4. Two positive point charges q_1 and q_2 produce the resultant electric field shown. The field is calculated at representative points and then smooth field lines drawn following the rules outlined in the text.

For example, the field is weaker between like charges, as shown by the lines being farther apart in that region. (This is because the fields from each charge exert opposing forces on any charge placed between them.) (See Figure 4 and Figure 5a.) Furthermore, at a great distance from two like charges, the field becomes identical to the field from a single, larger charge.Figure 5b shows the electric field of two unlike charges. The field is stronger between the charges. In that region, the fields from each charge are in the same direction, and so their strengths add. The field of two unlike charges is weak at large distances, because the fields of the individual charges are in opposite directions and so their strengths subtract. At very large distances, the field of two unlike charges looks like that of a smaller single charge.

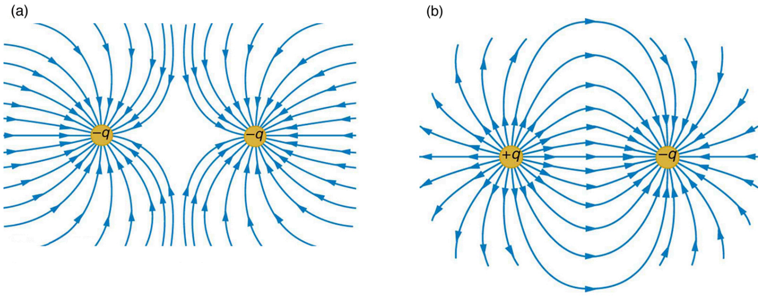


Figure 5. (a) Two negative charges produce the fields shown. It is very similar to the field produced by two positive charges, except that the directions are reversed. The field is clearly weaker between the charges. The individual forces on a test charge in that region are in opposite directions. (b) Two opposite charges produce the field shown, which is stronger in the region between the charges.

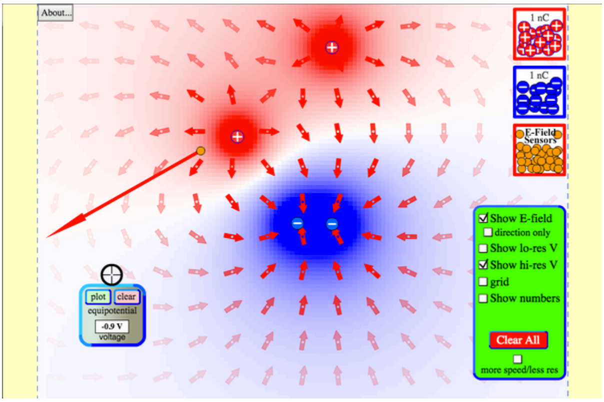
We use electric field lines to visualize and analyze electric fields (the lines are a pictorial tool, not a physical entity in themselves). The properties of electric field lines for any charge distribution can be summarized as follows:

1. Field lines must begin on positive charges and terminate on negative charges, or at infinity in the hypothetical case of isolated charges.
2. The number of field lines leaving a positive charge or entering a negative charge is proportional to the magnitude of the charge.
3. The strength of the field is proportional to the closeness of the field lines—more precisely, it is proportional to the number of lines per unit area perpendicular to the lines.
4. The direction of the electric field is tangent to the field line at any point in space.
5. Field lines can never cross.

The last property means that the field is unique at any point. The field line represents the direction of the field; so if they crossed, the field would have two directions at that location (an impossibility if the field is unique).

PhET Explorations: Charges and Fields

Move point charges around on the playing field and then view the electric field, voltages, equipotential lines, and more. It's colorful, it's dynamic, it's free.



Click to run the simulation.

Section Summary

- Drawings of electric field lines are useful visual tools. The properties of electric field lines for any charge distribution are that:

- Field lines must begin on positive charges and terminate on negative charges, or at infinity in the hypothetical case of isolated charges.
- The number of field lines leaving a positive charge or entering a negative charge is proportional to the magnitude of the charge.
- The strength of the field is proportional to the closeness of the field lines—more precisely, it is proportional to the number of lines per unit area perpendicular to the lines.
- The direction of the electric field is tangent to the field line at any point in space.
- Field lines can never cross.

Conceptual Questions

1. Compare and contrast the Coulomb force field and the electric field. To do this, make a list of five properties for the Coulomb force field analogous to the five properties listed for electric field lines. Compare each item in your list of Coulomb force field properties with those of the electric field—are they the same or different? (For example, electric field lines cannot cross. Is the same true for Coulomb field lines?)
2. [link] shows an electric field extending over three regions, labeled I, II, and III. Answer the following questions. (a) Are there any isolated charges? If so, in what region and what are their signs? (b) Where is the field strongest? (c) Where is it weakest? (d) Where is

the field the most uniform?

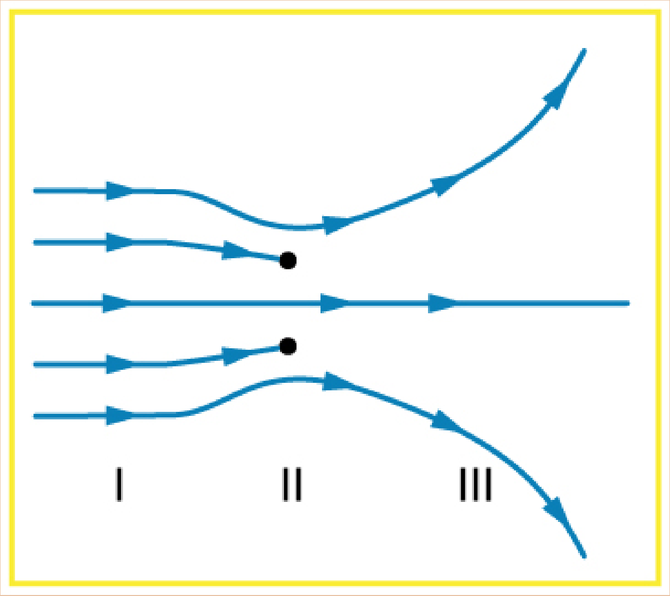


Figure 6.

Problems & Exercises

1. (a) Sketch the electric field lines near a point charge $+q$. (b) Do the same for a point charge $-3.00q$.
2. Sketch the electric field lines a long distance from the charge distributions shown in Figure 5a and 5b.

3. Figure 8 shows the electric field lines near two charges q_1 and q_2 . What is the ratio of their magnitudes? (b) Sketch the electric field lines a long distance from the charges shown in the figure.

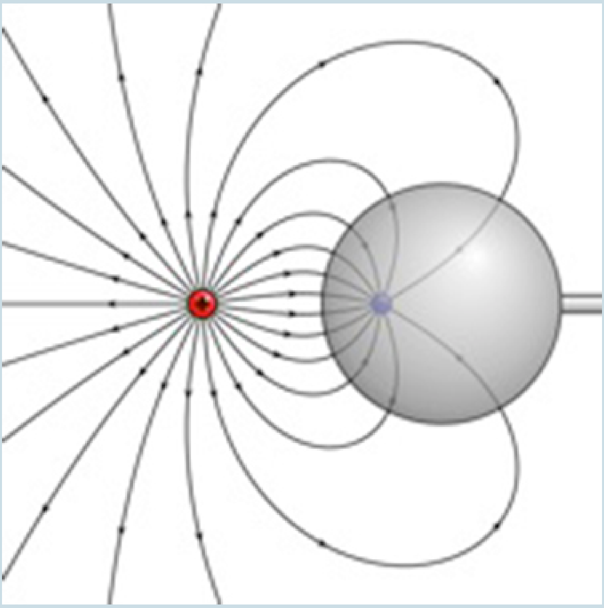


Figure 7. The electric field near two charges.

4. Sketch the electric field lines in the vicinity of two opposite charges, where the negative charge is three times greater in magnitude than the positive. (See Figure 7 for a similar situation).

Glossary

electric field: a three-dimensional map of the electric force extended out into space from a point charge

electric field lines: a series of lines drawn from a point charge representing the magnitude and direction of force exerted by that charge

vector: a quantity with both magnitude and direction

vector addition: mathematical combination of two or more vectors, including their magnitudes, directions, and positions

8. Electric Forces in Biology

Learning Objectives

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

- Describe how a water molecule is polar.
- Explain electrostatic screening by a water molecule within a living cell.

Classical electrostatics has an important role to play in modern molecular biology. Large molecules such as proteins, nucleic acids, and so on—so important to life—are usually electrically charged. DNA itself is highly charged; it is the electrostatic force that not only holds the molecule together but gives the molecule structure and strength. Figure 1 is a schematic of the DNA double helix.

The four nucleotide bases are given the symbols A (adenine), C (cytosine), G (guanine), and T (thymine). The order of the four bases varies in each strand, but the pairing between bases is always the same. C and G are always paired and A and T are always paired, which helps to preserve the order of bases in cell division (mitosis) so as to pass on the correct genetic information. Since the Coulomb force drops with distance $\left(F \propto \frac{1}{r^2}\right)$, the distances between the base pairs must be small enough that the electrostatic force is sufficient to hold them together.

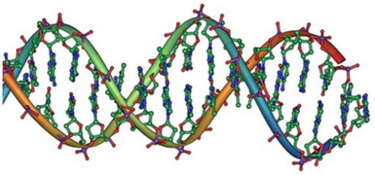


*Figure 1. DNA is a highly charged molecule. The DNA double helix shows the two coiled strands each containing a row of nitrogenous bases, which "code" the genetic information needed by a living organism. The strands are connected by bonds between pairs of bases. While pairing combinations between certain bases are fixed (*C-G *and* A-T*), the sequence of nucleotides in the strand varies. (credit: Jerome Walker)*

DNA is a highly charged molecule, with about $2q_e$ (fundamental charge) per 0.3×10^{-9} m. The distance separating the two strands that make up the DNA structure is about 1 nm, while the distance separating the individual atoms within each base is about 0.3 nm.

One might wonder why electrostatic forces do not play a larger role in biology than they do if we have so many charged molecules. The reason is that the electrostatic force is "diluted" due to *screening* between molecules. This is due to the presence of other charges in the cell.

Polarity of Water Molecules

The best example of this charge screening is the water molecule, represented as H_2O. Water is a strongly *polar molecule*. Its 10 electrons (8 from the oxygen atom and 2 from the two hydrogen atoms) tend to remain closer to the oxygen nucleus than the hydrogen nuclei. This creates two centers of equal and opposite charges—what is called a *dipole*, as illustrated in Figure 2. The magnitude of the dipole is called the dipole moment.

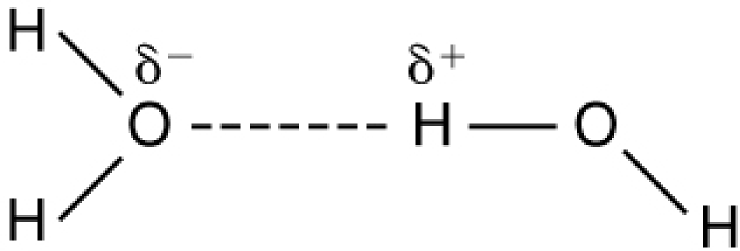


Figure 2. This schematic shows water (H_2O) as a polar molecule. Unequal sharing of electrons between the oxygen (O) and hydrogen (H) atoms leads to a net separation of positive and negative charge—forming a dipole. The symbols δ^- and δ^+ indicate that the oxygen side of the H_2O molecule tends to be more negative, while the hydrogen ends tend to be more positive. This leads to an attraction of opposite charges between molecules.

These two centers of charge will terminate some of the electric field lines coming from a free charge, as on a DNA molecule. This results in a reduction in the strength of the *Coulomb interaction*. One might say that screening makes the Coulomb force a short range force rather than long range.

Other ions of importance in biology that can reduce or screen Coulomb interactions are Na^+, and K^+, and Cl^-. These ions are located both inside and outside of living cells. The movement of these ions through cell membranes is crucial to the motion of nerve impulses through nerve axons.

Recent studies of electrostatics in biology seem to show that electric fields in cells can be extended over larger distances, in spite of screening, by “microtubules” within the cell. These microtubules are hollow tubes composed of proteins that guide the movement of chromosomes when cells divide, the motion of other organisms within the cell, and provide mechanisms for motion of some cells (as motors).

Section Summary

- Many molecules in living organisms, such as DNA, carry a charge.
- An uneven distribution of the positive and negative charges within a polar molecule produces a dipole.
- The effect of a Coulomb field generated by a charged object may be reduced or blocked by other nearby charged objects.
- Biological systems contain water, and because water molecules are polar, they have a strong effect on other molecules in living systems.

Conceptual Question

A cell membrane is a thin layer enveloping a cell. The thickness of the membrane is much less than the size of the cell. In a static situation the membrane has a charge distribution of $-2.5 \times 10^{-6}\ \mathrm{C/m^2}$ on its inner surface and $+2.5 \times 10^{-6}\ \mathrm{C/m^2}$ on its outer surface. Draw a diagram of the cell and the surrounding cell membrane. Include on this diagram the charge distribution and the corresponding electric field. Is there any electric field inside the cell? Is there any electric field outside the cell?

Glossary

dipole: a molecule's lack of symmetrical charge distribution, causing one side to be more positive and another to be more negative

polar molecule: a molecule with an asymmetrical distribution of positive and negative charge

screening: the dilution or blocking of an electrostatic force on a charged object by the presence of other charges nearby

Coulomb interaction: the interaction between two charged particles generated by the Coulomb forces they exert on one another

9. Conductors and Electric Fields in Static Equilibrium

Learning Objectives

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

- List the three properties of a conductor in electrostatic equilibrium.
- Explain the effect of an electric field on free charges in a conductor.
- Explain why no electric field may exist inside a conductor.
- Describe the electric field surrounding Earth.
- Explain what happens to an electric field applied to an irregular conductor.
- Describe how a lightning rod works.
- Explain how a metal car may protect passengers inside from the dangerous electric fields caused by a downed line touching the car.

Conductors contain *free charges* that move easily. When excess charge is placed on a conductor or the conductor is put into a static electric field, charges in the conductor quickly respond to reach a steady state called *electrostatic equilibrium*.

Figure 1 shows the effect of an electric field on free charges in a conductor. The free charges move until the field is perpendicular to the conductor's surface. There can be no component of the field parallel to the surface in electrostatic equilibrium, since, if there were, it would produce further movement of charge. A positive free charge is shown, but free charges can be either positive or negative and are, in fact, negative in metals. The motion of a positive charge is equivalent to the motion of a negative charge in the opposite direction.

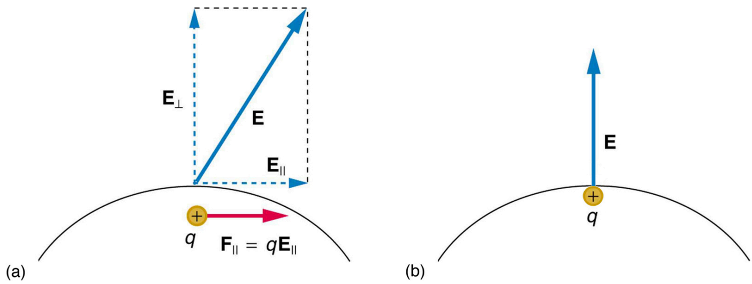


Figure 1. When an electric field E *is applied to a conductor, free charges inside the conductor move until the field is perpendicular to the surface. (a) The electric field is a vector quantity, with both parallel and perpendicular components. The parallel component (*E ∥*) exerts a force (*F ∥*) on the free charge q, which moves the charge until* F ∥=0*. (b) The resulting field is perpendicular to the surface. The free charge has been brought to the conductor's surface, leaving electrostatic forces in equilibrium.*

A conductor placed in an *electric field* will be *polarized*. Figure 2 shows the result of placing a neutral conductor in an originally uniform electric field. The field becomes stronger near the conductor but entirely disappears inside it.

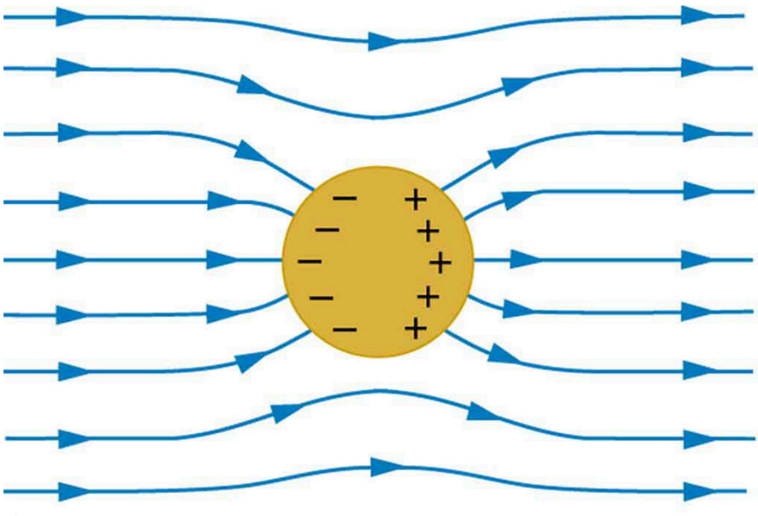


Figure 2. This illustration shows a spherical conductor in static equilibrium with an originally uniform electric field. Free charges move within the conductor, polarizing it, until the electric field lines are perpendicular to the surface. The field lines end on excess negative charge on one section of the surface and begin again on excess positive charge on the opposite side. No electric field exists inside the conductor, since free charges in the conductor would continue moving in response to any field until it was neutralized.

Misconception Alert: Electric Field inside a Conductor

Excess charges placed on a spherical conductor repel and move until they are evenly distributed, as shown in Figure 3. Excess charge is forced to the surface until the field inside the conductor is zero. Outside the conductor, the field is exactly the same as if the conductor were replaced by a point charge at its center equal to the excess charge.

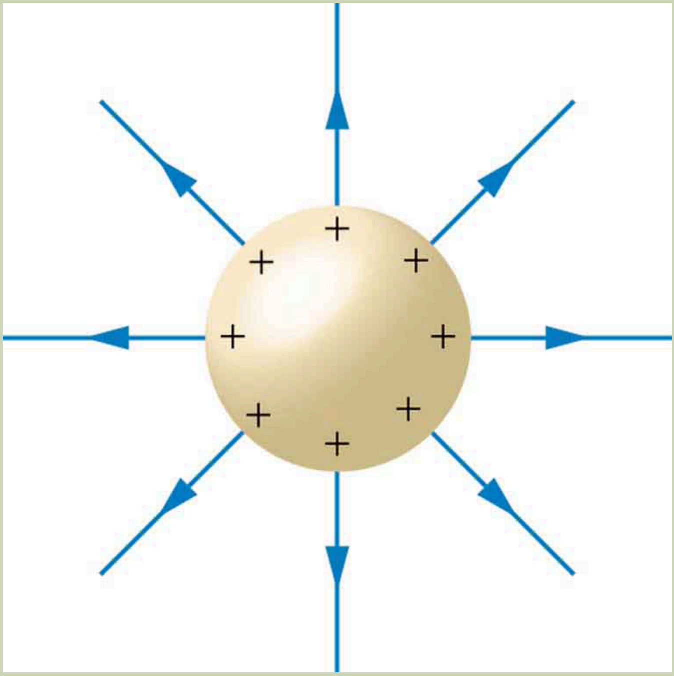


Figure 3. The mutual repulsion of excess positive charges on a spherical conductor distributes them uniformly on its surface. The resulting electric field is perpendicular to the surface and zero inside. Outside the conductor, the field is identical to that of a point charge at the center equal to the excess charge.

Properties of a Conductor in Electrostatic

Equilibrium

1. The electric field is zero inside a conducto

r.

2. Just outside a conductor, the electric field lines are perpendicular to its surface, ending or beginning

on charges on the surface.

3. Any excess charge resides entirely on the surface or surfaces of a

conducto
r.

The properties of a conductor are consistent with the situations already discussed and can be used to analyze any conductor in electrostatic equilibrium. This can lead to some interesting new insights, such as described below.

How can a very uniform electric field be created? Consider a system of two metal plates with opposite charges on them, as shown in Figure 4. The properties of conductors in electrostatic equilibrium indicate that the electric field between the plates will be uniform in strength and direction. Except near the edges, the excess charges distribute themselves uniformly, producing field lines that are uniformly spaced (hence uniform in strength) and perpendicular to the surfaces (hence uniform in direction, since the plates are flat). The edge effects are less important when the plates are close together.

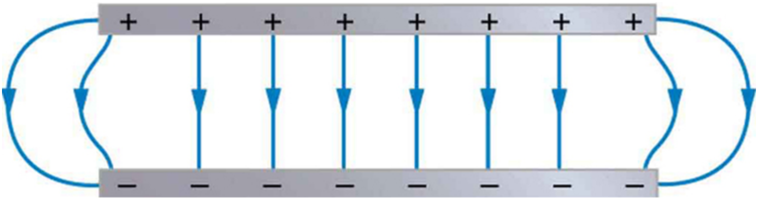


Figure 4. Two metal plates with equal, but opposite, excess charges. The field between them is uniform in strength and direction except near the edges. One use of such a field is to produce uniform acceleration of charges between the plates, such as in the electron gun of a TV tube.

Earth's Electric Field

A near uniform electric field of approximately 150 N/C, directed downward, surrounds Earth, with the magnitude increasing slightly as we get closer to the surface. What causes the electric field? At around 100 km above the surface of Earth we have a layer of charged particles, called the *ionosphere*. The ionosphere is responsible for a range of phenomena including the electric field surrounding Earth. In fair weather the ionosphere is positive and the Earth largely negative, maintaining the electric field (Figure 5a).

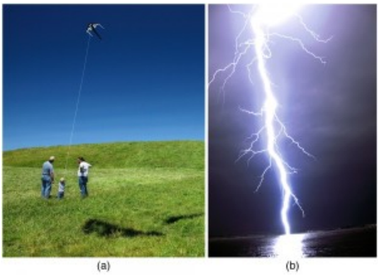


Figure 5. Earth's electric field. (a) Fair weather field. Earth and the ionosphere (a layer of charged particles) are both conductors. They produce a uniform electric field of about 150 N/C. (credit: D. H. Parks) (b) Storm fields. In the presence of storm clouds, the local electric fields can be larger. At very high fields, the insulating properties of the air break down and lightning can occur. (credit: Jan-Joost Verhoef)

In storm conditions clouds form and localized electric fields can be larger and reversed in direction (Figure 5b). The exact charge distributions depend on the local conditions, and variations of Figure 5b are possible.

If the electric field is sufficiently large, the insulating properties of the surrounding material break down and it becomes conducting. For air this occurs at around 3×10^6 N/C. Air ionizes ions and electrons recombine, and we get discharge in the form of lightning sparks and corona discharge.

Electric Fields on Uneven Surfaces

So far we have considered excess charges on a smooth, symmetrical conductor surface. What happens if a conductor has sharp corners or is pointed? Excess charges on a nonuniform conductor become concentrated at the sharpest points. Additionally, excess charge may move on or off the conductor at the sharpest points.

To see how and why this happens, consider the charged conductor in Figure 6. The electrostatic repulsion of like charges is most effective in moving them apart on the flattest surface, and so they become least concentrated there. This is because the forces between identical pairs of charges at either end of the conductor are identical, but the components of the forces parallel to the surfaces are different. The component parallel to the surface is greatest on the flattest surface and, hence, more effective in moving the charge.

The same effect is produced on a conductor by an externally applied electric field, as seen in Figure 6c. Since the field lines must be perpendicular to the surface, more of them are concentrated on the most curved parts.

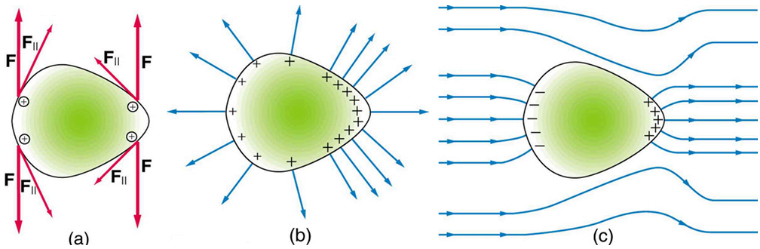


Figure 6. Excess charge on a nonuniform conductor becomes most concentrated at the location of greatest curvature. (a) The forces between identical pairs of charges at either end of the conductor are identical, but the components of the forces parallel to the surface are different. It is $\mathbf{F}_{\parallel}$ *that moves the charges apart once they have reached the surface. (b)* $\mathbf{F}_{\parallel}$ *is smallest at the more pointed end, the charges are left closer together, producing the electric field shown. (c) An uncharged conductor in an originally uniform electric field is polarized, with the most concentrated charge at its most pointed end.*

Applications of Conductors

On a very sharply curved surface, such as shown in Figure 7, the charges are so concentrated at the point that the resulting electric field can be great enough to remove them from the surface. This can be useful.

Lightning rods work best when they are most pointed. The large charges created in storm clouds induce an opposite charge on a building that can result in a lightning bolt hitting the building. The induced charge is bled away continually by a lightning rod, preventing the more dramatic lightning strike.

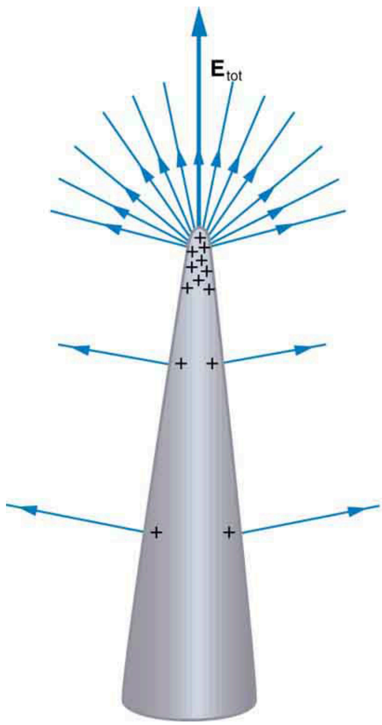


Figure 7. A very pointed conductor has a large charge concentration at the point. The electric field is very strong at the point and can exert a force large enough to transfer charge on or off the conductor. Lightning rods are used to prevent the buildup of large excess charges on structures and, thus, are pointed.

Of course, we sometimes wish to prevent the transfer of charge rather than to facilitate it. In that case, the conductor should be very smooth and have as large a radius of curvature as possible. (See Figure 8.) Smooth surfaces are used on high-voltage transmission lines, for example, to avoid leakage of charge into the air.

Another device that makes use of some of these principles is a *Faraday cage*. This is a metal shield that encloses a volume. All electrical charges will reside on the outside surface of this shield, and there will be no electrical field inside. A Faraday cage is used to prohibit stray electrical fields in the environment from interfering with sensitive measurements, such as the electrical signals inside a nerve cell.

During electrical storms if you are driving a car, it is best to stay inside the car as its metal body acts as a Faraday cage with zero electrical field inside. If in the vicinity of a lightning strike, its effect is felt on the outside of the car and the inside is unaffected, provided you remain totally inside. This is also true if an active ("hot") electrical wire was broken (in a storm or an accident) and fell on your car.

Figure 8. (a) A lightning rod is pointed to facilitate the transfer of charge. (credit: Romaine, Wikimedia Commons) (b) This Van de Graaff generator has a smooth surface with a large radius of curvature to prevent the transfer of charge and allow a large voltage to be generated. The mutual repulsion of like charges is evident in the person's hair while touching the metal sphere. (credit: Jon 'ShakataGaNai' Davis/Wikimedia Commons).

Section Summary

- A conductor allows free charges to move about within it.
- The electrical forces around a conductor will cause free charges to move around inside the conductor until static equilibrium is reached.
- Any excess charge will collect along the surface of a conductor.

- Conductors with sharp corners or points will collect more charge at those points.
- A lightning rod is a conductor with sharply pointed ends that collect excess charge on the building caused by an electrical storm and allow it to dissipate back into the air.
- Electrical storms result when the electrical field of Earth's surface in certain locations becomes more strongly charged, due to changes in the insulating effect of the air.
- A Faraday cage acts like a shield around an object, preventing electric charge from penetrating inside.

Conceptual Questions

1. Is the object in Figure 9 a conductor or an insulator? Justify your answer.

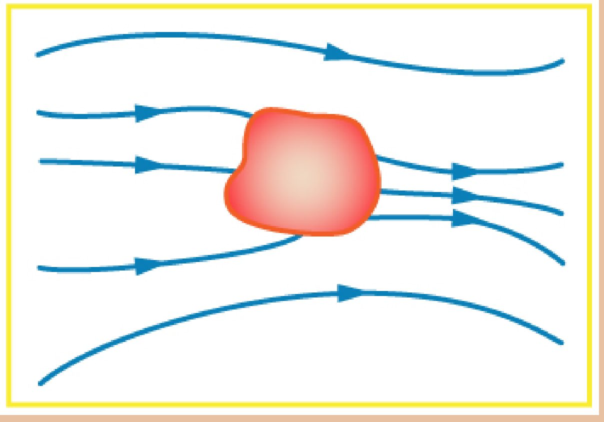


Figure 9.

2. External field lines entering the object from one end and emerging from another are shown by lines. If the electric field lines in the figure above were perpendicular to the object, would it necessarily be a conductor? Explain.
3. The discussion of the electric field between two parallel conducting plates, in this module states that edge effects are less important if the plates are close together. What does close mean? That is, is the actual plate separation crucial, or is the ratio of plate separation to plate area crucial?
4. Would the self-created electric field at the end of a pointed conductor, such as a lightning rod, remove positive or negative charge from the conductor? Would the same sign charge be removed from a neutral pointed conductor by the application of a similar externally created electric field? (The answers to both questions have implications for charge transfer utilizing points.)
5. Why is a golfer with a metal club over her shoulder vulnerable to lightning in an open fairway? Would she be any safer under a tree?
6. Can the belt of a Van de Graaff accelerator be a conductor? Explain.
7. Are you relatively safe from lightning inside an automobile? Give two reasons.
8. Discuss pros and cons of a lightning rod being grounded versus simply being attached to a building.
9. Using the symmetry of the arrangement, show that the net Coulomb force on the charge q at the center of the square below (Figure 10) is zero if the charges on the four corners are exactly equal.

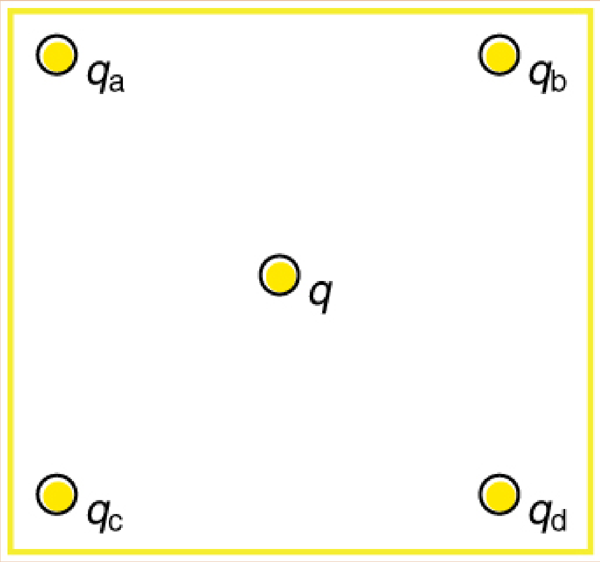


Figure 10. Four point charges q_a, q_b, q_c, and q_d lie on the corners of a square and q is located at its center.

10. (a) Using the symmetry of the arrangement, show that the electric field at the center of the square in Figure 10 is zero if the charges on the four corners are exactly equal. (b) Show that this is also true for any combination of charges in which $q_a = q_b$ and $q_b = q_c$
11. (a) What is the direction of the total Coulomb force on q in Figure 10 if q is negative, $q_a = q_c$ and both are negative, and $q_b = q_c$ and both are positive? (b) What is the direction of the electric field at the center of the square in this situation?
12. Considering Figure 10, suppose that $q_a = q_d$ and $q_b =$

q_c. First show that q is in static equilibrium. (You may neglect the gravitational force.) Then discuss whether the equilibrium is stable or unstable, noting that this may depend on the signs of the charges and the direction of displacement of q from the center of the square.

13. If $q_a = 0$ in Figure 10, under what conditions will there be no net Coulomb force on q?
14. In regions of low humidity, one develops a special "grip" when opening car doors, or touching metal door knobs. This involves placing as much of the hand on the device as possible, not just the ends of one's fingers. Discuss the induced charge and explain why this is done.
15. Tollbooth stations on roadways and bridges usually have a piece of wire stuck in the pavement before them that will touch a car as it approaches. Why is this done?
16. Suppose a woman carries an excess charge. To maintain her charged status can she be standing on ground wearing just any pair of shoes? How would you discharge her? What are the consequences if she simply walks away?

Problems & Exercises

1. Sketch the electric field lines in the vicinity of the conductor in Figure 11 given the field was originally

uniform and parallel to the object's long axis. Is the resulting field small near the long side of the object?

Figure 11

2. Sketch the electric field lines in the vicinity of the conductor in Figure 12 given the field was originally uniform and parallel to the object's long axis. Is the resulting field small near the long side of the object?

Figure 12.

3. Sketch the electric field between the two conducting plates shown in Figure 13, given the top plate is positive and an equal amount of negative charge is on the bottom plate. Be certain to indicate the distribution of charge on the plates.

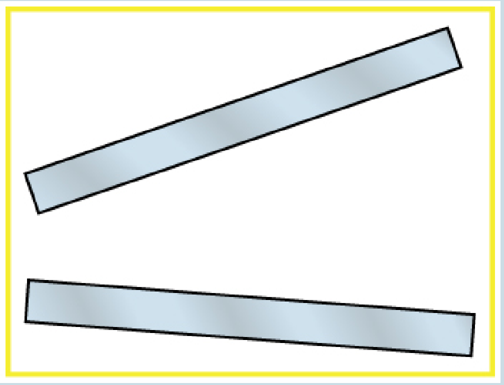


Figure 13.

4. Sketch the electric field lines in the vicinity of the charged insulator in Figure 14 noting its nonuniform charge distribution.

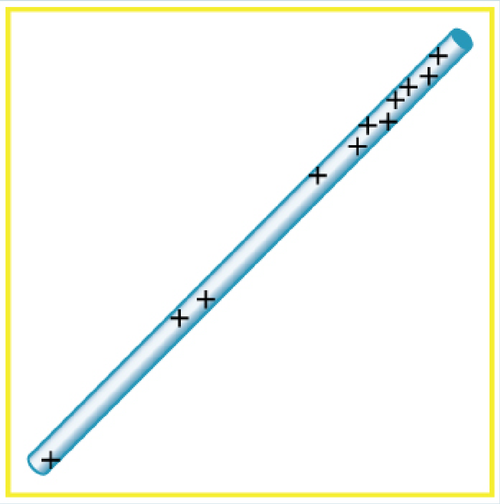


Figure 14. A charged insulating rod such as might be used in a classroom demonstration.

5. What is the force on the charge located at $x = 8.00$ cm in Figure 15a given that $q = 1.00\ \mu C$?

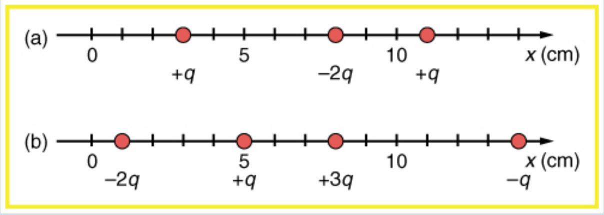


Figure 15. (a) Point charges located at 3.00, 8.00, and 11.0 cm along the x-axis. (b) Point charges located at 1.00, 5.00, 8.00, and 14.0 cm along the x-axis.

6. (a) Find the total electric field at x = 1.00 cm in Figure 15b given that q = 5.00 nC. (b) Find the total electric field at x = 11.00 cm in Figure 15b. (c) If the charges are allowed to move and eventually be brought to rest by friction, what will the final charge configuration be? (That is, will there be a single charge, double charge, etc., and what will its value(s) be?)
7. (a) Find the electric field at x = 5.00 cm in Figure 15a, given that q = 1.00 μC. (b) At what position between 3.00 and 8.00 cm is the total electric field the same as that for −2q alone? (c) Can the electric field be zero anywhere between 0.00 and 8.00 cm? (d) At very large positive or negative values of x, the electric field approaches zero in both (a) and (b). In which does it most rapidly approach zero and why? (e) At what position to the right of 11.0 cm is the total electric field zero, other than at infinity? (Hint: A graphing calculator can yield considerable insight in this problem.)
8. (a) Find the total Coulomb force on a charge of 2.00 nC located at x = 4.00 cm in Figure 15b, given that q = 1.00 μC. (b) Find the x-position at which the electric field is zero in Figure 15b.
9. Using the symmetry of the arrangement, determine the direction of the force on q in the figure below, given that $q_a = q_b$ = +7.50 μC and $q_c = q_d$ = −7.50 μC. (b) Calculate the magnitude of the force on the charge q, given that the square is 10.0 cm on a side and q = 2.00 μC.

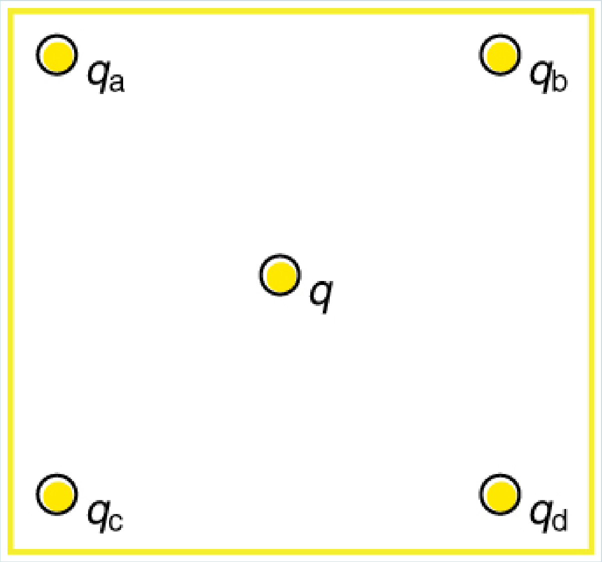


Figure 16.

10. (a) Using the symmetry of the arrangement, determine the direction of the electric field at the center of the square in Figure, given that $q_a = q_b = -1.00\ \mu C$ and $q_c = q_d = +1.00\ \mu C$. (b) Calculate the magnitude of the electric field at the location of q, given that the square is 5.00 cm on a side.
11. Find the electric field at the location of q_a in Figure 16 given that $q_b = q_c = q_d = +2.00$ nC, $q = -1.00$ nC, and the square is 20.0 cm on a side.
12. Find the total Coulomb force on the charge q in Figure 16, given that $q = 1.00\ \mu C$, $q_a = 2.00\ \mu C$, $q_b = -3.00\ \mu C$, $q_c = -4.00\ \mu C$, and $q_d = +1.00\ \mu C$. The square is 50.0 cm on a side.

13. (a) Find the electric field at the location of q_a in Figure 17, given that q_b = +10.00 μC and q_c = −5.00 μC. (b) What is the force on q_a, given that q_a = +1.50 nC?

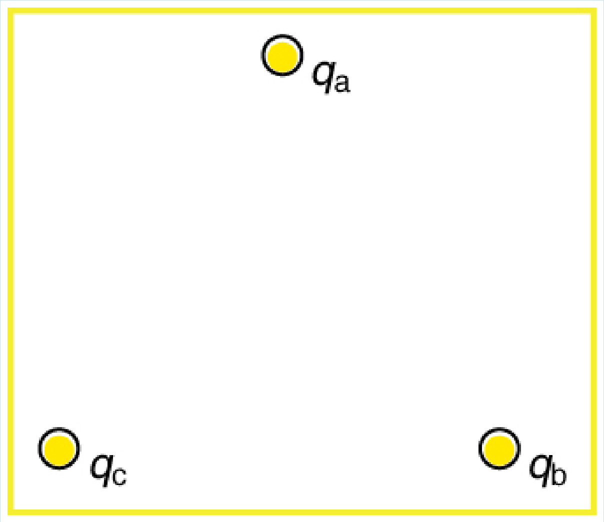


Figure 17. Point charges located at the corners of an equilateral triangle 25.0 cm on a side.

14. (a) Find the electric field at the center of the triangular configuration of charges in Figure 17, given that q_a = +2.50 nC, q_b = −8.00 nC, and q_c = +1.50 nC. (b) Is there any combination of charges, other than q_a = q_b = q_c, that will produce a zero strength electric field at the center of the triangular configuration?

Glossary

conductor: an object with properties that allow charges to move about freely within it

free charge: an electrical charge (either positive or negative) which can move about separately from its base molecule

electrostatic equilibrium: an electrostatically balanced state in which all free electrical charges have stopped moving about

polarized: a state in which the positive and negative charges within an object have collected in separate locations

ionosphere: a layer of charged particles located around 100 km above the surface of Earth, which is responsible for a range of phenomena including the electric field surrounding Earth

Faraday cage: a metal shield which prevents electric charge from penetrating its surface

Selected Solutions to Problems & Exercises

6. (a) $E_{x\,=\,1.00\text{ cm}} = -\infty$; (b) 2.12×10^5 N/C; (c) one charge of $+q$

8. (a) 0.252 N to the left; (b) x = 6.07 cm

10. (a) The electric field at the center of the square will be straight up, since q_a and q_b are positive and q_c and q_d are negative and all have the same magnitude; (b) 2.04×10^7 N/C (upward)

12. 0.102 N, in the $-y$ direction

14. (a) $\vec{E} = 4.36 \times 10^3 \text{ N/C}, 35.0^\circ$, below the horizontal; (b) No

10. Applications of Electrostatics

Learning Objective

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

- Name several real-world applications of the study of electrostatics.

The study of *electrostatics* has proven useful in many areas. This module covers just a few of the many applications of electrostatics.

The Van de Graaff Generator

Van de Graaff generators (or Van de Graaffs) are not only spectacular devices used to demonstrate high voltage due to static electricity—they are also used for serious research. The first was built by Robert Van de Graaff in 1931 (based on original suggestions by Lord Kelvin) for use in nuclear physics research. Figure 1 shows a schematic of a large research version. Van de Graaffs utilize both smooth and pointed surfaces, and conductors and insulators to generate large static charges and, hence, large voltages.

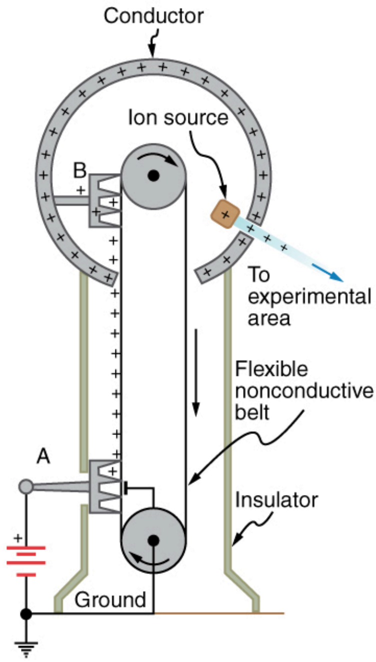


Figure 1. Schematic of Van de Graaff generator.

A battery (part A in Figure 1) supplies excess positive charge to a pointed conductor, the points of which spray the charge onto a moving insulating belt near the bottom. The pointed conductor (part B in Figure 1) on top in the large sphere picks up the charge. (The induced electric field at the points is so large that it removes the charge from the belt.) This can be done because the charge does not remain inside the conducting sphere but moves to its outside surface. An ion source inside the sphere produces positive ions, which are accelerated away from the positive sphere to high velocities.

A very large excess charge can be deposited on the sphere, because it moves quickly to the outer surface. Practical limits arise because the large electric fields polarize and eventually ionize surrounding materials, creating free charges that neutralize excess charge or allow it to escape. Nevertheless, voltages of 15 million volts are well within practical limits.

Take-Home Experiment: Electrostatics and Humidity

Rub a comb through your hair and use it to lift pieces of paper. It may help to tear the pieces of paper rather than cut them neatly. Repeat the exercise in your bathroom after you have had a long shower and the air in the bathroom is moist. Is it easier to get electrostatic effects in dry or moist air? Why would torn paper be more attractive to the comb than cut paper? Explain your observations.

Xerography

Most copy machines use an electrostatic process called *xerography*—a word coined from the Greek words *xeros* for dry and *graphos* for writing. The heart of the process is shown in simplified form in Figure 2.

A selenium-coated aluminum drum is sprayed with positive charge from points on a device called a corotron. Selenium is a substance with an interesting property—it is a *photoconductor*. That is, selenium is an insulator when in the dark and a conductor when exposed to light.

In the first stage of the xerography process, the conducting aluminum drum is *grounded* so that a negative charge is induced under the thin layer of uniformly positively charged selenium. In the second stage, the surface of the drum is exposed to the image of whatever is to be copied. Where the image is light, the selenium becomes conducting, and the positive charge is neutralized. In dark areas, the positive charge remains, and so the image has been transferred to the drum.

The third stage takes a dry black powder, called toner, and sprays it with a negative charge so that it will be attracted to the positive regions of the drum. Next, a blank piece of paper is given a greater positive charge than on the drum so that it will pull the toner from the drum. Finally, the paper and electrostatically held toner are passed through heated pressure rollers, which melt and permanently adhere the toner within the fibers of the paper.

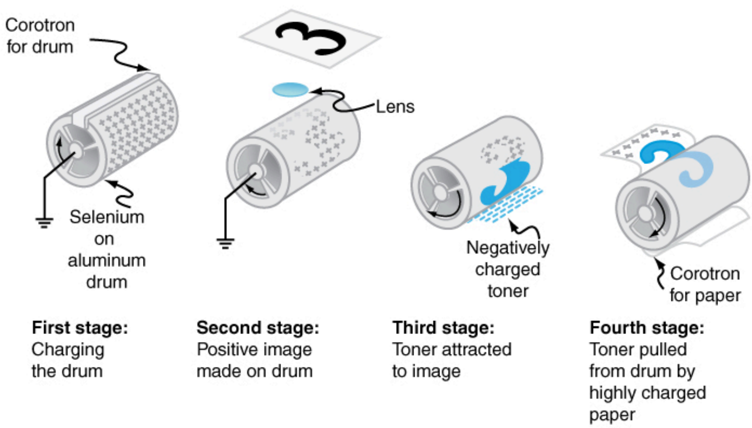


Figure 2. Xerography is a dry copying process based on electrostatics. The major steps in the process are the charging of the photoconducting drum, transfer of an image creating a positive charge duplicate, attraction of toner to the charged parts of the drum, and transfer of toner to the paper. Not shown are heat treatment of the paper and cleansing of the drum for the next copy.

Laser Printers

Laser printers use the xerographic process to make high-quality images on paper, employing a laser to produce an image on the photoconducting drum as shown in Figure 3. In its most common application, the laser printer receives output from a computer, and it can achieve high-quality output because of the precision with which laser light can be controlled. Many laser printers do significant information processing, such as making sophisticated letters or fonts, and may contain a computer more powerful than the one giving them the raw data to be printed.

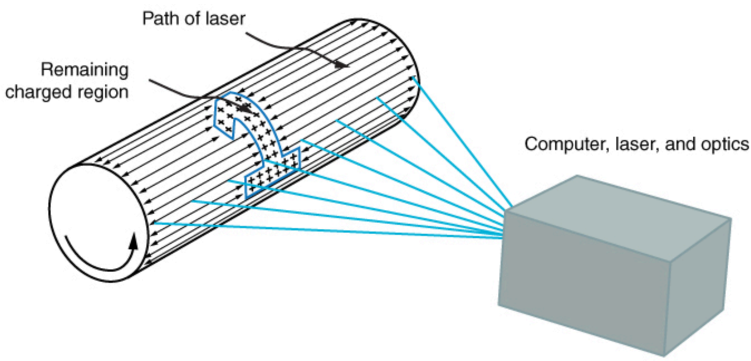


Figure 3. In a laser printer, a laser beam is scanned across a photoconducting drum, leaving a positive charge image. The other steps for charging the drum and transferring the image to paper are the same as in xerography. Laser light can be very precisely controlled, enabling laser printers to produce high-quality images.

Ink Jet Printers and Electrostatic Painting

The *ink jet printer*, commonly used to print computer-generated text and graphics, also employs electrostatics. A nozzle makes a fine spray of tiny ink droplets, which are then given an electrostatic charge. (See Figure 4.)

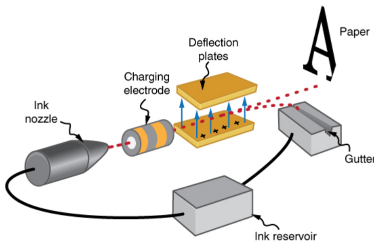


Figure 4. The nozzle of an ink-jet printer produces small ink droplets, which are sprayed with electrostatic charge. Various computer-driven devices are then used to direct the droplets to the correct positions on a page.

Once charged, the droplets can be directed, using pairs of charged plates, with great precision to form letters and images on paper. Ink jet printers can produce color images by using a black jet and three other jets with primary colors, usually cyan, magenta, and yellow, much as a color television produces color. (This is more difficult with xerography, requiring multiple drums and toners.)

Electrostatic painting employs electrostatic charge to spray paint onto odd-shaped surfaces. Mutual repulsion of like charges causes the paint to fly away from its source. Surface tension forms drops, which are then attracted by unlike charges to the surface to be painted. Electrostatic painting can reach those hard-to-get at places, applying an even coat in a controlled manner. If the object is a conductor, the electric field is perpendicular to the surface, tending to bring the drops in perpendicularly. Corners and points on conductors will receive extra paint. Felt can similarly be applied.

Smoke Precipitators and Electrostatic Air Cleaning

Another important application of electrostatics is found in air cleaners, both large and small. The electrostatic part of the process places excess (usually positive) charge on smoke, dust, pollen, and other particles in the air and then passes the air through an oppositely charged grid that attracts and retains the charged particles. (See Figure 5.)

Large *electrostatic precipitators* are used industrially to remove over 99% of the particles from stack gas emissions associated with the burning of coal and oil. Home precipitators, often in conjunction with the home heating and air conditioning system, are very effective in removing polluting particles, irritants, and allergens.

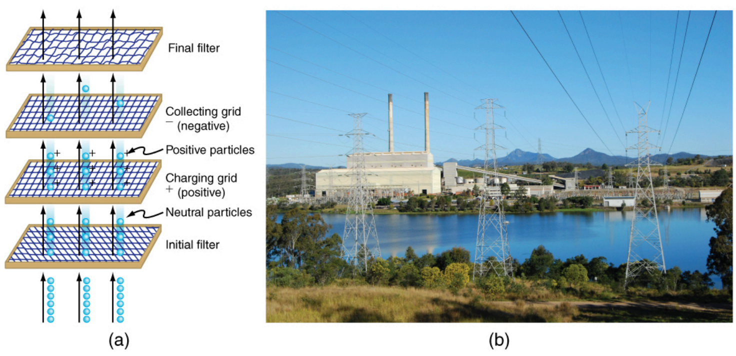


Figure 5. (a) Schematic of an electrostatic precipitator. Air is passed through grids of opposite charge. The first grid charges airborne particles, while the second attracts and collects them. (b) The dramatic effect of electrostatic precipitators is seen by the absence of smoke from this power plant. (credit: Cmdalgleish, Wikimedia Commons)

Problem-Solving Strategies for Electrostatics

1. Examine the situation to determine if static electricity is involved. This may concern separated stationary charges, the forces among them, and the electric fields they create.
2. Identify the system of interest. This includes noting the number, locations, and types of charges involved.
3. Identify exactly what needs to be determined in the problem (identify the unknowns). A written list is useful. Determine whether the Coulomb force is to be considered directly—if so, it may be useful to draw a free-body diagram, using electric field lines.
4. Make a list of what is given or can be inferred from the problem as stated (identify the knowns). It is important to distinguish the Coulomb force F from the electric field E, for example.
5. Solve the appropriate equation for the quantity to be determined (the unknown) or draw the field lines as requested.
6. Examine the answer to see if it is reasonable: Does it make sense? Are units correct and the numbers involved reasonable?

Integrated Concepts

The Integrated Concepts exercises for this module involve concepts such as electric charges, electric fields, and several other topics. Physics is most interesting when applied to general situations involving more than a narrow set of physical principles. The electric field exerts force on charges, for example, and hence the relevance of Dynamics: Force and Newton's Laws of Motion. The following topics are involved in some or all of the problems labeled "Integrated Concepts":

- Kinematics
- Two-Dimensional Kinematics
- Dynamics: Force and Newton's Laws of Motion
- Uniform Circular Motion and Gravitation
- Statics and Torque
- Fluid Statics

The following worked example illustrates how this strategy is applied to an Integrated Concept problem:

Example 1. Acceleration of a Charged Drop of Gasoline

If steps are not taken to ground a gasoline pump, static electricity can be placed on gasoline when filling your car's tank. Suppose a tiny drop of gasoline has a mass of 4.00×10^{-15} kg and is given a positive charge of 3.20×10^{-19} C.

1. Find the weight of the drop.
2. Calculate the electric force on the drop if there is an upward electric field of strength 3.00×10^{5} N/C due to other static electricity in the vicinity.
3. Calculate the drop's acceleration.

Strategy

To solve an integrated concept problem, we must first identify the physical principles involved and identify the chapters in which they are found. Part 1 of this example asks for weight. This is a topic of dynamics and is defined in Dynamics: Force and Newton's Laws of Motion. Part 2 deals with electric force on a charge, a topic of Electric Charge and Electric Field. Part 3 asks for acceleration, knowing forces and mass. These are part of Newton's laws, also found in Dynamics: Force and Newton's Laws of Motion.

The following solutions to each part of the example illustrate how the specific problem-solving strategies are applied. These involve identifying knowns and unknowns, checking to see if the answer is reasonable, and so on.

Solution for Part 1

Weight is mass times the acceleration due to gravity, as first expressed in $w = mg$. Entering the given mass and the average acceleration due to gravity yields

$$w = (4.00 \times 10^{-15}\ \text{kg})(9.80\ \text{m/s}^2) = 3.92 \times 10^{-14}\ \text{N}.$$

Discussion for Part 1

This is a small weight, consistent with the small mass of the drop.

Solution for Part 2

The force an electric field exerts on a charge is given by rearranging the following equation:

$$F = qE.$$

Here we are given the charge (3.20×10^{-19} C is twice the fundamental unit of charge) and the electric field strength, and so the electric force is found to be

$$F = (3.20 \times 10^{-19}\ \text{C})(3.00 \times 10^{5}\ \text{N/C}) = 9.60 \times 10^{-14}\ \text{N}.$$

Discussion for Part 2

While this is a small force, it is greater than the weight of the drop.

Solution for Part 3

The acceleration can be found using Newton's second law, provided we can identify all of the external forces acting on the drop. We assume only the drop's weight and the electric force are significant. Since the drop has a positive charge and the electric field is given to be upward, the electric force is upward. We thus have a one-dimensional (vertical direction) problem, and we can state Newton's second law as

$$a = \frac{F_{\text{net}}}{m}, \text{ where } F_{\text{net}} = F - w.$$

Entering this and the known values into the expression for Newton's second law yields

$$\begin{aligned} a &= \frac{F - w}{m} \\ &= \frac{9.60 \times 10^{-14}\ \text{N} - 3.92 \times 10^{-14}\ \text{N}}{4.00 \times 10^{-15}\ \text{kg}} \\ &= 14.2\ \text{m/s}^2 \end{aligned}$$

Discussion for Part 3

This is an upward acceleration great enough to carry the drop to places where you might not wish to have gasoline.

This worked example illustrates how to apply problem-solving strategies to situations that include topics in different chapters. The first step is to identify the physical principles involved in the problem. The second step is to solve for the unknown using familiar problem-solving strategies. These are found throughout the text, and many worked examples show how to use them for single topics. In this integrated concepts example, you can see how to apply them across several topics. You will find these techniques useful in applications of physics outside a physics course, such as in your profession, in other science disciplines, and in everyday life. The following problems will build your skills in the broad application of physical principles.

Unreasonable Results

The Unreasonable Results exercises for this module have results that are unreasonable because some premise is unreasonable or because certain of the premises are inconsistent with one another. Physical principles applied correctly then produce unreasonable results. The purpose of these problems is to give practice in assessing whether nature is being accurately described, and if it is not to trace the source of difficulty.

Problem-Solving Strategy

To determine if an answer is reasonable, and to determine the cause if it is not, do the following.

1. Solve the problem using strategies as outlined above. Use the format followed in the worked examples in the text to solve the problem as usual.
2. Check to see if the answer is reasonable. Is it too large or too small, or does it have the wrong sign, improper units, and so on?
3. If the answer is unreasonable, look for what specifically could cause the identified difficulty. Usually, the manner in which the answer is unreasonable is an indication of the difficulty. For example, an extremely large Coulomb force could be due to the assumption of an excessively large separated charge.

Section Summary

- Electrostatics is the study of electric fields in static equilibrium.
- In addition to research using equipment such as a Van de Graaff generator, many practical applications of electrostatics exist, including photocopiers, laser printers, ink-jet printers and electrostatic air filters.

Problems & Exercises

1. (a) What is the electric field 5.00 m from the center of the terminal of a Van de Graaff with a 3.00 mC charge, noting that the field is equivalent to that of a point charge at the center of the terminal? (b) At this distance, what force does the field exert on a 2.00 μC charge on the Van de Graaff's belt?
2. (a) What is the direction and magnitude of an electric field that supports the weight of a free electron near the surface of Earth? (b) Discuss what the small value for this field implies regarding the relative strength of the gravitational and electrostatic forces.
3. A simple and common technique for accelerating electrons is shown in Figure, where there is a uniform electric field between two plates. Electrons are released, usually from a hot filament, near the negative plate, and there is a small hole in the positive plate that allows the electrons to continue moving. (a) Calculate the acceleration of the electron if the field

strength is 2.50×10^4 N/C. (b) Explain why the electron will not be pulled back to the positive plate once it moves through the hole.

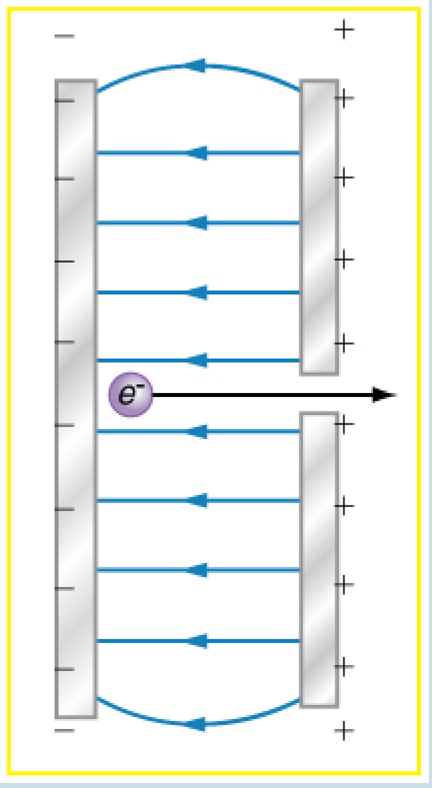


Figure 6. Parallel conducting plates with opposite charges on them create a relatively uniform electric field used to accelerate electrons to the right. Those that go through the hole can be used to make a TV or computer screen glow or to produce X-rays.

4. Earth has a net charge that produces an electric field of approximately 150 N/C downward at its surface. (a) What is the magnitude and sign of the excess charge, noting the electric field of a conducting sphere is equivalent to a point charge at its center? (b) What acceleration will the field produce on a free electron near Earth's surface? (c) What mass object with a single extra electron will have its weight supported by this field?
5. Point charges of 25.0 μC and 45.0μC are placed 0.500 m apart. (a) At what point along the line between them is the electric field zero? (b) What is the electric field halfway between them?
6. What can you say about two charges q_1 and q_2, if the electric field one-fourth of the way from q_1 to q_2 is zero?
7. **Integrated Concepts.** Calculate the angular velocity ω of an electron orbiting a proton in the hydrogen atom, given the radius of the orbit is 0.530×10^{-10} m. You may assume that the proton is stationary and the centripetal force is supplied by Coulomb attraction.
8. **Integrated Concepts.** An electron has an initial velocity of 5.00×10^6 m/s in a uniform 2.00×10^5 N/C strength electric field. The field accelerates the electron in the direction opposite to its initial velocity. (a) What is the direction of the electric field? (b) How far does the electron travel before coming to rest? (c) How long does it take the electron to come to rest? (d) What is the electron's velocity when it returns to its starting point?
9. **Integrated Concepts.** The practical limit to an electric field in air is about 3.00×10^6 N/C. Above this

strength, sparking takes place because air begins to ionize and charges flow, reducing the field. (a) Calculate the distance a free proton must travel in this field to reach 3.00% of the speed of light, starting from rest. (b) Is this practical in air, or must it occur in a vacuum?

10. **Integrated Concepts.** A 5.00 g charged insulating ball hangs on a 30.0 cm long string in a uniform horizontal electric field as shown in Figure 7. Given the charge on the ball is 1.00 μC, find the strength of the field.

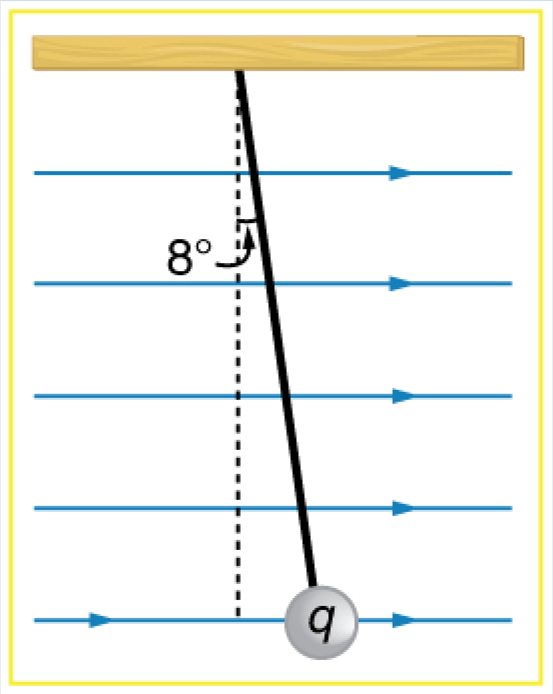


Figure 7. A horizontal electric field causes the charged ball to hang at an angle of 8.00º.

11. **Integrated Concepts.** Figure 8 shows an electron passing between two charged metal plates that create an 100 N/C vertical electric field perpendicular to the electron's original horizontal velocity. (These can be used to change the electron's direction, such as in an oscilloscope.) The initial speed of the electron is 3.00 $\times 10^6$ m/s, and the horizontal distance it travels in the uniform field is 4.00 cm. (a) What is its vertical

deflection? (b) What is the vertical component of its final velocity? (c) At what angle does it exit? Neglect any edge effects.

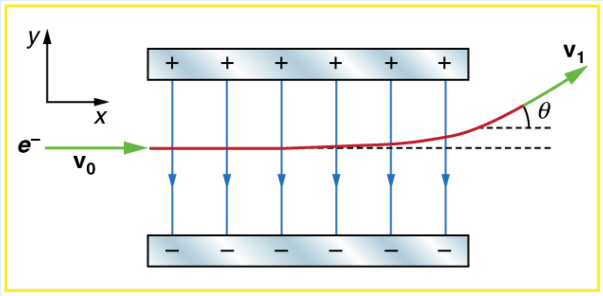


Figure 8.

12. **Integrated Concepts.** The classic Millikan oil drop experiment was the first to obtain an accurate measurement of the charge on an electron. In it, oil drops were suspended against the gravitational force by a vertical electric field. (See Figure 9.) Given the oil drop to be 1.00 μm in radius and have a density of 920 kg/m^3: (a) Find the weight of the drop. (b) If the drop has a single excess electron, find the electric field strength needed to balance its weight.

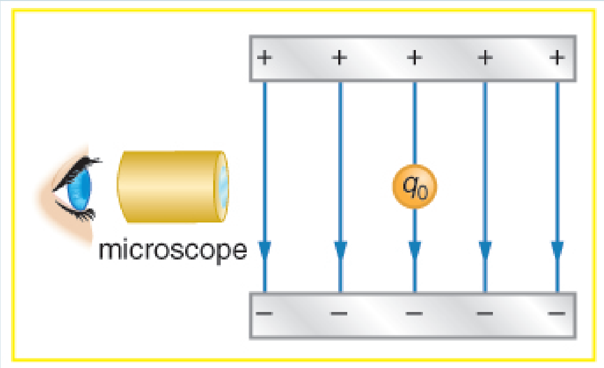


Figure 9. In the Millikan oil drop experiment, small drops can be suspended in an electric field by the force exerted on a single excess electron. Classically, this experiment was used to determine the electron charge q_e by measuring the electric field and mass of the drop.

13. **Integrated Concepts.** (a) In Figure 10, four equal charges *q* lie on the corners of a square. A fifth charge Q is on a mass m directly above the center of the square, at a height equal to the length d of one side of the square. Determine the magnitude of *q* in terms of Q, *m*, and *d*, if the Coulomb force is to equal the weight of *m*. (b) Is this equilibrium stable or unstable? Discuss.

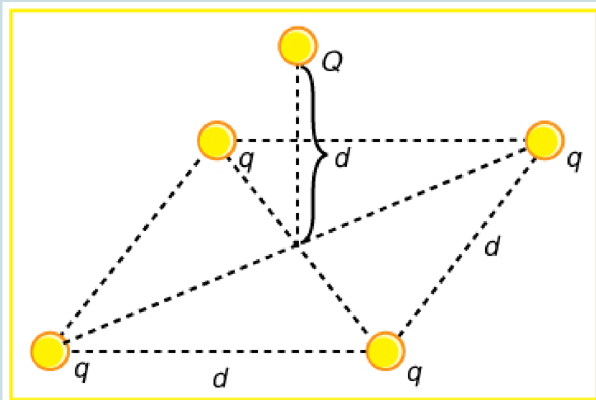


Figure 10. Four equal charges on the corners of a horizontal square support the weight of a fifth charge located directly above the center of the square.

14. **Unreasonable Results.** (a) Calculate the electric field strength near a 10.0 cm diameter conducting sphere that has 1.00 C of excess charge on it. (b) What is unreasonable about this result? (c) Which assumptions are responsible?
15. **Unreasonable Results.** (a) Two 0.500 g raindrops in a thunderhead are 1.00 cm apart when they each acquire 1.00 mC charges. Find their acceleration. (b) What is unreasonable about this result? (c) Which premise or assumption is responsible?
16. **Unreasonable Results.** A wrecking yard inventor wants to pick up cars by charging a 0.400 m diameter ball and inducing an equal and opposite charge on the car. If a car has a 1000 kg mass and the ball is to be able to lift it from a distance of 1.00 m: (a) What

minimum charge must be used? (b) What is the electric field near the surface of the ball? (c) Why are these results unreasonable? (d) Which premise or assumption is responsible?

17. **Construct Your Own Problem.** Consider two insulating balls with evenly distributed equal and opposite charges on their surfaces, held with a certain distance between the centers of the balls. Construct a problem in which you calculate the electric field (magnitude and direction) due to the balls at various points along a line running through the centers of the balls and extending to infinity on either side. Choose interesting points and comment on the meaning of the field at those points. For example, at what points might the field be just that due to one ball and where does the field become negligibly small? Among the things to be considered are the magnitudes of the charges and the distance between the centers of the balls. Your instructor may wish for you to consider the electric field off axis or for a more complex array of charges, such as those in a water molecule.
18. **Construct Your Own Problem.** Consider identical spherical conducting space ships in deep space where gravitational fields from other bodies are negligible compared to the gravitational attraction between the ships. Construct a problem in which you place identical excess charges on the space ships to exactly counter their gravitational attraction. Calculate the amount of excess charge needed. Examine whether that charge depends on the distance between the centers of the ships, the masses

of the ships, or any other factors. Discuss whether this would be an easy, difficult, or even impossible thing to do in practice.

Glossary

Van de Graaff generator: a machine that produces a large amount of excess charge, used for experiments with high voltage

electrostatics: the study of electric forces that are static or slow-moving

photoconductor: a substance that is an insulator until it is exposed to light, when it becomes a conductor

xerography: a dry copying process based on electrostatics

grounded: connected to the ground with a conductor, so that charge flows freely to and from the Earth to the grounded object

laser printer: uses a laser to create a photoconductive image on a drum, which attracts dry ink particles that are then rolled onto a sheet of paper to print a high-quality copy of the image

ink-jet printer: small ink droplets sprayed with an electric charge are controlled by electrostatic plates to create images on paper

electrostatic precipitators: filters that apply charges to particles in the air, then attract those charges to a filter, removing them from the airstream

Selected Solutions to Problems & Exercises

2. (a) 5.58×10^{-11} N/C; (b)the coulomb force is extraordinarily stronger than gravity

4. (a) -6.76×10^{5} C; (b) 2.63×10^{13} m/s^{2} (upward); (c) 2.45×10^{-18} kg

6. The charge q_2 is 9 times greater than q_1.

PART II

ELECTRIC POTENTIAL AND ELECTRIC FIELD

11. Introduction to Electric Potential and Electric Energy

Figure 1. Automated external defibrillator unit (AED) *(credit:* U.S. *Defense Department photo/Tech. Sgt. Suzanne M. Day)*

In Electric Charge and Electric Field, we just scratched the surface (or at least rubbed it) of electrical phenomena. Two of the most familiar aspects of electricity are its energy and *voltage*. We know, for example, that great amounts of electrical energy can be stored in batteries, are transmitted cross-country through power lines, and may jump from clouds to explode the sap of trees. In a similar manner, at molecular levels, *ions* cross cell membranes and transfer information. We also know about voltages associated with electricity. Batteries are typically a few volts, the outlets in your home produce 120 volts, and power lines can be as high as hundreds of thousands of volts. But energy and voltage are not the same thing. A motorcycle battery, for example, is small and would not be very successful in replacing the much larger car battery, yet each has the same voltage. In this chapter, we shall examine the relationship between voltage and electrical energy and begin to explore some of the many applications of electricity.

12. Video: Electric Potential

Watch the following Physics Concept Trailer to see how impressive new technologies are able to measure and utilize electric potentials.

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://library.achievingthedream.org/austinccphysics2/?p=31#oembed-1

13. Electric Potential Energy: Potential Difference

Learning Objectives

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

- Define electric potential and electric potential energy.
- Describe the relationship between potential difference and electrical potential energy.
- Explain electron volt and its usage in submicroscopic process.
- Determine electric potential energy given potential difference and amount of charge.

When a free positive charge q is accelerated by an electric field, such as shown in Figure 1, it is given kinetic energy. The process is analogous to an object being accelerated by a gravitational field. It is as if the charge is going down an electrical hill where its electric potential energy is converted to kinetic energy. Let us explore the work done on a charge q by the electric field in this process, so that we may develop a definition of electric potential energy.

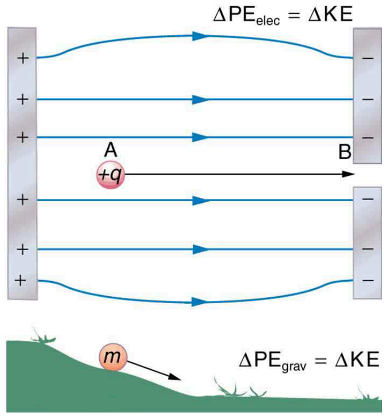


Figure 1. A charge accelerated by an electric field is analogous to a mass going down a hill. In both cases potential energy is converted to another form. Work is done by a force, but since this force is conservative, we can write W = −ΔPE.

The electrostatic or Coulomb force is conservative, which means that the work done on q is independent of the path taken. This is exactly analogous to the gravitational force in the absence of dissipative forces such as friction. When a force is conservative, it is possible to define a potential energy associated with the force, and it is usually easier to deal with the potential energy (because it depends only on position) than to calculate the work directly.

We use the letters PE to denote electric potential energy, which has units of joules (J). The change in potential energy, ΔPE, is crucial, since the work done by a conservative force is the negative of the change in potential energy; that is, W = −ΔPE. For example, work W done to accelerate a positive charge from rest is positive and results from a loss in PE, or a negative ΔPE. There must be a minus sign in front of ΔPE to make W positive. PE can be found at any point by taking one point as a reference and calculating the work needed to move a charge to the other point.

Potential Energy

$W = -\Delta PE$. For example, work W done to accelerate a positive charge from rest is positive and results from a loss in PE, or a negative ΔPE. There must be a minus sign in front of ΔPE to make W positive. PE can be found at any point by taking one point as a reference and calculating the work needed to move a charge to the other point.

Gravitational potential energy and electric potential energy are quite analogous. Potential energy accounts for work done by a conservative force and gives added insight regarding energy and energy transformation without the necessity of dealing with the force directly. It is much more common, for example, to use the concept of voltage (related to electric potential energy) than to deal with the Coulomb force directly.

Calculating the work directly is generally difficult, since $W = Fd \cos\theta$ and the direction and magnitude of F can be complex for multiple charges, for odd-shaped objects, and along arbitrary paths. But we do know that, since $F = qE$, the work, and hence ΔPE, is proportional to the test charge q. To have a physical quantity that is independent of test charge, we define *electric potential* V (or simply potential, since electric is understood) to be the potential energy per unit charge $V = \frac{\text{PE}}{q}$.

Electric Potential

This is the electric potential energy per unit charge.

$$V = \frac{\text{PE}}{q}$$

Since PE is proportional to *q* , the dependence on *q* cancels. Thus V does not depend on *q*. The change in potential energy ΔPE is crucial, and so we are concerned with the difference in potential or potential difference ΔV between two points, where

$$\Delta V = V_{\text{B}} - V_{\text{A}} = \frac{\Delta\text{PE}}{q}$$

The *potential difference* between points A and B, $V_B - V_A$, is thus defined to be the change in potential energy of a charge *q* moved from A to B, divided by the charge. Units of potential difference are joules per coulomb, given the name volt (V) after Alessandro Volta.

$$1\text{V} = 1\frac{\text{J}}{\text{C}}$$

Potential Difference

The potential difference between points A and B,

$V_B - V_A$, is defined to be the change in potential energy of a charge q moved from A to B, divided by the charge. Units of potential difference are joules per coulomb, given the name volt (V) after Alessandro Volta.

$$1\mathrm{V} = 1\frac{\mathrm{J}}{\mathrm{C}}$$

The familiar term *voltage* is the common name for potential difference. Keep in mind that whenever a voltage is quoted, it is understood to be the potential difference between two points. For example, every battery has two terminals, and its voltage is the potential difference between them. More fundamentally, the point you choose to be zero volts is arbitrary. This is analogous to the fact that gravitational potential energy has an arbitrary zero, such as sea level or perhaps a lecture hall floor.

In summary, the relationship between potential difference (or voltage) and electrical potential energy is given by $\Delta V = \dfrac{\Delta \mathrm{PE}}{q}$ and $\Delta\mathrm{PE} = q\Delta\mathrm{V}$.

Potential Difference and Electrical Potential Energy

The relationship between potential difference (or voltage) and electrical potential energy is given by

$$\Delta V = \frac{\Delta \text{PE}}{q} \text{ and } \Delta\text{PE} = q\Delta\text{V}$$

The second equation is equivalent to the first.

Voltage is not the same as energy. Voltage is the energy per unit charge. Thus a motorcycle battery and a car battery can both have the same voltage (more precisely, the same potential difference between battery terminals), yet one stores much more energy than the other since $\Delta\text{PE} = q\Delta\text{V}$. The car battery can move more charge than the motorcycle battery, although both are 12 V batteries.

Example 1. Calculating Energy

Suppose you have a 12.0 V motorcycle battery that can move 5000 C of charge, and a 12.0 V car battery that can move 60,000 C of charge. How much energy does each deliver? (Assume that the numerical value of each charge is accurate to three significant figures.)

Strategy

To say we have a 12.0 V battery means that its terminals have a 12.0 V potential difference. When such a battery moves charge, it puts the charge through a potential difference of 12.0 V, and the charge is given a change in potential energy equal to $\Delta\text{PE} = q\Delta\text{V}$.

So to find the energy output, we multiply the charge moved by the potential difference.

Solution

For the motorcycle battery, q = 5000 C and ΔV = 12.0 V. The total energy delivered by the motorcycle battery is

$$\begin{aligned}\Delta \mathrm{PE}_{\mathrm{cycle}} &= (5000\ \mathrm{C})\,(12.0\ \mathrm{V}) \\ &= (5000\ \mathrm{C})\,(12.0\ \mathrm{J/C}) \\ &= 6.00 \times 10^4\ \mathrm{J}\end{aligned}$$

Similarly, for the car battery, q = 60,000 C and

$$\begin{aligned}\Delta \mathrm{PE}_{\mathrm{car}} &= (60,000\ \mathrm{C})\,(12.0\ \mathrm{V}) \\ &= 7.20 \times 10^5\ \mathrm{J}\end{aligned}$$

Discussion

While voltage and energy are related, they are not the same thing. The voltages of the batteries are identical, but the energy supplied by each is quite different. Note also that as a battery is discharged, some of its energy is used internally and its terminal voltage drops, such as when headlights dim because of a low car battery. The energy supplied by the battery is still calculated as in this example, but not all of the energy is available for external use.

Note that the energies calculated in the previous example are absolute values. The change in potential energy for the battery is negative, since it loses energy. These batteries, like many electrical systems, actually move negative charge—electrons in particular. The batteries repel electrons from their negative terminals (A) through whatever circuitry is involved and attract them to their positive terminals (B) as shown in Figure 2. The change in potential is $\Delta V = V_B - V_A = +12\text{ V}$ and the charge q is negative, so that $\Delta PE = q\Delta V$ is negative, meaning the potential energy of the battery has decreased when q has moved from A to B.

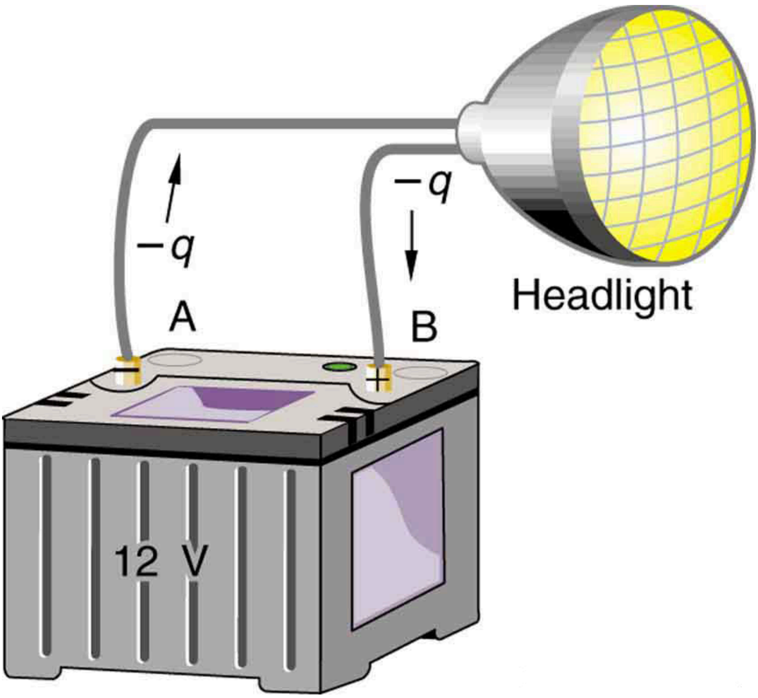


Figure 2. A battery moves negative charge from its negative terminal through a headlight to its positive terminal. Appropriate combinations of chemicals in the battery separate charges so that the negative terminal has an excess of negative charge, which is repelled by it and attracted to the excess positive charge on the other terminal. In terms of potential, the positive terminal is at a higher voltage than the negative. Inside the battery, both positive and negative charges move.

Example 2. How Many Electrons Move through a Headlight Each Second?

When a 12.0 V car battery runs a single 30.0 W headlight, how many electrons pass through it each second?

Strategy

To find the number of electrons, we must first find the charge that moved in 1.00 s. The charge moved is related to voltage and energy through the equation ΔPE = *q*ΔV. A 30.0 W lamp uses 30.0 joules per second. Since the battery loses energy, we have ΔPE = −30.0 J and, since the electrons are going from the negative terminal to the positive, we see that ΔV = +12.0V.

Solution

To find the charge *q* moved, we solve the equation ΔPE = qΔV: $q=\frac{\Delta\text{PE}}{\Delta V}$.

Entering the values for ΔPE and ΔV, we get

$$q=\frac{-30.0\text{ J}}{+12.0\text{ V}}=\frac{-30.0\text{ J}}{+12.0\text{ J/C}}-2.50\text{ C}$$

The number of electrons n_e is the total charge divided by the charge per electron. That is,

$$n_e = \frac{-2.50\ \text{C}}{-1.60 \times 10^{-19}\ \text{C/e}^-} = 1.56 \times 10^{19}\ \text{electrons}$$

Discussion

This is a very large number. It is no wonder that we do not ordinarily observe individual electrons with so many being present in ordinary systems. In fact, electricity had been in use for many decades before it was determined that the moving charges in many circumstances were negative. Positive charge moving in the opposite direction of negative charge often produces identical effects; this makes it difficult to determine which is moving or whether both are moving.

The Electron Volt

The energy per electron is very small in macroscopic situations like that in the previous example—a tiny fraction of a joule. But on a submicroscopic scale, such energy per particle (electron, proton, or ion) can be of great importance. For example, even a tiny fraction of a joule can be great enough for these particles to destroy organic molecules and harm living tissue. The particle may do its damage by direct collision, or it may create harmful x rays, which can also inflict damage. It is useful to have an energy unit related to submicroscopic effects. Figure 3 shows a situation related to the definition of such an energy unit. An electron is accelerated between two charged metal plates as it might be in an old-model television tube or oscilloscope. The electron is given kinetic energy that is later converted to another form—light in the television tube, for example. (Note that downhill for the electron is uphill for a positive charge.) Since energy is related to voltage by ΔPE = qΔV, we can think of the joule as a coulomb-volt.

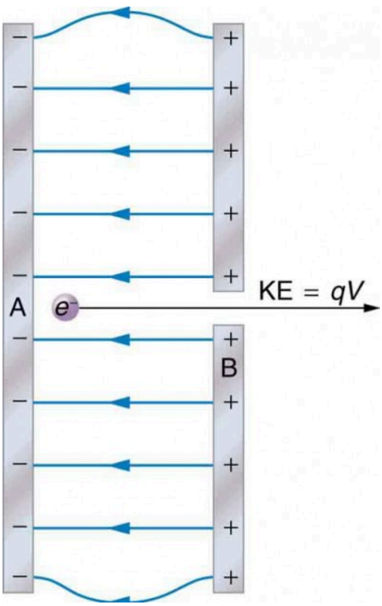


Figure 3. A typical electron gun accelerates electrons using a potential difference between two metal plates. The energy of the electron in electron volts is numerically the same as the voltage between the plates. For example, a 5000 V potential difference produces 5000 eV electrons.

On the submicroscopic scale, it is more convenient to define an energy unit called the *electron volt* (eV), which is the energy given to a fundamental charge accelerated through a potential difference of 1 V. In equation form,

$$\begin{aligned} 1\text{eV} &= \left(1.60 \times 10^{-19}\ \text{C}\right)(1\ \text{V}) = \left(1.60 \times 10^{-19}\ \text{C}\right)(1\ \text{J/C}) \\ &= 1.60 \times 10^{-19}\ \text{J} \end{aligned}$$

Electron Volt

On the submicroscopic scale, it is more convenient to define an energy unit called the electron volt (eV), which is the energy given to a fundamental charge accelerated through a potential difference of 1 V. In equation form,

$$\begin{aligned} 1\text{eV} &= \left(1.60 \times 10^{-19}\ \text{C}\right)(1\ \text{V}) = \left(1.60 \times 10^{-19}\ \text{C}\right)(1\ \text{J/C}) \\ &= 1.60 \times 10^{-19}\ \text{J} \end{aligned}$$

An electron accelerated through a potential difference of 1 V is given an energy of 1 eV. It follows that an electron accelerated through 50 V is given 50 eV. A potential difference of 100,000 V (100 kV) will give an electron an energy of 100,000 eV (100 keV), and so on. Similarly, an ion with a double positive charge accelerated through 100 V will be given 200 eV of energy. These simple relationships between accelerating voltage and particle charges make the electron volt a simple and convenient energy unit in such circumstances.

Making Connections: Energy Units

The electron volt (eV) is the most common energy unit for submicroscopic processes. This will be particularly noticeable in the chapters on modern physics. Energy is so important to so many subjects that there is a tendency to define a special energy unit for each major topic. There are, for example, calories for food energy, kilowatt-hours for electrical energy, and therms for natural gas energy.

The electron volt is commonly employed in submicroscopic processes—chemical valence energies and molecular and nuclear binding energies are among the quantities often expressed in electron volts. For example, about 5 eV of energy is required to break up certain organic molecules. If a proton is accelerated from rest through a potential difference of 30 kV, it is given an energy of 30 keV (30,000 eV) and it can break up as many as 6000 of these molecules (30,000 eV ÷ 5 eV per molecule= 6000 molecules). Nuclear decay energies are on the order of 1 MeV (1,000,000 eV) per event and can, thus, produce significant biological damage.

Conservation of Energy

The total energy of a system is conserved if there is no net addition (or subtraction) of work or heat transfer. For conservative forces, such as the electrostatic force, conservation of energy states that mechanical energy is a constant.

Mechanical energy is the sum of the kinetic energy and potential energy of a system; that is, KE+PE = constant. A loss of PE of a charged particle becomes an increase in its KE. Here PE is the electric potential energy. Conservation of energy is stated in equation form as KE + PE = constant or $KE_i + PE_i = KE_f + PE_f$, where i and f stand for initial and final conditions. As we have found many times before, considering energy can give us insights and facilitate problem solving.

Example 3. Electrical Potential Energy Converted to Kinetic Energy

Calculate the final speed of a free electron accelerated from rest through a potential difference of 100 V. (Assume that this numerical value is accurate to three significant figures.)

Strategy

We have a system with only conservative forces. Assuming the electron is accelerated in a vacuum, and neglecting the gravitational force (we will check on this assumption later), all of the electrical potential energy is converted into kinetic energy. We can identify the initial and final forms of energy to be $KE_i = 0$, $KE_f = \frac{1}{2}mv^2$, $PE_i = qV$, and $PE_f = 0$.

Solution

Conservation of energy states that $KE_i + PE_i = KE_f + PE_f$.

Entering the forms identified above, we obtain

$$qV = \frac{mv^2}{2}.$$

We solve this for v:

$$v = \sqrt{\frac{2qV}{m}}$$

Entering values for q, V, and m gives

$$\begin{aligned} v &= \sqrt{\frac{2\left(-1.60\times10^{-19}\text{ C}\right)\left(-100\text{ J/C}\right)}{9.11\times10^{-31}\text{kg}}} \\ &= 5.93\times10^{6}\text{ m/s} \end{aligned}$$

Discussion

Note that both the charge and the initial voltage are negative, as in Figure 3. From the discussions in Electric Charge and Electric Field, we know that electrostatic forces on small particles are generally very large compared with the gravitational force. The large final speed confirms that the gravitational force is indeed negligible here. The large speed also indicates how easy it is to accelerate electrons with small voltages because of their very small mass. Voltages much higher than the 100 V in this problem are typically used in electron guns. Those higher voltages produce electron speeds so great that relativistic effects must be taken into account. That is why a low voltage is considered (accurately) in this example.

Section Summary

- Electric potential is potential energy per unit charge.
- The potential difference between points A and B, $V_B - V_A$, defined to be the change in potential energy of a charge q moved from A to B, is equal to the change in potential energy divided by the charge, Potential difference is commonly called voltage, represented by the symbol ΔV: $\Delta V = \frac{\Delta \text{PE}}{q}$ and $\Delta\text{PE} = q\Delta\text{V}$.
- An electron volt is the energy given to a fundamental charge accelerated through a potential difference of 1 V. In equation form,
$$\begin{aligned}1\text{ eV} &= \left(1.60\times10^{-19}\text{C}\right)(1V) = \left(1.60\times10^{-19}\text{C}\right)(1J/C)\\ &= 1.60\times10^{-19}\text{J}.\end{aligned}$$
- Mechanical energy is the sum of the kinetic energy and potential energy of a system, that is, KE + PE. This sum is a constant.

Conceptual Questions

1. Voltage is the common word for potential difference. Which term is more descriptive, voltage or potential difference?
2. If the voltage between two points is zero, can a test charge be moved between them with zero net work being done? Can this necessarily be done without exerting a force? Explain.
3. What is the relationship between voltage and energy? More precisely, what is the relationship

between potential difference and electric potential energy?
4. Voltages are always measured between two points. Why?
5. How are units of volts and electron volts related? How do they differ?

Problems & Exercises

1. Find the ratio of speeds of an electron and a negative hydrogen ion (one having an extra electron) accelerated through the same voltage, assuming non-relativistic final speeds. Take the mass of the hydrogen ion to be 1.67×10^{-27} kg.
2. An evacuated tube uses an accelerating voltage of 40 kV to accelerate electrons to hit a copper plate and produce x rays. Non-relativistically, what would be the maximum speed of these electrons?
3. A bare helium nucleus has two positive charges and a mass of 6.64×10^{-27} kg. (a) Calculate its kinetic energy in joules at 2.00% of the speed of light. (b) What is this in electron volts? (c) What voltage would be needed to obtain this energy?
4. **Integrated Concepts.** Singly charged gas ions are accelerated from rest through a voltage of 13.0 V. At what temperature will the average kinetic energy of gas molecules be the same as that given these ions?
5. **Integrated Concepts.** The temperature near the

center of the Sun is thought to be 15 million degrees Celsius (1.5×10^7 °C). Through what voltage must a singly charged ion be accelerated to have the same energy as the average kinetic energy of ions at this temperature?

6. **Integrated Concepts.** (a) What is the average power output of a heart defibrillator that dissipates 400 J of energy in 10.0 ms? (b) Considering the high-power output, why doesn't the defibrillator produce serious burns?
7. **Integrated Concepts.** A lightning bolt strikes a tree, moving 20.0 C of charge through a potential difference of 1.00×10^2 MV. (a) What energy was dissipated? (b) What mass of water could be raised from 15°C to the boiling point and then boiled by this energy? (c) Discuss the damage that could be caused to the tree by the expansion of the boiling steam.
8. **Integrated Concepts.** A 12.0 V battery-operated bottle warmer heats 50.0 g of glass, 2.50×10^2 g of baby formula, and 2.00×10^2 g of aluminum from 20.0°C to 90.0°C. (a) How much charge is moved by the battery? (b) How many electrons per second flow if it takes 5.00 min to warm the formula? (Hint: Assume that the specific heat of baby formula is about the same as the specific heat of water.)
9. **Integrated Concepts.** A battery-operated car utilizes a 12.0 V system. Find the charge the batteries must be able to move in order to accelerate the 750 kg car from rest to 25.0 m/s, make it climb a 2.00×10^2 m high hill, and then cause it to travel at a constant 25.0 m/s by exerting a 5.00×10^2 N force for an hour.

10. **Integrated Concepts.** Fusion probability is greatly enhanced when appropriate nuclei are brought close together, but mutual Coulomb repulsion must be overcome. This can be done using the kinetic energy of high-temperature gas ions or by accelerating the nuclei toward one another. (a) Calculate the potential energy of two singly charged nuclei separated by 1.00×10^{-12} m by finding the voltage of one at that distance and multiplying by the charge of the other. (b) At what temperature will atoms of a gas have an average kinetic energy equal to this needed electrical potential energy?
11. **Unreasonable Results.** (a) Find the voltage near a 10.0 cm diameter metal sphere that has 8.00 C of excess positive charge on it. (b) What is unreasonable about this result? (c) Which assumptions are responsible?
12. **Construct Your Own Problem.** Consider a battery used to supply energy to a cellular phone. Construct a problem in which you determine the energy that must be supplied by the battery, and then calculate the amount of charge it must be able to move in order to supply this energy. Among the things to be considered are the energy needs and battery voltage. You may need to look ahead to interpret manufacturer's battery ratings in ampere-hours as energy in joules.

Glossary

electric potential: potential energy per unit charge

potential difference (or voltage): change in potential energy of a charge moved from one point to another, divided by the charge; units of potential difference are joules per coulomb, known as volt

electron volt: the energy given to a fundamental charge accelerated through a potential difference of one volt

mechanical energy: sum of the kinetic energy and potential energy of a system; this sum is a constant

Selected Solutions to Problems & Exercises

1. 42.8

4. 1.00×10^5 K

6. (a) 4×10^4 W; (b) A defibrillator does not cause serious burns because the skin conducts electricity well at high voltages, like those used in defibrillators. The gel used aids in the transfer of energy to the body, and the skin doesn't absorb the energy, but rather lets it pass through to the heart.

8. (a) 7.40×10^3 C; (b) 1.54×10^{20} electrons per second

9. 3.89×10^6 C

11. (a) 1.44×10^{12} V; (b) This voltage is very high. A 10.0 cm diameter sphere could never maintain this voltage; it would discharge; (c) An 8.00 C charge is more charge than can reasonably be accumulated on a sphere of that size.

14. Electric Potential in a Uniform Electric Field

Learning Objectives

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

- Describe the relationship between voltage and electric field.
- Derive an expression for the electric potential and electric field.
- Calculate electric field strength given distance and voltage.

In the previous section, we explored the relationship between voltage and energy. In this section, we will explore the relationship between voltage and electric field. For example, a uniform electric field **E** is produced by placing a potential difference (or voltage) ΔV across two parallel metal plates, labeled A and B. (See Figure 1.)

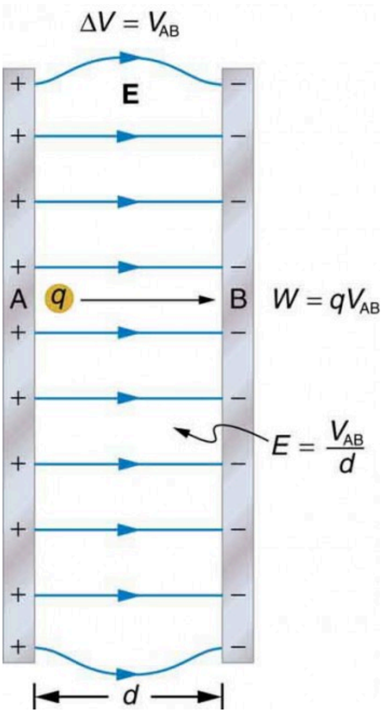


Figure 1. The relationship between V *and* E *for parallel conducting plates is [latex]E=\frac{V}{d}\\[/latex]. (Note that* $\Delta V = V_{AB}$ *in magnitude. For a charge that is moved from plate* A *at higher potential to plate* B *at lower potential, a minus sign needs to be included as follows:* $-\Delta V = V_A - V_B = V_{AB}$*. See the text for details.)*

Examining this will tell us what voltage is needed to produce a certain electric field strength; it will also reveal a more fundamental relationship between electric potential and electric field. From a physicist's point of view, either ΔV or **E** can be used to describe any charge distribution. ΔV is most closely tied to energy, whereas **E** is most closely related to force. ΔV is a *scalar* quantity and has no direction, while **E** is a *vector* quantity, having both magnitude and direction. (Note that the magnitude of the electric field strength, a scalar quantity, is represented by E below.) The relationship between ΔV and **E** is revealed by calculating the work done by the force in moving a charge from point A to point B.

But, as noted in [Electric Potential Energy: Potential Difference](), this is complex for arbitrary charge distributions, requiring calculus. We therefore look at a uniform electric field as an interesting special case.

The work done by the electric field in Figure 1 to move a positive charge q from A, the positive plate, higher potential, to B, the negative plate, lower potential, is

$$W = -\Delta PE = -q\Delta V.$$

The potential difference between points A and B is

$$-\Delta V = -(V_B - V_A) = V_A - V_B = V_{AB}.$$

Entering this into the expression for work yields $W = qV_{AB}$.

Work is $W = Fd \cos \theta$; here $\cos \theta = 1$, since the path is parallel to the field, and so W = Fd. Since F = qE, we see that $W = qEd$. Substituting this expression for work into the previous equation gives $qEd = qV_{AB}$.

The charge cancels, and so the voltage between points A and B is seen to be

$$\begin{cases} V_{\mathrm{AB}} & = & Ed \\ E & = & \frac{V_{\mathrm{AB}}}{d} \end{cases} \text{(uniform E − field only)}$$

where d is the distance from A to B, or the distance between the plates in Figure 1. Note that the above equation implies the units for electric field are volts per meter. We already know the units for electric field are newtons per coulomb; thus the following relation among units is valid: 1 N/C = 1 V/m.

Voltage between Points A and B

$$\begin{cases} V_{\mathrm{AB}} & = & Ed \\ E & = & \frac{V_{\mathrm{AB}}}{d} \end{cases} \text{(uniform E − field only)}$$

where d is the distance from A to B, or the distance between the plates.

Example 1. What Is the Highest Voltage Possible between Two Plates?

Dry air will support a maximum electric field strength of about 3.0×10^6 V/m. Above that value, the field creates enough ionization in the air to make the air a conductor. This allows a discharge or spark that reduces the field. What, then, is the maximum voltage between two parallel conducting plates separated by 2.5 cm of dry air?

Strategy

We are given the maximum electric field E between the plates and the distance d between them. The equation $V_{AB} = Ed$ can thus be used to calculate the maximum voltage.

Solution

The potential difference or voltage between the plates is

$$V_{AB} = Ed.$$

Entering the given values for E and d gives

$$V_{AB} = (3.0 \times 10^6 \text{ V/m})(0.025 \text{ m}) \; 7.5 \times 10^4 \text{ V or } V_{AB} = 75 \text{ kV}.$$

(The answer is quoted to only two digits, since the maximum field strength is approximate.)

Discussion

One of the implications of this result is that it takes about 75 kV to make a spark jump across a 2.5 cm (1 in.) gap, or 150 kV for a 5 cm spark. This limits the voltages that can exist between conductors, perhaps on a power transmission line. A smaller voltage will cause a spark if there are points on the surface, since points create greater fields than smooth surfaces. Humid air breaks down at a lower field strength, meaning that a smaller voltage will make a spark jump through humid air. The largest voltages can be built up, say with static electricity, on dry days.

Figure 2. A spark chamber is used to trace the paths of high-energy particles. Ionization created by the particles as they pass through the gas between the plates allows a spark to jump. The sparks are perpendicular to the plates, following electric field lines between them. The potential difference between adjacent plates is not high enough to cause sparks without the ionization produced by particles from accelerator experiments (or cosmic rays). (credit: Daderot, Wikimedia Commons)

Example 2. Field and Force inside an Electron Gun

1. An electron gun has parallel plates separated by 4.00 cm and gives electrons 25.0 keV of energy. What is the electric field strength between the plates?
2. What force would this field exert on a piece of plastic with a 0.500 µC charge that gets between the plates?

Strategy

Since the voltage and plate separation are given, the electric field strength can be calculated directly from the expression $E=\frac{V_{\text{AB}}}{d}$. Once the electric field strength is known, the force on a charge is found using **F** = q**E**. Since the electric field is in only one direction, we can write this equation in terms of the magnitudes, $F = qE$.

Solution for Part 1

The expression for the magnitude of the electric field between two uniform metal plates is

$$E=\frac{V_{\text{AB}}}{d}.$$

Since the electron is a single charge and is given 25.0 keV of energy, the potential difference must be 25.0 kV. Entering this value for V_{AB} and the plate separation of 0.0400 m, we obtain

$$E = \frac{25.0\ \text{kV}}{0.0400\ \text{m}} = 6.25 \times 10^5\ \text{V/m}$$

Solution for Part 2

The magnitude of the force on a charge in an electric field is obtained from the equation $F = qE$.

Substituting known values gives

$$F = (0.500 \times 10^{-6}\ \text{C})(6.25 \times 10^5\ \text{V/m}) = 0.313\ \text{N}.$$

Discussion

Note that the units are newtons, since 1 V/m = 1 N/C. The force on the charge is the same no matter where the charge is located between the plates. This is because the electric field is uniform between the plates.

In more general situations, regardless of whether the electric field is uniform, it points in the direction of decreasing potential, because the force on a positive charge is in the direction of **E** and also in the direction of lower potential V. Furthermore, the magnitude of **E** equals the rate of decrease of V with distance. The faster V decreases over distance, the greater the electric field. In equation form, the general relationship between voltage and electric field is

$$E = -\frac{\Delta V}{\Delta s},$$

where Δs is the distance over which the change in potential, ΔV, takes place. The minus sign tells us that **E** points in the direction of decreasing potential. The electric field is said to be the *gradient* (as in grade or slope) of the electric potential.

Relationship between Voltage and Electric Field

In equation form, the general relationship between voltage and electric field is

$$E = -\frac{\Delta V}{\Delta s},$$

where Δs is the distance over which the change in potential, ΔV, takes place. The minus sign tells us that **E** points in the direction of decreasing potential. The electric field is said to be the *gradient* (as in grade or slope) of the electric potential.

For continually changing potentials, ΔV and Δs become infinitesimals and differential calculus must be employed to determine the electric field.

Section Summary

- The voltage between points A and B is

$$\begin{cases} V_{\mathrm{AB}} & = & Ed \\ E & = & \frac{V_{\mathrm{AB}}}{d} \end{cases} \text{(uniform E − field only)}$$

where d is the distance from A to B, or the distance between the plates.

- In equation form, the general relationship between voltage and electric field is

$$E = -\frac{\Delta V}{\Delta s},$$

where Δs is the distance over which the change in potential, ΔV, takes place. The minus sign tells us that **E** points in the direction of decreasing potential.) The electric field is said to be the gradient (as in grade or slope) of the electric potential.

Conceptual Questions

1. Discuss how potential difference and electric field strength are related. Give an example.
2. What is the strength of the electric field in a region where the electric potential is constant?
3. Will a negative charge, initially at rest, move toward higher or lower potential? Explain why.

Problems & Exercises

1. Show that units of V/m and N/C for electric field

strength are indeed equivalent.

2. What is the strength of the electric field between two parallel conducting plates separated by 1.00 cm and having a potential difference (voltage) between them of 1.50×10^4 V?
3. The electric field strength between two parallel conducting plates separated by 4.00 cm is 7.50×10^4 V/m. (a) What is the potential difference between the plates? (b) The plate with the lowest potential is taken to be at zero volts. What is the potential 1.00 cm from that plate (and 3.00 cm from the other)?
4. How far apart are two conducting plates that have an electric field strength of 4.50×10^3 V/m between them, if their potential difference is 15.0 kV?
5. (a) Will the electric field strength between two parallel conducting plates exceed the breakdown strength for air (3.0×10^6 V/m) if the plates are separated by 2.00 mm and a potential difference of 5.0×10^3 V is applied? (b) How close together can the plates be with this applied voltage?
6. The voltage across a membrane forming a cell wall is 80.0 mV and the membrane is 9.00 nm thick. What is the electric field strength? (The value is surprisingly large, but correct. Membranes are discussed in Capacitors and Dielectrics and Nerve Conduction—Electrocardiograms.) You may assume a uniform electric field.
7. Membrane walls of living cells have surprisingly large electric fields across them due to separation of ions. (Membranes are discussed in some detail in Nerve Conduction—Electrocardiograms.) What is the voltage across an 8.00 nm–thick membrane if the

electric field strength across it is 5.50 MV/m? You may assume a uniform electric field.

8. Two parallel conducting plates are separated by 10.0 cm, and one of them is taken to be at zero volts. (a) What is the electric field strength between them, if the potential 8.00 cm from the zero volt plate (and 2.00 cm from the other) is 450 V? (b) What is the voltage between the plates?
9. Find the maximum potential difference between two parallel conducting plates separated by 0.500 cm of air, given the maximum sustainable electric field strength in air to be 3.0×10^6 V/m.
10. A doubly charged ion is accelerated to an energy of 32.0 keV by the electric field between two parallel conducting plates separated by 2.00 cm. What is the electric field strength between the plates?
11. An electron is to be accelerated in a uniform electric field having a strength of 2.00×10^6 V/m. (a) What energy in keV is given to the electron if it is accelerated through 0.400 m? (b) Over what distance would it have to be accelerated to increase its energy by 50.0 GeV?

Glossary

scalar: physical quantity with magnitude but no direction

vector: physical quantity with both magnitude and direction

Selected Solutions to Problems & Exercises

3. (a) 3.00 kV; (b) 750 V

5. (a) No. The electric field strength between the plates is 2.5×10^6 V/m, which is lower than the breakdown strength for air (3.0×10^6 V/m}); (b) 1.7 mm

7. 44.0 mV

9. 15 kV

11. (a) 800 KeV; (b) 25.0 km

15. Electrical Potential Due to a Point Charge

Learning Objectives

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

- Explain point charges and express the equation for electric potential of a point charge.
- Distinguish between electric potential and electric field.
- Determine the electric potential of a point charge given charge and distance.

Point charges, such as electrons, are among the fundamental building blocks of matter. Furthermore, spherical charge distributions (like on a metal sphere) create external electric fields exactly like a point charge. The electric potential due to a point charge is, thus, a case we need to consider. Using calculus to find the work needed to move a test charge q from a large distance away to a distance of r from a point charge Q, and noting the connection between work and potential ($W = -q\Delta V$), it can be shown that the *electric potential* V *of a point charge* is $V = \frac{kQ}{r}$ (Point Charge), where k is a constant equal to $9.0 \times 10^9\ \mathrm{N \cdot m^2/C^2}$.

Electric Potential *V* of a Point Charge

The electric

potential V of a point charge is given by

$$V = \frac{kQ}{r} \text{ (Point Charge)}$$

The potential at infinity is chosen to be zero. Thus V for a point charge decreases with distance, whereas E for a point charge decreases with distance squared:

$$E = \frac{F}{q} = \frac{kQ}{r^2}.$$

Recall that the electric potential V is a scalar and has no direction, whereas the electric field E is a vector. To find the voltage due to a combination of point charges, you add the individual voltages as numbers. To find the total electric field, you must add the individual fields as *vectors*, taking magnitude and direction into account. This is consistent with the fact that V is closely associated with energy, a scalar, whereas E is closely associated with force, a vector.

Example 1. What Voltage Is Produced by a Small Charge on a Metal Sphere?

Charges in static electricity are typically in the nanocoulomb (nC) to microcoulomb (µC) range. What is the voltage 5.00 cm away from the center of a 1-cm diameter metal sphere that has a −3.00 nC static charge?

Strategy

As we have discussed in Electric Charge and Electric Field, charge on a metal sphere spreads out uniformly and produces a field like that of a point charge located at its center. Thus we can find the voltage using the equation

$$V = k\frac{Q}{r}.$$

Solution

Entering known values into the expression for the potential of a point charge, we obtain

$$\begin{aligned} V &= k\frac{Q}{r} \\ &= \left(8.99 \times 10^9\ \mathrm{N \cdot m^2/C^2}\right)\left(\frac{-3.00\times10^{-9}\ \mathrm{C}}{5.00\times10^{-2}\ \mathrm{m}}\right) \\ &= -539\ \mathrm{V} \end{aligned}$$

Discussion

The negative value for voltage means a positive charge would be attracted from a larger distance, since the potential is lower (more negative) than at larger distances. Conversely, a negative charge would be repelled, as expected.

Example 2. What Is the Excess Charge on a Van de Graaff Generator

A demonstration Van de Graaff generator has a 25.0 cm diameter metal sphere that produces a voltage of 100 kV near its surface. (See Figure 1.) What excess charge resides on the sphere? (Assume that each numerical value here is shown with three significant figures.)

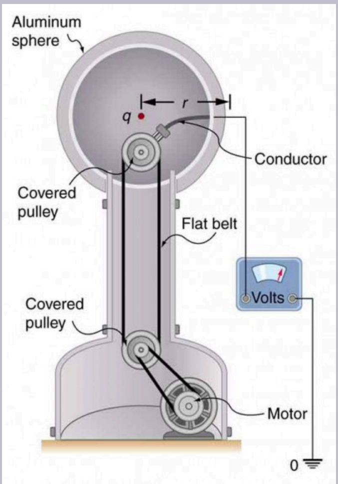


Figure 1. The voltage of this demonstration Van de Graaff generator is measured between the charged sphere and ground. Earth's potential is taken to be zero as a reference. The potential of the charged conducting sphere is the same as that of an equal point charge at its center.

Strategy

The potential on the surface will be the same as that of a point charge at the center of the sphere, 12.5 cm away. (The radius of the sphere is 12.5 cm.) We can thus determine the excess charge using the equation $V = \frac{kQ}{r}$.

Solution

Solving for Q and entering known values gives

$$\begin{aligned} Q &= \frac{rV}{k} \\ &= \frac{(0.125\ \text{m})(100\times10^{3}\ \text{V})}{8.99\times10^{9}\ \text{N}\cdot\text{m}^{2}/\text{C}^{2}} \\ &= 1.39\times10^{-6}\ \text{C} = 1.39\mu\text{C} \end{aligned}$$

Discussion

This is a relatively small charge, but it produces a rather large voltage. We have another indication here that it is difficult to store isolated charges.

The voltages in both of these examples could be measured with a meter that compares the measured potential with ground potential. Ground potential is often taken to be zero (instead of taking the potential at infinity to be zero). It is the potential difference between two points that is of importance, and very often there is a tacit assumption that some reference point, such as Earth or a very distant point, is at zero potential. As noted in Electric Potential Energy: Potential Difference, this is analogous to taking sea level as $h = 0$ when considering gravitational potential energy, $\text{PE}_g = mgh$.

Section Summary

- Electric potential of a point charge is $V = \frac{kQ}{r}$.
- Electric potential is a scalar, and electric field is a vector. Addition of voltages as numbers gives the voltage due to a combination of point charges, whereas addition of individual fields as vectors gives the total electric field.

Conceptual Questions

1. In what region of space is the potential due to a uniformly charged sphere the same as that of a point charge? In what region does it differ from that of a point charge?
2. Can the potential of a non-uniformly charged sphere be the same as that of a point charge? Explain.

Problems & Exercises

1. A 0.500 cm diameter plastic sphere, used in a static electricity demonstration, has a uniformly distributed 40.0 pC charge on its surface. What is the potential near its surface?
2. What is the potential 0.530×10^{-10} m from a proton (the average distance between the proton and

electron in a hydrogen atom)?

3. (a) A sphere has a surface uniformly charged with 1.00 C. At what distance from its center is the potential 5.00 MV? (b) What does your answer imply about the practical aspect of isolating such a large charge?
4. How far from a 1.00 μC point charge will the potential be 100 V? At what distance will it be 2.00×10^2 V?
5. What are the sign and magnitude of a point charge that produces a potential of −2.00 V at a distance of 1.00 mm?
6. If the potential due to a point charge is 5.00×10^2 V at a distance of 15.0 m, what are the sign and magnitude of the charge?
7. In nuclear fission, a nucleus splits roughly in half. (a) What is the potential 2.00×10^{-14} m from a fragment that has 46 protons in it? (b) What is the potential energy in MeV of a similarly charged fragment at this distance?
8. A research Van de Graaff generator has a 2.00-m-diameter metal sphere with a charge of 5.00 mC on it. (a) What is the potential near its surface? (b) At what distance from its center is the potential 1.00 MV? (c) An oxygen atom with three missing electrons is released near the Van de Graaff generator. What is its energy in MeV at this distance?
9. An electrostatic paint sprayer has a 0.200-m-diameter metal sphere at a potential of 25.0 kV that repels paint droplets onto a grounded object. (a) What charge is on the sphere? (b) What charge must a 0.100-mg drop of paint have to arrive at the object

with a speed of 10.0 m/s?

10. In one of the classic nuclear physics experiments at the beginning of the 20th century, an alpha particle was accelerated toward a gold nucleus, and its path was substantially deflected by the Coulomb interaction. If the energy of the doubly charged alpha nucleus was 5.00 MeV, how close to the gold nucleus (79 protons) could it come before being deflected?
11. (a) What is the potential between two points situated 10 cm and 20 cm from a 3.0 µC point charge? (b) To what location should the point at 20 cm be moved to increase this potential difference by a factor of two?
12. **Unreasonable Results.** (a) What is the final speed of an electron accelerated from rest through a voltage of 25.0 MV by a negatively charged Van de Graaff terminal? (b) What is unreasonable about this result? (c) Which assumptions are responsible?

Selected Solutions to Problems & Exercises

1. 144 V

3. (a) 1.80 km; (b) A charge of 1 C is a very large amount of charge; a sphere of radius 1.80 km is not practical.

5. -2.22×10^{-13} C

7. (a) 3.31×10^{6} V; (b) 152 MeV

9. (a) 2.78×10^{-7} C; (b) 2.00×10^{-10} C

12. (a) 2.96×10^{9} m/s; (b) This velocity is far too great. It is faster than the speed of light; (c) The assumption that the speed of the electron is far less than that of light and that the problem does not require a relativistic treatment produces an answer greater than the speed of light.

16. Equipotential Lines

Learning Objectives

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

- Explain equipotential lines and equipotential surfaces.
- Describe the action of grounding an electrical appliance.
- Compare electric field and equipotential lines.

We can represent electric potentials (voltages) pictorially, just as we drew pictures to illustrate electric fields. Of course, the two are related. Consider Figure 1, which shows an isolated positive point charge and its electric field lines. Electric field lines radiate out from a positive charge and terminate on negative charges. While we use blue arrows to represent the magnitude and direction of the electric field, we use green lines to represent places where the electric potential is constant. These are called *equipotential lines* in two dimensions, or *equipotential surfaces* in three dimensions. The term *equipotential* is also used as a noun, referring to an equipotential line or surface. The potential for a point charge is the same anywhere on an imaginary sphere of radius *r* surrounding the charge. This is true since the potential for a point charge is given by $V = \dfrac{kQ}{r}$ and, thus, has the same value at any point that is a given distance *r* from the charge. An equipotential sphere is a circle in the two-dimensional view of Figure 1. Since the electric field lines point radially away from the charge, they are perpendicular to the equipotential lines.

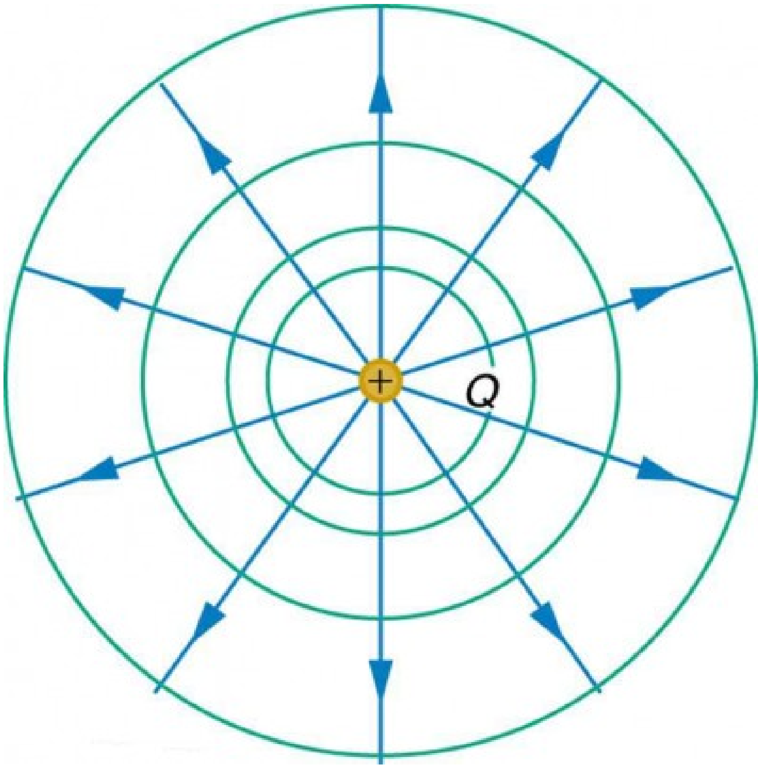


Figure 1. An isolated point charge Q with its electric field lines in blue and equipotential lines in green. The potential is the same along each equipotential line, meaning that no work is required to move a charge anywhere along one of those lines. Work is needed to move a charge from one equipotential line to another. Equipotential lines are perpendicular to electric field lines in every case.

It is important to note that *equipotential lines are always perpendicular to electric field lines.* No work is required to move a charge along an equipotential, since $\Delta V = 0$. Thus the work is

$$W = -\Delta PE = -q\Delta V = 0.$$

Work is zero if force is perpendicular to motion. Force is in the same direction as E, so that motion along an equipotential must be perpendicular to E. More precisely, work is related to the electric field by

$$W = Fd \cos\theta = qEd \cos\theta = 0.$$

Note that in the above equation, E and F symbolize the magnitudes of the electric field strength and force, respectively. Neither q nor **E** nor d is zero, and so $\cos\theta$ must be 0, meaning θ must be 90°. In other words, motion along an equipotential is perpendicular to E.

One of the rules for static electric fields and conductors is that the electric field must be perpendicular to the surface of any conductor. This implies that a *conductor is an equipotential surface in static situations*. There can be no voltage difference across the surface of a conductor, or charges will flow. One of the uses of this fact is that a conductor can be fixed at zero volts by connecting it to the earth with a good conductor—a process called *grounding*. Grounding can be a useful safety tool. For example, grounding the metal case of an electrical appliance ensures that it is at zero volts relative to the earth.

Grounding

A conductor can be fixed at zero volts by connecting it to the earth with a good conductor—a process called grounding.

Because a conductor is an equipotential, it can replace any equipotential surface. For example, in Figure 1 a charged spherical conductor can replace the point charge, and the electric field and potential surfaces outside of it will be unchanged, confirming the contention that a spherical charge distribution is equivalent to a point charge at its center.

Figure 2 shows the electric field and equipotential lines for two equal and opposite charges. Given the electric field lines, the equipotential lines can be drawn simply by making them perpendicular to the electric field lines. Conversely, given the equipotential lines, as in Figure 3a, the electric field lines can be drawn by making them perpendicular to the equipotentials, as in Figure 3b.

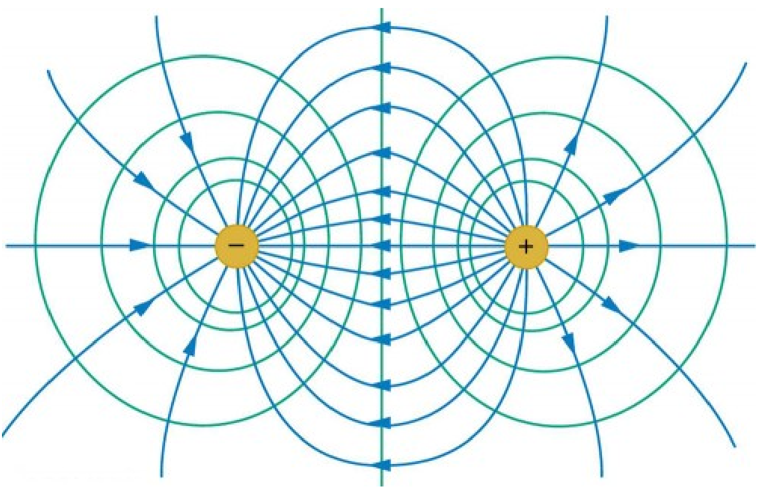


Figure 2. The electric field lines and equipotential lines for two equal but opposite charges. The equipotential lines can be drawn by making them perpendicular to the electric field lines, if those are known. Note that the potential is greatest (most positive) near the positive charge and least (most negative) near the negative charge.

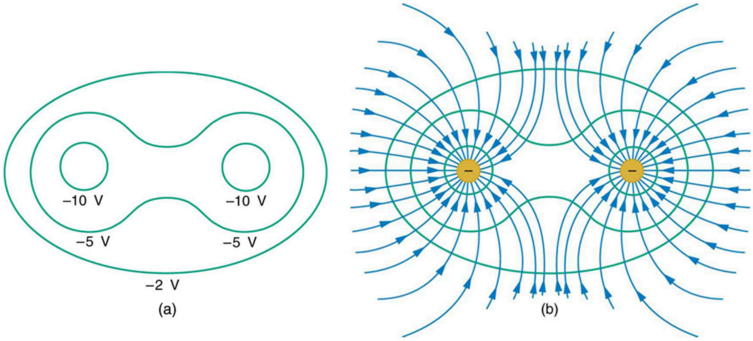


Figure 3. (a) These equipotential lines might be measured with a voltmeter in a laboratory experiment. (b) The corresponding electric field lines are found by drawing them perpendicular to the equipotentials. Note that these fields are consistent with two equal negative charges.

One of the most important cases is that of the familiar parallel conducting plates shown in Figure 4. Between the plates, the equipotentials are evenly spaced and parallel. The same field could be maintained by placing conducting plates at the equipotential lines at the potentials shown.

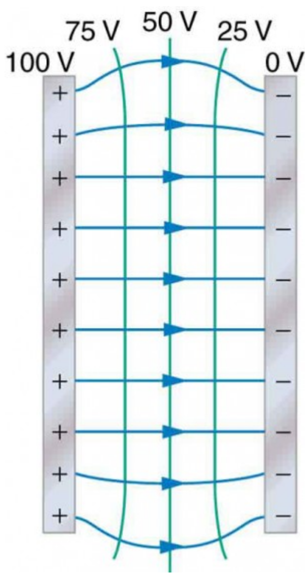
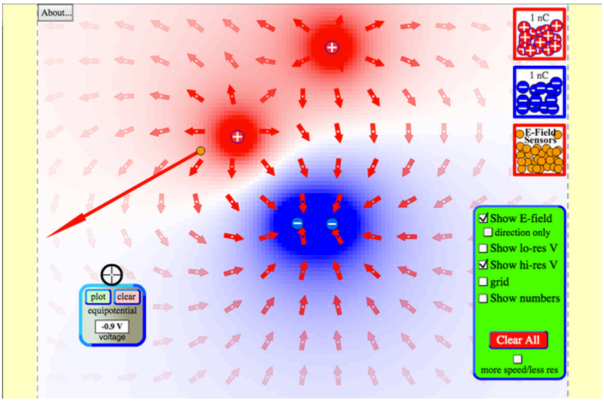


Figure 4. The electric field and equipotential lines between two metal plates.

An important application of electric fields and equipotential lines involves the heart. The heart relies on electrical signals to maintain its rhythm. The movement of electrical signals causes the chambers of the heart to contract and relax. When a person has a heart attack, the movement of these electrical signals may be disturbed. An artificial pacemaker and a defibrillator can be used to initiate the rhythm of electrical signals. The equipotential lines around the heart, the thoracic region, and the axis of the heart are useful ways of monitoring the structure and functions of the heart. An electrocardiogram (ECG) measures the small electric signals being generated during the activity of the heart. More about the relationship between electric fields and the heart is discussed in Energy Stored in Capacitors.

PhET Explorations: Charges and Fields

Move point charges around on the playing field and then view the electric field, voltages, equipotential lines, and more. It's colorful, it's dynamic, it's free.



Click to run the simulation.

Section Summary

- An equipotential line is a line along which the electric potential is constant.
- An equipotential surface is a three-dimensional version of equipotential lines.
- Equipotential lines are always perpendicular to electric field lines.
- The process by which a conductor can be fixed at zero volts by connecting it to the earth with a good conductor is called grounding.

Conceptual Questions

1. What is an equipotential line? What is an equipotential surface?
2. Explain in your own words why equipotential lines and surfaces must be perpendicular to electric field lines.
3. Can different equipotential lines cross? Explain.

Problems & Exercises

1. (a) Sketch the equipotential lines near a point charge $+q$. Indicate the direction of increasing potential. (b) Do the same for a point charge $-3q$.

2. Sketch the equipotential lines for the two equal positive charges shown in Figure 5. Indicate the direction of increasing potential.

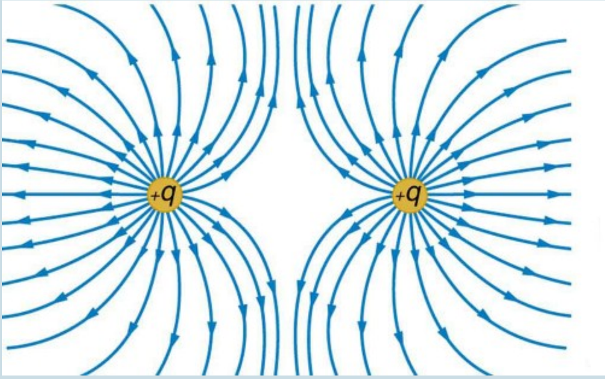


Figure 5. The electric field near two equal positive charges is directed away from each of the charges.

3. Figure 6 shows the electric field lines near two charges q_1 and q_2, the first having a magnitude four times that of the second. Sketch the equipotential lines for these two charges, and indicate the direction of increasing potential.
4. Sketch the equipotential lines a long distance from the charges shown in Figure 6. Indicate the direction of increasing potential.

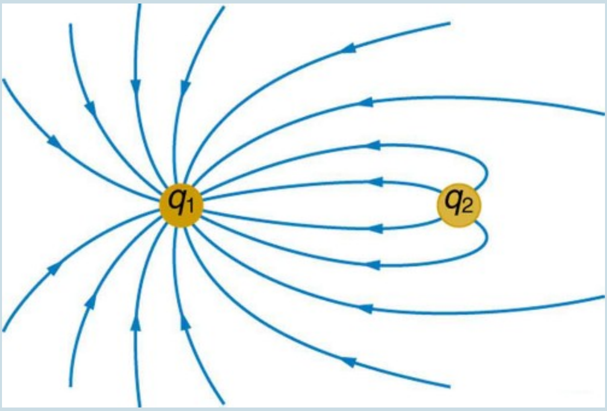


Figure 6. The electric field near two charges.

5. Sketch the equipotential lines in the vicinity of two opposite charges, where the negative charge is three times as great in magnitude as the positive. See Figure 6 for a similar situation. Indicate the direction of increasing potential.
6. Sketch the equipotential lines in the vicinity of the negatively charged conductor in Figure 7. How will these equipotentials look a long distance from the object?

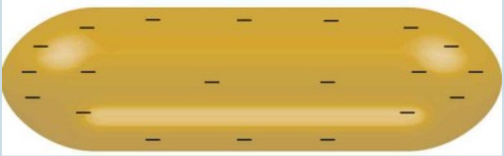


Figure 7. A negatively charged conductor.

7. Sketch the equipotential lines surrounding the two conducting plates shown in Figure 8, given the top plate is positive and the bottom plate has an equal amount of negative charge. Be certain to indicate the distribution of charge on the plates. Is the field strongest where the plates are closest? Why should it be?

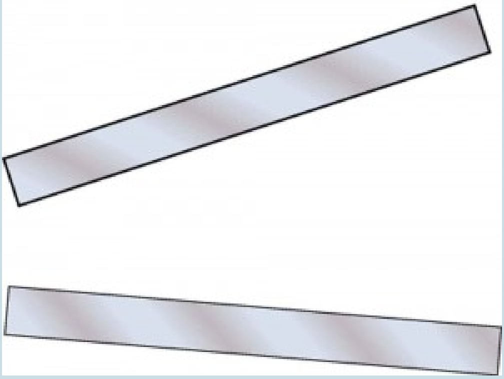


Figure 8.

8. (a) Sketch the electric field lines in the vicinity of the charged insulator in Figure 9. Note its non-uniform charge distribution. (b) Sketch equipotential lines surrounding the insulator. Indicate the direction of increasing potential.

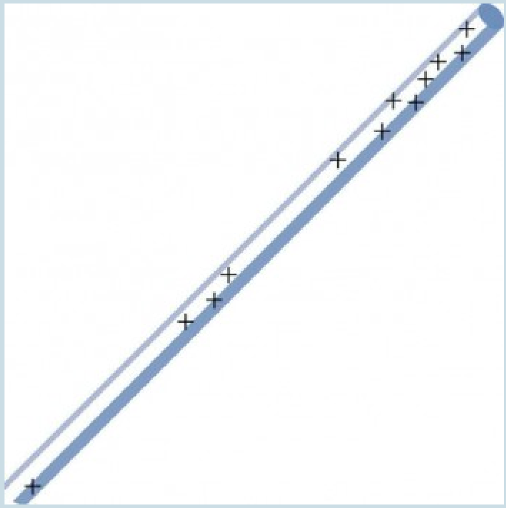


Figure 9. A charged insulating rod such as might be used in a classroom demonstration.

9. The naturally occurring charge on the ground on a fine day out in the open country is $-1.00\ nC/m^2$. (a) What is the electric field relative to ground at a height of 3.00 m? (b) Calculate the electric potential at this height. (c) Sketch electric field and equipotential lines for this scenario.
10. The lesser electric ray (Narcine bancroftii) maintains an incredible charge on its head and a charge equal in magnitude but opposite in sign on its tail (Figure 10). (a) Sketch the equipotential lines surrounding the ray. (b) Sketch the equipotentials when the ray is near a ship with a conducting surface. (c) How could this charge distribution be of use to the ray?

Figure 10. Lesser electric ray (Narcine bancroftii) (credit: National Oceanic and Atmospheric Administration, NOAA's Fisheries Collection).

Glossary

equipotential line: a line along which the electric potential is constant

grounding: fixing a conductor at zero volts by connecting it to the earth or ground

17. 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 1. (Most of the time an insulator is used between the two plates to provide separation—see the discussion on dielectrics below.) 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.

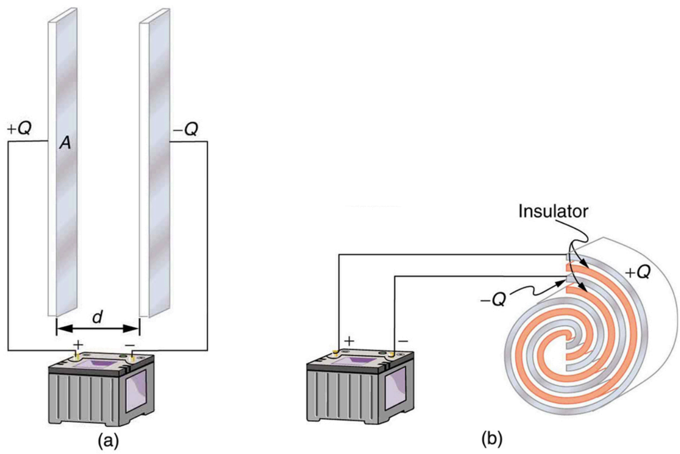


Figure 1. Both capacitors shown here were initially uncharged before being connected to a battery. They now have separated charges of +Q and −Q on their two halves. (a) A parallel plate capacitor. (b) A rolled capacitor with an insulating material between its two conducting sheets.

Capacitor

A capacitor is a device used to store electric 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 Amount of Charge *Q* a Capacitor Can Store

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 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 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.

The field is proportional to the charge:

$$E \propto Q,$$

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, $V \propto E$. It follows, then, that $V \propto Q$, and conversely,

$$Q \propto V.$$

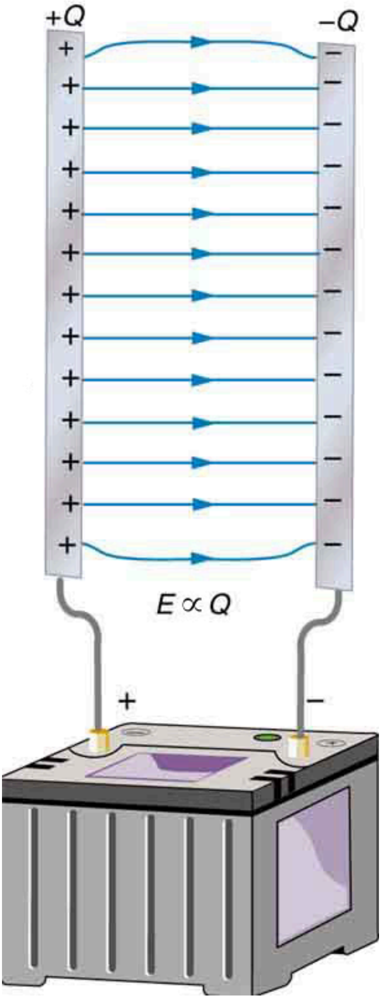


Figure 2. Electric field lines in this parallel plate capacitor, as always, start on positive charges and end on negative charges. Since the electric field strength is proportional to the density of field lines, it is also proportional to the amount of charge on the capacitor.

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

$$Q = CV.$$

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 see that *capacitance C is the amount of charge stored per volt*, or

$$C = \frac{Q}{V}.$$

Capacitance

Capacitance C is the amount of charge stored per volt, or

$$C = \frac{Q}{V}$$

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

$$1\ \text{F} = \frac{1\ \text{C}}{1\ \text{V}}.$$

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 ($1\ \text{pF} = 10^{-12}\ \text{F}$) to millifarads ($1\ \text{mF} = 10^{-3}\ \text{F}$).

Figure 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.

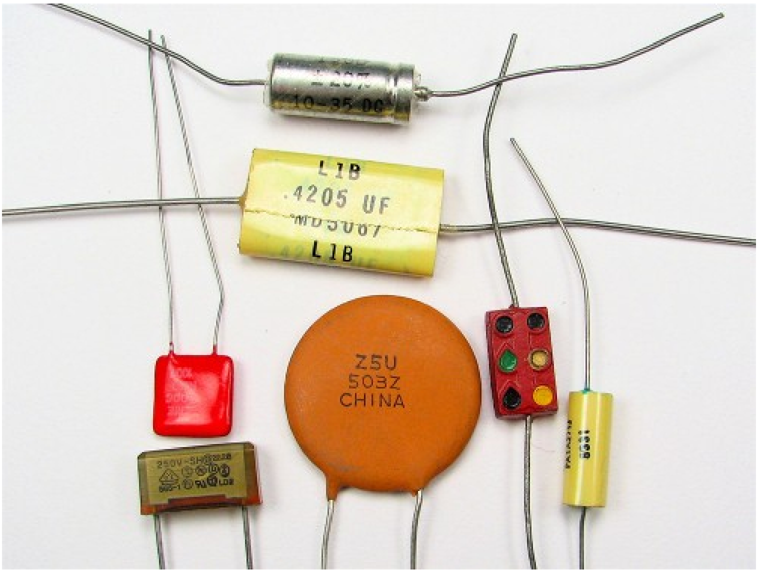


Figure 3. Some typical capacitors. Size and value of capacitance are not necessarily related. (credit: Windell Oskay)

Parallel Plate Capacitor

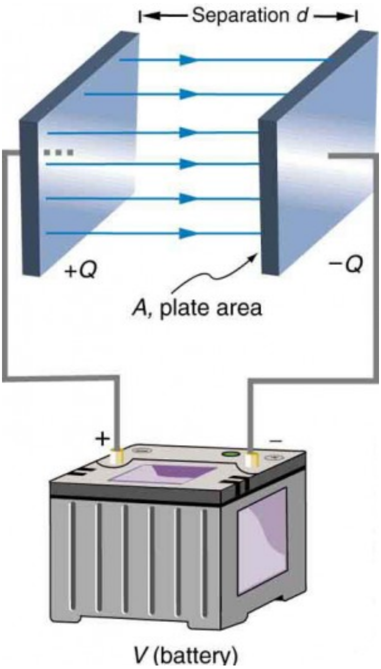


Figure 4. Parallel plate capacitor with plates separated by a distance d. Each plate has an area A.

The parallel plate capacitor shown in Figure 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*.

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 is given by

$$C = \epsilon_o \frac{A}{d}.$$

Capacitance of a Parallel Plate Capacitor

$$C = \epsilon_o \frac{A}{d}$$

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 $\varepsilon_0 = 8.85 \times 10^{-12}$ F/m. The units of F/m are equivalent to $C^2/N \cdot m^2$. 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 1. Capacitance and Charge Stored in a Parallel Plate Capacitor

1. What is the capacitance of a parallel plate capacitor with metal plates, each of area $1.00\ m^2$, separated by 1.00 mm?
2. What charge is stored in this capacitor if a voltage of 3.00×10^3 V is applied to it?

Strategy

Finding the capacitance C is a straightforward application of the equation $C = \epsilon_o \dfrac{A}{d}$. Once C is found, the charge stored can be found using the equation Q = CV.

Solution for Part 1

Entering the given values into the equation for the capacitance of a parallel plate capacitor yields

$$\begin{aligned} C &= \epsilon_o \frac{A}{d} = \left(8.85 \times 10^{-12}\,\frac{\text{F}}{\text{m}}\right)\frac{1.00\ \text{m}^2}{1.00{\times}10^{-3}\ \text{m}} \\ &= 8.85 \times 10^{-9}\ \text{F} = 8.85\ \text{nF} \end{aligned}$$

Discussion for Part 1

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 Part 2

The charge stored in any capacitor is given by the equation Q = CV. Entering the known values into this equation gives

$$\begin{aligned} Q &= CV = \left(8.85 \times 10^{-9}\ \mathrm{F}\right)\left(3.00 \times 10^{3}\ \mathrm{V}\right) \\ &= 26.6\mu\mathrm{C} \end{aligned}$$

Discussion for Part 2

This charge is only slightly greater than those found in typical static electricity. Since air breaks down at about 3.00×10^6 V/m, more charge cannot be stored on this capacitor by increasing the voltage.

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 −70 mV . 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

$$E = \frac{V}{d} = \frac{-70 \times 10^{-3}\ \mathrm{V}}{8 \times 10^{-9}\ \mathrm{m}} = -9 \times 10^{6}\ \mathrm{V/m}$$

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 $E = \frac{V}{d}$). 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 $C = \kappa\epsilon_0 \frac{A}{d}$ by a factor κ, called the *dielectric constant*. A parallel plate capacitor with a dielectric between its plates has a capacitance given by $C = \kappa\epsilon_0 \frac{A}{d}$ (parallel plate capacitor with dielectric).

Values of the dielectric constant κ for various materials are given in Table 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 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 1. Dielectric Constants and Dielectric Strengths for Various Materials at 20°C

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 $E = \frac{V}{d}$ for a parallel plate capacitor.) Also shown in Table 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 1, the separation is 1.00 mm, and so the voltage limit for air is

$$\begin{aligned} V &= E \cdot d \\ &= \left(3 \times 10^6 \text{ V/m}\right)\left(1.00 \times 10^{-3} \text{ m}\right) \\ &= 3000 \text{ V} \end{aligned}$$

However, the limit for a 1.00 mm separation filled with Teflon is 60,000 V, since the dielectric strength of Teflon is 60×10^6 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

$$\begin{aligned} Q &= CV \\ &= \kappa C_{\text{air}} V \\ &= (2.1)(8.85 \text{ nF})(6.0 \times 10^4 \text{ V}) \\ &= 1.1 \text{ mC} \end{aligned}$$

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 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.

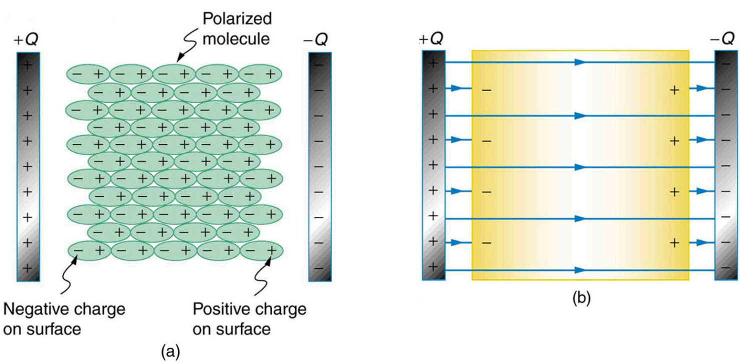


Figure 5. (a) The molecules in the insulating material between the plates of a capacitor are polarized by the charged plates. This produces a layer of opposite charge on the surface of the dielectric that attracts more charge onto the plate, increasing its capacitance. (b) The dielectric reduces the 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 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 $C = \frac{Q}{V}$, the capacitance C is greater.

The dielectric constant is generally defined to be $\kappa = \frac{E_0}{E}$, 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 Atomic Physics. The submicroscopic origin of polarization can be modeled as shown in Figure 6.

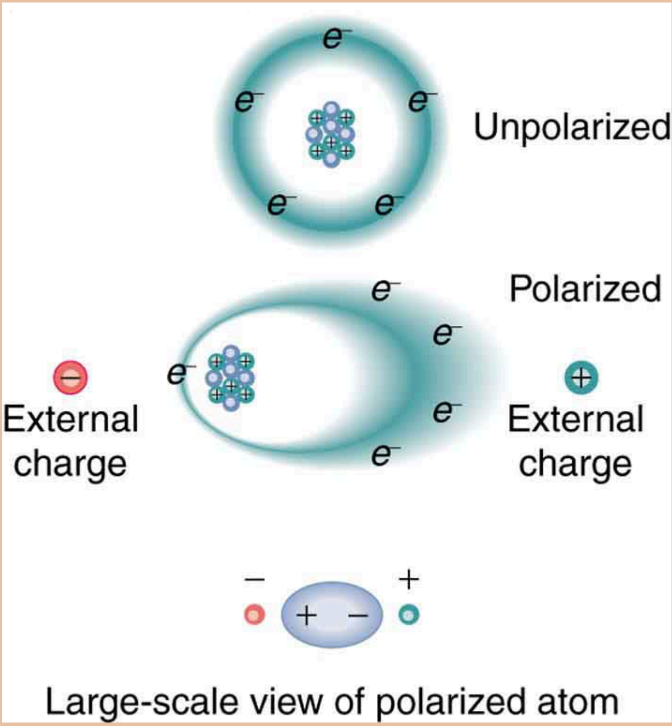


Figure 6. Artist's conception of a polarized atom. The orbits of electrons around the nucleus are shifted slightly by the external charges (shown exaggerated). The resulting separation of charge within the atom means that it is polarized. Note that the unlike charge is now closer to the external charges, 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 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.

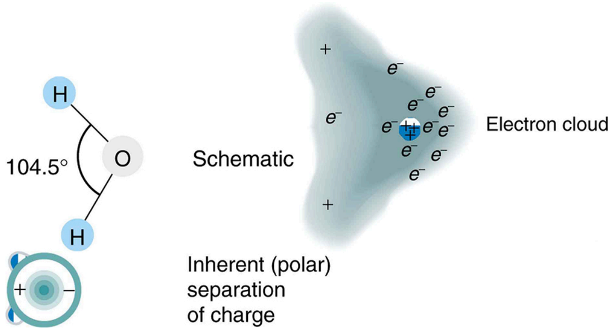
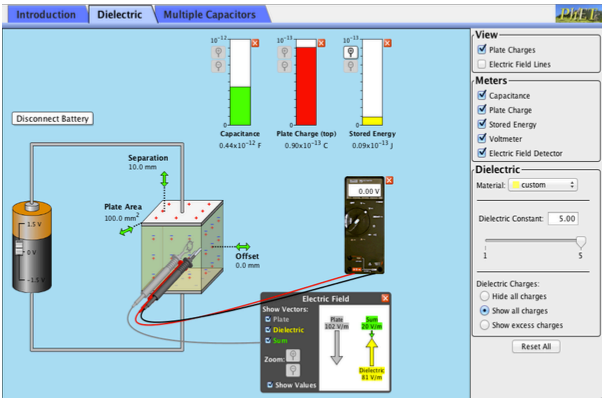


Figure 7. Artist's conception of a water molecule. There is an inherent separation of charge, and so water is a polar molecule. Electrons in the molecule are attracted to the oxygen nucleus and leave an excess of positive charge near the two hydrogen nuclei. (Note that the schematic on the right is a 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.



Click to download the simulation. Run using Java.

Section 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 $C = \dfrac{Q}{V}$.
- The capacitance of a parallel plate capacitor is $C = \epsilon_0 \dfrac{A}{d}$, 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 $C = \kappa\epsilon_0 \dfrac{A}{d}$, 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.

Conceptual Questions

1. Does the capacitance of a device depend on the applied voltage? What about the charge stored in it?
2. Use the characteristics of the Coulomb force to explain why capacitance should be proportional to the plate area of a capacitor. Similarly, explain why capacitance should be inversely proportional to the separation between plates.
3. Give the reason why a dielectric material increases capacitance compared with what it would be with air between the plates of a capacitor. What is the independent reason that a dielectric material also allows a greater voltage to be applied to a capacitor?

(The dielectric thus increases C and permits a greater V.)

4. How does the polar character of water molecules help to explain water's relatively large dielectric constant? (See Figure 7.)
5. Sparks will occur between the plates of an air-filled capacitor at lower voltage when the air is humid than when dry. Explain why, considering the polar character of water molecules.
6. Water has a large dielectric constant, but it is rarely used in capacitors. Explain why.
7. Membranes in living cells, including those in humans, are characterized by a separation of charge across the membrane. Effectively, the membranes are thus charged capacitors with important functions related to the potential difference across the membrane. Is energy required to separate these charges in living membranes and, if so, is its source the metabolization of food energy or some other source?

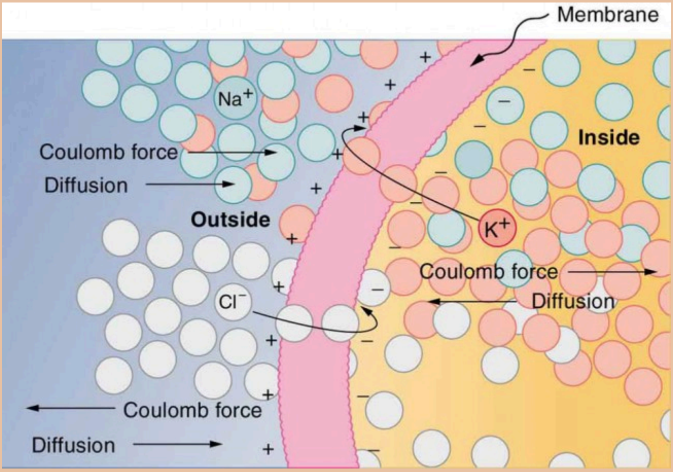


Figure 8. The semipermeable membrane of a cell has different concentrations of ions inside and out. Diffusion moves the K^+ (potassium) and Cl^- (chloride) ions in the directions shown, until the Coulomb force halts further transfer. This results in a layer of positive charge on the outside, a layer of negative charge on the inside, and thus a voltage across the cell membrane. The membrane is normally impermeable to Na^+ (sodium ions).

Problems & Exercises

1. What charge is stored in a 180 μF capacitor when 120 V is applied to it?
2. Find the charge stored when 5.50 V is applied to an 8.00 pF capacitor.
3. What charge is stored in the capacitor in Example

1?

4. Calculate the voltage applied to a 2.00 μF capacitor when it holds 3.10 μC of charge.
5. What voltage must be applied to an 8.00 nF capacitor to store 0.160 mC of charge?
6. What capacitance is needed to store 3.00 μC of charge at a voltage of 120 V?
7. What is the capacitance of a large Van de Graaff generator's terminal, given that it stores 8.00 mC of charge at a voltage of 12.0 MV?
8. Find the capacitance of a parallel plate capacitor having plates of area 5.00 m^2 that are separated by 0.100 mm of Teflon.
9. (a)What is the capacitance of a parallel plate capacitor having plates of area 1.50 m^2 that are separated by 0.0200 mm of neoprene rubber? (b) What charge does it hold when 9.00 V is applied to it?
10. **Integrated Concepts.** A prankster applies 450 V to an 80.0 μF capacitor and then tosses it to an unsuspecting victim. The victim's finger is burned by the discharge of the capacitor through 0.200 g of flesh. What is the temperature increase of the flesh? Is it reasonable to assume no phase change?
11. **Unreasonable Results.** (a) A certain parallel plate capacitor has plates of area 4.00 m^2, separated by 0.0100 mm of nylon, and stores 0.170 C of charge. What is the applied voltage? (b) What is unreasonable about this result? (c) Which assumptions are responsible or inconsistent?

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

Selected Solutions to Problems & Exercises

1. 21.6 mC

3. 80.0 mC

5. 20.0 kV

7. 667 pF

9. (a) 4.4 μF; (b) 4.0×10^{-5} C

11. (a) 14.2 kV; (b) The voltage is unreasonably large, more than 100 times the breakdown voltage of nylon; (c) The assumed charge is unreasonably large and cannot be stored in a capacitor of these dimensions.

18. Video: Capacitors

Watch the following Physics Concept Trailer to learn about how capacitors send electric charges to revive a heart after a heart attack.

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://library.achievingthedream.org/austinccphysics2/?p=37#oembed-1

19. Capacitors in Series and Parallel

Learning Objectives

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

- Derive expressions for total capacitance in series and in parallel.
- Identify series and parallel parts in the combination of connection of capacitors.
- Calculate the effective capacitance in series and parallel given individual capacitances.

Several capacitors may be connected together in a variety of applications. Multiple connections of capacitors act like a single equivalent capacitor. The total capacitance of this equivalent single capacitor depends both on the individual capacitors and how they are connected. There are two simple and common types of connections, called *series* and *parallel*, for which we can easily calculate the total capacitance. Certain more complicated connections can also be related to combinations of series and parallel.

Capacitance in Series

Figure 1a shows a series connection of three capacitors with a voltage applied. As for any capacitor, the capacitance of the combination is related to charge and voltage by $C=\frac{Q}{V}$.

Note in Figure 1 that opposite charges of magnitude Q flow to either side of the originally uncharged combination of capacitors when the voltage V is applied. Conservation of charge requires that equal-magnitude charges be created on the plates of the individual capacitors, since charge is only being separated in these originally neutral devices. The end result is that the combination resembles a single capacitor with an effective plate separation greater than that of the individual capacitors alone. (See Figure 1b.) Larger plate separation means smaller capacitance. It is a general feature of series connections of capacitors that the total capacitance is less than any of the individual capacitances.

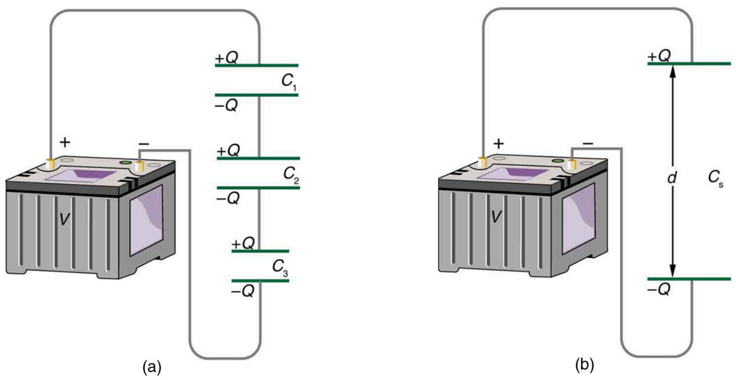


Figure 1. (a) Capacitors connected in series. The magnitude of the charge on each plate is Q. (b) An equivalent capacitor has a larger plate separation d. Series connections produce a total capacitance that is less than that of any of the individual capacitors.

We can find an expression for the total capacitance by considering the voltage across the individual capacitors shown in Figure 1. Solving $C=\frac{Q}{V}$ for V gives $V=\frac{Q}{C}$. The voltages across the individual capacitors are thus $V_1 = \frac{Q}{C_1}, V_2 = \frac{Q}{C_2}$, and $V_3 = \frac{Q}{C_3}$.

The total voltage is the sum of the individual voltages:

$$V = V_1 + V_2 + V_3.$$

Now, calling the total capacitance C_S for series capacitance, consider that

$$V = \frac{Q}{C_S} = V_1 + V_2 + V_3.$$

Entering the expressions for V_1, V_2, and V_3, we get

$$\frac{Q}{C_S} = \frac{Q}{C_1} + \frac{Q}{C_2} + \frac{Q}{C_3}.$$

Canceling the Qs, we obtain the equation for the total capacitance in series C_S to be

$$\frac{1}{C_S} = \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3} + \ldots,$$

where "…" indicates that the expression is valid for any number of capacitors connected in series. An expression of this form always results in a total capacitance C_S that is less than any of the individual capacitances C_1, C_2, …, as Example 1 illustrates.

Total Capacitance in Series, C_s

Total capacitance in series:

$$\frac{1}{C_S} = \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3} + \ldots$$

Example 1. What Is the Series Capacitance?

Find the total capacitance for three capacitors connected in series, given their individual capacitances are 1.000, 5.000, and 8.000 μF.

Strategy

With the given information, the total capacitance can be found using the equation for capacitance in series.

Solution

Entering the given capacitances into the expression for $\frac{1}{C_\text{S}}$ gives $\frac{1}{C_\text{S}} = \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3}$.

$$\frac{1}{C_\text{S}} = \frac{1}{1.000 \mu\text{F}} + \frac{1}{5.000 \mu\text{F}} + \frac{1}{8.000 \mu\text{F}} = \frac{1.325}{\mu\text{F}}$$

Inverting to find C_S yields

$$C_\text{S} = \frac{1.325}{\mu\text{F}} = 0.755 \mu\text{F}.$$

Discussion

The total series capacitance C_s is less than the smallest individual capacitance, as promised. In series connections of capacitors, the sum is less than the parts. In fact, it is less than any individual. Note that it is sometimes possible, and more convenient, to solve an equation like the above by finding the least common denominator, which in this case (showing only whole-number calculations) is 40. Thus,

$$\frac{1}{C_\mathrm{S}} = \frac{40}{40\mu\mathrm{F}} + \frac{8}{40\mu\mathrm{F}} + \frac{5}{40\mu\mathrm{F}} = \frac{53}{40\mu\mathrm{F}}$$

so that

$$C_\mathrm{S} = \frac{40\mu\mathrm{F}}{53} = 0.755\mu\mathrm{F}$$

Capacitors in Parallel

Figure 2a shows a parallel connection of three capacitors with a voltage applied. Here the total capacitance is easier to find than in the series case. To find the equivalent total capacitance C_p, we first note that the voltage across each capacitor is V, the same as that of the source, since they are connected directly to it through a conductor. (Conductors are equipotentials, and so the voltage across the capacitors is the same as that across the voltage source.) Thus the capacitors have the same charges on them as they would have if connected individually to the voltage source. The total charge Q is the sum of the individual charges: $Q = Q_1 + Q_2 + Q_3$.

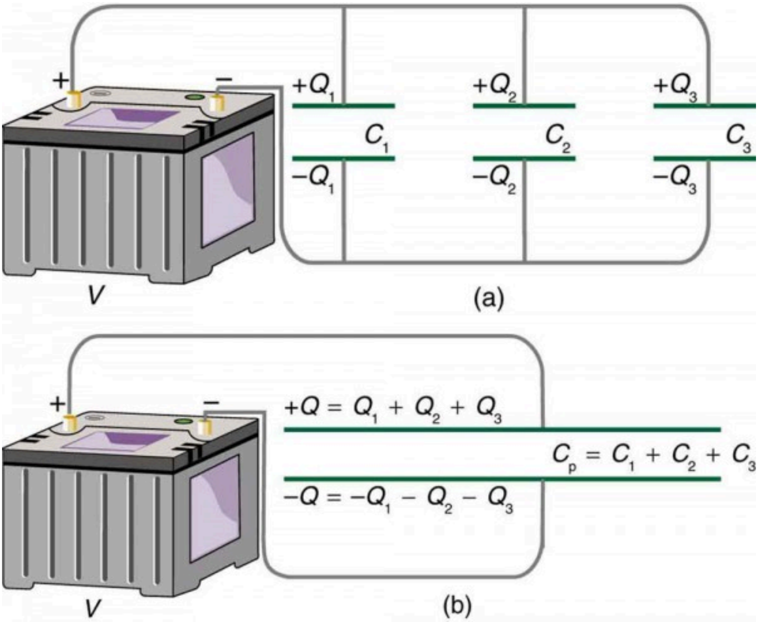


Figure 2. (a) Capacitors in parallel. Each is connected directly to the voltage source just as if it were all alone, and so the total capacitance in parallel is just the sum of the individual capacitances. (b) The equivalent capacitor has a larger plate area and can therefore hold more charge than the individual capacitors.

Using the relationship $Q = CV$, we see that the total charge is $Q = C_pV$, and the individual charges are $Q_1 = C_1V$, $Q_2 = C_2V$, and $Q_3 = C_3V$. Entering these into the previous equation gives

$$C_pV = C_1V + C_2V + C_3V.$$

Canceling V from the equation, we obtain the equation for the total capacitance in parallel

$$C_p\text{: } C_p = C_1 + C_2 + C_3 + \ldots.$$

Total capacitance in parallel is simply the sum of the individual capacitances. (Again the "…" indicates the expression is valid for any number of capacitors connected in parallel.) So, for example, if the capacitors in Example 1 were connected in parallel, their capacitance would be

$$C_p = 1.000\ \mu F + 5.000\ \mu F + 8.000\ \mu F = 14.000\ \mu F.$$

The equivalent capacitor for a parallel connection has an effectively larger plate area and, thus, a larger capacitance, as illustrated in Figure 2b.

Total Capacitance in Parallel, C_p

Total capacitance in parallel $C_p = C_1 + C_2 + C_3 + \ldots$

More complicated connections of capacitors can sometimes be combinations of series and parallel. (See Figure 3.) To find the total capacitance of such combinations, we identify series and parallel parts, compute their capacitances, and then find the total.

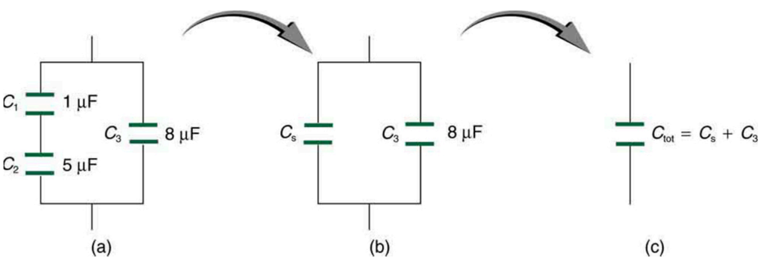


Figure 3. (a) This circuit contains both series and parallel connections of capacitors. See Example 2 for the calculation of the overall capacitance of the circuit. (b) C_1 and C_2 are in series; their equivalent capacitance C_S is less than either of them. (c) Note that C_S is in parallel with C_3. The total capacitance is, thus, the sum of C_S and C_3.

Example 2. A Mixture of Series and Parallel Capacitance

Find the total capacitance of the combination of capacitors shown in Figure 3. Assume the capacitances in Figure 3 are known to three decimal places (C_1 = 1.000 µF, C_2 = 3.000 µF, and C_3 = 8.000 µF), and round your answer to three decimal places.

Strategy

To find the total capacitance, we first identify which capacitors are in series and which are in parallel. Capacitors C_1 and C_2 are in series. Their combination, labeled C_S in the figure, is in parallel with C_3.

Solution

Since C_1 and C_2 are in series, their total capacitance is given by $\frac{1}{C_\mathrm{S}}=\frac{1}{C_1}+\frac{1}{C_2}+\frac{1}{C_3}$. Entering their values into the equation gives

$$\frac{1}{C_\mathrm{S}}=\frac{1}{C_1}+\frac{1}{C_2}=\frac{1}{1.000\mu\mathrm{F}}+\frac{1}{5.000\mu\mathrm{F}}=\frac{1.200}{\mu\mathrm{F}}.$$

Inverting gives C_S = 0.833 µF.

This equivalent series capacitance is in parallel with the third capacitor; thus, the total is the sum

$$\begin{aligned} C_{\text{tot}} &= C_{\text{S}} + C_{\text{S}} \\ &= 0.833\mu\text{F} + 8.000\mu\text{F} \\ &= 8.833\mu\text{F} \end{aligned}$$

Discussion

This technique of analyzing the combinations of capacitors piece by piece until a total is obtained can be applied to larger combinations of capacitors.

Section Summary

- Total capacitance in series $$\frac{1}{C_{\text{S}}} = \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3} + \dots$$
- Total capacitance in parallel $C_p = C_1 + C_2 + C_3 + \dots$
- If a circuit contains a combination of capacitors in series and parallel, identify series and parallel parts, compute their capacitances, and then find the total.

Conceptual Questions

1. If you wish to store a large amount of energy in a capacitor bank, would you connect capacitors in series or parallel? Explain.

Problems & Exercises

1. Find the total capacitance of the combination of capacitors in Figure 4.

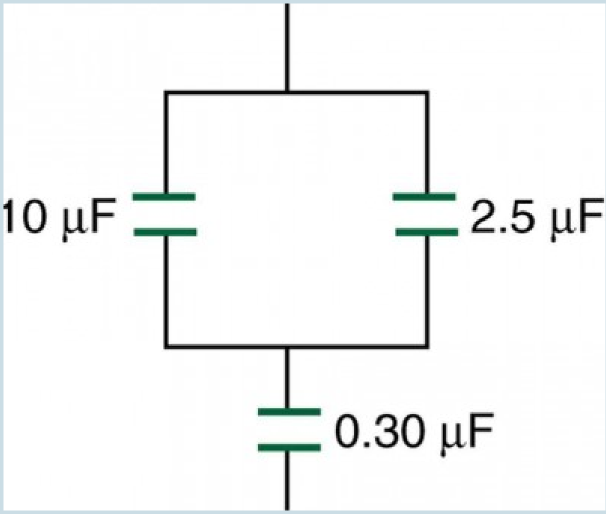


Figure 4. A combination of series and parallel connections of capacitors.

2. Suppose you want a capacitor bank with a total capacitance of 0.750 F and you possess numerous 1.50 mF capacitors. What is the smallest number you could hook together to achieve your goal, and how would you connect them?
3. What total capacitances can you make by connecting a 5.00 μF and an 8.00 μF capacitor together?
4. Find the total capacitance of the combination of capacitors shown in Figure 5.

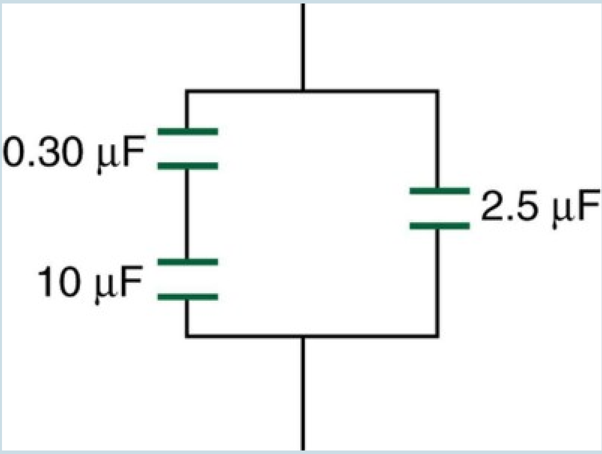


Figure 5. A combination of series and parallel connections of capacitors.

5. Find the total capacitance of the combination of capacitors shown in Figure 6.

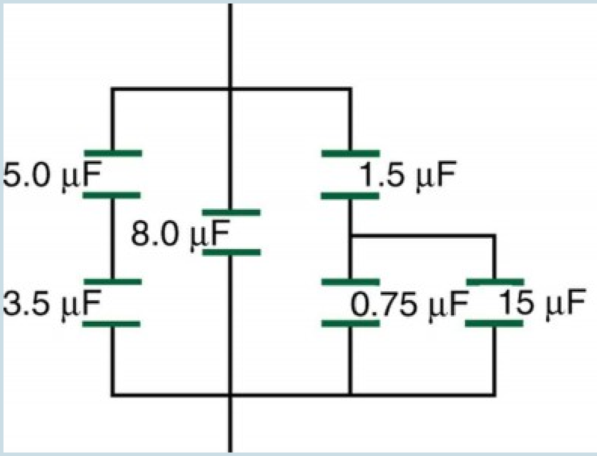


Figure 6. A combination of series and parallel connections of capacitors.

6. **Unreasonable Results.**(a) An 8.00 μF capacitor is connected in parallel to another capacitor, producing a total capacitance of 5.00 μF. What is the capacitance of the second capacitor? (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

Selected Solutions to Problems & Exercises

1. 0.293 μF

3. 3.08 μF in series combination, 13.0 μF in parallel combination

4. 2.79 μF

6. (a) −3.00 μF; (b) You cannot have a negative value of capacitance; (c) The assumption that the capacitors were hooked up in parallel, rather than in series, was incorrect. A parallel connection always produces a greater capacitance, while here a smaller capacitance was assumed. This could happen only if the capacitors are connected in series.

20. Energy Stored in Capacitors

Learning Objectives

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

- List some uses of capacitors.
- Express in equation form the energy stored in a capacitor.
- Explain the function of a defibrillator.

Most of us have seen dramatizations in which medical personnel use a *defibrillator* to pass an electric current through a patient's heart to get it to beat normally. (Review Figure 1.) Often realistic in detail, the person applying the shock directs another person to "make it 400 joules this time." The energy delivered by the defibrillator is stored in a capacitor and can be adjusted to fit the situation. SI units of joules are often employed. Less dramatic is the use of capacitors in microelectronics, such as certain handheld calculators, to supply energy when batteries are charged. (See Figure 1.) Capacitors are also used to supply energy for flash lamps on cameras.

Energy stored in a capacitor is electrical potential energy, and it is thus related to the charge Q and voltage V on the capacitor. We must be careful when applying the equation for electrical potential energy $\Delta PE = q\Delta V$ to a capacitor. Remember that ΔPE is the potential energy of a charge q going through a voltage ΔV. But the capacitor starts with zero voltage and gradually comes up to its full voltage as it is charged. The first charge placed on a capacitor experiences a change in voltage $\Delta V = 0$, since the capacitor has zero voltage when uncharged. The final charge placed on a capacitor experiences $\Delta V = V$, since the capacitor now has its full voltage V on it. The average voltage on the capacitor during the charging process is $\frac{V}{2}$, and so the average voltage experienced by the full charge q is $\frac{V}{2}$. Thus the energy stored in a capacitor, E_{cap}, is $E_{\text{cap}} = \frac{QV}{2}$, where Q is the charge on a capacitor with a voltage V applied. (Note that the energy is not QV, but $\frac{QV}{2}$.) Charge and voltage are related to the capacitance C of a capacitor by $Q = CV$, and so the expression for E_{cap} can be algebraically manipulated into three equivalent expressions:

Figure 1. Energy stored in the large capacitor is used to preserve the memory of an electronic calculator when its batteries are charged. (credit: Kucharek, Wikimedia Commons)

$$E_{\text{cap}} = \frac{QV}{2} = \frac{CV^2}{2} = \frac{Q^2}{2C},$$

where Q is the charge and V the voltage on a capacitor C. The energy is in joules for a charge in coulombs, voltage in volts, and capacitance in farads.

Energy Stored in Capacitors

The energy stored in a capacitor can be expressed in three ways:

$$E_{\text{cap}} = \frac{QV}{2} = \frac{CV^2}{2} = \frac{Q^2}{2C},$$

where Q is the charge, V is the voltage, and C is the capacitance of the capacitor. The energy is in joules for a charge in coulombs, voltage in volts, and capacitance in farads.

In a defibrillator, the delivery of a large charge in a short burst to a set of paddles across a person's chest can be a lifesaver. The person's heart attack might have arisen from the onset of fast, irregular beating of the heart—cardiac or ventricular fibrillation. The application of a large shock of electrical energy can terminate the arrhythmia and allow the body's pacemaker to resume normal patterns. Today it is common for ambulances to carry a defibrillator, which also uses an electrocardiogram to analyze the patient's heartbeat pattern. Automated external defibrillators (AED) are found in many public places (Figure 2). These are designed to be used by lay persons. The device automatically diagnoses the patient's heart condition and then applies the shock with appropriate energy and waveform. CPR is recommended in many cases before use of an AED.

Figure 2. Automated external defibrillators are found in many public places. These portable units provide verbal instructions for use in the important first few minutes for a person suffering a cardiac attack. (credit: Owain Davies, Wikimedia Commons)

Example 1. Capacitance in a Heart Defibrillator

A heart defibrillator delivers 4.00×10^2 J of energy by discharging a capacitor initially at 1.00×10^4 V. What is its capacitance?

Strategy

We are given E_{cap} and V, and we are asked to find the capacitance C. Of the three expressions in the equation for E_{cap}, the most convenient relationship is $E_{\text{cap}} = \frac{CV^2}{2}$.

Solution

Solving this expression for C and entering the given values yields

$$\begin{aligned} C &= \frac{2E_{\text{cap}}}{V^2} = \frac{2(4.00\times 10^2\text{ J})}{(1.00\times 10^4\text{ V})^2} = 8.00\times 10^{-6}\text{ F} \\ &= 8.00\mu\text{F} \end{aligned}$$

Discussion

This is a fairly large, but manageable, capacitance at 1.00×10^4 V.

Section Summary

- Capacitors are used in a variety of devices, including defibrillators, microelectronics such as calculators, and flash lamps, to supply energy.
- The energy stored in a capacitor can be expressed in three ways: $E_{\text{cap}} = \frac{\text{QV}}{2} = \frac{\text{CV}^2}{2} = \frac{Q^2}{2C}$, where Q is the charge, V is the voltage, and C is the capacitance of the capacitor. The energy is in joules when the charge is in coulombs, voltage is in volts, and capacitance is in farads.

Conceptual Questions

1. How does the energy contained in a charged capacitor change when a dielectric is inserted, assuming the capacitor is isolated and its charge is constant? Does this imply that work was done?
2. What happens to the energy stored in a capacitor connected to a battery when a dielectric is inserted? Was work done in the process?

Problems & Exercises

1. (a) What is the energy stored in the 10.0 μF capacitor of a heart defibrillator charged to

9.00×10^3 V? (b) Find the amount of stored charge.

2. In open heart surgery, a much smaller amount of energy will defibrillate the heart. (a) What voltage is applied to the 8.00 μF capacitor of a heart defibrillator that stores 40.0 J of energy? (b) Find the amount of stored charge.
3. A 165 μF capacitor is used in conjunction with a motor. How much energy is stored in it when 119 V is applied?
4. Suppose you have a 9.00 V battery, a 2.00 μF capacitor, and a 7.40 μF capacitor. (a) Find the charge and energy stored if the capacitors are connected to the battery in series. (b) Do the same for a parallel connection.
5. A nervous physicist worries that the two metal shelves of his wood frame bookcase might obtain a high voltage if charged by static electricity, perhaps produced by friction. (a) What is the capacitance of the empty shelves if they have area 1.00×10^2 m^2 and are 0.200 m apart? (b) What is the voltage between them if opposite charges of magnitude 2.00 nC are placed on them? (c) To show that this voltage poses a small hazard, calculate the energy stored.
6. Show that for a given dielectric material the maximum energy a parallel plate capacitor can store is directly proportional to the volume of dielectric (Volume = $A \cdot d$). Note that the applied voltage is limited by the dielectric strength.
7. **Construct Your Own Problem.** Consider a heart defibrillator similar to that discussed in Example 1. Construct a problem in which you examine the charge stored in the capacitor of a defibrillator as a

function of stored energy. Among the things to be considered are the applied voltage and whether it should vary with energy to be delivered, the range of energies involved, and the capacitance of the defibrillator. You may also wish to consider the much smaller energy needed for defibrillation during open-heart surgery as a variation on this problem.

8. **Unreasonable Results.**(a) On a particular day, it takes 9.60×10^3 J of electric energy to start a truck's engine. Calculate the capacitance of a capacitor that could store that amount of energy at 12.0 V. (b) What is unreasonable about this result? (c) Which assumptions are responsible?

Glossary

defibrillator: a machine used to provide an electrical shock to a heart attack victim's heart in order to restore the heart's normal rhythmic pattern

Selected Solutions to Problems & Exercises

1. (a) 405 J; (b) 90.0 mC

2. (a) 3.16 kV; (b) 25.3 mC

4. (a) 1.42×10^{-5} C, 6.38×10^{-5} J; (b) 8.46×10^{-5} C, 3.81×10^{-4} J

5. (a) 4.43×10^{-12} F; (b) 452 V; (c) 4.52×10^{-7} J

8. (a) 133 F; (b) Such a capacitor would be too large to carry with a truck. The size of the capacitor would be enormous; (c) It is unreasonable to assume that a capacitor can store the amount of energy needed.

PART III

ELECTRIC CURRENT, RESISTANCE, AND OHM'S LAW

21. Introduction to Electric Current, Resistance, and Ohm's Law

Figure 1. Electric energy in massive quantities is transmitted from this hydroelectric facility, the Srisailam power station located along the Krishna River in India, by the movement of charge—that is, by electric current. (credit: Chintohere, Wikimedia Commons)

The flicker of numbers on a handheld calculator, nerve impulses carrying signals of vision to the brain, an ultrasound device sending a signal to a computer screen, the brain sending a message for a baby to twitch its toes, an electric train pulling its load over a mountain pass, a hydroelectric plant sending energy to metropolitan and rural users—these and many other examples of electricity involve *electric current, the movement of charge*. Humankind has indeed harnessed electricity, the basis of technology, to improve our quality of life. Whereas the previous two chapters concentrated on static electricity and the fundamental force underlying its behavior, the next few chapters will be devoted to electric and magnetic phenomena involving current. In addition to exploring applications of electricity, we shall gain new insights into nature—in particular, the fact that all magnetism results from electric current.

22. Current

Learning Objectives

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

- Define electric current, ampere, and drift velocity
- Describe the direction of charge flow in conventional current.
- Use drift velocity to calculate current and vice versa.

Electric Current

Electric current is defined to be the rate at which charge flows. A large current, such as that used to start a truck engine, moves a large amount of charge in a small time, whereas a small current, such as that used to operate a hand-held calculator, moves a small amount of charge over a long period of time. In equation form, *electric current* I is defined to be

$$I = \frac{\Delta Q}{\Delta t},$$

where ΔQ is the amount of charge passing through a given area in time Δt. (As in previous chapters, initial time is often taken to be zero, in which case $\Delta t = t$.) (See Figure 1.) The SI unit for current is the *ampere* (A), named for the French physicist André-Marie Ampère (1775–1836). Since $I = \Delta Q/\Delta t$, we see that an ampere is one coulomb per second:

$$1\ \text{A} = 1\ \text{C/s}$$

Not only are fuses and circuit breakers rated in amperes (or amps), so are many electrical appliances.

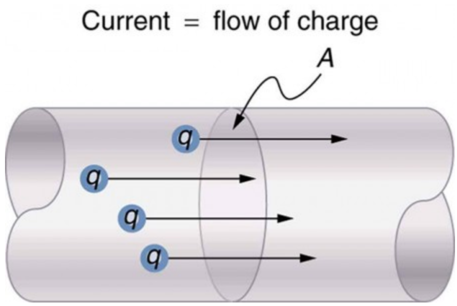


Figure 1. The rate of flow of charge is current. An ampere is the flow of one coulomb through an area in one second.

Example 1. Calculating Currents: Current in a Truck Battery and a Handheld Calculator

(a) What is the current involved when a truck battery sets in motion 720 C of charge in 4.00 s while starting an engine? (b) How long does it take 1.00 C of charge to flow through a handheld calculator if a 0.300-mA current is flowing?

Strategy

We can use the definition of current in the equation $I = \Delta Q/\Delta t$ to find the current in part (a), since charge and time are given. In part (b), we rearrange the definition of current and use the given values of charge and current to find the time required.

Solution for (a)

Entering the given values for charge and time into the definition of current gives

$$\begin{aligned} I &= \frac{\Delta Q}{\Delta t} = \frac{720\ \text{C}}{4.00\ \text{s}} = 180\ \text{C/s} \\ &= 180\ \text{A}. \end{aligned}$$

Discussion for (a)

This large value for current illustrates the fact that a large charge is moved in a small amount of time. The currents in these "starter motors" are fairly large because large frictional forces need to be overcome when setting something in motion.

Solution for (b)

Solving the relationship I = ΔQ/Δt for time Δ*t*, and entering the known values for charge and current gives

$$\begin{aligned} \Delta t &= \frac{\Delta Q}{I} = \frac{1.00\ \text{C}}{0.300 \times 10^{-3}\text{C/s}} \\ &= 3.33 \times 10^{3}\text{s}. \end{aligned}$$

Discussion for (b)

This time is slightly less than an hour. The small current used by the hand-held calculator takes a much longer time to move a smaller charge than the large current of the truck starter. So why can we operate our calculators only seconds after turning them on? It's because calculators require very little energy. Such small current and energy demands allow handheld calculators to operate from solar cells or to get many hours of use out of small batteries. Remember, calculators do not have moving parts in the same way that a truck engine has with cylinders and pistons, so the technology requires smaller currents.

Figure 2 shows a simple circuit and the standard schematic representation of a battery, conducting path, and load (a resistor). Schematics are very useful in visualizing the main features of a circuit. A single schematic can represent a wide variety of situations. The schematic in Figure 2 (b), for example, can represent anything from a truck battery connected to a headlight lighting the street in front of the truck to a small battery connected to a penlight lighting a keyhole in a door. Such schematics are useful because the analysis is the same for a wide variety of situations. We need to understand a few schematics to apply the concepts and analysis to many more situations.

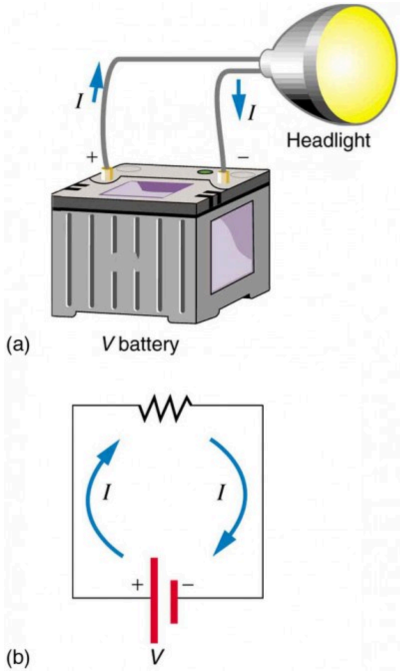


Figure 2. (a) A simple electric circuit. A closed path for current to flow through is supplied by conducting wires connecting a load to the terminals of a battery. (b) In this schematic, the battery is represented by the two parallel red lines, conducting wires are shown as straight lines, and the zigzag represents the load. The schematic represents a wide variety of similar circuits.

Note that the direction of current flow in Figure 2 is from positive to negative. *The direction of conventional current is the direction that positive charge would flow*. Depending on the situation, positive charges, negative charges, or both may move. In metal wires, for example, current is carried by electrons—that is, negative charges move. In ionic solutions, such as salt water, both positive and negative charges move. This is also true in nerve cells. A Van de

Graaff generator used for nuclear research can produce a current of pure positive charges, such as protons. Figure 3 illustrates the movement of charged particles that compose a current. The fact that conventional current is taken to be in the direction that positive charge would flow can be traced back to American politician and scientist Benjamin Franklin in the 1700s. He named the type of charge associated with electrons negative, long before they were known to carry current in so many situations. Franklin, in fact, was totally unaware of the small-scale structure of electricity. It is important to realize that there is an electric field in conductors responsible for producing the current, as illustrated in Figure 3. Unlike static electricity, where a conductor in equilibrium cannot have an electric field in it, conductors carrying a current have an electric field and are not in static equilibrium. An electric field is needed to supply energy to move the charges.

Making Connections: Take-Home Investigation—Electric Current Illustration

Find a straw and little peas that can move freely in the straw. Place the straw flat on a table and fill the straw with peas. When you pop one pea in at one end, a different pea should pop out the other end. This demonstration is an analogy for an electric current. Identify what compares to the electrons and what compares to the supply of energy. What other analogies can you find for an electric current? Note that the flow of peas is based on the peas physically bumping into each other; electrons flow due to mutually repulsive electrostatic forces.

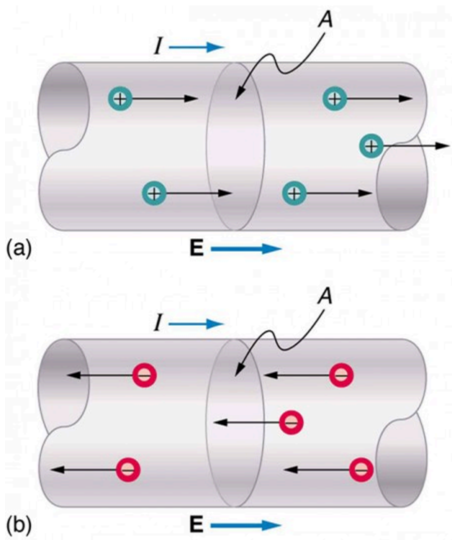


Figure 3. Current I is the rate at which charge moves through an area A, such as the cross-section of a wire. Conventional current is defined to move in the direction of the electric field. (a) Positive charges move in the direction of the electric field and the same direction as conventional current. (b) Negative charges move in the direction opposite to the electric field. Conventional current is in the direction opposite to the movement of negative charge. The flow of electrons is sometimes referred to as electronic flow.

Example 2. Calculating the Number of Electrons that Move through a Calculator

If the 0.300-mA current through the calculator mentioned in Example 1 is carried by electrons, how many electrons per second pass through it?

Strategy

The current calculated in the previous example was defined for the flow of positive charge. For electrons, the magnitude is the same, but the sign is opposite, $I_{\text{electrons}} = -0.300 \times 10^{-3}\,\text{C/s}$. Since each electron ($e^-$) has a charge of -1.60×10^{-19} C, we can convert the current in coulombs per second to electrons per second.

Solution

Starting with the definition of current, we have

$$I_{\text{electrons}} = \frac{\Delta Q_{\text{electrons}}}{\Delta t} = \frac{-0.300 \times 10^{-3}\text{C}}{\text{s}}$$

We divide this by the charge per electron, so that

$$\begin{aligned}\frac{e^-}{\text{s}} &= \frac{-0.300\times10^{-3}\text{C}}{\text{s}} \times \frac{1e^-}{-1.60\times10^{-19}\text{C}}. \\ &= 1.88 \times 10^{15}\,\frac{e^-}{\text{s}}\end{aligned}$$

Discussion

There are so many charged particles moving, even in small currents, that individual charges are not noticed, just as individual water molecules are not noticed in water flow. Even more amazing is that they do not always keep moving forward like soldiers in a parade. Rather they are like a crowd of people with movement in different directions but a general trend to move forward. There are lots of collisions with atoms in the metal wire and, of course, with other electrons.

Drift Velocity

Electrical signals are known to move very rapidly. Telephone conversations carried by currents in wires cover large distances without noticeable delays. Lights come on as soon as a switch is flicked. Most electrical signals carried by currents travel at speeds on the order of 10^8 m/s, a significant fraction of the speed of light. Interestingly, the individual charges that make up the current move *much* more slowly on average, typically drifting at speeds on the order of 10^{-4} m/s. How do we reconcile these two speeds, and what does it tell us about standard conductors? The high speed of electrical signals results from the fact that the force between charges acts rapidly at a distance. Thus, when a free charge is forced into a wire, as in Figure 4, the incoming charge pushes other charges ahead of it, which in turn push on charges farther down the line. The density of charge in a system cannot easily be increased, and so the signal is passed on rapidly. The resulting electrical shock wave moves through the system at nearly the speed of light. To be precise, this rapidly moving signal or shock wave is a rapidly propagating change in electric field.

Good conductors have large numbers of free charges in them. In metals, the free charges are free electrons. Figure 5 shows how free electrons move through an ordinary conductor. The distance that an individual electron can move between collisions with atoms or other electrons is quite small. The electron paths thus appear nearly random, like the motion of atoms in a gas. But there is an electric field in the conductor that causes the electrons to drift in the direction shown (opposite to the field, since they are negative). The *drift velocity* v_d is the average velocity of the free charges. Drift velocity is quite small, since there are so many free charges. If we have an estimate of the density of free electrons in a conductor, we can calculate the drift velocity for a given current. The larger the density, the lower the velocity required for a given current.

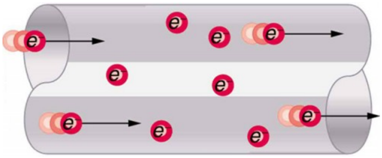


Figure 4. When charged particles are forced into this volume of a conductor, an equal number are quickly forced to leave. The repulsion between like charges makes it difficult to increase the number of charges in a volume. Thus, as one charge enters, another leaves almost immediately, carrying the signal rapidly forward.

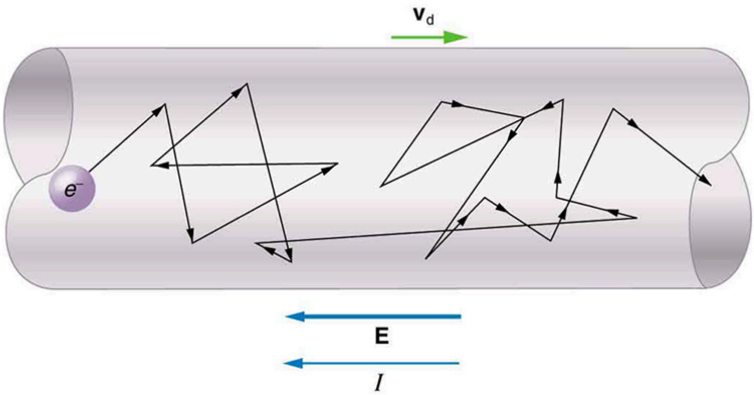


Figure 5. Free electrons moving in a conductor make many collisions with other electrons and atoms. The path of one electron is shown. The average velocity of the free charges is called the drift velocity, v_d, and it is in the direction opposite to the electric field for electrons. The collisions normally transfer energy to the conductor, requiring a constant supply of energy to maintain a steady current.

Conduction of Electricity and Heat

Good electrical conductors are often good heat conductors, too. This is because large numbers of free electrons can carry electrical current and can transport thermal energy.

The free-electron collisions transfer energy to the atoms of the conductor. The electric field does work in moving the electrons through a distance, but that work does not increase the kinetic energy (nor speed, therefore) of the electrons. The work is transferred to the conductor's atoms, possibly increasing temperature. Thus a continuous power input is required to keep a current flowing. An exception, of course, is found in superconductors, for reasons we shall explore in a later chapter. Superconductors can have a steady current without a continual supply of energy—a great energy savings. In contrast, the supply of energy can be useful, such as in a lightbulb filament. The supply of energy is necessary to increase the temperature of the tungsten filament, so that the filament glows.

Making Connections: Take-Home Investigation—Filament Observations

Find a lightbulb with a filament. Look carefully at the filament and describe its structure. To what points is the filament connected?

We can obtain an expression for the relationship between current and drift velocity by considering the number of free charges in a segment of wire, as illustrated in Figure 6. *The number of free charges per unit volume* is given the symbol n and depends on the material. The shaded segment has a volume Ax, so that the number of free charges in it is nAx. The charge ΔQ in this segment is thus $qnAx$, where q is the amount of charge on each carrier. (Recall that for electrons, q is -1.60×10^{-19} C.) Current is charge moved per unit time; thus, if all the original charges move out of this segment in time Δt, the current is

$$I = \frac{\Delta Q}{\Delta t} = \frac{qnAx}{\Delta t}.$$

Note that $x/\Delta t$ is the magnitude of the drift velocity, v_d, since the charges move an average distance x in a time Δt. Rearranging terms gives

$$I = nqAv_{\mathrm{d}},$$

where I is the current through a wire of cross-sectional area A made of a material with a free charge density n. The carriers of the current each have charge q and move with a drift velocity of magnitude v_{d}.

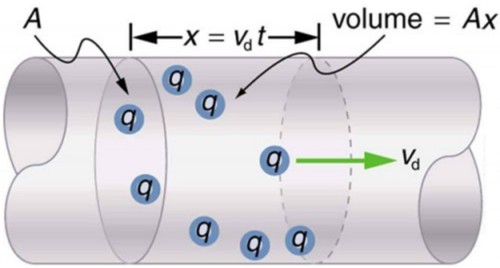


Figure 6. All the charges in the shaded volume of this wire move out in a time t, having a drift velocity of magnitude $v_{\mathrm{d}} = x/t$*. See text for further discussion.*

Note that simple drift velocity is not the entire story. The speed of an electron is much greater than its drift velocity. In addition, not all of the electrons in a conductor can move freely, and those that do might move somewhat faster or slower than the drift velocity. So what do we mean by free electrons? Atoms in a metallic conductor are packed in the form of a lattice structure. Some electrons are far enough away from the atomic nuclei that they do not experience the attraction of the nuclei as much as the inner electrons do. These are the free electrons. They are not bound to a single atom but can instead move freely among the atoms in a “sea” of electrons. These free electrons respond by accelerating when an electric field is applied. Of course as they move they collide with the atoms in the lattice and other electrons, generating thermal energy, and the conductor gets warmer. In an insulator, the organization of the atoms and the structure do not allow for such free electrons.

Example 3. Calculating Drift Velocity in a Common Wire

Calculate the drift velocity of electrons in a 12-gauge copper wire (which has a diameter of 2.053 mm) carrying a 20.0-A current, given that there is one free electron per copper atom. (Household wiring often contains 12-gauge copper wire, and the maximum current allowed in such wire is usually 20 A.) The density of copper is $8.80 \times 10^3 \text{ kg/m}^3$.

Strategy

We can calculate the drift velocity using the equation $I = nqAv_d$. The current I = 20.0 A is given, and $q = -1.60 \times 10^{-19}$ C is the charge of an electron. We can calculate the area of a cross-section of the wire using the formula $A = \pi r^2$, where r is one-half the given diameter, 2.053 mm. We are given the density of copper, $8.80 \times 10^3 \text{ kg/m}^3$ and the periodic table shows that the atomic mass of copper is 63.54 g/mol. We can use these two quantities along with Avogadro's number, 6.02×10^{23} atoms/mol, to determine n, the number of free electrons per cubic meter.

Solution

First, calculate the density of free electrons in copper.

There is one free electron per copper atom. Therefore, is the same as the number of copper atoms per m^3. We can now find n as follows:

$$\begin{aligned} n &= \frac{1e^-}{\text{atom}} \times \frac{6.02\times10^{23}\text{atoms}}{\text{mol}} \times \frac{1\text{ mol}}{63.54\text{ g}} \times \frac{1000\text{ g}}{\text{kg}} \times \frac{8.80\times10^3\text{kg}}{1\text{ m}^3} \\ &= 8.342\times10^{28}e^-/\text{m}^3 \end{aligned}$$

.

The cross-sectional area of the wire is

$$\begin{aligned} A &= \pi r^2 \\ &= \pi\left(\frac{2.053\times10^{-3}\text{m}}{2}\right)^2 \\ &= 3.310\times10^{-6}\text{m}^2. \end{aligned}$$

Rearranging $I = nqAv_d$ to isolate drift velocity gives

$$v_d = \frac{I}{\text{nqA}}$$

$$= \frac{20.0\text{ A}}{\left(8.342\times10^{28}/\text{m}^3\right)\left(-1.60\times10^{-19}\text{C}\right)\left(3.310\times10^{-6}\text{m}^2\right)}$$

$$= -4.53\times10^{-4}\text{m/s}.$$

Discussion

The minus sign indicates that the negative charges are moving in the direction opposite to conventional current. The small value for drift velocity (on the order of 10^{-4} m/s) confirms that the signal moves on the order of 10^{12} times faster (about 10^8 m/s) than the charges that carry it.

Section Summary

- Electric current I is the rate at which charge flows, given by
$$I = \frac{\Delta Q}{\Delta t},$$
where ΔQ is the amount of charge passing through an area in time Δt.
- The direction of conventional current is taken as the direction in which positive charge moves.
- The SI unit for current is the ampere (A), where 1 A = 1 C/s.
- Current is the flow of free charges, such as electrons and ions.
- Drift velocity v_d is the average speed at which these charges move.
- Current I is proportional to drift velocity v_d, as expressed in the relationship $I = \mathrm{nqAv_d}$. Here, I is the current through a wire of cross-sectional area A. The wire's material has a free-charge density n, and each carrier has charge q and a drift velocity v_d.
- Electrical signals travel at speeds about 10^{12} times greater than the drift velocity of free electrons.

Conceptual Questions

1. Can a wire carry a current and still be neutral—that is, have a total charge of zero? Explain.
2. Car batteries are rated in ampere-hours (A · h). To what physical quantity do ampere-hours correspond (voltage, charge, . . .), and what relationship do ampere-hours have to energy content?

3. If two different wires having identical cross-sectional areas carry the same current, will the drift velocity be higher or lower in the better conductor? Explain in terms of the equation $v_{\mathrm{d}} = \frac{I}{\mathrm{nqA}}$, by considering how the density of charge carriers n relates to whether or not a material is a good conductor.
4. Why are two conducting paths from a voltage source to an electrical device needed to operate the device?
5. In cars, one battery terminal is connected to the metal body. How does this allow a single wire to supply current to electrical devices rather than two wires?
6. Why isn't a bird sitting on a high-voltage power line electrocuted? Contrast this with the situation in which a large bird hits two wires simultaneously with its wings.

Problems & Exercises

1. What is the current in milliamperes produced by the solar cells of a pocket calculator through which 4.00 C of charge passes in 4.00 h?

2. A total of 600 C of charge passes through a flashlight in 0.500 h. What is the average current?

3. What is the current when a typical static charge of 0.250 μC moves from your finger to a metal doorknob in 1.00 μs?

4. Find the current when 2.00 nC jumps between your comb and hair over a 0.500 −μs time interval.

5. A large lightning bolt had a 20,000-A current and moved 30.0 C of charge. What was its duration?

6. The 200-A current through a spark plug moves 0.300 mC of charge. How long does the spark last?

7. (a) A defibrillator sends a 6.00-A current through the chest of a patient by applying a 10,000-V potential as in the figure below. What is the resistance of the path? (b) The defibrillator paddles make contact with the patient through a conducting gel that greatly reduces the path resistance. Discuss the difficulties that would ensue if a larger voltage were used to produce the same current through the patient, but with the path having perhaps 50 times the resistance. (Hint: The current must be about the same, so a higher voltage would imply greater power. Use this equation for power: $P = I^2R$.)

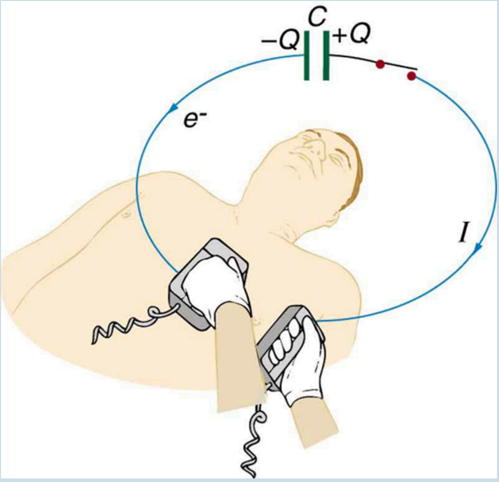


Figure 7. The capacitor in a defibrillation unit drives a current through the heart of a patient.

8. During open-heart surgery, a defibrillator can be used to bring a patient out of cardiac arrest. The resistance of the path is 500 Ω and a 10.0-mA current is needed. What voltage should be applied?

9. (a) A defibrillator passes 12.0 A of current through the torso of a person for 0.0100 s. How much charge moves? (b) How many electrons pass through the wires connected to the patient? (See Figure 7.)

10. A clock battery wears out after moving 10,000 C of charge through the clock at a rate of 0.500 mA. (a) How long did the clock run? (b) How many electrons per second flowed?

11. The batteries of a submerged non-nuclear submarine supply 1000 A at full speed ahead. How long does it take to move Avogadro's number (6.02×10^{23}) of electrons at this rate?

12. Electron guns are used in X-ray tubes. The electrons are accelerated through a relatively large voltage and directed onto a metal target, producing X-rays. (a) How many electrons per second strike the target if the current is 0.500 mA? (b) What charge strikes the target in 0.750 s?

13. A large cyclotron directs a beam of He^{++} nuclei onto a target with a beam current of 0.250 mA. (a) How many He^{++} nuclei per second is this? (b) How long does it take for 1.00 C to strike the target? (c) How long before 1.00 mol of He^{++} nuclei strike the target?

14. Repeat the above *Example 3: Calculating Drift and Velocity in a Common Wire*, but for a wire made of silver and given there is one free electron per silver atom.

15. Using the results of the above example, *Example 3: Calculating Drift and Velocity in a Common Wire*, find the drift velocity in a copper wire of twice the diameter and carrying 20.0 A.

16. A 14-gauge copper wire has a diameter of 1.628 mm. What magnitude current flows when the drift velocity is 1.00 mm/s? (See above *Example 3: Calculating Drift and Velocity in a Common Wire* for useful information.)

17. SPEAR, a storage ring about 72.0 m in diameter at the Stanford Linear Accelerator (closed in 2009), has a 20.0-A

circulating beam of electrons that are moving at nearly the speed of light. (See Figure 8.) How many electrons are in the beam?

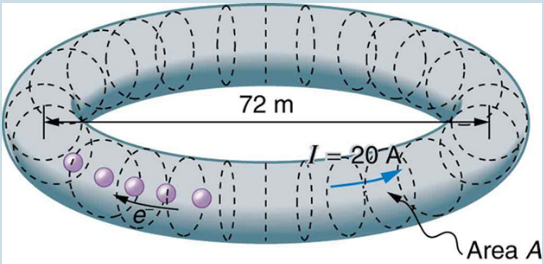


Figure 8. Electrons circulating in the storage ring called SPEAR constitute a 20.0-A current. Because they travel close to the speed of light, each electron completes many orbits in each second.

Glossary

electric current:
: the rate at which charge flows, $I = \Delta Q/\Delta t$

ampere:
: (amp) the SI unit for current; $1\ \text{A} = 1\ \text{C/s}$

drift velocity:
: the average velocity at which free charges flow in response to an electric field

Selected Solutions to Problems & Exercises

1. 0.278 mA

3. 0.250 A

5. 1.50 ms

7. (a) 1.67 kΩ (b) If a 50 times larger resistance existed, keeping the current about the same, the power would be increased by a factor of about 50 (based on the equation $P = I^2R$), causing much more energy to be transferred to the skin, which could cause serious burns. The gel used reduces the resistance, and therefore reduces the power transferred to the skin.

9. (a) 0.120 C (b) 7.50×10^{17} electrons

11. 96.3 s

13.(a) 7.81×10^{14} He^{++} nuclei/s
(b) 4.00×10^{3} s
(c) 7.71×10^{8} s

15. -1.13×10^{-4} m/s

17. 9.42×10^{13} electrons

23. Ohm's Law: Resistance and Simple Circuits

Learning Objectives

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

- 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 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.

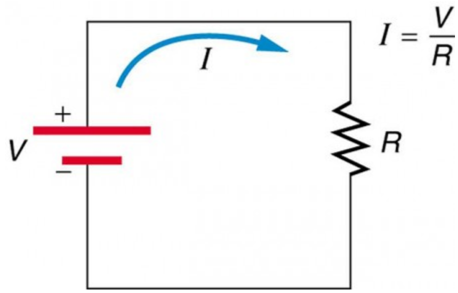


Figure 1. A simple electric circuit in which a closed path for current to flow is supplied by conductors (usually metal wires) connecting a load to the terminals of a battery, represented by the red parallel lines. The zigzag symbol represents the single resistor and includes any resistance in the connections to the voltage source.

Example 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

$$R=\frac{V}{I}=\frac{\text{12}\text{.}\text{0 V}}{2\text{.}\text{50 A}}=\text{4}\text{.}\text{80 }\Omega\\ .$$

Discussion

This is a relatively small resistance, but it is larger than the cold resistance of the headlight. As we shall see in Resistance and Resistivity, 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, as will be seen in Resistance and Resistivity. 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 1* above 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 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 2.)

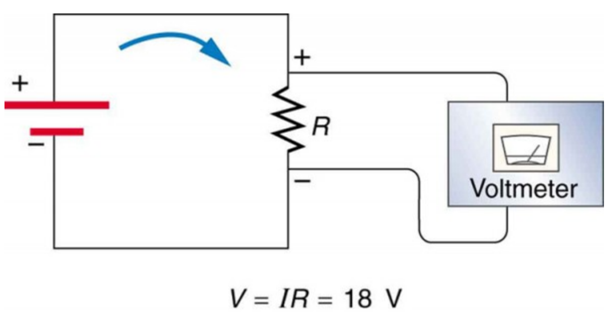


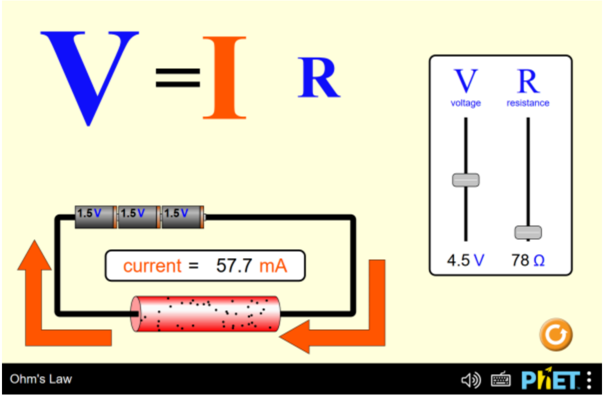
Figure 2. The voltage drop across a resistor in a simple circuit equals the voltage output of the battery.

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.

PhET Explorations: Ohm's Law

See how the equation form of Ohm's law relates to a simple circuit. Adjust the voltage and resistance, and see the current change according to Ohm's law. The sizes of the symbols in the equation change to match the circuit diagram.



Click to run the simulation.

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 Ω = 1 V/A.
- There is a voltage or *IR* drop across a resistor, caused by the current flowing through it, given by V = *IR*.

Conceptual Questions

1. The *IR* drop across a resistor means that there is a change in potential or voltage across the resistor. Is there any change in current as it passes through a resistor? Explain.
2. How is the *IR* drop in a resistor similar to the pressure drop in a fluid flowing through a pipe?

Problems & Exercises

1. What current flows through the bulb of a 3.00-V flashlight when its hot resistance is 3.60 Ω?

2. Calculate the effective resistance of a pocket calculator that has a 1.35-V battery and through which 0.200 mA flows.

3. What is the effective resistance of a car's starter motor when 150 A flows through it as the car battery applies 11.0 V to the motor?

4. How many volts are supplied to operate an indicator light on a DVD player that has a resistance of 140 Ω, given that 25.0 mA passes through it?

5. (a) Find the voltage drop in an extension cord having a 0.0600-Ω resistance and through which 5.00 A is flowing.

(b) A cheaper cord utilizes thinner wire and has a resistance of 0.300 Ω. What is the voltage drop in it when 5.00 A flows? (c) Why is the voltage to whatever appliance is being used reduced by this amount? What is the effect on the appliance?

6. A power transmission line is hung from metal towers with glass insulators having a resistance of 1.00×10^9 Ω. What current flows through the insulator if the voltage is 200 kV? (Some high-voltage lines are DC.)

Glossary

Ohm's law:
: an empirical relation stating that the current I is proportional to the potential difference V, ∝ 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 1Ω = 1 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

Selected Solutions to Problems & Exercises

1. 0.833 A

3. $7.33 \times 10^{-2}\ \Omega$

5. (a) 0.300 V

(b) 1.50 V

(c) The voltage supplied to whatever appliance is being used is reduced because the total voltage drop from the wall to the final output of the appliance is fixed. Thus, if the voltage drop across the extension cord is large, the voltage drop across the appliance is significantly decreased, so the power output by the appliance can be significantly decreased, reducing the ability of the appliance to work properly.

24. Resistance and Resistivity

Learning Objectives

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

- Explain the concept of resistivity.
- Use resistivity to calculate the resistance of specified configurations of material.
- Use the thermal coefficient of resistivity to calculate the change of resistance with temperature.

Material and Shape Dependence of Resistance

The resistance of an object depends on its shape and the material of which it is composed. The cylindrical resistor in Figure 1 is easy to analyze, and, by so doing, we can gain insight into the resistance of more complicated shapes. As you might expect, the cylinder's electric resistance R is directly proportional to its length L, similar to the resistance of a pipe to fluid flow. The longer the cylinder, the more collisions charges will make with its atoms. The greater the diameter of the cylinder, the more current it can carry (again similar to the flow of fluid through a pipe). In fact, R is inversely proportional to the cylinder's cross-sectional area A.

For a given shape, the resistance depends on the material of which the object is composed. Different materials offer different resistance to the flow of charge. We define the *resistivity*ρ of a substance so that the **resistance R** of an object is directly proportional to ρ. Resistivity ρ is an *intrinsic* property of a material, independent of its shape or size. The resistance R of a uniform cylinder of length L, of cross-sectional area A, and made of a material with resistivity ρ, is

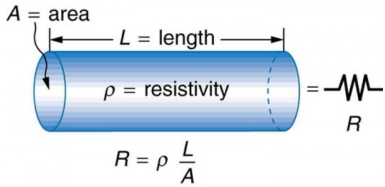


Figure 1. A uniform cylinder of length L *and cross-sectional area* A*. Its resistance to the flow of current is similar to the resistance posed by a pipe to fluid flow. The longer the cylinder, the greater its resistance. The larger its cross-sectional area* A*, the smaller its resistance.*

$$R = \frac{\rho L}{A}.$$

Table 1 gives representative values of ρ. The materials listed in the table are separated into categories of conductors, semiconductors, and insulators, based on broad groupings of resistivities. Conductors have the smallest resistivities, and insulators have the largest; semiconductors have intermediate resistivities. Conductors have varying but large free charge densities, whereas most charges in insulators are bound to atoms and are not free to move. Semiconductors are intermediate, having far fewer free charges than conductors, but having properties that make the number of free charges depend strongly on the type and amount of impurities in the semiconductor. These unique properties of semiconductors are put to use in modern electronics, as will be explored in later chapters.

Table 1. Resistivities ρ of Various materials at 20° C

Material	Resistivity ρ ($\Omega \cdot m$)
Conductors	
Silver	1.59×10^{-8}
Copper	1.72×10^{-8}
Gold	2.44×10^{-8}
Aluminum	2.65×10^{-8}
Tungsten	5.6×10^{-8}
Iron	9.71×10^{-8}
Platinum	10.6×10^{-8}
Steel	20×10^{-8}
Lead	22×10^{-8}
Manganin (Cu, Mn, Ni alloy)	44×10^{-8}
Constantan (Cu, Ni alloy)	49×10^{-8}
Mercury	96×10^{-8}
Nichrome (Ni, Fe, Cr alloy)	100×10^{-8}
Semiconductors[1]	
Carbon (pure)	3.5×10^{5}
Carbon	$(3.5 - 60) \times 10^{5}$
Germanium (pure)	600×10^{-3}
Germanium	$(1–600) \times 10^{-3}$
Silicon (pure)	2300
Silicon	0.1–2300
Insulators	
Amber	5×10^{14}
Glass	$10^{9} - 10^{14}$
Lucite	$>10^{13}$
Mica	$10^{11} - 10^{15}$
Quartz (fused)	75×10^{16}

Material	Resistivity ρ ($\Omega \cdot m$)
Rubber (hard)	$10^{13} - 10^{16}$
Sulfur	10^{15}
Teflon	$>10^{13}$
Wood	$10^{8} - 10^{11}$

Example 1. Calculating Resistor Diameter: A Headlight Filament

A car headlight filament is made of tungsten and has a cold resistance of 0.350 Ω. If the filament is a cylinder 4.00 cm long (it may be coiled to save space), what is its diameter?

Strategy

We can rearrange the equation $R = \frac{\rho L}{A}$ to find the cross-sectional area A of the filament from the given information. Then its diameter can be found by assuming it has a circular cross-section.

1. Values depend strongly on amounts and types of impurities

Solution

The cross-sectional area, found by rearranging the expression for the resistance of a cylinder given in $R = \frac{\rho L}{A}$, is

$$A = \frac{\rho L}{R}.$$

Substituting the given values, and taking ρ from Table 1, yields

$$\begin{aligned} A &= \frac{\left(5.6\times10^{-8}\Omega\cdot\text{m}\right)\left(4.00\times10^{-2}\text{m}\right)}{0.350\Omega} \\ &= 6.40 \times 10^{-9}\text{m}^2 \end{aligned}.$$

The area of a circle is related to its diameter D by

$$A = \frac{\pi D^2}{4}.$$

Solving for the diameter D, and substituting the value found for A, gives

$$\begin{aligned} D &= 2\left(\frac{A}{p}\right)^{\frac{1}{2}} = 2\left(\frac{6.40\times10^{-9}\text{m}^2}{3.14}\right)^{\frac{1}{2}} \\ &= 9.0 \times 10^{-5}\text{m} \end{aligned}.$$

Discussion

The diameter is just under a tenth of a millimeter. It is quoted to only two digits, because ρ is known to only two digits.

Temperature Variation of Resistance

The resistivity of all materials depends on temperature. Some even become superconductors (zero resistivity) at very low temperatures. (See Figure 2.)

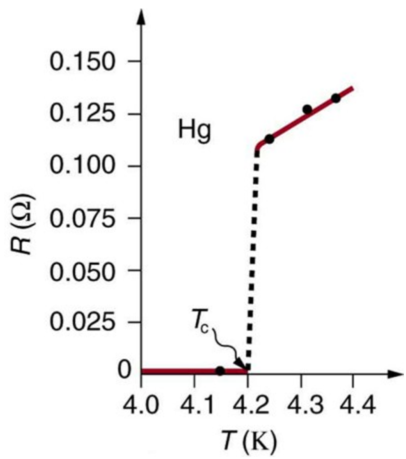


Figure 2. The resistance of a sample of mercury is zero at very low temperatures—it is a superconductor up to about 4.2 K. Above that critical temperature, its resistance makes a sudden jump and then increases nearly linearly with temperature.

Conversely, the resistivity of conductors increases with increasing temperature. Since the atoms vibrate more rapidly and over larger distances at higher temperatures, the electrons moving through a metal make more collisions, effectively making the resistivity higher. Over relatively small temperature changes (about 100°C or less), resistivity ρ varies with temperature change ΔT as expressed in the following equation

$$\rho = \rho_0 (1 + \alpha \Delta T),$$

where ρ_0 is the original resistivity and α is the *temperature coefficient of resistivity*. (See the values of α in Table 2 below.) For larger temperature changes, α may vary or a nonlinear equation may be needed to find ρ. Note that α is positive for metals, meaning their resistivity increases with temperature. Some alloys have been developed specifically to have a small temperature dependence. Manganin (which is made of copper, manganese and nickel), for example, has α close to zero (to three digits on the scale in Table 2), and so its resistivity varies only slightly with temperature. This is useful for making a temperature-independent resistance standard, for example.

Table 2. Tempature Coefficients of Resistivity α

Material	Coefficient α (1/°C)[2]
Conductors	
Silver	3.8×10^{-3}
Copper	3.9×10^{-3}
Gold	3.4×10^{-3}
Aluminum	3.9×10^{-3}
Tungsten	4.5×10^{-3}
Iron	5.0×10^{-3}
Platinum	3.93×10^{-3}
Lead	3.9×10^{-3}
Manganin (Cu, Mn, Ni alloy)	0.000×10^{-3}
Constantan (Cu, Ni alloy)	0.002×10^{-3}
Mercury	0.89×10^{-3}
Nichrome (Ni, Fe, Cr alloy)	0.4×10^{-3}
Semiconductors	
Carbon (pure)	-0.5×10^{-3}
Germanium (pure)	-50×10^{-3}
Silicon (pure)	-70×10^{-3}

Note also that α is negative for the semiconductors listed in Table 2, meaning that their resistivity decreases with increasing temperature. They become better conductors at higher temperature, because increased thermal agitation increases the number of free charges available to carry current. This property of decreasing ρ with temperature is also related to the type and amount of impurities present in the semiconductors. The resistance

2. Values at 20°C.

of an object also depends on temperature, since R_0 is directly proportional to ρ. For a cylinder we know $R = \rho L/A$, and so, if L and A do not change greatly with temperature, R will have the same temperature dependence as ρ. (Examination of the coefficients of linear expansion shows them to be about two orders of magnitude less than typical temperature coefficients of resistivity, and so the effect of temperature on L and A is about two orders of magnitude less than on ρ.) Thus,

$$R = R_0(1 + \alpha\Delta T)$$

is the temperature dependence of the resistance of an object, where R_0 is the original resistance and R is the resistance after a temperature change ΔT. Numerous thermometers are based on the effect of temperature on resistance. (See Figure 3.) One of the most common is the thermistor, a semiconductor crystal with a strong temperature dependence, the resistance of which is measured to obtain its temperature. The device is small, so that it quickly comes into thermal equilibrium with the part of a person it touches.

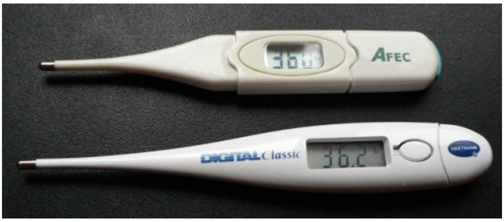


Figure 3. These familiar thermometers are based on the automated measurement of a thermistor's temperature-dependent resistance. (credit: Biol, Wikimedia Commons)

Example 2. Calculating Resistance: Hot-Filament Resistance

Although caution must be used in applying $\rho = \rho_0(1 + \alpha\Delta T)$ and $R = R_0(1 + \alpha\Delta T)$ for temperature changes greater than 100°C, for tungsten the equations work reasonably well for very large temperature changes. What, then, is the resistance of the tungsten filament in the previous example if its temperature is increased from room temperature (20°C) to a typical operating temperature of 2850°C?

Strategy

This is a straightforward application of $R = R_0(1 + \alpha\Delta T)$, since the original resistance of the filament was given to be $R_0 = 0.350\ \Omega$, and the temperature change is $\Delta T = 2830°C$.

Solution

The hot resistance R is obtained by entering known values into the above equation:

$$\begin{aligned} R &= R_0\left(1 + \alpha\Delta T\right) \\ &= (0.350\Omega)\left[1 + \left(4.5 \times 10^{-3}/^{\circ}\ \mathrm{C}\right)\left(2830^{\circ}\ \mathrm{C}\right)\right] \\ &= 4.8\Omega \end{aligned}$$

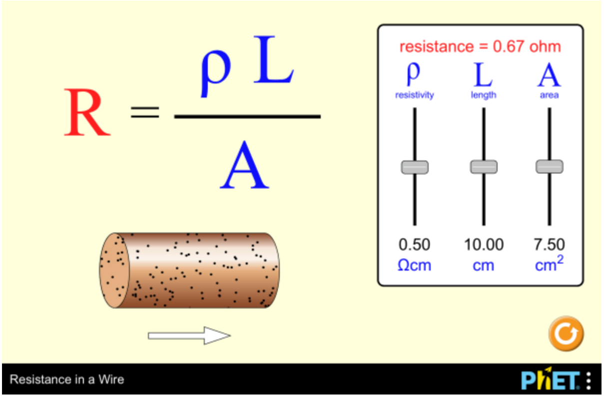
.

Discussion

This value is consistent with the headlight resistance example in Ohm's Law: Resistance and Simple Circuits.

PhET Explorations: Resistance in a Wire

Learn about the physics of resistance in a wire. Change its resistivity, length, and area to see how they affect the wire's resistance. The sizes of the symbols in the equation change along with the diagram of a wire.



Click to run the simulation.

Section Summary

- The resistance R of a cylinder of length L and cross-sectional area A is $R = \frac{\rho L}{A}$, where ρ is the resistivity of the material.
- Values of ρ in Table 1 show that materials fall into three groups—*conductors, semiconductors, and insulators*.
- Temperature affects resistivity; for relatively small temperature changes ΔT, resistivity is $\rho = \rho_0 \left(1 + \alpha \Delta T\right)$, where ρ_0 is the original resistivity and \text{\alpha } is the temperature coefficient of resistivity.
- Table 2 gives values for α, the temperature coefficient of resistivity.
- The resistance R of an object also varies with temperature:

$R = R_0\left(1 + \alpha\Delta T\right)$, where R_0 is the original resistance, and R is the resistance after the temperature change.

Conceptual Questions

1. In which of the three semiconducting materials listed in Table 1 do impurities supply free charges? (Hint: Examine the range of resistivity for each and determine whether the pure semiconductor has the higher or lower conductivity.)

2. Does the resistance of an object depend on the path current takes through it? Consider, for example, a rectangular bar—is its resistance the same along its length as across its width? (See Figure 5.)

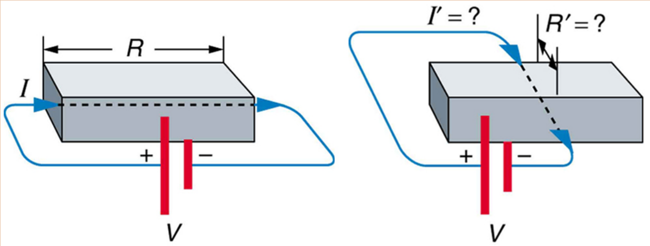


Figure 5. Does current taking two different paths through the same object encounter different resistance?

3. If aluminum and copper wires of the same length have the same resistance, which has the larger diameter? Why?

4. Explain why $R = R_0\left(1 + \alpha\Delta T\right)$ for the temperature variation of the resistance R of an object is not

as accurate as $\rho = \rho_0\left(1 + \alpha\Delta T\right)$, which gives the temperature variation of resistivity ρ.

Problems & Exercises

1. What is the resistance of a 20.0-m-long piece of 12-gauge copper wire having a 2.053-mm diameter?

2. The diameter of 0-gauge copper wire is 8.252 mm. Find the resistance of a 1.00-km length of such wire used for power transmission.

3. If the 0.100-mm diameter tungsten filament in a light bulb is to have a resistance of 0.200 Ω at 20°C, how long should it be?

4. Find the ratio of the diameter of aluminum to copper wire, if they have the same resistance per unit length (as they might in household wiring).

5. What current flows through a 2.54-cm-diameter rod of pure silicon that is 20.0 cm long, when 1.00×10^3V is applied to it? (Such a rod may be used to make nuclear-particle detectors, for example.)

6. (a) To what temperature must you raise a copper wire, originally at 20.0°C, to double its resistance, neglecting any changes in dimensions? (b) Does this happen in household wiring under ordinary circumstances?

7. A resistor made of Nichrome wire is used in an application where its resistance cannot change more than 1.00% from its value at 20.0°C. Over what temperature range can it be used?

8. Of what material is a resistor made if its resistance is 40.0% greater at 100°C than at 20.0°C?

9. An electronic device designed to operate at any temperature in the range from −10.0°C to 55.0°C contains pure carbon resistors. By what factor does their resistance increase over this range?

10. (a) Of what material is a wire made, if it is 25.0 m long with a 0.100 mm diameter and has a resistance of 77.7 Ω at 20.0°C? (b) What is its resistance at 150°C?

11. Assuming a constant temperature coefficient of resistivity, what is the maximum percent decrease in the resistance of a constantan wire starting at 20.0°C?

12. A wire is drawn through a die, stretching it to four times its original length. By what factor does its resistance increase?

13. A copper wire has a resistance of 0.500 Ω at 20.0°C, and an iron wire has a resistance of 0.525 Ω at the same temperature. At what temperature are their resistances equal?

14. (a) Digital medical thermometers determine temperature by measuring the resistance of a semiconductor device called a thermistor (which has $\alpha = -0.0600/°C$) when it is at the same temperature as the

patient. What is a patient's temperature if the thermistor's resistance at that temperature is 82.0% of its value at 37.0°C (normal body temperature)? (b) The negative value for α may not be maintained for very low temperatures. Discuss why and whether this is the case here. (Hint: Resistance can't become negative.)

15. **Integrated Concepts** (a) Redo Exercise 2 taking into account the thermal expansion of the tungsten filament. You may assume a thermal expansion coefficient of $12 \times 10^{-6}/°C$. (b) By what percentage does your answer differ from that in the example?

16. **Unreasonable Results** (a) To what temperature must you raise a resistor made of constantan to double its resistance, assuming a constant temperature coefficient of resistivity? (b) To cut it in half? (c) What is unreasonable about these results? (d) Which assumptions are unreasonable, or which premises are inconsistent?

Footnotes

1. 1 Values depend strongly on amounts and types of impurities
2. 2 Values at 20°C.

Glossary

resistivity:
an intrinsic property of a material, independent of its shape or

size, directly proportional to the resistance, denoted by ρ

temperature coefficient of resistivity:
an empirical quantity, denoted by α, which describes the change in resistance or resistivity of a material with temperature

Selected Solutions to Problems & Exercises

1. $0.104\ \Omega$

3. $2.8 \times 10^{-2}\text{m}$

5. $1.10 \times 10^{-3}\text{A}$

7. -5°C to 45°C

9. 1.03

11. 0.06%

13. -17°C

15. (a) $4.7\ \Omega$ (total) (b) 3.0% decrease

25. Electric Power and Energy

Learning Objectives

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

- Calculate the power dissipated by a resistor and power supplied by a power supply.
- Calculate the cost of electricity under various circumstances.

Power in Electric Circuits

Power is associated by many people with electricity. Knowing that power is the rate of energy use or energy conversion, what is the expression for *electric power*? Power transmission lines might come to mind. We also think of lightbulbs in terms of their power ratings in watts. Let us compare a 25-W bulb with a 60-W bulb. (See Figure 1(a).) Since both operate on the same voltage, the 60-W bulb must draw more current to have a greater power rating. Thus the 60-W bulb's resistance must be lower than that of a 25-W bulb. If we increase voltage, we also increase power. For example, when a 25-W bulb that is designed to operate on 120 V is connected to 240 V, it briefly glows very brightly and then burns out. Precisely how are voltage, current, and resistance related to electric power?

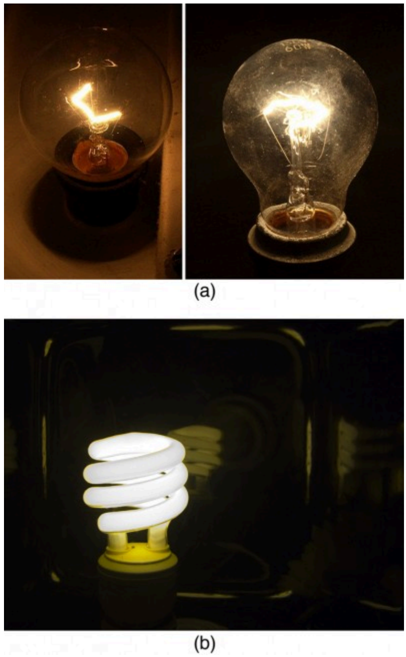


Figure 1. (a) Which of these lightbulbs, the 25-W bulb (upper left) or the 60-W bulb (upper right), has the higher resistance? Which draws more current? Which uses the most energy? Can you tell from the color that the 25-W filament is cooler? Is the brighter bulb a different color and if so why? (credits: Dickbauch, Wikimedia Commons; Greg Westfall, Flickr) (b) This compact fluorescent light (CFL) puts out the same intensity of light as the 60-W bulb, but at 1/4 to 1/10 the input power. (credit: dbgg1979, Flickr)

Electric energy depends on both the voltage involved and the charge moved. This is expressed most simply as PE = qV, where q is the charge moved and V is the voltage (or more precisely, the potential difference the charge moves through). Power is the rate at which energy is moved, and so electric power is

$$P = \frac{PE}{t} = \frac{qV}{t}.$$

Recognizing that current is $I = q/t$ (note that $\Delta t = t$ here), the expression for power becomes

$$P = IV$$

Electric power (P) is simply the product of current times voltage. Power has familiar units of watts. Since the SI unit for potential energy (PE) is the joule, power has units of joules per second, or watts. Thus, 1 A ·V= 1 W. For example, cars often have one or more auxiliary power outlets with which you can charge a cell phone or other electronic devices. These outlets may be rated at 20 A, so that the circuit can deliver a maximum power $P = IV = (20\text{ A})(12\text{ V}) = 240\text{ W}$. In some applications, electric power may be expressed as volt-amperes or even kilovolt-amperes (1 kA ·V = 1 kW). To see the relationship of power to resistance, we combine Ohm's law with $P = IV$. Substituting $I = V/R$ gives $P = (V/R)V = V^2/R$. Similarly, substituting $V = IR$ gives $P = I(IR) = I^2R$. Three expressions for electric power are listed together here for convenience:

$$P = \mathrm{IV}$$

$$P = \frac{V^2}{R}$$

$$P = I^2R.$$

Note that the first equation is always valid, whereas the other two can be used only for resistors. In a simple circuit, with one voltage source and a single resistor, the power supplied by the voltage source and that dissipated by the resistor are identical. (In more complicated circuits, P can be the power dissipated by a single device and not the total power in the circuit.) Different insights can be gained from the three different expressions for electric power. For example, $P = V^2/R$ implies that the lower the resistance connected to a given voltage source, the greater the power delivered. Furthermore, since voltage is squared in $P = V^2/R$, the effect of applying a higher voltage is perhaps greater than expected. Thus, when the voltage is doubled to a 25-W bulb, its power nearly quadruples to about 100 W, burning it out. If the bulb's resistance remained constant, its power would be exactly 100 W, but at the higher temperature its resistance is higher, too.

Example 1. Calculating Power Dissipation and Current: Hot and Cold Power

(a) Consider the examples given in Ohm's Law: Resistance and Simple Circuits and Resistance and Resistivity. Then find the power dissipated by the car headlight in these examples, both when it is hot and when it is cold. (b) What current does it draw when cold?

Strategy for (a)

For the hot headlight, we know voltage and current, so we can use $P = IV$ to find the power. For the cold headlight, we know the voltage and resistance, so we can use $P = V^2/R$ to find the power.

Solution for (a)

Entering the known values of current and voltage for the hot headlight, we obtain

$$P = IV = (2.50\text{ A})(12.0\text{ V}) = 30.0\text{ W}.$$

The cold resistance was 0.350 Ω, and so the power it uses when first switched on is

$$P = \frac{V^2}{R} = \frac{(12.0\text{ V})^2}{0.350\ \Omega} = 411\text{ W}.$$

Discussion for (a)

The 30 W dissipated by the hot headlight is typical. But the 411 W when cold is surprisingly higher. The initial power quickly decreases as the bulb's temperature increases and its resistance increases.

Strategy and Solution for (b)

The current when the bulb is cold can be found several different ways. We rearrange one of the power equations, $P = I^2R$, and enter known values, obtaining

$$I = \sqrt{\frac{P}{R}} = \sqrt{\frac{411\text{ W}}{0.350\ \Omega}} = 34.3\text{ A}.$$

Discussion for (b)

The cold current is remarkably higher than the steady-state value of 2.50 A, but the current will quickly decline to that value as the bulb's temperature increases. Most fuses and circuit breakers (used to limit the current in a circuit) are designed to tolerate very high currents briefly as a device comes on. In some cases, such as with electric motors, the current remains high for several seconds, necessitating special “slow blow” fuses.

The Cost of Electricity

The more electric appliances you use and the longer they are left on, the higher your electric bill. This familiar fact is based on the relationship between energy and power. You pay for the energy used. Since $P = E/t$, we see that

$$E = Pt$$

is the energy used by a device using power P for a time interval t. For example, the more lightbulbs burning, the greater P used; the longer they are on, the greater t is. The energy unit on electric bills is the kilowatt-hour ($\text{kW} \cdot \text{h}$), consistent with the relationship $E = Pt$. It is easy to estimate the cost of operating electric appliances if you have some idea of their power consumption rate in watts or kilowatts, the time they are on in hours, and the cost per kilowatt-hour for your electric utility. Kilowatt-hours, like all other specialized energy units such as food calories, can be converted to joules. You can prove to yourself that $1\ \text{kW} \cdot \text{h} = 3.6 \times 10^6\ \text{J}$.

The electrical energy (E) used can be reduced either by reducing the time of use or by reducing the power consumption of that appliance or fixture. This will not only reduce the cost, but it will also result in a reduced impact on the environment. Improvements to lighting are some of the fastest ways to reduce the electrical energy used in a home or business. About 20% of a home’s use of energy goes to lighting, while the number for commercial establishments is closer to 40%. Fluorescent lights are about four times more efficient than incandescent lights—this is true for both the long tubes and the compact fluorescent lights (CFL). (See Figure 1(b).) Thus, a 60-W incandescent bulb can be replaced by a 15-W

CFL, which has the same brightness and color. CFLs have a bent tube inside a globe or a spiral-shaped tube, all connected to a standard screw-in base that fits standard incandescent light sockets. (Original problems with color, flicker, shape, and high initial investment for CFLs have been addressed in recent years.) The heat transfer from these CFLs is less, and they last up to 10 times longer. The significance of an investment in such bulbs is addressed in the next example. New white LED lights (which are clusters of small LED bulbs) are even more efficient (twice that of CFLs) and last 5 times longer than CFLs. However, their cost is still high.

Making Connections: Energy, Power, and Time

The relationship $E = Pt$ is one that you will find useful in many different contexts. The energy your body uses in exercise is related to the power level and duration of your activity, for example. The amount of heating by a power source is related to the power level and time it is applied. Even the radiation dose of an X-ray image is related to the power and time of exposure.

Example 2. Calculating the Cost Effectiveness of Compact Fluorescent Lights (CFL)

If the cost of electricity in your area is 12 cents per kWh, what is the total cost (capital plus operation) of using a 60-W incandescent bulb for 1000 hours (the lifetime of that bulb) if the bulb cost 25 cents? (b) If we replace this bulb with a compact fluorescent light that provides the same light output, but at one-quarter the wattage, and which costs $1.50 but lasts 10 times longer (10,000 hours), what will that total cost be?

Strategy

To find the operating cost, we first find the energy used in kilowatt-hours and then multiply by the cost per kilowatt-hour.

Solution for (a)

The energy used in kilowatt-hours is found by entering the power and time into the expression for energy:

$$E = Pt = (60\text{ W})(1000\text{ h}) = 60{,}000\text{ W}\cdot\text{h}$$

In kilowatt-hours, this is

$$E = 60.0\text{ kW}\cdot\text{h}.$$

Now the electricity cost is

$$\text{cost} = (60.0\text{ kW}\cdot\text{h})\,(\$0.12/\text{kW}\cdot\text{h}) = \$\,7.20.$$

The total cost will be $7.20 for 1000 hours (about one-half year at 5 hours per day).

Solution for (b)

Since the CFL uses only 15 W and not 60 W, the electricity cost will be \$7.20/4 = \$1.80. The CFL will last 10 times longer than the incandescent, so that the investment cost will be 1/10 of the bulb cost for that time period of use, or 0.1(\$1.50) = \$0.15. Therefore, the total cost will be \$1.95 for 1000 hours.

Discussion

Therefore, it is much cheaper to use the CFLs, even though the initial investment is higher. The increased cost of labor that a business must include for replacing the incandescent bulbs more often has not been figured in here.

Making Connections: Take-Home Experiment—Electrical Energy Use Inventory

1) Make a list of the power ratings on a range of appliances in your home or room. Explain why something like a toaster has a higher rating than a digital clock. Estimate the energy consumed by these appliances in an average day (by estimating their time of use). Some

appliances might only state the operating current. If the household voltage is 120 V, then use $P = IV$. 2) Check out the total wattage used in the rest rooms of your school's floor or building. (You might need to assume the long fluorescent lights in use are rated at 32 W.) Suppose that the building was closed all weekend and that these lights were left on from 6 p.m. Friday until 8 a.m. Monday. What would this oversight cost? How about for an entire year of weekends?

Section Summary

- Electric power P is the rate (in watts) that energy is supplied by a source or dissipated by a device.
- Three expressions for electrical power are

$$P = \text{IV}$$
$$P = \frac{V^2}{R}$$
$$P = I^2 R.$$

- The energy used by a device with a power P over a time t is $E = Pt$.

Conceptual Questions

1. Why do incandescent lightbulbs grow dim late in their lives, particularly just before their filaments break?

The power dissipated in a resistor is given by $P = V^2/R$ which means power decreases if resistance increases. Yet this power is also given by $P = I^2R$, which means power increases if resistance increases. Explain why there is no contradiction here.

Problems & Exercises

1. What is the power of a 1.00×10^2 MV lightning bolt having a current of 2.00×10^4 A?

2. What power is supplied to the starter motor of a large truck that draws 250 A of current from a 24.0-V battery hookup?

3. A charge of 4.00 C of charge passes through a pocket calculator's solar cells in 4.00 h. What is the power output, given the calculator's voltage output is 3.00 V? (See Figure 2.)

Figure 2. The strip of solar cells just above the keys of this calculator convert light to electricity to supply its energy needs. (credit: Evan-Amos, Wikimedia Commons)

4. How many watts does a flashlight that has $6.00×10^2$ pass through it in 0.500 h use if its voltage is 3.00 V?

5. Find the power dissipated in each of these extension cords: (a) an extension cord having a 0.0600 Ω resistance and through which 5.00 A is flowing; (b) a cheaper cord utilizing thinner wire and with a resistance of 0.300 Ω.

6. Verify that the units of a volt-ampere are watts, as implied by the equation P = IV.

7. Show that the units $1\text{V}^2/\Omega = 1\text{W}$ as implied by the equation $P = V^2/R$.

8. Show that the units $1\ \text{A}^2 \cdot \Omega = 1\text{W}$, as implied by the equation $P = I^2R$.

9. Verify the energy unit equivalence that 1 kW · h = 3.60 × 10^6 J.

10. Electrons in an X-ray tube are accelerated through 1.00 × 10^2 kV and directed toward a target to produce X-rays. Calculate the power of the electron beam in this tube if it has a current of 15.0 mA.

11. An electric water heater consumes 5.00 kW for 2.00 h per day. What is the cost of running it for one year if electricity costs 12.0 cents/kW · h? See Figure 3.

Figure 3. On-demand electric hot water heater. Heat is supplied to water only when needed. (credit: aviddavid, Flickr)

12. With a 1200-W toaster, how much electrical energy is needed to make a slice of toast (cooking time = 1 minute)? At 9.0 cents/kW · h, how much does this cost?

13. What would be the maximum cost of a CFL such that the total cost (investment plus operating) would be the same for both CFL and incandescent 60-W bulbs? Assume the cost of the incandescent bulb is 25 cents and that

electricity costs 10 cents/kWh. Calculate the cost for 1000 hours, as in the cost effectiveness of CFL example.

14. Some makes of older cars have 6.00-V electrical systems. (a) What is the hot resistance of a 30.0-W headlight in such a car? (b) What current flows through it?

15. Alkaline batteries have the advantage of putting out constant voltage until very nearly the end of their life. How long will an alkaline battery rated at $1.00\ A \cdot h$ and 1.58 V keep a 1.00-W flashlight bulb burning?

16. A cauterizer, used to stop bleeding in surgery, puts out 2.00 mA at 15.0 kV. (a) What is its power output? (b) What is the resistance of the path?

17. The average television is said to be on 6 hours per day. Estimate the yearly cost of electricity to operate 100 million TVs, assuming their power consumption averages 150 W and the cost of electricity averages $12.0\ cents/kW \cdot h$.

18. An old lightbulb draws only 50.0 W, rather than its original 60.0 W, due to evaporative thinning of its filament. By what factor is its diameter reduced, assuming uniform thinning along its length? Neglect any effects caused by temperature differences.

19. 00-gauge copper wire has a diameter of 9.266 mm. Calculate the power loss in a kilometer of such wire when it carries $1.00 \times 10^2 A$.

20. **Integrated Concepts**

Cold vaporizers pass a current through water, evaporating it with only a small increase in temperature. One such home device is rated at 3.50 A and utilizes 120 V AC with 95.0% efficiency. (a) What is the vaporization rate in grams per minute? (b) How much water must you put into the vaporizer for 8.00 h of overnight operation? (See Figure 4.)

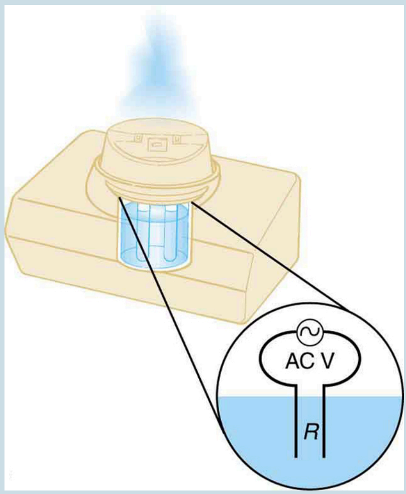


Figure 4. This cold vaporizer passes current directly through water, vaporizing it directly with relatively little temperature increase.

21. **Integrated Concepts** (a) What energy is dissipated by a lightning bolt having a 20,000-A current, a voltage of 1.00×10^2 MV and a length of 1.00 ms? (b) What mass of tree sap could be raised from 18ºC to its boiling point and then evaporated by this energy, assuming sap has the same thermal characteristics as water?

22. **Integrated Concepts** What current must be produced by a 12.0-V battery-operated bottle warmer in order to heat 75.0 g of glass, 250 g of baby formula, and 3.00×10^2 of aluminum from 20° C to 90° in 5.00 min?

23. **Integrated Concepts** How much time is needed for a surgical cauterizer to raise the temperature of 1.00 g of tissue from 37° to 100and then boil away 0.500 g of water, if it puts out 2.00 mA at 15.0 kV? Ignore heat transfer to the surroundings.

24. **Integrated Concepts** Hydroelectric generators (see Figure 5) at Hoover Dam produce a maximum current of 8.00×10^3 A at 250 kV. (a) What is the power output? (b) The water that powers the generators enters and leaves the system at low speed (thus its kinetic energy does not change) but loses 160 m in altitude. How many cubic meters per second are needed, assuming 85.0% efficiency?

Figure 5. Hydroelectric generators at the Hoover dam. (credit: Jon Sullivan)

25. **Integrated Concepts** (a) Assuming 95.0% efficiency for the conversion of electrical power by the motor, what current must the 12.0-V batteries of a 750-kg electric car be able to supply: (a) To accelerate from rest to 25.0 m/s in 1.00 min? (b) To climb a 2.00×10^2-m-high-hill in 2.00 min at a constant 25.0-m/s speed while exerting 5.00×10^2 N of force to overcome air resistance and friction? (c) To travel at a constant 25.0-m/s speed, exerting a 5.00×10^2 N force to overcome air resistance and friction? See Figure 6.

Figure 6. This REVAi, an electric car, gets recharged on a street in London. (credit: Frank Hebbert)

26. **Integrated Concepts** A light-rail commuter train draws 630 A of 650-V DC electricity when accelerating. (a) What is its power consumption rate in kilowatts? (b) How long does it take to reach 20.0 m/s starting from rest if its loaded mass is 5.30×10^4 kg, assuming 95.0% efficiency and constant power? (c) Find its average acceleration. (d) Discuss how the acceleration you found for the light-rail train compares to what might be typical for an automobile.

27. **Integrated Concepts** (a) An aluminum power transmission line has a resistance of 0.0580 Ω/km. What is its mass per kilometer? (b) What is the mass per kilometer of a copper line having the same resistance? A lower resistance would shorten the heating time. Discuss the practical limits to speeding the heating by lowering the resistance.

28. **Integrated Concepts** (a) An immersion heater utilizing 120 V can raise the temperature of a 1.00×10^{2}-g aluminum cup containing 350 g of water from 20° C to 95° C in 2.00 min. Find its resistance, assuming it is constant during the process. (b) A lower resistance would shorten the heating time. Discuss the practical limits to speeding the heating by lowering the resistance.

29. **Integrated Concepts** (a) What is the cost of heating a hot tub containing 1500 kg of water from 10° C to 40° C, assuming 75.0% efficiency to account for heat transfer to the surroundings? The cost of electricity is 9 cents/kW · h. (b) What current was used by the 220-V AC electric heater, if this took 4.00 h?

30**. Unreasonable Results** (a) What current is needed to transmit 1.00×10^{2} MW of power at 480 V? (b) What power is dissipated by the transmission lines if they have a 1.00 − Ω resistance? (c) What is unreasonable about this result? (d) Which assumptions are unreasonable, or which premises are inconsistent?

31. **Unreasonable Results** (a) What current is needed to transmit 1.00×10^{2} MW of power at 10.0 kV? (b) Find the resistance of 1.00 km of wire that would cause a 0.0100%

power loss. (c) What is the diameter of a 1.00-km-long copper wire having this resistance? (d) What is unreasonable about these results? (e) Which assumptions are unreasonable, or which premises are inconsistent?

32. Construct Your Own Problem Consider an electric immersion heater used to heat a cup of water to make tea. Construct a problem in which you calculate the needed resistance of the heater so that it increases the temperature of the water and cup in a reasonable amount of time. Also calculate the cost of the electrical energy used in your process. Among the things to be considered are the voltage used, the masses and heat capacities involved, heat losses, and the time over which the heating takes place. Your instructor may wish for you to consider a thermal safety switch (perhaps bimetallic) that will halt the process before damaging temperatures are reached in the immersion unit.

Glossary

electric power:
: the rate at which electrical energy is supplied by a source or dissipated by a device; it is the product of current times voltage

Selected Solutions to Problems & Exercises

1. 2.00×10^{12} W

5. (a) 1.50 W (b) 7.50 W

7.

$$\frac{V^2}{\Omega} = \frac{V^2}{\mathrm{V/A}} = \mathrm{AV} = \left(\frac{C}{s}\right)\left(\frac{J}{C}\right) = \frac{J}{s} = 1\mathrm{W}$$

9.

$$1\mathrm{kW}\cdot\mathrm{h} = \left(\frac{1\times10^{3}\mathrm{J}}{1\ \mathrm{s}}\right)(1h)\left(\frac{3600\ \mathrm{s}}{1\ \mathrm{h}}\right) = 3.60\times10^{6}\mathrm{J}$$

11. $438/y

13. $6.25

15. 1.58 h

17. $3.94 billion/year

19. 25.5 W

21.(a) 2.00×10^{9} J (b) 769 kg

23. 45.0 s

25. (a) 343 A (b) 2.17×10^{3} A (c) 1.10×10^{3} A

27. (a) 1.23×10^{3}kg (b) 2.64×10^{3}kg

29. (a) 2.08×10^{5} A
(b) 4.33×10^{4} MW
(c) The transmission lines dissipate more power than they are supposed to transmit.
(d) A voltage of 480 V is unreasonably low for a transmission voltage. Long-distance transmission lines are kept at much higher voltages (often hundreds of kilovolts) to reduce power losses.

26. Alternating Current versus Direct Current

Learning Objectives

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

- Explain the differences and similarities between AC and DC current.
- Calculate rms voltage, current, and average power.
- Explain why AC current is used for power transmission.

Alternating Current

Most of the examples dealt with so far, and particularly those utilizing batteries, have constant voltage sources. Once the current is established, it is thus also a constant. *Direct current* (DC) is the flow of electric charge in only one direction. It is the steady state of a constant-voltage circuit. Most well-known applications, however, use a time-varying voltage source. *Alternating current* (AC) is the flow of electric charge that periodically reverses direction. If the source varies periodically, particularly sinusoidally, the circuit is known as an alternating current circuit. Examples include the commercial and residential power that serves so many of our needs. Figure 1 shows graphs of voltage and current versus time for typical

DC and AC power. The AC voltages and frequencies commonly used in homes and businesses vary around the world.

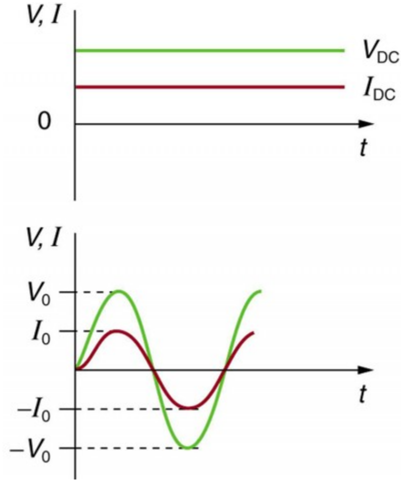


Figure 1. (a) DC voltage and current are constant in time, once the current is established. (b) A graph of voltage and current versus time for 60-Hz AC power. The voltage and current are sinusoidal and are in phase for a simple resistance circuit. The frequencies and peak voltages of AC sources differ greatly.

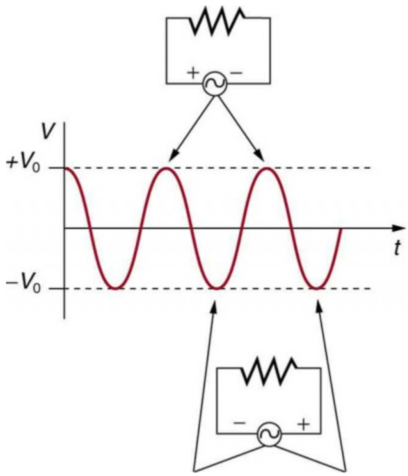


Figure 2. The potential difference V *between the terminals of an AC voltage source fluctuates as shown. The mathematical expression for* V *is given by [latex]V={V}_{0}\sin\text{ 2}\pi {ft}\\[/latex].*

Figure 2 shows a schematic of a simple circuit with an AC voltage source. The voltage between the terminals fluctuates as shown, with the AC *voltage* given by

$$V = V_0 \sin 2\pi ft,$$

where V is the voltage at time *t*, V_0 is the peak voltage, and *f* is the frequency in hertz. For this simple resistance circuit, I=V/R, and so the AC *current* is

$$I = I_0 \sin 2\pi ft,$$

where *I* is the current at time *t*, and $I_0 = V_0/R$ is the peak current. For this example, the voltage and current are said to be in phase, as seen in Figure 1(b).

Current in the resistor alternates back and forth just like the driving voltage, since I = V/R. If the resistor is a fluorescent light bulb, for example, it brightens and dims 120 times per second as the current repeatedly goes through zero. A 120-Hz flicker is too rapid for your eyes to detect, but if you wave your hand back and forth between your face and a fluorescent light, you will see a stroboscopic effect evidencing AC. The fact that the light output fluctuates means that the power is fluctuating. The power supplied is $P = IV$. Using the expressions for I and V above, we see that the time dependence of power is $P = I_0 V_0 \sin^2 2\pi ft$, as shown in Figure 3.

Making Connections: Take-Home Experiment—AC/DC Lights

Wave your hand back and forth between your face and a fluorescent light bulb. Do you observe the same thing with the headlights on your car? Explain what you observe. *Warning: Do not look directly at very bright light.*

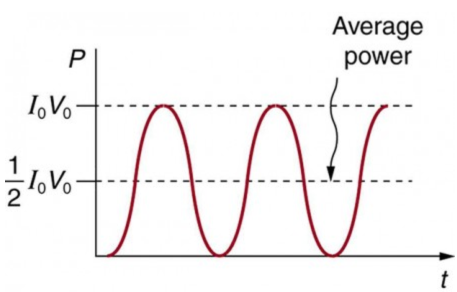


Figure 3. AC power as a function of time. Since the voltage and current are in phase here, their product is non-negative and fluctuates between zero and I_0V_0. *Average power is* $(1/2)I_0V_0$.

We are most often concerned with average power rather than its fluctuations—that 60-W light bulb in your desk lamp has an average power consumption of 60 W, for example. As illustrated in Figure 3, the average power P_{ave} is

$$P_{\text{ave}} = \frac{1}{2} I_0 V_0.$$

This is evident from the graph, since the areas above and below the $(1/2)I_0V_0$ line are equal, but it can also be proven using trigonometric identities. Similarly, we define an average or *rms current* I_{rms} and average or *rms voltage* V_{rms} to be, respectively,

$$I_{\text{rms}} = \frac{I_0}{\sqrt{2}}$$

and

$$V_{\text{rms}} = \frac{V_0}{\sqrt{2}}.$$

where rms stands for root mean square, a particular kind of average. In general, to obtain a root mean square, the particular quantity is squared, its mean (or average) is found, and the square root is taken. This is useful for AC, since the average value is zero. Now,

$$P_{ave} = I_{rms}V_{rms},$$

which gives

$$P_{\text{ave}} = \frac{I_0}{\sqrt{2}} \cdot \frac{V_0}{\sqrt{2}} = \frac{1}{2} I_0 V_0,$$

as stated above. It is standard practice to quote I_{rms}, V_{rms}, and P_{ave} rather than the peak values. For example, most household electricity is 120 V AC, which means that V_{rms} is 120 V. The common 10-A circuit breaker will interrupt a sustained I_{rms} greater than 10 A. Your 1.0-kW microwave oven consumes P_{ave} = 1.0 kW, and so on. You can think of these rms and average values as the equivalent DC values for a simple resistive circuit. To summarize, when dealing with AC, Ohm's law and the equations for power are completely analogous to those for DC, but rms and average values are used for AC. Thus, for AC, Ohm's law is written

$$I_{\text{rms}} = \frac{V_{\text{rms}}}{R}.$$

The various expressions for AC power P_{ave} are

$$P_{\text{ave}} = I_{\text{rms}} V_{\text{rms}},$$

$$P_{\text{ave}} = \frac{V_{\text{rms}}^{\ 2}}{R},$$

and

$$P_{\text{ave}} = I_{\text{rms}}^{\ 2} R.$$

Example 1. Peak Voltage and Power for AC

(a) What is the value of the peak voltage for 120-V AC power? (b) What is the peak power consumption rate of a 60.0-W AC light bulb?

Strategy

We are told that V_{rms} is 120 V and P_{ave} is 60.0 W. We can use $V_{\text{rms}} = \frac{V_0}{\sqrt{2}}$ to find the peak voltage, and we can manipulate the definition of power to find the peak power from the given average power.

Solution for (a)

Solving the equation $V_{\text{rms}} = \frac{V_0}{\sqrt{2}}$ for the peak voltage V_0 and substituting the known value for V_{rms} gives

$$V_0 = \sqrt{2}V_{\text{rms}} = 1.414(120\text{ V}) = 170\text{V}$$

Discussion for (a)

This means that the AC voltage swings from 170 V to −170 V and back 60 times every second. An equivalent DC voltage is a constant 120 V.

Solution for (b)

Peak power is peak current times peak voltage. Thus,

$$P_0 = I_0V_0 = 2\left(\frac{1}{2}I_0V_0\right) = 2P_{\text{ave}}.$$

We know the average power is 60.0 W, and so

$$P_0 = 2(60.0\text{ W}) = 120\text{ W}.$$

Discussion

So the power swings from zero to 120 W one hundred twenty times per second (twice each cycle), and the power averages 60 W.

Why Use AC for Power Distribution?

Most large power-distribution systems are AC. Moreover, the power is transmitted at much higher voltages than the 120-V AC (240 V in most parts of the world) we use in homes and on the job. Economies of scale make it cheaper to build a few very large electric power-generation plants than to build numerous small ones. This necessitates sending power long distances, and it is obviously important that energy losses en route be minimized. High voltages can be transmitted with much smaller power losses than low voltages, as we shall see. (See Figure 4.) For safety reasons, the voltage at the user is reduced to familiar values. The crucial factor is that it is much easier to increase and decrease AC voltages than DC, so AC is used in most large power distribution systems.

Figure 4. Power is distributed over large distances at high voltage to reduce power loss in the transmission lines. The voltages generated at the power plant are stepped up by passive devices called transformers (see [Transformers](#)) to 330,000 volts (or more in some places worldwide). At the point of use, the transformers reduce the voltage transmitted for safe residential and commercial use. (Credit: GeorgHH, Wikimedia Commons)

Example 2. Power Losses Are Less for High-Voltage Transmission

(a) What current is needed to transmit 100 MW of power at 200 kV? (b) What is the power dissipated by the transmission lines if they have a resistance of 1.00 Ω(c) What percentage of the power is lost in the transmission lines?

Strategy

We are given P_{ave} = 100 MW, V_{rms} = 200 kV, and the resistance of the lines is R = 1.00 Ω. Using these givens, we can find the current flowing (from P = IV) and then the power dissipated in the lines ($P = I^2R$), and we take the ratio to the total power transmitted.

Solution

To find the current, we rearrange the relationship $P_{ave} = I_{rms}V_{rms}$ and substitute known values. This gives

$$I_{\text{rms}} = \frac{P_{\text{ave}}}{V_{\text{rms}}} = \frac{100 \times 10^6\ \text{W}}{200 \times 10^3\ \text{V}} = 500\ \text{A}.$$

Solution

Knowing the current and given the resistance of the lines, the power dissipated in them is found from $P_{\text{ave}} = {I_{\text{rms}}}^2 R$ Substituting the known values gives

$$P_{\text{ave}} = {I_{\text{rms}}}^2 R = (500\text{ A})^2 (1.00\ \Omega) = 250\text{ kW}$$

Solution

The percent loss is the ratio of this lost power to the total or input power, multiplied by 100:

\text{% loss=}\frac{\text{250 kW}}{\text{100 MW}}\times \text{100}=0\text{.}\text{250%}\\

.

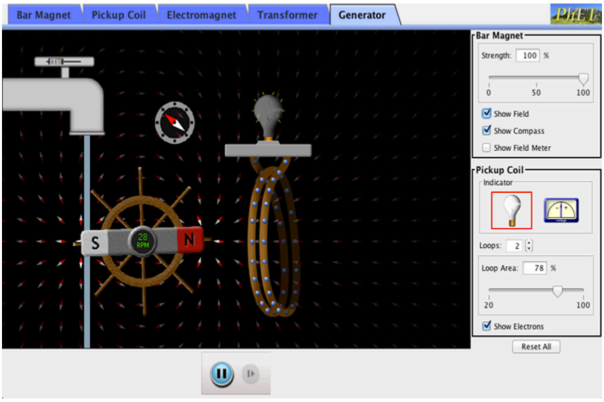
Discussion

One-fourth of a percent is an acceptable loss. Note that if 100 MW of power had been transmitted at 25 kV, then a current of 4000 A would have been needed. This would result in a power loss in the lines of 16.0 MW, or 16.0% rather than 0.250%. The lower the voltage, the more current is needed, and the greater the power loss in the fixed-resistance transmission lines. Of course, lower-resistance lines can be built, but this requires larger and more expensive wires. If superconducting lines could be economically produced, there would be no loss in the transmission lines at all. But, as we shall see in a later chapter, there is a limit to current in superconductors, too. In short, high voltages are more economical for transmitting power, and AC voltage is much easier to raise and lower, so that AC is used in most large-scale power distribution systems.

It is widely recognized that high voltages pose greater hazards than low voltages. But, in fact, some high voltages, such as those associated with common static electricity, can be harmless. So it is not voltage alone that determines a hazard. It is not so widely recognized that AC shocks are often more harmful than similar DC shocks. Thomas Edison thought that AC shocks were more harmful and set up a DC power-distribution system in New York City in the late 1800s. There were bitter fights, in particular between Edison and George Westinghouse and Nikola Tesla, who were advocating the use of AC in early power-distribution systems. AC has prevailed largely due to transformers and lower power losses with high-voltage transmission.

PhET Explorations: Generator

Generate electricity with a bar magnet! Discover the physics behind the phenomena by exploring magnets and how you can use them to make a bulb light.



Click to download the simulation. Run using Java.

Section Summary

- Direct current (DC) is the flow of electric current in only one direction. It refers to systems where the source voltage is constant.
- The voltage source of an alternating current (AC) system puts out $V = V_0 \sin 2\pi ft$, where V is the voltage at time *t*, V_0 is the peak voltage, and *f* is the frequency in hertz.
- In a simple circuit, I = V/R and AC current is $I = I_0 \sin 2\pi ft$, where I is the current at time *t*, and $I_0 = V_0/\mathrm{R}$ is the peak current.
- The average AC power is $P_{\text{ave}} = \frac{1}{2} I_0 V_0$.
- Average (rms) current I_{rms} and average (rms) voltage V_{rms} are

$I_{\text{rms}} = \dfrac{I_0}{\sqrt{2}}$ and $V_{\text{rms}} = \dfrac{V_0}{\sqrt{2}}$, where rms stands for root mean square.

- Thus, $P_{\text{ave}} = I_{\text{rms}} V_{\text{rms}}$.
- Ohm's law for AC is $I_{\text{rms}} = \dfrac{V_{\text{rms}}}{R}$.
- Expressions for the average power of an AC circuit are $P_{\text{ave}} = I_{\text{rms}} V_{\text{rms}}$, $P_{\text{ave}} = \dfrac{V_{\text{rms}}^2}{R}$, and $P_{\text{ave}} = {I_{\text{rms}}}^2 R$, analogous to the expressions for DC circuits.

Conceptual Questions

1. Give an example of a use of AC power other than in the household. Similarly, give an example of a use of DC power other than that supplied by batteries.

2. Why do voltage, current, and power go through zero 120 times per second for 60-Hz AC electricity?

3. You are riding in a train, gazing into the distance through its window. As close objects streak by, you notice that the nearby fluorescent lights make *dashed* streaks. Explain.

Problem & Exercises

1. (a) What is the hot resistance of a 25-W light bulb that runs on 120-V AC? (b) If the bulb's operating temperature is 2700°C, what is its resistance at 2600°C?

2. Certain heavy industrial equipment uses AC power that has a peak voltage of 679 V. What is the rms voltage?

3. A certain circuit breaker trips when the rms current is 15.0 A. What is the corresponding peak current?

4. Military aircraft use 400-Hz AC power, because it is possible to design lighter-weight equipment at this higher frequency. What is the time for one complete cycle of this power?

5. A North American tourist takes his 25.0-W, 120-V AC razor to Europe, finds a special adapter, and plugs it into 240 V AC. Assuming constant resistance, what power does the razor consume as it is ruined?

6. In this problem, you will verify statements made at the end of the power losses for Example 2 above. (a) What current is needed to transmit 100 MW of power at a voltage of 25.0 kV? (b) Find the power loss in a 1.00-Ω transmission line. (c) What percent loss does this represent?

7. A small office-building air conditioner operates on 408-V AC and consumes 50.0 kW. (a) What is its effective resistance? (b) What is the cost of running the air conditioner during a hot summer month when it is on 8.00

h per day for 30 days and electricity costs 9.00 cents/kW · h?

8. What is the peak power consumption of a 120-V AC microwave oven that draws 10.0 A?

9. What is the peak current through a 500-W room heater that operates on 120-V AC power?

10. Two different electrical devices have the same power consumption, but one is meant to be operated on 120-V AC and the other on 240-V AC. (a) What is the ratio of their resistances? (b) What is the ratio of their currents? (c) Assuming its resistance is unaffected, by what factor will the power increase if a 120-V AC device is connected to 240-V AC?

11. Nichrome wire is used in some radiative heaters. (a) Find the resistance needed if the average power output is to be 1.00 kW utilizing 120-V AC. (b) What length of Nichrome wire, having a cross-sectional area of 500 mm^2, is needed if the operating temperature is 500°C (c) What power will it draw when first switched on?

12. Find the time after $t = 0$ when the instantaneous voltage of 60-Hz AC first reaches the following values: (a) $V_0/2$ (b)V_0 (c) 0.

13. (a) At what two times in the first period following $t = 0$ does the instantaneous voltage in 60-Hz AC equal V_{rms}? (b) $-V_{rms}$?

Glossary

direct current:
(DC) the flow of electric charge in only one direction

alternating current:
(AC) the flow of electric charge that periodically reverses direction

AC voltage:
voltage that fluctuates sinusoidally with time, expressed as $V = V_0 \sin 2\pi f t$, where V is the voltage at time *t*, V_0 is the peak voltage, and *f* is the frequency in hertz

AC current:
current that fluctuates sinusoidally with time, expressed as $I = I_0 \sin 2\pi f t$, where *I* is the current at time *t*, I_0 is the peak current, and *f* is the frequency in hertz

rms current:
the root mean square of the current, $I_{\text{rms}} = I_0/\sqrt{2}$, where I_0 is the peak current, in an AC system

rms voltage:
the root mean square of the voltage, $V_{\text{rms}} = V_0/\sqrt{2}$, where V_0 is the peak voltage, in an AC system

Selected Solutions to Problems & Exercises

2. 480 V

4. 2.50 ms

6. (a) 4.00 kA (b) 16.0 MW (c) 16.0%

8. 2.40 kW

10. (a) 4.0 (b) 0.50 (c) 4.0

12. (a) 1.39 ms (b) 4.17 ms (c) 8.33 ms

27. Electric Hazards and the Human Body

Learning Objectives

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

- Define thermal hazard, shock hazard, and short circuit.
- Explain what effects various levels of current have on the human body.

There are two known hazards of electricity—thermal and shock. A *thermal hazard* is one where excessive electric power causes undesired thermal effects, such as starting a fire in the wall of a house. A *shock hazard* occurs when electric current passes through a person. Shocks range in severity from painful, but otherwise harmless, to heart-stopping lethality. This section considers these hazards and the various factors affecting them in a quantitative manner. Electrical Safety: Systems and Devices will consider systems and devices for preventing electrical hazards.

Thermal Hazards

Electric power causes undesired heating effects whenever electric energy is converted to thermal energy at a rate faster than it can be safely dissipated. A classic example of this is the *short circuit*, a low-resistance path between terminals of a voltage source. An example of a short circuit is shown in Figure 1. Insulation on wires leading to an appliance has worn through, allowing the two wires to come into contact. Such an undesired contact with a high voltage is called a *short*. Since the resistance of the short, r, is very small, the power dissipated in the short, $P = V^2/r$, is very large. For example, if V is 120 V and r is 0.100 Ω, then the power is 144 kW, *much* greater than that used by a typical household appliance. Thermal energy delivered at this rate will very quickly raise the temperature of surrounding materials, melting or perhaps igniting them.

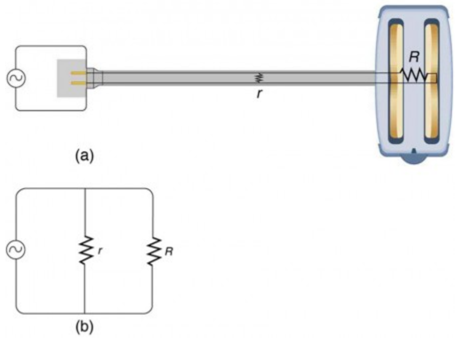


Figure 1. A short circuit is an undesired low-resistance path across a voltage source. (a) Worn insulation on the wires of a toaster allow them to come into contact with a low resistance r. Since $P = V^2/r$*, thermal power is created so rapidly that the cord melts or burns. (b) A schematic of the short circuit.*

One particularly insidious aspect of a short circuit is that its resistance may actually be decreased due to the increase in temperature. This can happen if the short creates ionization. These charged atoms and molecules are free to move and, thus, lower the resistance r. Since $P = V^2/r$, the power dissipated in the short rises, possibly causing more ionization, more power, and so on. High voltages, such as the 480-V AC used in some industrial applications, lend themselves to this hazard, because higher voltages create higher initial power production in a short.

Another serious, but less dramatic, thermal hazard occurs when wires supplying power to a user are overloaded with too great a current. As discussed in the previous section, the power dissipated in the supply wires is $P = I^2R_W$, where R_W is the resistance of the wires and I the current flowing through them. If either I or R_W is too large, the wires overheat. For example, a worn appliance cord (with some of its braided wires broken) may have $R_W = 2.00\ \Omega$ rather than the $0.100\ \Omega$ it should be. If 10.0 A of current passes through the cord, then $P = I^2R_W = 200$ W is dissipated in the cord—much more than is safe. Similarly, if a wire with a 0.100-Ω resistance is meant to carry a few amps, but is instead carrying 100 A, it will severely overheat. The power dissipated in the wire will in that case be $P = 1000$ W. Fuses and circuit breakers are used to limit excessive currents. (See Figure 1 and Figure 2.) Each device opens the circuit automatically when a sustained current exceeds safe limits.

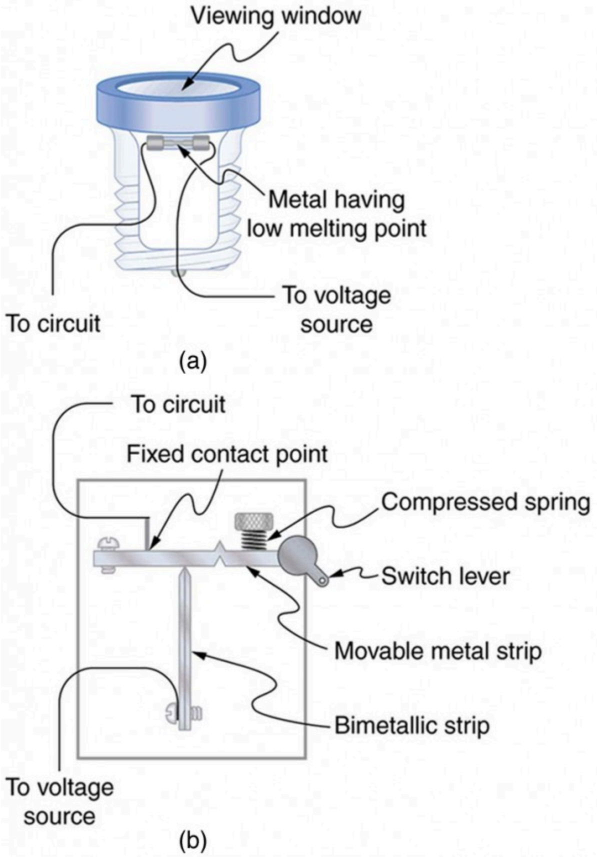


Figure 1. (a) A fuse has a metal strip with a low melting point that, when overheated by an excessive current, permanently breaks the connection of a circuit to a voltage source. (b) A circuit breaker is an automatic but restorable electric switch. The one shown here has a bimetallic strip that bends to the right and into the notch if overheated. The spring then forces the metal strip downward, breaking the electrical connection at the points.

Fuses and circuit breakers for typical household voltages and currents are relatively simple to produce, but those for large voltages and currents experience special problems. For example, when a circuit breaker tries to interrupt the flow of high-voltage electricity, a spark can jump across its points that ionizes the air in the gap and allows the current to continue flowing. Large circuit breakers found in power-distribution systems employ insulating gas and even use jets of gas to blow out such sparks. Here AC is safer than DC, since AC current goes through zero 120 times per second, giving a quick opportunity to extinguish these arcs.

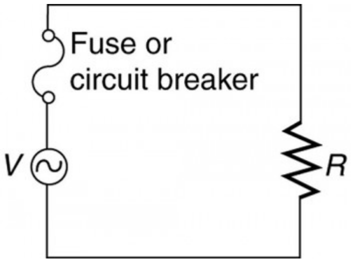


Figure 2. Schematic of a circuit with a fuse or circuit breaker in it. Fuses and circuit breakers act like automatic switches that open when sustained current exceeds desired limits.

Shock Hazards

Electrical currents through people produce tremendously varied effects. An electrical current can be used to block back pain. The possibility of using electrical current to stimulate muscle action in paralyzed limbs, perhaps allowing paraplegics to walk, is under study. TV dramatizations in which electrical shocks are used to bring a heart attack victim out of ventricular fibrillation (a massively irregular, often fatal, beating of the heart) are more than common. Yet most electrical shock fatalities occur because a current put the heart into fibrillation. A pacemaker uses electrical shocks to stimulate the heart to beat properly. Some fatal shocks do not produce burns, but warts can be safely burned off with electric current (though freezing using liquid nitrogen is now more common). Of course, there are consistent explanations for these disparate effects. The major factors upon which the effects of electrical shock depend are

1. The amount of current I
2. The path taken by the current
3. The duration of the shock
4. The frequency f of the current (f=0 for DC)

Table 1 gives the effects of electrical shocks as a function of current for a typical accidental shock. The effects are for a shock that passes through the trunk of the body, has a duration of 1 s, and is caused by 60-Hz power.

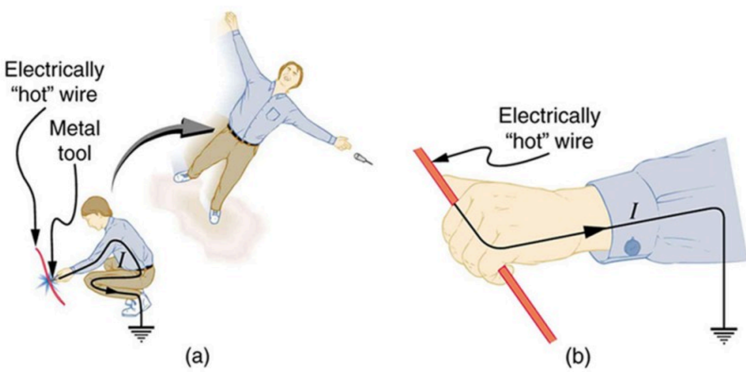


Figure 3. An electric current can cause muscular contractions with varying effects. (a) The victim is "thrown" backward by involuntary muscle contractions that extend the legs and torso. (b) The victim can't let go of the wire that is stimulating all the muscles in the hand. Those that close the fingers are stronger than those that open them.

Table 1. Effects of Electrical Shock as a Function of Current[1]

Current (mA)	Effect
1	Threshold of sensation
5	Maximum harmless current
10–20	Onset of sustained muscular contraction; cannot let go for duration of shock; contraction of chest muscles may stop breathing during shock
50	Onset of pain
100–300+	Ventricular fibrillation possible; often fatal
300	Onset of burns depending on concentration of current
6000 (6 A)	Onset of sustained ventricular contraction and respiratory paralysis; both cease when shock ends; heartbeat may return to normal; used to defibrillate the heart

Our bodies are relatively good conductors due to the water in our bodies. Given that larger currents will flow through sections with lower resistance (to be further discussed in the next chapter), electric currents preferentially flow through paths in the human body that have a minimum resistance in a direct path to earth. The earth is a natural electron sink. Wearing insulating shoes, a requirement in many professions, prohibits a pathway for electrons by providing a large resistance in that path. Whenever working with high-power tools (drills), or in risky situations, ensure that you do not provide a pathway for current flow (especially through the heart).

Very small currents pass harmlessly and unfelt through the body. This happens to you regularly without your knowledge. The threshold of sensation is only 1 mA and, although unpleasant, shocks

1. For an average male shocked through trunk of body for 1 s by 60-Hz AC. Values for females are 60–80% of those listed.

are apparently harmless for currents less than 5 mA. A great number of safety rules take the 5-mA value for the maximum allowed shock. At 10 to 20 mA and above, the current can stimulate sustained muscular contractions much as regular nerve impulses do. People sometimes say they were knocked across the room by a shock, but what really happened was that certain muscles contracted, propelling them in a manner not of their own choosing. (See Figure 3(a).) More frightening, and potentially more dangerous, is the “can't let go” effect illustrated in Figure 3(b). The muscles that close the fingers are stronger than those that open them, so the hand closes involuntarily on the wire shocking it. This can prolong the shock indefinitely. It can also be a danger to a person trying to rescue the victim, because the rescuer's hand may close about the victim's wrist. Usually the best way to help the victim is to give the fist a hard knock/blow/jar with an insulator or to throw an insulator at the fist. Modern electric fences, used in animal enclosures, are now pulsed on and off to allow people who touch them to get free, rendering them less lethal than in the past.

Greater currents may affect the heart. Its electrical patterns can be disrupted, so that it beats irregularly and ineffectively in a condition called “ventricular fibrillation.” This condition often lingers after the shock and is fatal due to a lack of blood circulation. The threshold for ventricular fibrillation is between 100 and 300 mA. At about 300 mA and above, the shock can cause burns, depending on the concentration of current—the more concentrated, the greater the likelihood of burns.

Very large currents cause the heart and diaphragm to contract for the duration of the shock. Both the heart and breathing stop. Interestingly, both often return to normal following the shock. The electrical patterns on the heart are completely erased in a manner that the heart can start afresh with normal beating, as opposed to the permanent disruption caused by smaller currents that can put the heart into ventricular fibrillation. The latter is something like scribbling on a blackboard, whereas the former completely erases it. TV dramatizations of electric shock used to bring a heart attack victim out of ventricular fibrillation also show large paddles. These are used to spread out current passed through the victim to reduce the likelihood of burns.

Current is the major factor determining shock severity (given that other conditions such as path, duration, and frequency are fixed, such as in the table and preceding discussion). A larger voltage is more hazardous, but since $I = V/R$, the severity of the shock depends on the combination of voltage and resistance. For example, a person with dry skin has a resistance of about 200 kΩ. If he comes into contact with 120-V AC, a current $I = (120\ \text{V})/(200\ \text{k}\ \Omega) = 0.6$ mA passes harmlessly through him. The same person soaking wet may have a resistance of 10.0 kΩ and the same 120 V will produce a current of 12 mA—above the "can't let go" threshold and potentially dangerous.

Most of the body's resistance is in its dry skin. When wet, salts go into ion form, lowering the resistance significantly. The interior of the body has a much lower resistance than dry skin because of all the ionic solutions and fluids it contains. If skin resistance is bypassed, such as by an intravenous infusion, a catheter, or exposed pacemaker leads, a person is rendered *microshock sensitive*. In this condition, currents about 1/1000 those listed in Table 1 produce similar effects. During open-heart surgery, currents as small as 20 μA can be used to still the heart. Stringent electrical safety requirements in hospitals, particularly in surgery and intensive care, are related to the doubly disadvantaged microshock-sensitive patient. The break in the skin has reduced his resistance, and so the same voltage causes a greater current, and a much smaller current has a greater effect.

Factors other than current that affect the severity of a shock are its path, duration, and AC frequency. Path has obvious consequences. For example, the heart is unaffected by an electric shock through the brain, such as may be used to treat manic depression. And it is a general truth that the longer the duration of a shock, the greater its effects. Figure 4 presents a graph that illustrates the effects of frequency on a shock. The curves show the minimum current for two different effects, as a function of frequency. The lower the current needed, the more sensitive the body is at that frequency. Ironically, the body is most sensitive to frequencies near the 50- or 60-Hz frequencies in common use. The body is slightly less sensitive for DC (f = 0), mildly confirming Edison's claims that AC presents a greater hazard. At higher and higher frequencies, the body becomes progressively less sensitive to any effects that involve nerves. This is related to the maximum rates at which nerves can fire or be stimulated. At very high frequencies, electrical current travels only on the surface of a person. Thus a wart can be burned off with very high frequency current without causing the heart to stop. (Do not try this at home with 60-Hz AC!) Some of the spectacular demonstrations of electricity, in which high-voltage arcs are passed through the air and over people's bodies, employ high frequencies and low currents. (See Figure 5.) Electrical safety devices and techniques are discussed in detail in Electrical Safety: Systems and Devices.

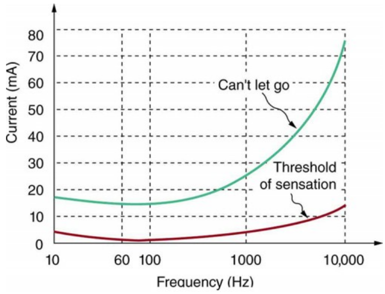


Figure 4. Graph of average values for the threshold of sensation and the "can't let go" current as a function of frequency. The lower the value, the more sensitive the body is at that frequency.

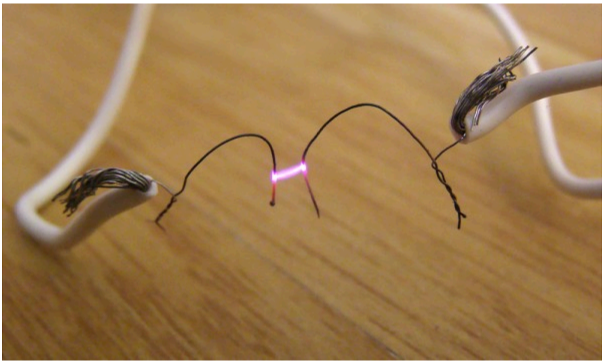


Figure 5 Is this electric arc dangerous? The answer depends on the AC frequency and the power involved. (credit: Khimich Alex, Wikimedia Commons)

Section Summary

- The two types of electric hazards are thermal (excessive power) and shock (current through a person).
- Shock severity is determined by current, path, duration, and AC frequency.
- Table 1 lists shock hazards as a function of current.
- Figure 5 graphs the threshold current for two hazards as a function of frequency.

Conceptual Questions

1. Using an ohmmeter, a student measures the resistance between various points on his body. He finds that the resistance between two points on the

same finger is about the same as the resistance between two points on opposite hands—both are several hundred thousand ohms. Furthermore, the resistance decreases when more skin is brought into contact with the probes of the ohmmeter. Finally, there is a dramatic drop in resistance (to a few thousand ohms) when the skin is wet. Explain these observations and their implications regarding skin and internal resistance of the human body.

2. What are the two major hazards of electricity?
3. Why isn't a short circuit a shock hazard?
4. What determines the severity of a shock? Can you say that a certain voltage is hazardous without further information?
5. An electrified needle is used to burn off warts, with the circuit being completed by having the patient sit on a large butt plate. Why is this plate large?
6. Some surgery is performed with high-voltage electricity passing from a metal scalpel through the tissue being cut. Considering the nature of electric fields at the surface of conductors, why would you expect most of the current to flow from the sharp edge of the scalpel? Do you think high- or low-frequency AC is used?
7. Some devices often used in bathrooms, such as hairdryers, often have safety messages saying "Do not use when the bathtub or basin is full of water." Why is this so?
8. We are often advised to not flick electric switches with wet hands, dry your hand first. We are also advised to never throw water on an electric fire. Why is this so?

9. Before working on a power transmission line, linemen will touch the line with the back of the hand as a final check that the voltage is zero. Why the back of the hand?
10. Why is the resistance of wet skin so much smaller than dry, and why do blood and other bodily fluids have low resistances?
11. Could a person on intravenous infusion (an IV) be microshock sensitive?
12. In view of the small currents that cause shock hazards and the larger currents that circuit breakers and fuses interrupt, how do they play a role in preventing shock hazards?

Problems & Exercises

1. (a) How much power is dissipated in a short circuit of 240-V AC through a resistance of 0.250 Ω? (b) What current flows?

2. What voltage is involved in a 1.44-kW short circuit through a 0.100-Ω resistance?

3. Find the current through a person and identify the likely effect on her if she touches a 120-V AC source: (a) if she is standing on a rubber mat and offers a total resistance of 300 kΩ; (b) if she is standing barefoot on wet grass and has a resistance of only 4000 kΩ.

4. While taking a bath, a person touches the metal case of a radio. The path through the person to the drainpipe and ground has a resistance of 4000 Ω. What is the smallest voltage on the case of the radio that could cause ventricular fibrillation?

5. Foolishly trying to fish a burning piece of bread from a toaster with a metal butter knife, a man comes into contact with 120-V AC. He does not even feel it since, luckily, he is wearing rubber-soled shoes. What is the minimum resistance of the path the current follows through the person?

6. (a) During surgery, a current as small as 20.0 μA applied directly to the heart may cause ventricular fibrillation. If the resistance of the exposed heart is 300 Ω, what is the smallest voltage that poses this danger? (b) Does your answer imply that special electrical safety precautions are needed?

7. (a) What is the resistance of a 220-V AC short circuit that generates a peak power of 96.8 kW? (b) What would the average power be if the voltage was 120 V AC?

8. A heart defibrillator passes 10.0 A through a patient's torso for 5.00 ms in an attempt to restore normal beating. (a) How much charge passed? (b) What voltage was applied if 500 J of energy was dissipated? (c) What was the path's resistance? (d) Find the temperature increase caused in the 8.00 kg of affected tissue.

9. **Integrated Concepts** A short circuit in a 120-V appliance cord has a 0.500-Ω resistance. Calculate the

temperature rise of the 2.00 g of surrounding materials, assuming their specific heat capacity is $0.200 \text{ cal/g} \cdot {}^\circ\text{C}$ and that it takes 0.0500 s for a circuit breaker to interrupt the current. Is this likely to be damaging?

10. Temperature increases 860°C. It is very likely to be damaging.

11. **Construct Your Own Problem** Consider a person working in an environment where electric currents might pass through her body. Construct a problem in which you calculate the resistance of insulation needed to protect the person from harm. Among the things to be considered are the voltage to which the person might be exposed, likely body resistance (dry, wet, …), and acceptable currents (safe but sensed, safe and unfelt, …).

Glossary

thermal hazard:
: a hazard in which electric current causes undesired thermal effects

shock hazard:
: when electric current passes through a person

short circuit:
: also known as a "short," a low-resistance path between terminals of a voltage source

microshock sensitive:
: a condition in which a person's skin resistance is bypassed, possibly by a medical procedure, rendering the person

vulnerable to electrical shock at currents about 1/1000 the normally required level

Selected Solutions to Problems & Exercises

1. (a) 230 kW (b) 960 A

3. (a) 0.400 mA, no effect (b) 26.7 mA, muscular contraction for duration of the shock (can't let go)

5. $1.20 \times 10^5\ \Omega$

7. (a) $1.00\ \Omega$ (b) 14.4 kW

28. Nerve Conduction–Electrocardiograms

Learning Objectives

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

- Explain the process by which electric signals are transmitted along a neuron.
- Explain the effects myelin sheaths have on signal propagation.
- Explain what the features of an ECG signal indicate.

Nerve Conduction

Electric currents in the vastly complex system of billions of nerves in our body allow us to sense the world, control parts of our body, and think. These are representative of the three major functions of nerves. First, nerves carry messages from our sensory organs and others to the central nervous system, consisting of the brain and spinal cord. Second, nerves carry messages from the central nervous system to muscles and other organs. Third, nerves transmit and process signals within the central nervous system. The sheer number of nerve cells and the incredibly greater number of connections between them makes this system the subtle wonder that it is. *Nerve conduction* is a general term for electrical signals carried by nerve cells. It is one aspect of *bioelectricity*, or electrical effects in and created by biological systems. Nerve cells, properly called *neurons*, look different from other cells—they have tendrils, some of them many centimeters long, connecting them with other cells. (See Figure 1.) Signals arrive at the cell body across *synapses* or through *dendrites*, stimulating the neuron to generate its own signal, sent along its long *axon* to other nerve or muscle cells. Signals may arrive from many other locations and be transmitted to yet others, conditioning the synapses by use, giving the system its complexity and its ability to learn.

The method by which these electric currents are generated and transmitted is more complex than the simple movement of free charges in a conductor, but it can be understood with principles already discussed in this text. The most important of these are the Coulomb force and diffusion. Figure 2 illustrates how a voltage (potential difference) is created across the cell membrane of a neuron in its resting state. This thin membrane separates electrically neutral fluids having differing concentrations of ions, the most important varieties being Na^+, K^+, and Cl^- (these are sodium, potassium, and chlorine ions with single plus or minus charges as

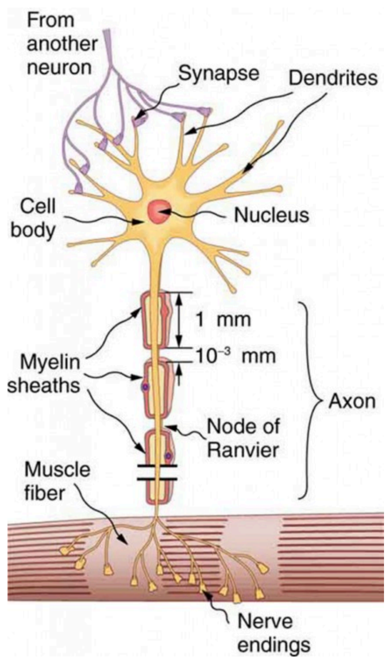


Figure 1. A neuron with its dendrites and long axon. Signals in the form of electric currents reach the cell body through dendrites and across synapses, stimulating the neuron to generate its own signal sent down the axon. The number of interconnections can be far greater than shown here.

indicated). As discussed in Molecular Transport Phenomena: Diffusion, Osmosis, and Related Processes, free ions will diffuse from a region of high concentration to one of low concentration. But the cell membrane is *semipermeable*, meaning that some ions may cross it while others cannot. In its resting state, the cell membrane is permeable to K^+ and Cl^-, and impermeable to Na^+. Diffusion of K^+ and Cl^- thus creates the layers of positive and negative charge on the outside and inside of the membrane. The Coulomb force prevents the ions from diffusing across in their entirety. Once the charge layer has built up, the repulsion of like charges prevents more from moving across, and the attraction of unlike charges prevents more from leaving either side. The result is two layers of charge right on the membrane, with diffusion being balanced by the Coulomb force. A tiny fraction of the charges move across and the fluids remain neutral (other ions are present), while a separation of charge and a voltage have been created across the membrane.

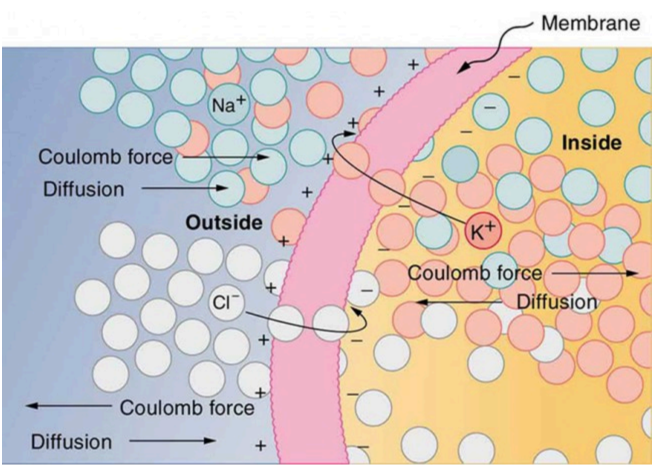


Figure 2. The semipermeable membrane of a cell has different concentrations of ions inside and out. Diffusion moves the K^+ and Cl^- ions in the direction shown, until the Coulomb force halts further transfer. This results in a layer of positive charge on the outside, a layer of negative charge on the inside, and thus a voltage across the cell membrane. The membrane is normally impermeable to Na^+.

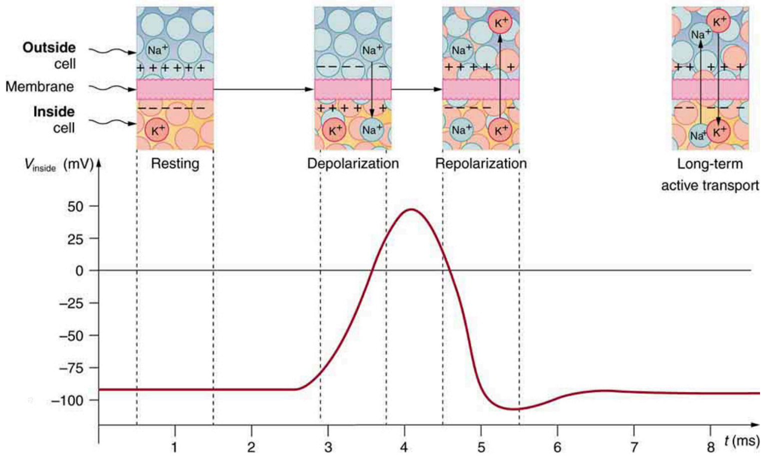


Figure 3. An action potential is the pulse of voltage inside a nerve cell graphed here. It is caused by movements of ions across the cell membrane as shown. Depolarization occurs when a stimulus makes the membrane permeable to Na+ ions. Repolarization follows as the membrane again becomes impermeable to Na^+, and K^+ moves from high to low concentration. In the long term, active transport slowly maintains the concentration differences, but the cell may fire hundreds of times in rapid succession without seriously depleting them.

The separation of charge creates a potential difference of 70 to 90 mV across the cell membrane. While this is a small voltage, the resulting electric field ($E = V/d$) across the only 8-nm-thick membrane is immense (on the order of 11 MV/m!) and has fundamental effects on its structure and permeability. Now, if the exterior of a neuron is taken to be at 0 V, then the interior has a *resting potential* of about −90 mV. Such voltages are created across the membranes of almost all types of animal cells but are largest in nerve and muscle cells. In fact, fully 25% of the energy used by cells goes toward creating and maintaining these potentials.

Electric currents along the cell membrane are created by any stimulus that changes the membrane's permeability. The membrane thus temporarily becomes permeable to Na^+, which then rushes in, driven both by diffusion and the Coulomb force. This inrush of Na^+

first neutralizes the inside membrane, or *depolarizes* it, and then makes it slightly positive. The depolarization causes the membrane to again become impermeable to Na^+, and the movement of K^+ quickly returns the cell to its resting potential, or *repolarizes* it. This sequence of events results in a voltage pulse, called the *action potential*. (See Figure 3.) Only small fractions of the ions move, so that the cell can fire many hundreds of times without depleting the excess concentrations of Na^+ and K^+. Eventually, the cell must replenish these ions to maintain the concentration differences that create bioelectricity. This sodium-potassium pump is an example of *active transport*, wherein cell energy is used to move ions across membranes against diffusion gradients and the Coulomb force.

The action potential is a voltage pulse at one location on a cell membrane. How does it get transmitted along the cell membrane, and in particular down an axon, as a nerve impulse? The answer is that the changing voltage and electric fields affect the permeability of the adjacent cell membrane, so that the same process takes place there. The adjacent membrane depolarizes, affecting the membrane further down, and so on, as illustrated in Figure 4. Thus the action potential stimulated at one location triggers a *nerve impulse* that moves slowly (about 1 m/s) along the cell membrane.

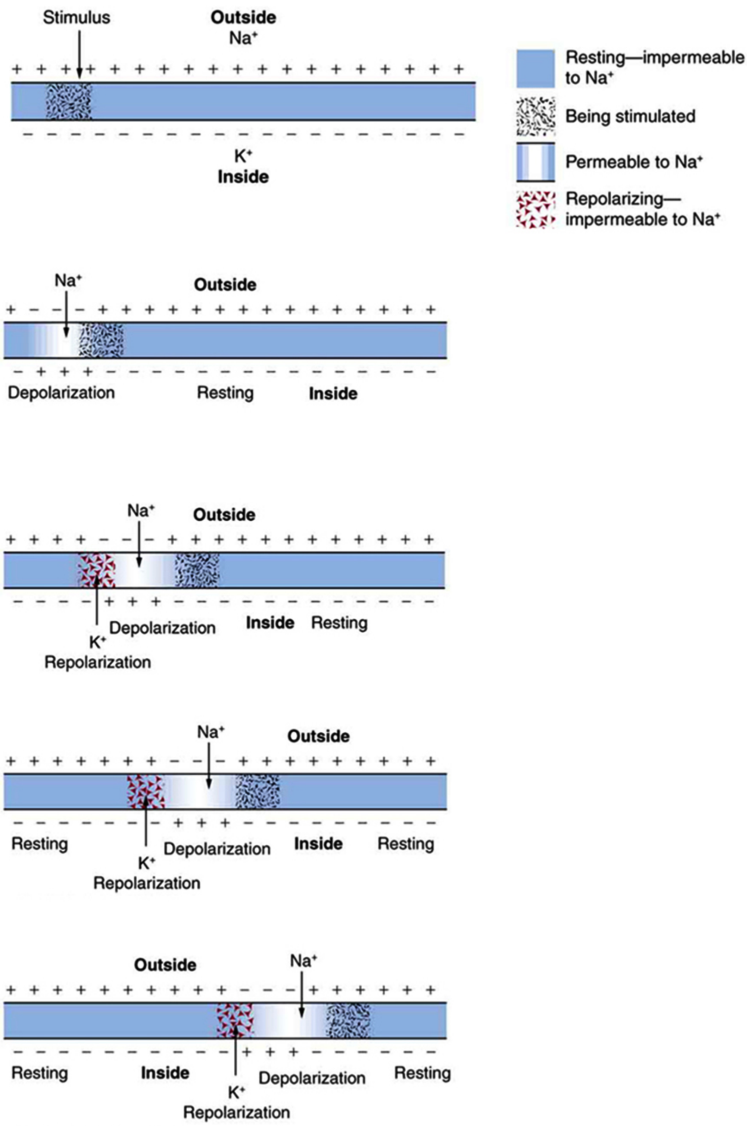


Figure 4. A nerve impulse is the propagation of an action potential along a cell membrane. A stimulus causes an action potential at one location, which changes the permeability of the adjacent membrane, causing an action potential there. This in turn affects the membrane further down, so that the action potential moves slowly (in electrical terms) along the cell membrane.

Although the impulse is due to Na^+ *and* K^+ *going across the membrane, it is equivalent to a wave of charge moving along the outside and inside of the membrane.*

Some axons, like that in Figure 1, are sheathed with *myelin*, consisting of fat-containing cells. Figure 5 shows an enlarged view of an axon having myelin sheaths characteristically separated by unmyelinated gaps (called nodes of Ranvier). This arrangement gives the axon a number of interesting properties. Since myelin is an insulator, it prevents signals from jumping between adjacent nerves (cross talk). Additionally, the myelinated regions transmit electrical signals at a very high speed, as an ordinary conductor or resistor would. There is no action potential in the myelinated regions, so that no cell energy is used in them. There is an IR signal loss in the myelin, but the signal is regenerated in the gaps, where the voltage pulse triggers the action potential at full voltage. So a myelinated axon transmits a nerve impulse faster, with less energy consumption, and is better protected from cross talk than an unmyelinated one. Not all axons are myelinated, so that cross talk and slow signal transmission are a characteristic of the normal operation of these axons, another variable in the nervous system.

The degeneration or destruction of the myelin sheaths that surround the nerve fibers impairs signal transmission and can lead to numerous neurological effects. One of the most prominent of these diseases comes from the body's own immune system attacking the myelin in the central nervous system—multiple sclerosis. MS symptoms include fatigue, vision problems, weakness of arms and legs, loss of balance, and tingling or numbness in one's extremities (neuropathy). It is more apt to strike younger adults, especially females. Causes might come from infection, environmental or geographic affects, or genetics. At the moment there is no known cure for MS.

Most animal cells can fire or create their own action potential. Muscle cells contract when they fire and are often induced to do so by a nerve impulse. In fact, nerve and muscle cells are physiologically similar, and there are even hybrid cells, such as in the heart, that have characteristics of both nerves and muscles. Some animals, like the infamous electric eel (see Figure 6), use muscles ganged so that their voltages add in order to create a shock great enough to stun prey.

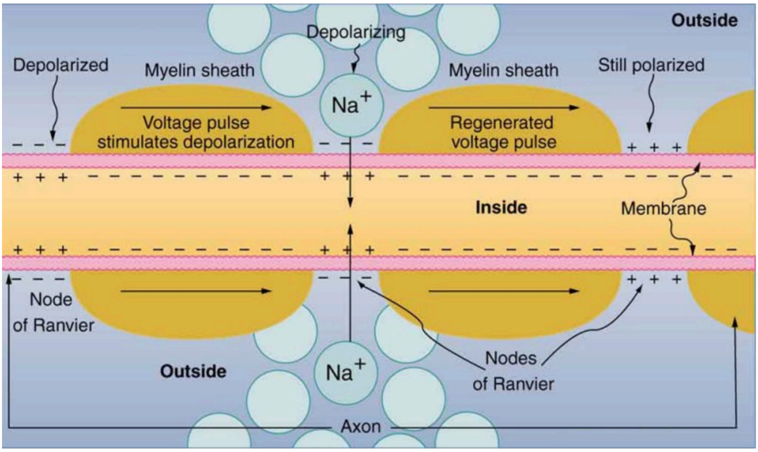


Figure 5. Propagation of a nerve impulse down a myelinated axon, from left to right. The signal travels very fast and without energy input in the myelinated regions, but it loses voltage. It is regenerated in the gaps. The signal moves faster than in unmyelinated axons and is insulated from signals in other nerves, limiting cross talk.

Figure 6. An electric eel flexes its muscles to create a voltage that stuns prey. (credit: chrisbb, Flickr)

Electrocardiograms

Just as nerve impulses are transmitted by depolarization and repolarization of adjacent membrane, the depolarization that causes muscle contraction can also stimulate adjacent muscle cells to depolarize (fire) and contract. Thus, a depolarization wave can be sent across the heart, coordinating its rhythmic contractions and enabling it to perform its vital function of propelling blood through the circulatory system. Figure 7 is a simplified graphic of a depolarization wave spreading across the heart from the *sinoarterial* (SA) *node*, the heart's natural pacemaker.

An *electrocardiogram* (ECG) is a record of the voltages created by the wave of depolarization and subsequent repolarization in the heart. Voltages between pairs of electrodes placed on the chest are vector components of the voltage wave on the heart. Standard ECGs have 12 or more electrodes, but only three are shown in Figure 7 for clarity. Decades ago, three-electrode ECGs were performed by placing electrodes on the left and right arms and the left leg. The voltage between the right arm and the left leg is called the *lead II potential* and is the most often graphed. We shall examine the lead II potential as an indicator of heart-muscle function and see that it is coordinated with arterial blood pressure as well.

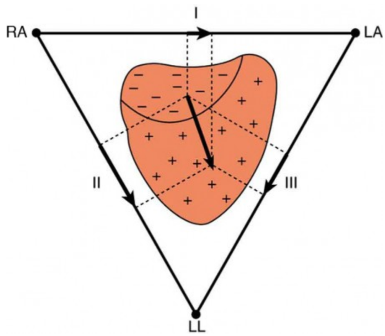


Figure 7. The outer surface of the heart changes from positive to negative during depolarization. This wave of depolarization is spreading from the top of the heart and is represented by a vector pointing in the direction of the wave. This vector is a voltage (potential difference) vector. Three electrodes, labeled RA, LA, and LL, are placed on the patient. Each pair (called leads I, II, and III) measures a component of the depolarization vector and is graphed in an ECG.

Heart function and its four-chamber action are explored in Viscosity and Laminar Flow; Poiseuille's Law. Basically, the right and left atria receive blood from the body and lungs, respectively, and pump the blood into the ventricles. The right and left ventricles, in turn, pump blood through the lungs and the rest of the body, respectively. Depolarization of the heart muscle causes it to contract. After contraction it is repolarized to ready it for the next beat. The ECG measures components of depolarization and repolarization of the heart muscle and can yield significant information on the functioning and malfunctioning of the heart.

Figure 8 shows an ECG of the lead II potential and a graph of the corresponding arterial blood pressure. The major features are labeled P, Q, R, S, and T. The *P wave* is generated by the depolarization and contraction of the atria as they pump blood into the ventricles. The QRS *complex* is created by the depolarization of the ventricles as they pump blood to the lungs and body. Since the shape of the heart and the path of the depolarization wave are not simple, the QRS complex has this typical shape and time span. The lead II QRS signal also masks the repolarization of the atria, which occur at the same time. Finally, the *T wave* is generated by the repolarization of the ventricles and is followed by the next P wave in the next heartbeat. Arterial blood pressure varies with each part of the heartbeat, with systolic (maximum) pressure occurring closely after the QRS complex, which signals contraction of the ventricles.

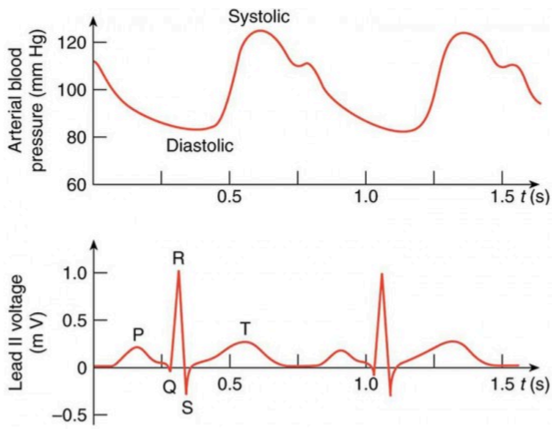


Figure 8. A lead II ECG with corresponding arterial blood pressure. The QRS complex is created by the depolarization and contraction of the ventricles and is followed shortly by the maximum or systolic blood pressure. See text for further description.

Taken together, the 12 leads of a state-of-the-art ECG can yield a wealth of information about the heart. For example, regions of damaged heart tissue, called infarcts, reflect electrical waves and are apparent in one or more lead potentials. Subtle changes due to slight or gradual damage to the heart are most readily detected by comparing a recent ECG to an older one. This is particularly the case since individual heart shape, size, and orientation can cause variations in ECGs from one individual to another. ECG technology has advanced to the point where a portable ECG monitor with a liquid crystal instant display and a printer can be carried to patients' homes or used in emergency vehicles. See Figure 9.

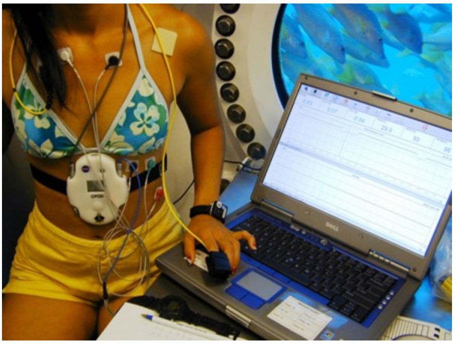
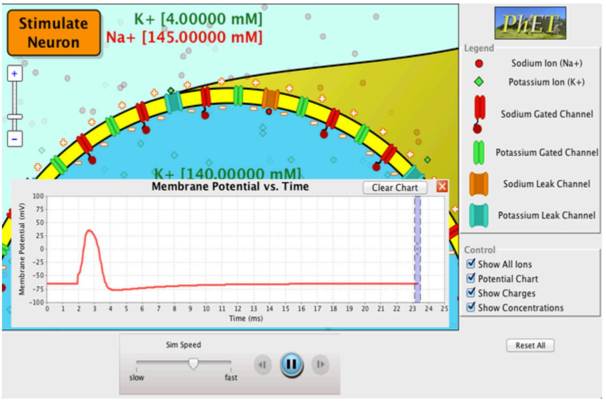


Figure 9. This NASA scientist and NEEMO 5 *aquanaut's heart rate and other vital signs are being recorded by a portable device while living in an underwater habitat. (credit:* NASA, *Life Sciences Data Archive at Johnson Space Center, Houston, Texas)*

PhET Explorations: Neuron

Stimulate a neuron and monitor what happens. Pause, rewind, and move forward in time in order to observe the ions as they move across the neuron membrane.



Click to download the simulation. Run using Java.

Section Summary

- Electric potentials in neurons and other cells are created by ionic concentration differences across semipermeable membranes.
- Stimuli change the permeability and create action potentials that propagate along neurons.
- Myelin sheaths speed this process and reduce the needed energy input.
- This process in the heart can be measured with an electrocardiogram (ECG).

Conceptual Questions

1. Note that in Figure 2, both the concentration gradient and the Coulomb force tend to move Na^+ ions into the cell. What prevents this?

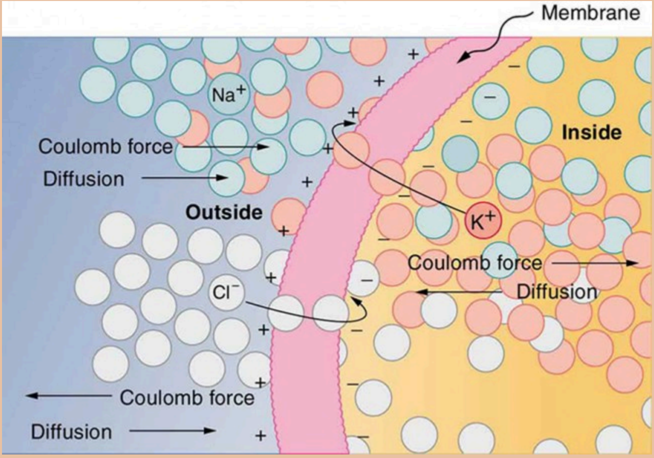


Figure 2. The semipermeable membrane of a cell has different concentrations of ions inside and out. Diffusion moves the K+ and Cl− ions in the direction shown, until the Coulomb force halts further transfer. This results in a layer of positive charge on the outside, a layer of negative charge on the inside, and thus a voltage across the cell membrane. The membrane is normally impermeable to Na+.

2. Define depolarization, repolarization, and the action potential.

3. Explain the properties of myelinated nerves in terms of the insulating properties of myelin.

Problems & Exercises

1. **Integrated Concepts** Use the ECG in Figure 8 to determine the heart rate in beats per minute assuming a constant time between beats.

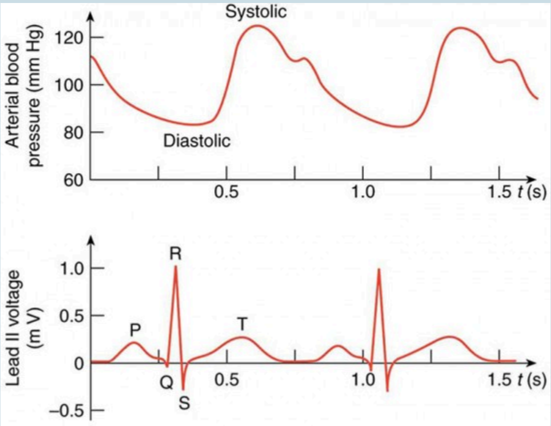


Figure 8. A lead II ECG with corresponding arterial blood pressure. The QRS complex is created by the depolarization and contraction of the ventricles and is followed shortly by the maximum or systolic blood pressure. See text for further description.

Integrated Concepts

(a) Referring to Figure 8, find the time systolic pressure lags behind the middle of the QRS complex. (b) Discuss the reasons for the time lag.

Glossary

nerve conduction:
: the transport of electrical signals by nerve cells

bioelectricity:
: electrical effects in and created by biological systems

semipermeable:
: property of a membrane that allows only certain types of ions to cross it

electrocardiogram (ECG):
: usually abbreviated ECG, a record of voltages created by depolarization and repolarization, especially in the heart

Selected Solutions to Problems & Exercises

1. 80 beats/minute

PART IV

CIRCUITS AND DC INSTRUMENTS

29. Introduction to Circuits and DC Instruments

Figure 1. Electric circuits in a computer allow large amounts of data to be quickly and accurately analyzed.. (credit: Airman 1st Class Mike Meares, United States Air Force)

Electric circuits are commonplace. Some are simple, such as those in flashlights. Others, such as those used in supercomputers, are extremely complex.

This collection of modules takes the topic of electric circuits a step beyond simple circuits. When the circuit is purely resistive, everything in this module applies to both DC and AC. Matters become more complex when capacitance is involved. We do consider what happens when capacitors are connected to DC voltage sources, but the interaction of capacitors and other nonresistive devices with AC is left for a later chapter. Finally, a number of important DC instruments, such as meters that measure voltage and current, are covered in this chapter.

30. 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.

Most 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 1. The total resistance of a combination of resistors depends on both their individual values and how they are connected.

Resistors in Series

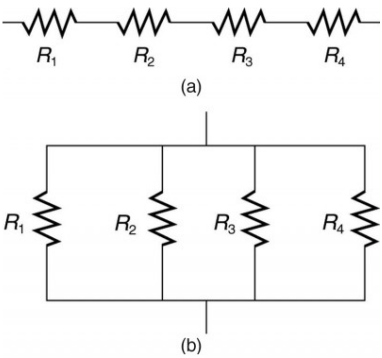


Figure 1. (a) A series connection of resistors. (b) A parallel connection of resistors.

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 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 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.)

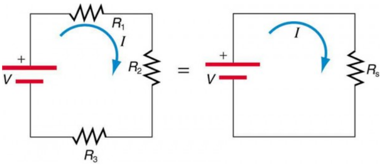


Figure 2. Three resistors connected in series to a battery (left) and the equivalent single or series resistance (right).

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 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 R_2 is $V_2 = IR_2$, and that across R_3 is $V_3 = IR_3$. The sum of these voltages equals the voltage output of the source; that is,

$$V = V_1 + V_2 + V_3.$$

This equation is based on the conservation of energy and conservation of charge. Electrical potential energy can be described by the equation $PE = qV$, 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

$$qV_1 + qV_2 + qV_3.$$

Making 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

$$V = IR_1 + IR_2 + IR_3 = I(R_1 + R_2 + R_3).$$

Note that for the equivalent single series resistance R_s, we have

$$V = IR_s.$$

This implies that the total or equivalent series resistance R_s of three resistors is $R_s = R_1 + R_2 + R_3$. This logic is valid in general for any number of resistors in series; thus, the total resistance R_s of a series connection is

$$R_s = R_1 + R_2 + R_3 + \ldots,$$

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 1. Calculating Resistance, Current, Voltage Drop, and Power Dissipation: Analysis of a Series Circuit

Suppose the voltage output of the battery in Figure 2 is 12.0 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:

$$\begin{aligned} R_s &= R_1 + R_2 + R_3 \\ &= 1.00\ \Omega + 6.00\ \Omega + 13.0\ \Omega. \\ &= 20.0\ \Omega \end{aligned}$$

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:

$$I = \frac{V}{R_s} = \frac{12.0\ \text{V}}{20.0\ \Omega} = 0.60\ \text{A}.$$

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

$$V_1 = IR_1 = (0.600\text{A})(1.0\ \Omega) = 0.600\ \text{V}.$$

Similarly,

$$V_2 = IR_2 = (0.600\text{A})(6.0\ \Omega) = 3.60\ \text{V}$$

and

$$V3 = IR_3 = (0.600\text{A})(13.0\ \Omega) = 7.80\ \text{V}.$$

Discussion for (c)

The three IR drops add to 12.0 V, as predicted:

$$V_1 + V_2 + V_3 = (0.600 + 3.60 + 7.80)\text{V} = 12.0\ \text{V}.$$

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

$$P_1 = I^2R_1 = (0.600 \text{ A})^2(1.00 \ \Omega) = 0.360 \text{ W}.$$

Similarly,

$$P_2 = I^2R_2 = (0.600 \text{ A})^2(6.00 \ \Omega) = 2.16 \text{ W}.$$

and

$$P_3 = I^2R_3 = (0.600 \text{ A})^2(13.0 \ \Omega) = 4.68 \text{ W}.$$

Discussion for (d)

Power can also be calculated using either $P = IV$ or $P = \frac{V^2}{R}$, 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

$$P = (0.600\ A)(12.0\ V) = 7.20\ W.$$

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,

$$P_1 + P_2 + P_3 = (0.360 + 2.16 + 4.68)\ W = 7.20\ W.$$

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: $R_s = R_1 + R_2 + R_3 + ...$
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 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(b).)

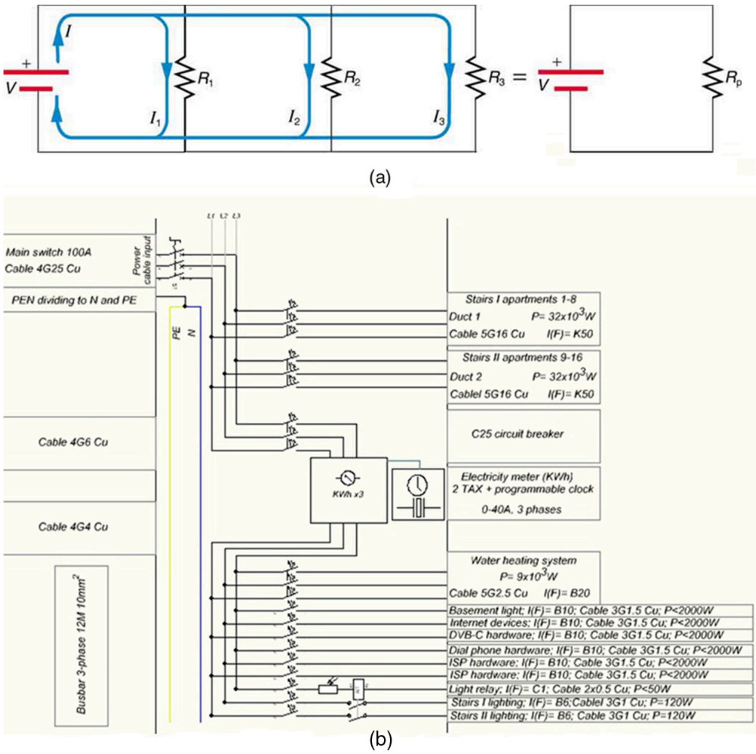


Figure 3. (a) Three resistors connected in parallel to a battery and the equivalent single or parallel resistance. (b) Electrical power setup in a house. (credit: Dmitry G, Wikimedia Commons)

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 $I_1 = \frac{V}{R_1}$,

$I_2 = \dfrac{V}{R_2}$, and $I_3 = \dfrac{V}{R_3}$. Conservation of charge implies that the total current I produced by the source is the sum of these currents:

$$I = I_1 + I_2 + I_3.$$

Substituting the expressions for the individual currents gives

$$I = \frac{V}{R_1} + \frac{V}{R_2} + \frac{V}{R_3} = V\left(\frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3}\right)$$

.

Note that Ohm's law for the equivalent single resistance gives

$$I = \frac{V}{R_p} = V\left(\frac{1}{R_p}\right).$$

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

$$\frac{1}{R_p} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_{.3}} + ...$$

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 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 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

$$\frac{1}{R_p} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} = \frac{1}{1.00\ \Omega} + \frac{1}{6.00\ \Omega} + \frac{1}{13.0\ \Omega}.$$

Thus,

$$\frac{1}{R_p} = \frac{1.00}{\Omega} + \frac{0.1667}{\Omega} + \frac{0.07692}{\Omega} = \frac{1.2436}{\Omega}.$$

(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

$$R_\mathrm{p} = \frac{1}{1.2436}\ \Omega = 0.8041\ \Omega.$$

The total resistance with the correct number of significant digits is $R_p = 0.804\ \Omega$

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

$$I = \frac{V}{R_{\mathrm{p}}} = \frac{12.0 \text{ V}}{0.8041\ \Omega} = 14.92 \text{ A}.$$

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,

$$I_1 = \frac{V}{R_1} = \frac{12.0 \text{ V}}{1.00\ \Omega} = 12.0 \text{ A}.$$

Similarly,

$$I_2 = \frac{V}{R_2} = \frac{12.0\ \text{V}}{6.00\ \Omega} = 2.00\ \text{A}$$

and

$$I_3 = \frac{V}{R_3} = \frac{12.0\ \text{V}}{13.0\ \Omega} = 0.92\ \text{A}.$$

Discussion for (c)

The total current is the sum of the individual currents:

$$I_1 + I_2 + I_3 = 14.92\ \text{A}.$$

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,

$$P_1 = \frac{V^2}{R_1} = \frac{(12.0\ \text{V})^2}{1.00\ \Omega} = 144\ \text{W}.$$

Similarly,

$$P_2 = \frac{V^2}{R_2} = \frac{(12.0\ \text{V})^2}{6.00\ \Omega} = 24.0\ \text{W}.$$

and

$$P_3 = \frac{V^2}{R_3} = \frac{(12.0\text{ V})^2}{13.0\ \Omega} = 11.1\text{ W}.$$

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

$$P = IV = (14.92\text{ A})(12.0\text{ V}) = 179\text{ W}.$$

Discussion for (e)

Total power dissipated by the resistors is also 179 W:

$$P_1 + P_2 + P_3 = 144\text{ W} + 24.0\text{ W} + 11.1\text{ W} = 179\text{ W}.$$

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 $\frac{1}{R_\mathrm{p}}=\frac{1}{R_1}+\frac{1}{R_2}+\frac{1}{R_3}+\ldots$, 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 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.

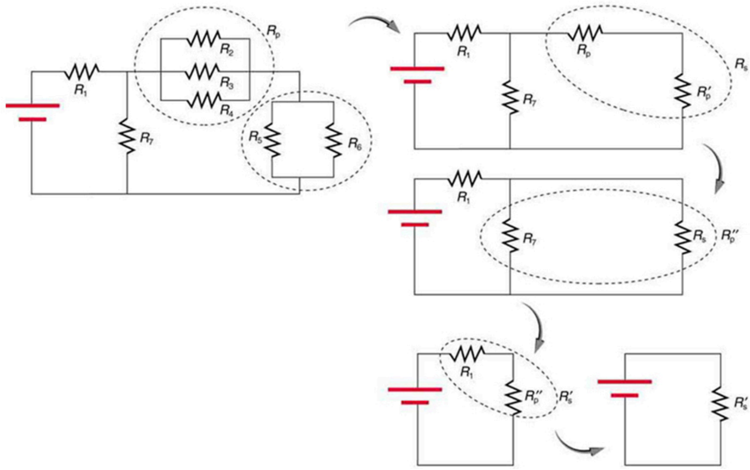


Figure 4. This combination of seven resistors has both series and parallel parts. Each is identified and reduced to an equivalent resistance, and these are further reduced until a single equivalent resistance is reached.

The simplest combination of series and parallel resistance, shown in Figure 4, 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. Calculating Resistance, IR Drop,

Current, and Power Dissipation: Combining Series and Parallel Circuits

Figure 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?

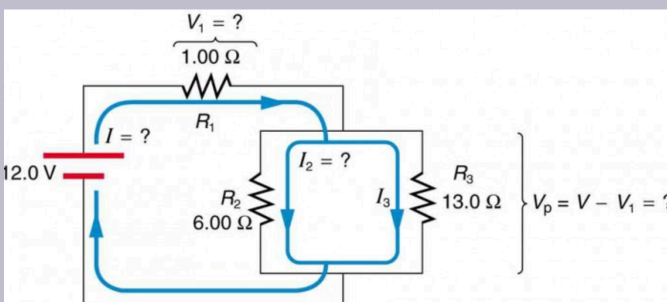


Figure 5. These three resistors are connected to a voltage source so that R_2 and R_3 are in parallel with one another and that combination is in series with R_1.

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

$$R_{tot} = R_1 + R_p.$$

First, we find R_p using the equation for resistors in parallel and entering known values:

$$\frac{1}{R_\text{p}} = \frac{1}{R_2} + \frac{1}{R_3} = \frac{1}{6.00\ \Omega} + \frac{1}{13.0\ \Omega} = \frac{0.2436}{\Omega}$$

.

Inverting gives

$$R_\text{p} = \frac{1}{0.2436}\ \Omega = 4.11\ \Omega.$$

So the total resistance is

$$R_{tot} = R_1 + R_p = 1.00\ \Omega + 4.11\ \Omega = 5.11\ \Omega.$$

Discussion for (a)

The total resistance of this combination is intermediate between the pure series and pure parallel values (20.0 Ω and 0.804 Ω, 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

$$V_1 = IR_1$$

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,

$$I = \frac{V}{R_\text{tot}} = \frac{12.0\ \text{V}}{5.11\ \Omega} = 2.35\ \text{A}.$$

Entering this into the expression above, we get

$$V_1 = IR_1 = (2.35\text{ A})(1.00\ \Omega) = 2.35\text{ V}.$$

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

$$V_p = V - V_1 = 12.0\text{ V} - 2.35\text{ V} = 9.65\text{ V}.$$

Now the current I_2 through resistance R_2 is found using Ohm's law:

$$I_2 = \frac{V_\text{p}}{R_2} = \frac{9.65\text{ V}}{6.00\ \Omega} = 1.61\text{A}.$$

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

$$P_2 = (I_2)^2 R_2 = (1.61\ \text{A})^2 (6.00\ \Omega) = 15.5\ \text{W}$$

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 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.

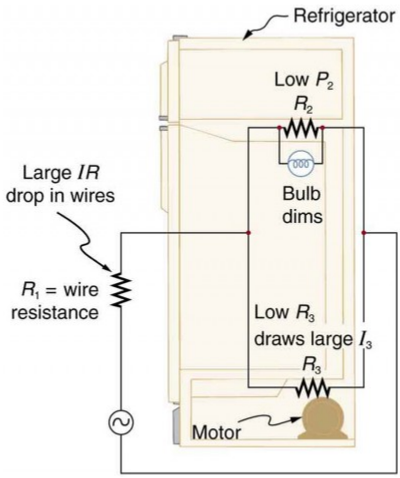


Figure 6. Why do lights dim when a large appliance is switched on? The answer is that the large current the appliance motor draws causes a significant drop in the wires and reduces the voltage across the light.

Check Your Understanding

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.

Solution

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 , 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.

Section Summary

- The total resistance of an electrical circuit with resistors wired in a series is the sum of the individual resistances: $R_s = R_1 + R_2 + R_3 + \ldots$.
- 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:

$$\frac{1}{R_{\mathrm{p}}} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} + \ldots.$$

- 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.

Conceptual Questions

1. A switch has a variable resistance that is nearly zero when closed and extremely large when open, and it is placed in series with the device it controls. Explain the effect the switch in Figure 7 has on current when open and when closed.

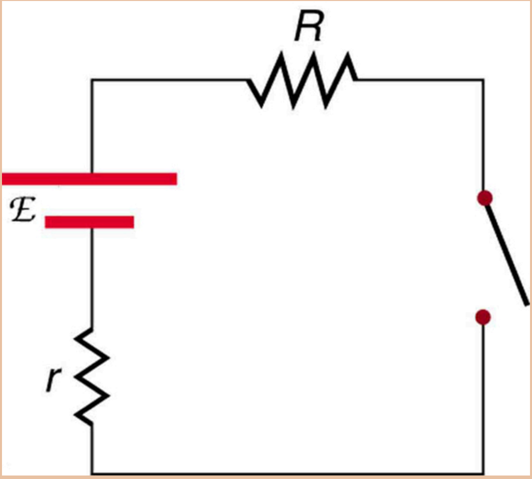


Figure 7. A switch is ordinarily in series with a resistance and voltage source. Ideally, the switch has nearly zero resistance when closed but has an extremely large resistance when open. (Note that in this diagram, the script E represents the voltage (or electromotive force) of the battery.)

2. What is the voltage across the open switch in Figure 7?

3. There is a voltage across an open switch, such as in Figure 7. Why, then, is the power dissipated by the open switch small?

4. Why is the power dissipated by a closed switch, such as in Figure 7, small?

5. A student in a physics lab mistakenly wired a light bulb, battery, and switch as shown in Figure 8. Explain why the bulb is on when the switch is open, and off when the switch is closed. (Do not try this—it is hard on the battery!)

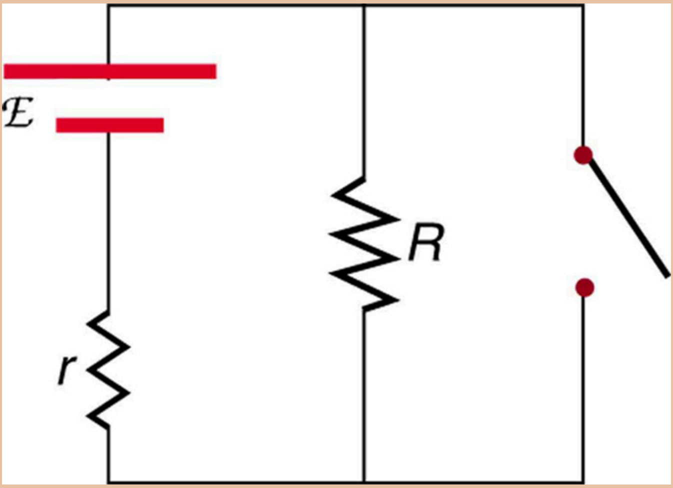


Figure 8. A wiring mistake put this switch in parallel with the device represented by [latex]R[/latex] . (Note that in this diagram, the script E represents the voltage (or electromotive force) of the battery.)

6. Knowing that the severity of a shock depends on the magnitude of the current through your body, would you prefer to be in series or parallel with a resistance, such as the heating element of a toaster, if shocked by it? Explain.

7. Would your headlights dim when you start your car's engine if the wires in your automobile were superconductors? (Do not neglect the battery's internal resistance.) Explain.

8. Some strings of holiday lights are wired in series to save wiring costs. An old version utilized bulbs that break the electrical connection, like an open switch, when they burn out. If one such bulb burns out, what happens to the

others? If such a string operates on 120 V and has 40 identical bulbs, what is the normal operating voltage of each? Newer versions use bulbs that short circuit, like a closed switch, when they burn out. If one such bulb burns out, what happens to the others? If such a string operates on 120 V and has 39 remaining identical bulbs, what is then the operating voltage of each?

9. If two household lightbulbs rated 60 W and 100 W are connected in series to household power, which will be brighter? Explain.

10. Suppose you are doing a physics lab that asks you to put a resistor into a circuit, but all the resistors supplied have a larger resistance than the requested value. How would you connect the available resistances to attempt to get the smaller value asked for?

11. Before World War II, some radios got power through a "resistance cord" that had a significant resistance. Such a resistance cord reduces the voltage to a desired level for the radio's tubes and the like, and it saves the expense of a transformer. Explain why resistance cords become warm and waste energy when the radio is on.

12. Some light bulbs have three power settings (not including zero), obtained from multiple filaments that are individually switched and wired in parallel. What is the minimum number of filaments needed for three power settings?

Problems & Exercises

Note: Data taken from figures can be assumed to be accurate to three significant digits.

1. (a) What is the resistance of ten 275-Ω resistors connected in series? (b) In parallel?

2. (a) What is the resistance of a 1.00×10^{2}-Ω, a 2.50-kΩ, and a 4.00-kΩ resistor connected in series? (b) In parallel?

3. What are the largest and smallest resistances you can obtain by connecting a 36.0-Ω, a 50.0-Ω, and a 700-Ω resistor together?

4. An 1800-W toaster, a 1400-W electric frying pan, and a 75-W lamp are plugged into the same outlet in a 15-A, 120-V circuit. (The three devices are in parallel when plugged into the same socket.). (a) What current is drawn by each device? (b) Will this combination blow the 15-A fuse?

5. Your car's 30.0-W headlight and 2.40-kW starter are ordinarily connected in parallel in a 12.0-V system. What power would one headlight and the starter consume if connected in series to a 12.0-V battery? (Neglect any other resistance in the circuit and any change in resistance in the two devices.)

6. (a) Given a 48.0-V battery and 24.0-Ω and 96.0-Ω resistors, find the current and power for each when connected in series. (b) Repeat when the resistances are in parallel.

7. Referring to the example combining series and parallel circuits and Figure 5, calculate I_3 in the following two different ways: (a) from the known values of I and I_2 ; (b) using Ohm's law for R_3. In both parts explicitly show how you follow the steps in the *Problem-Solving Strategies for Series and Parallel Resistors* above.

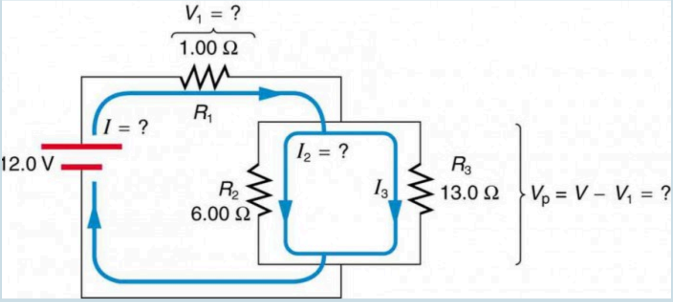


Figure 5. These three resistors are connected to a voltage source so that R_2 and R_3 are in parallel with one another and that combination is in series with R_1.

8. Referring to Figure 5: (a) Calculate P_3 and note how it compares with P_3 found in the first two example problems in this module. (b) Find the total power supplied by the source and compare it with the sum of the powers dissipated by the resistors.

9. Refer to Figure 6 and the discussion of lights dimming when a heavy appliance comes on. (a) Given the voltage source is 120 V, the wire resistance is 0.400 Ω, and the bulb is nominally 75.0 W, what power will the bulb dissipate if a total of 15.0 A passes through the wires when the motor

comes on? Assume negligible change in bulb resistance. (b) What power is consumed by the motor?

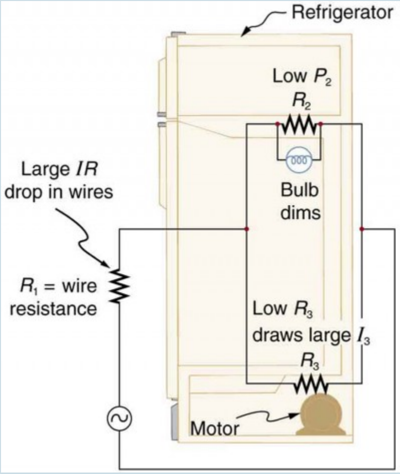


Figure 6. Why do lights dim when a large appliance is switched on? The answer is that the large current the appliance motor draws causes a significant drop in the wires and reduces the voltage across the light.

10. A 240-kV power transmission line carrying 5.00×10^2 is hung from grounded metal towers by ceramic insulators, each having a 1.00×10^9-Ω resistance (Figure 9(a)). What is the resistance to ground of 100 of these insulators? (b) Calculate the power dissipated by 100 of them. (c) What fraction of the power carried by the line is this? Explicitly show how you follow the steps in the *Problem-Solving Strategies for Series and Parallel Resistors* above.

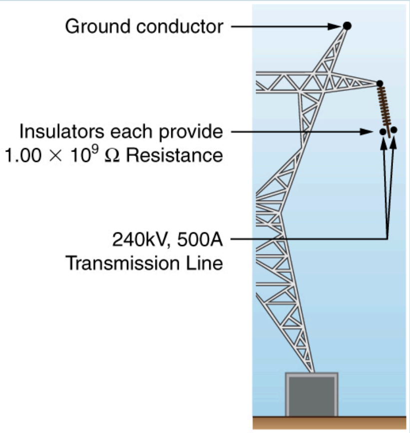


Figure 9. High-voltage (240-kV) transmission line carrying 5.00×10^2 is hung from a grounded metal transmission tower. The row of ceramic insulators provide 1.00×10^9 Ω of resistance each.

11. Show that if two resistors R_1 and R_2 are combined and one is much greater than the other ($R_1 >> R_2$): (a) Their series resistance is very nearly equal to the greater resistance R_1 . (b) Their parallel resistance is very nearly equal to smaller resistance R_2.

12. **Unreasonable Results** Two resistors, one having a resistance of 145 Ω, are connected in parallel to produce a total resistance of 150 Ω. (a) What is the value of the second resistance? (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

13. **Unreasonable Results** Two resistors, one having a resistance of 900 kΩ, are connected in series to produce a

total resistance of 0.500 MΩ. (a) What is the value of the second resistance? (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

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:
: the relationship between potential electrical power, voltage, and resistance in an electrical circuit, given by: $P_e = \mathrm{IV}$

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

Selected Solutions to Problems & Exercises

1. (a) 2.75 kΩ (b) 27.5 Ω

3.(a) 786 Ω (b) 20.3 Ω

5. 29.6 W

7. (a) 0.74 A (b) 0.742 A

9. (a) 60.8 W (b) 3.18 kW

11. (a) $$\begin{array}{c} R_{\text{s}} = R_1 + R_2 \\ \Rightarrow R_{\text{s}} \approx R_1 \; (R_1 >> R_2) \end{array}$$

(b) $$\frac{1}{R_p} = \frac{1}{R_1} + \frac{1}{R_2} = \frac{R_1 + R_2}{R_1 R_2},$$

so that

$$R_p = \frac{R_1 R_2}{R_1 + R_2} \approx \frac{R_1 R_2}{R_1} = R_2 \; (R_1 >> R_2).$$

13.(a) −400 kΩ (b) Resistance cannot be negative. (c) Series resistance is said to be less than one of the resistors, but it must be greater than any of the resistors.

31. Electromotive Force: Terminal Voltage

Learning Objectives

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

- Compare and contrast the voltage and the electromagnetic force of an electric power source.
- Describe what happens to the terminal voltage, current, and power delivered to a load as internal resistance of the voltage source increases (due to aging of batteries, for example).
- Explain why it is beneficial to use more than one voltage source connected in parallel.

When you forget to turn off your car lights, they slowly dim as the battery runs down. Why don't they simply blink off when the battery's energy is gone? Their gradual dimming implies that battery output voltage decreases as the battery is depleted. Furthermore, if you connect an excessive number of 12-V lights in parallel to a car battery, they will be dim even when the battery is fresh and even if the wires to the lights have very low resistance. This implies that the battery's output voltage is reduced by the overload. The reason for the decrease in output voltage for depleted or overloaded batteries is that all voltage sources have two fundamental parts—a source of electrical energy and an *internal resistance*. Let us examine both.

Electromotive Force

You can think of many different types of voltage sources. Batteries themselves come in many varieties. There are many types of mechanical/electrical generators, driven by many different energy sources, ranging from nuclear to wind. Solar cells create voltages directly from light, while thermoelectric devices create voltage from temperature differences. A few voltage sources are shown in Figure 1. All such devices create a *potential difference* and can supply current if connected to a resistance. On the small scale, the potential difference creates an electric field that exerts force on charges, causing current. We thus use the name *electromotive force*, abbreviated emf. Emf is not a force at all; it is a special type of potential difference. To be precise, the electromotive force (emf) is the potential difference of a source when no current is flowing. Units of emf are volts.

Figure 1. A variety of voltage sources (clockwise from top left): the Brazos Wind Farm in Fluvanna, Texas (credit: Leaflet, Wikimedia Commons); the Krasnoyarsk Dam in Russia (credit: Alex Polezhaev); a solar farm (credit: U.S. Department of Energy); and a group of nickel metal hydride batteries (credit: Tiaa Monto). The voltage output of each depends on its construction and load, and equals emf only if there is no load.

Electromotive force is directly related to the source of potential difference, such as the particular combination of chemicals in a battery. However, emf differs from the voltage output of the device when current flows. The voltage across the terminals of a battery, for example, is less than the emf when the battery supplies current, and it declines further as the battery is depleted or loaded down. However, if the device's output voltage can be measured without drawing current, then output voltage will equal emf (even for a very depleted battery).

Internal Resistance

As noted before, a 12-V truck battery is physically larger, contains more charge and energy, and can deliver a larger current than a 12-V motorcycle battery. Both are lead-acid batteries with identical emf, but, because of its size, the truck battery has a smaller internal resistance r. Internal resistance is the inherent resistance to the flow of current within the source itself. Figure 2 is a schematic representation of the two fundamental parts of any voltage source. The emf (represented by a script E in the figure) and internal resistance r are in series. The smaller the internal resistance for a given emf, the more current and the more power the source can supply.

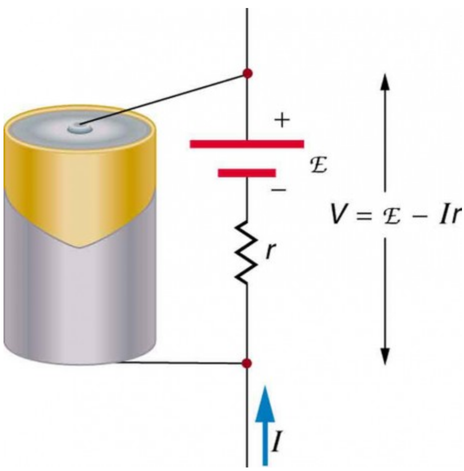


Figure 2. Any voltage source (in this case, a carbon-zinc dry cell) has an emf related to its source of potential difference, and an internal resistance r related to its construction. (Note that the script E stands for emf.). Also shown are the output terminals across which the terminal voltage V is measured. Since V = emf – Ir, terminal voltage equals emf only if there is no current flowing.

The internal resistance r can behave in complex ways. As noted, r increases as a battery is depleted. But internal resistance may also depend on the magnitude and direction of the current through a voltage source, its temperature, and even its history. The internal resistance of rechargeable nickel-cadmium cells, for example, depends on how many times and how deeply they have been depleted.

Things Great and Small: The Submicroscopic Origin of Battery Potential

Various types of batteries are available, with emfs determined by the combination of chemicals involved. We can view this as a molecular reaction (what much of chemistry is about) that separates charge. The lead-acid battery used in cars and other vehicles is one of the most common types. A single cell (one of six) of this battery is seen in Figure 3. The cathode (positive) terminal of the cell is connected to a lead oxide plate, while the anode (negative) terminal is connected to a lead plate. Both plates are immersed in sulfuric acid, the electrolyte for the system.

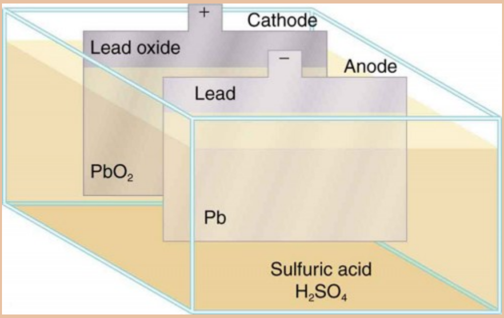


Figure 3. Artist's conception of a lead-acid cell. Chemical reactions in a lead-acid cell separate charge, sending negative charge to the anode, which is connected to the lead plates. The lead oxide plates are connected to the positive or cathode terminal of the cell. Sulfuric acid conducts the charge as well as participating in the chemical reaction.

The details of the chemical reaction are left to the reader to pursue in a chemistry text, but their results at the molecular level help explain the potential created by the battery. Figure 4 shows the result of a single chemical reaction. Two electrons are placed on the anode, making it negative, provided that the cathode supplied two electrons. This leaves the cathode positively charged, because it has lost two electrons. In short, a separation of charge has been driven by a chemical reaction. Note that the reaction will not take place unless there is a complete circuit to allow two electrons to be supplied to the cathode. Under many circumstances, these electrons come from the anode, flow through a resistance, and return to the cathode. Note also that since the chemical reactions involve substances with resistance, it is not possible to create the emf without an internal resistance.

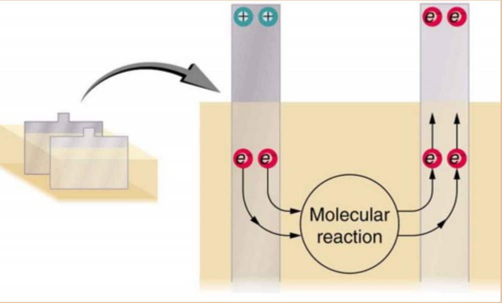


Figure 4. Artist's conception of two electrons being forced onto the anode of a cell and two electrons being removed from the cathode of the cell. The chemical reaction in a lead-acid battery places two electrons on the anode and removes two from the cathode. It requires a closed circuit to proceed, since the two electrons must be supplied to the cathode.

Why are the chemicals able to produce a unique potential difference? Quantum mechanical descriptions of molecules, which take into account the types of atoms and numbers of electrons in them, are able to predict the energy states they can have and the energies of reactions between them.

In the case of a lead-acid battery, an energy of 2 eV is given to each electron sent to the anode. Voltage is defined as the electrical potential energy divided by charge

$$V = \frac{P_{\mathrm{E}}}{q}.$$

An electron volt is the energy given to a single electron by a voltage of 1 V. So the voltage here is 2 V, since 2 eV is given to each electron. It is the energy produced in each molecular reaction that produces the voltage. A different reaction produces a different energy and, hence, a different voltage.

Terminal Voltage

The voltage output of a device is measured across its terminals and, thus, is called its *terminal voltage* V. Terminal voltage is given by

$$V = \text{emf} - Ir,$$

where r is the internal resistance and I is the current flowing at the time of the measurement. I is positive if current flows away from the positive terminal, as shown in Figure 2. You can see that the larger the current, the smaller the terminal voltage. And it is likewise true that the larger the internal resistance, the smaller the terminal voltage. Suppose a load resistance R_{load} is connected to a voltage source, as in Figure 5. Since the resistances are in series, the total resistance in the circuit is $R_{\text{load}} + r$. Thus the current is given by Ohm's law to be

$$I = \frac{\text{emf}}{R_{\text{load}} + r}.$$

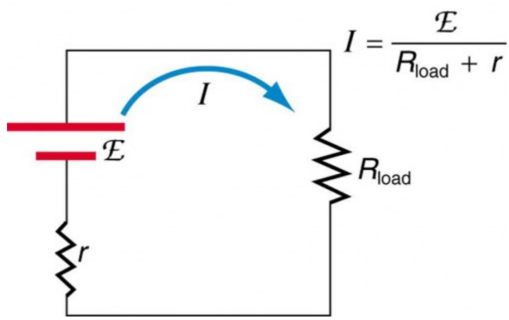


Figure 5. Schematic of a voltage source and its load Rload. Since the internal resistance r is in series with the load, it can significantly affect the terminal voltage and current delivered to the load. (Note that the script E stands for emf.)

We see from this expression that the smaller the internal resistance r, the greater the current the voltage source supplies to its load R_{load}. As batteries are depleted, r increases. If r becomes a significant fraction of the load resistance, then the current is significantly reduced, as the following example illustrates.

Example 1. Calculating Terminal Voltage, Power Dissipation, Current, and Resistance: Terminal Voltage and Load

A certain battery has a 12.0-V emf and an internal resistance of 0.100 Ω. (a) Calculate its terminal voltage when connected to a 10.0-Ω load. (b) What is the terminal voltage when connected to a 0.500-Ω load? (c) What power does the 0.500-Ω load dissipate? (d) If the internal resistance grows to0.500 Ω, find the current, terminal voltage, and power dissipated by a 0.500-Ω load.

Strategy

The analysis above gave an expression for current when internal resistance is taken into account. Once the current is found, the terminal voltage can be calculated using the equation V = emf − Ir. Once current is found, the power dissipated by a resistor can also be found.

Solution for (a)

Entering the given values for the emf, load resistance, and internal resistance into the expression above yields

$$I = \frac{\text{emf}}{R_{\text{load}} + r} = \frac{12.0\text{ V}}{10.1\ \Omega} = 1.188\text{ A}.$$

Enter the known values into the equation V = emf − *Ir* to get the terminal voltage:

$$\begin{aligned} V &= \text{emf} - Ir = 12.0\text{ V} - (1.188\text{ A})(0.100\ \Omega) \\ &= 11.9\text{ V} \end{aligned}$$

.

Discussion for (a)

The terminal voltage here is only slightly lower than the emf, implying that 10.0 Ω is a light load for this particular battery.

Solution for (b)

Similarly, with R_{load} = 0.500 Ω, the current is

$$I = \frac{\text{emf}}{R_{\text{load}} + r} = \frac{12.0\text{ V}}{0.600\ \Omega} = 20.0\text{ A}.$$

The terminal voltage is now

$$\begin{aligned} V &= \text{emf} - Ir = 12.0\text{ V} - (20.0\text{ A})(0.100\ \Omega) \\ &= 10.0\text{ V} \end{aligned}$$

Discussion for (b)

This terminal voltage exhibits a more significant reduction compared with emf, implying 0.500 Ω is a heavy load for this battery.

Solution for (c)

The power dissipated by the 0.500-Ω load can be found using the formula $P = I^2R$. Entering the known values gives

$$P_{load} = I^2R_{load} = (20.0\ \text{A})^2(0.500\ \Omega) = 2.00 \times 10^2\ \text{W}.$$

Discussion for (c)

Note that this power can also be obtained using the expressions $\frac{V^2}{R}$ or IV, where V is the terminal voltage (10.0 V in this case).

Solution for (d)

Here the internal resistance has increased, perhaps due to the depletion of the battery, to the point where it is as great as the load resistance. As before, we first find the current by entering the known values into the expression, yielding

$$I = \frac{\text{emf}}{R_{\text{load}} + r} = \frac{12.0\text{ V}}{1.00\ \Omega} = 12.0\text{ A}.$$

Now the terminal voltage is

$$V = \text{emf} - Ir = 12.0\text{ V} - (12.0\text{ A})(0.500\Omega) = 6.00\text{ V}.$$

and the power dissipated by the load is

$$P_{\text{load}} = I^2 R_{\text{load}} = (12.0\text{ A})^2 (0.500\ \Omega) = 72.0\text{ W}.$$

Discussion for (d)

We see that the increased internal resistance has significantly decreased terminal voltage, current, and power delivered to a load.

Battery testers, such as those in Figure 6, use small load resistors to intentionally draw current to determine whether the terminal voltage drops below an acceptable level. They really test the internal resistance of the battery. If internal resistance is high, the battery is weak, as evidenced by its low terminal voltage.

Figure 6. These two battery testers measure terminal voltage under a load to determine the condition of a battery. The large device is being used by a U.S. Navy electronics technician to test large batteries aboard the aircraft carrier USS Nimitz and has a small resistance that can dissipate large amounts of power. (credit: U.S. Navy photo by Photographer's Mate Airman Jason A. Johnston) The small device is used on small batteries and has a digital display to indicate the acceptability of their terminal voltage. (credit: Keith Williamson)

Some batteries can be recharged by passing a current through them in the direction opposite to the current they supply to a resistance. This is done routinely in cars and batteries for small electrical appliances and electronic devices, and is represented pictorially in Figure 7. The voltage output of the battery charger must be greater than the emf of the battery to reverse current through it. This will cause the terminal voltage of the battery to be greater than the emf, since V = emf − Ir, and I is now negative.

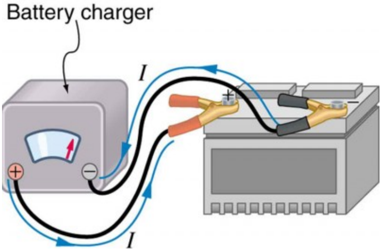


Figure 7. A car battery charger reverses the normal direction of current through a battery, reversing its chemical reaction and replenishing its chemical potential.

Multiple Voltage Sources

There are two voltage sources when a battery charger is used. Voltage sources connected in series are relatively simple. When voltage sources are in series, their internal resistances add and their emfs add algebraically. (See Figure 8.) Series connections of voltage sources are common—for example, in flashlights, toys, and other appliances. Usually, the cells are in series in order to produce a larger total emf. But if the cells oppose one another, such as when one is put into an appliance backward, the total emf is less, since it is the algebraic sum of the individual emfs. A battery is a multiple connection of voltaic cells, as shown in Figure 9. The disadvantage of series connections of cells is that their internal resistances add. One of the authors once owned a 1957 MGA that had two 6-V batteries in series, rather than a single 12-V battery. This arrangement produced a large internal resistance that caused him many problems in starting the engine.

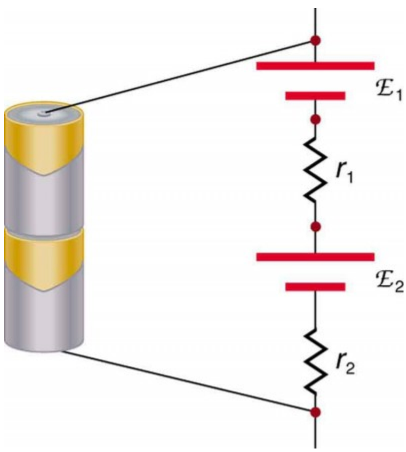


Figure 8. A series connection of two voltage sources. The emfs (each labeled with a script E) and internal resistances add, giving a total emf of emf$_1$ + emf$_2$ and a total internal resistance of r_1+r_2.

Figure 9. Batteries are multiple connections of individual cells, as shown in this modern rendition of an old print. Single cells, such as AA or C cells, are commonly called batteries, although this is technically incorrect.

If the *series* connection of two voltage sources is made into a complete circuit with the emfs in opposition, then a current of magnitude $I = \dfrac{(\text{emf}_1 - \text{emf}_2)}{r_1 + r_2}$ flows. See Figure 10, for example, which shows a circuit exactly analogous to the battery charger discussed above. If two voltage sources in series with emfs in the same sense are connected to a load R_{load}, as in Figure 11, then the following flows:

$$I = \frac{(\text{emf}_1 + \text{emf}_2)}{r_1 + r_2 + R_{\text{load}}}$$

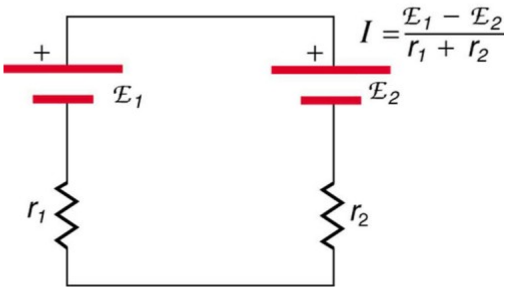


Figure 10. These two voltage sources are connected in series with their emfs in opposition. Current flows in the direction of the greater emf and is limited to [latex]I=\frac{\left({\text{emf}}_{1}-{\text{emf}}_{2}\right)}{{r}_{1}+{r}_{2}}\\[/latex] by the sum of the internal resistances. (Note that each emf is represented by script E in the figure.) A battery charger connected to a battery is an example of such a connection. The charger must have a larger emf than the battery to reverse current through it.

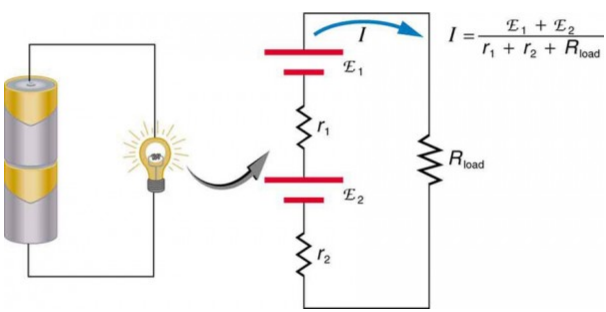


Figure 11. This schematic represents a flashlight with two cells (voltage sources) and a single bulb (load resistance) in series. The current that flows is [latex]I=\frac{\left({\text{emf}}_{1}+{\text{emf}}_{2}\right)}{{r}_{1}+{r}_{2}+{R}_{\text{load}}}\\[/latex]. (Note that each emf is represented by script E in the figure.)

Take-Home Experiment: Flashlight Batteries

Find a flashlight that uses several batteries and find new and old batteries. Based on the discussions in this module, predict the brightness of the flashlight when different combinations of batteries are used. Do your predictions match what you observe? Now place new batteries in the flashlight and leave the flashlight switched on for several hours. Is the flashlight still quite bright? Do the same with the old batteries. Is the flashlight as bright when left on for the same length of time with old and new batteries? What does this say for the case when you are limited in the number of available new batteries?

Figure 12 shows two voltage sources with identical emfs in parallel and connected to a load resistance. In this simple case, the total emf is the same as the individual emfs. But the total internal resistance is reduced, since the internal resistances are in parallel. The parallel connection thus can produce a larger current. Here, $I = \dfrac{\text{emf}}{(r_{\text{tot}} + R_{\text{load}})}$ flows through the load, and r_{tot} is less than those of the individual batteries. For example, some diesel-powered cars use two 12-V batteries in parallel; they produce a total emf of 12 V but can deliver the larger current needed to start a diesel engine.

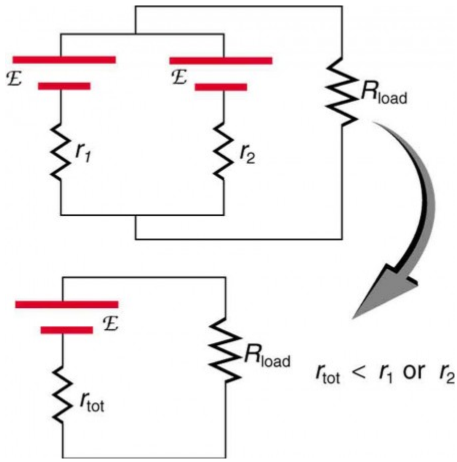


Figure 12. Two voltage sources with identical emfs (each labeled by script E) connected in parallel produce the same emf but have a smaller total internal resistance than the individual sources. Parallel combinations are often used to deliver more current. Here [latex]I=\frac{\text{emf}}{\left({r}_{\text{tot }}+{R}_{\text{load}}\right)}\\[/latex] flows through the load.

Animals as Electrical Detectors

A number of animals both produce and detect electrical signals. Fish, sharks, platypuses, and echidnas (spiny anteaters) all detect electric fields generated by nerve activity in prey. Electric eels produce their own emf through biological cells (electric organs) called electroplaques, which are arranged in both series and parallel as a set of batteries. Electroplaques are flat, disk-like cells; those of the electric eel have a voltage of 0.15 V across each one. These cells are usually located toward the head or tail of the animal, although in the case of the electric eel, they are found along the entire body. The electroplaques in the South American eel are arranged in 140 rows, with each row stretching horizontally along the body and containing

5,000 electroplaques. This can yield an emf of approximately 600 V, and a current of 1 A—deadly. The mechanism for detection of external electric fields is similar to that for producing nerve signals in the cell through depolarization and repolarization—the movement of ions across the cell membrane. Within the fish, weak electric fields in the water produce a current in a gel-filled canal that runs from the skin to sensing cells, producing a nerve signal. The Australian platypus, one of the very few mammals that lay eggs, can detect fields of $30\frac{\text{mV}}{\text{m}}$, while sharks have been found to be able to sense a field in their snouts as small as $100\frac{\text{mV}}{\text{m}}$ (Figure 13). Electric eels use their own electric fields produced by the electroplaques to stun their prey or enemies.

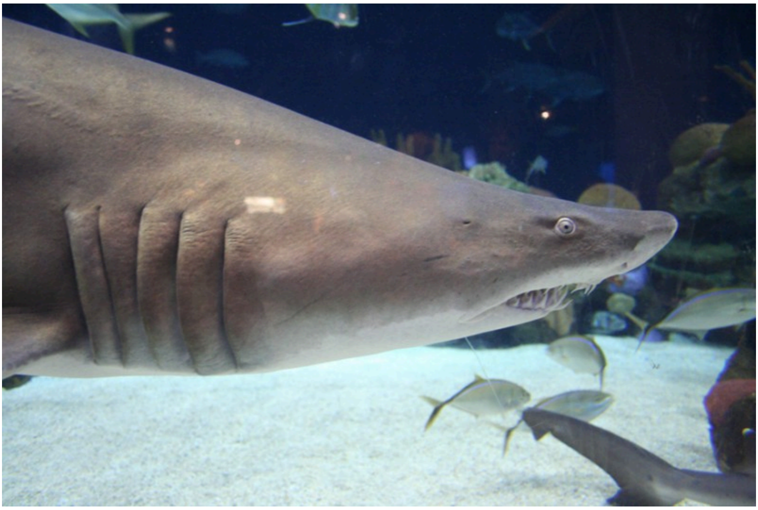


Figure 13. Sand tiger sharks (Carcharias taurus), like this one at the Minnesota Zoo, use electroreceptors in their snouts to locate prey. (credit: Jim Winstead, Flickr)

Solar Cell Arrays

Another example dealing with multiple voltage sources is that of combinations of solar cells—wired in both series and parallel combinations to yield a desired voltage and current. Photovoltaic generation (PV), the conversion of sunlight directly into electricity, is based upon the photoelectric effect, in which photons hitting the surface of a solar cell create an electric current in the cell. Most solar cells are made from pure silicon—either as single-crystal silicon, or as a thin film of silicon deposited upon a glass or metal backing. Most single cells have a voltage output of about 0.5 V, while the current output is a function of the amount of sunlight upon the cell (the incident solar radiation—the insolation). Under bright noon sunlight, a current of about $100\ \text{mA/cm}^2$ of cell surface area is produced by typical single-crystal cells. Individual solar cells are connected electrically in modules to meet electrical-energy needs. They can be wired together in series or in parallel—connected like the batteries discussed earlier. A solar-cell array or module usually consists of between 36 and 72 cells, with a power output of 50 W to 140 W. The output of the solar cells is direct current. For most uses in a home, AC is required, so a device called an inverter must be used to convert the DC to AC. Any extra output can then be passed on to the outside electrical grid for sale to the utility.

Take-Home Experiment: Virtual Solar Cells

One can assemble a “virtual” solar cell array by using playing cards, or business or index cards, to represent a solar cell. Combinations of these cards in series and/or parallel can model the required array output. Assume each card has an output of 0.5 V and a current (under bright light) of 2 A. Using your cards, how would you arrange them to produce an output of 6 A at 3 V (18 W)? Suppose you were told that you needed only 18 W (but no required voltage). Would you need more cards to make this arrangement?

Section Summary

- All voltage sources have two fundamental parts—a source of electrical energy that has a characteristic electromotive force (emf), and an internal resistance r.
- The emf is the potential difference of a source when no current is flowing.
- The numerical value of the emf depends on the source of potential difference.
- The internal resistance r of a voltage source affects the output voltage when a current flows.
- The voltage output of a device is called its terminal voltage V and is given by V = emf − Ir, where I is the electric current and is positive when flowing away from the positive terminal of the voltage source.
- When multiple voltage sources are in series, their internal resistances add and their emfs add algebraically.
- Solar cells can be wired in series or parallel to provide increased voltage or current, respectively.

Conceptual Questions

1. Is every emf a potential difference? Is every potential difference an emf? Explain.

2. Explain which battery is doing the charging and which is being charged in Figure 14.

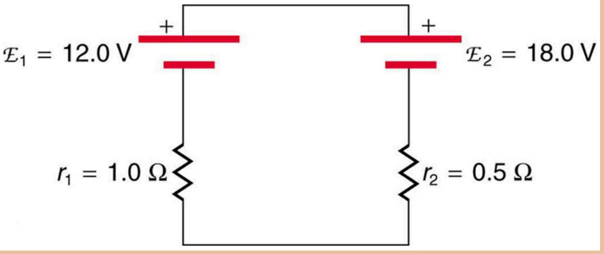


Figure 14.

3. Given a battery, an assortment of resistors, and a variety of voltage and current measuring devices, describe how you would determine the internal resistance of the battery.

4. Two different 12-V automobile batteries on a store shelf are rated at 600 and 850 "cold cranking amps." Which has the smallest internal resistance?

5. What are the advantages and disadvantages of connecting batteries in series? In parallel?

6. Semitractor trucks use four large 12-V batteries. The starter system requires 24 V, while normal operation of the

truck's other electrical components utilizes 12 V. How could the four batteries be connected to produce 24 V? To produce 12 V? Why is 24 V better than 12 V for starting the truck's engine (a very heavy load)?

Problems & Exercises

1. Standard automobile batteries have six lead-acid cells in series, creating a total emf of 12.0 V. What is the emf of an individual lead-acid cell?

2. Carbon-zinc dry cells (sometimes referred to as non-alkaline cells) have an emf of 1.54 V, and they are produced as single cells or in various combinations to form other voltages. (a) How many 1.54-V cells are needed to make the common 9-V battery used in many small electronic devices? (b) What is the actual emf of the approximately 9-V battery? (c) Discuss how internal resistance in the series connection of cells will affect the terminal voltage of this approximately 9-V battery.

3. What is the output voltage of a 3.0000-V lithium cell in a digital wristwatch that draws 0.300 mA, if the cell's internal resistance is 2.00 Ω?

4. (a) What is the terminal voltage of a large 1.54-V carbon-zinc dry cell used in a physics lab to supply 2.00 A to a circuit, if the cell's internal resistance is 0.100 Ω? (b)

How much electrical power does the cell produce? (c) What power goes to its load?

5. What is the internal resistance of an automobile battery that has an emf of 12.0 V and a terminal voltage of 15.0 V while a current of 8.00 A is charging it?

6. (a) Find the terminal voltage of a 12.0-V motorcycle battery having a 0.600-Ω internal resistance, if it is being charged by a current of 10.0 A. (b) What is the output voltage of the battery charger?

7. A car battery with a 12-V emf and an internal resistance of 0.050 Ω is being charged with a current of 60 A. Note that in this process the battery is being charged. (a) What is the potential difference across its terminals? (b) At what rate is thermal energy being dissipated in the battery? (c) At what rate is electric energy being converted to chemical energy? (d) What are the answers to (a) and (b) when the battery is used to supply 60 A to the starter motor?

8. The hot resistance of a flashlight bulb is 2.30 Ω, and it is run by a 1.58-V alkaline cell having a 0.100-Ω internal resistance. (a) What current flows? (b) Calculate the power supplied to the bulb using I^2R_{bulb}. (c) Is this power the same as calculated using $\frac{V^2}{R_{\text{bulb}}}$?

9. The label on a portable radio recommends the use of rechargeable nickel-cadmium cells (nicads), although they have a 1.25-V emf while alkaline cells have a 1.58-V emf. The radio has a 3.20-Ω resistance. (a) Draw a circuit diagram of the radio and its batteries. Now, calculate the power

delivered to the radio. (b) When using Nicad cells each having an internal resistance of 0.0400 Ω. (c) When using alkaline cells each having an internal resistance of 0.200 Ω. (d) Does this difference seem significant, considering that the radio's effective resistance is lowered when its volume is turned up?

10. An automobile starter motor has an equivalent resistance of 0.0500 Ω and is supplied by a 12.0-V battery with a 0.0100-Ω internal resistance. (a) What is the current to the motor? (b) What voltage is applied to it? (c) What power is supplied to the motor? (d) Repeat these calculations for when the battery connections are corroded and add 0.0900 Ω to the circuit. (Significant problems are caused by even small amounts of unwanted resistance in low-voltage, high-current applications.)

11. A child's electronic toy is supplied by three 1.58-V alkaline cells having internal resistances of 0.0200 Ω in series with a 1.53-V carbon-zinc dry cell having a 0.100-Ω internal resistance. The load resistance is 10.0 Ω. (a) Draw a circuit diagram of the toy and its batteries. (b) What current flows? (c) How much power is supplied to the load? (d) What is the internal resistance of the dry cell if it goes bad, resulting in only 0.500 W being supplied to the load?

12. (a) What is the internal resistance of a voltage source if its terminal voltage drops by 2.00 V when the current supplied increases by 5.00 A? (b) Can the emf of the voltage source be found with the information supplied?

13. A person with body resistance between his hands of 10.0 kΩ accidentally grasps the terminals of a 20.0-kV

power supply. (Do NOT do this!) (a) Draw a circuit diagram to represent the situation. (b) If the internal resistance of the power supply is 2000 Ω, what is the current through his body? (c) What is the power dissipated in his body? (d) If the power supply is to be made safe by increasing its internal resistance, what should the internal resistance be for the maximum current in this situation to be 1.00 mA or less? (e) Will this modification compromise the effectiveness of the power supply for driving low-resistance devices? Explain your reasoning.

14. Electric fish generate current with biological cells called electroplaques, which are physiological emf devices. The electroplaques in the South American eel are arranged in 140 rows, each row stretching horizontally along the body and each containing 5000 electroplaques. Each electroplaque has an emf of 0.15 V and internal resistance of 0.25 Ω. If the water surrounding the fish has resistance of 800 Ω, how much current can the eel produce in water from near its head to near its tail?

15. **Integrated Concepts** A 12.0-V emf automobile battery has a terminal voltage of 16.0 V when being charged by a current of 10.0 A. (a) What is the battery's internal resistance? (b) What power is dissipated inside the battery? (c) At what rate (in ºC/min) will its temperature increase if its mass is 20.0 kg and it has a specific heat of 0.300kcal/kg · ºC assuming no heat escapes?

16. **Unreasonable Results** A 1.58-V alkaline cell with a 0.200-Ω internal resistance is supplying 8.50 A to a load. (a) What is its terminal voltage? (b) What is the value of the load resistance? (c) What is unreasonable about these

results? (d) Which assumptions are unreasonable or inconsistent?

17. **Unreasonable Results** (a) What is the internal resistance of a 1.54-V dry cell that supplies 1.00 W of power to a 15.0-Ω bulb? (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

Glossary

electromotive force (emf):
the potential difference of a source of electricity when no current is flowing; measured in volts

internal resistance:
the amount of resistance within the voltage source

potential difference:
the difference in electric potential between two points in an electric circuit, measured in volts

terminal voltage:
the voltage measured across the terminals of a source of potential difference

Selected Solutions to Problems & Exercises

1. 2.00 V

3. 2.9994 V

5. 0.375 Ω

8. (a) 0.658 A (b) 0.997 W (c) 0.997 W; yes

10. (a) 200 A (b) 10.0 V (c) 2.00 kW (d) 0.1000 Ω;80.0 A, 4.0 V, 320 W

12. (a) 0.400 Ω (b) No, there is only one independent equation, so only *r* can be found.

16. (a) −0.120 V (b) -1.41×10^{-2} Ω (c) Negative terminal voltage; negative load resistance. (d) The assumption that such a cell could provide 8.50 A is inconsistent with its internal resistance.

32. Kirchhoff's Rules

Learning Objectives

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

- Analyze a complex circuit using Kirchhoff's rules, using the conventions for determining the correct signs of various terms.

Many complex circuits, such as the one in Figure 1, cannot be analyzed with the series-parallel techniques developed in Resistors in Series and Parallel and Electromotive Force: Terminal Voltage. There are, however, two circuit analysis rules that can be used to analyze any circuit, simple or complex. These rules are special cases of the laws of conservation of charge and conservation of energy. The rules are known as *Kirchhoff's rules*, after their inventor Gustav Kirchhoff (1824–1887).

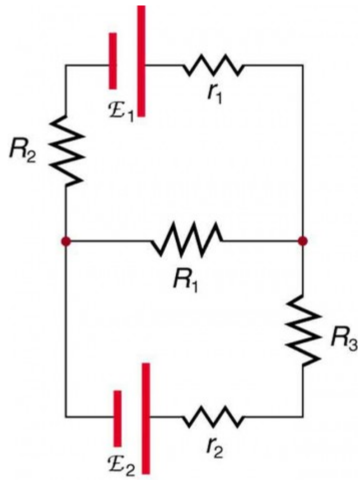


Figure 1. This circuit cannot be reduced to a combination of series and parallel connections. Kirchhoff's rules, special applications of the laws of conservation of charge and energy, can be used to analyze it. (Note: The script E in the figure represents electromotive force, emf.)

Kirchhoff's Rules

- Kirchhoff's first rule—the junction rule. The sum of all currents entering a junction must equal the sum of all currents leaving the junction.
- Kirchhoff's second rule—the loop rule. The

algebraic sum of changes in potential around any closed circuit path (loop) must be zero.

Explanations of the two rules will now be given, followed by problem-solving hints for applying Kirchhoff's rules, and a worked example that uses them.

Kirchhoff's First Rule

Kirchhoff's first rule (the *junction rule*) is an application of the conservation of charge to a junction; it is illustrated in Figure 2. Current is the flow of charge, and charge is conserved; thus, whatever charge flows into the junction must flow out. Kirchhoff's first rule requires that $I_1 = I_2 + I_3$ (see figure). Equations like this can and will be used to analyze circuits and to solve circuit problems.

Making Connections: Conservation Laws

Kirchhoff's rules for circuit analysis are applications of *conservation laws* to circuits. The first rule is the application of conservation of charge, while the second rule is the application of conservation of energy. Conservation laws, even used in a specific application, such as circuit analysis, are so basic as to form the foundation of that application.

Kirchhoff's Second Rule

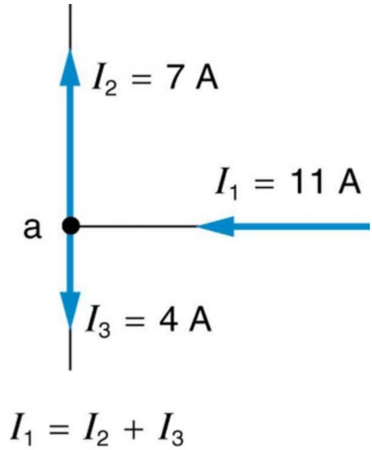


Figure 22. The junction rule. The diagram shows an example of Kirchhoff's first rule where the sum of the currents into a junction equals the sum of the currents out of a junction. In this case, the current going into the junction splits and comes out as two currents, so that $I_1 = I_2 + I_3$*. Here* I_1 *must be* 11 A*, since* I_2 *is* 7 A *and* I_3 *is* 4 A.

Kirchhoff's second rule (the *loop rule*) is an application of conservation of energy. The loop rule is stated in terms of potential, V, rather than potential energy, but the two are related since $PE_{elec} = qV$. Recall that *emf* is the potential difference of a source when no current is flowing. In a closed loop, whatever energy is supplied by emf must be transferred into other forms by devices in the loop, since there are no other ways in which energy can be transferred into or out of the circuit. Figure 3 illustrates the changes in potential in a simple series circuit loop. Kirchhoff's second rule requires emf − Ir − IR_1 − IR_2 = 0. Rearranged, this is emf = $Ir + IR_1 + IR_2$ = 0, which means the emf equals the sum of the *IR* (voltage) drops in the loop.

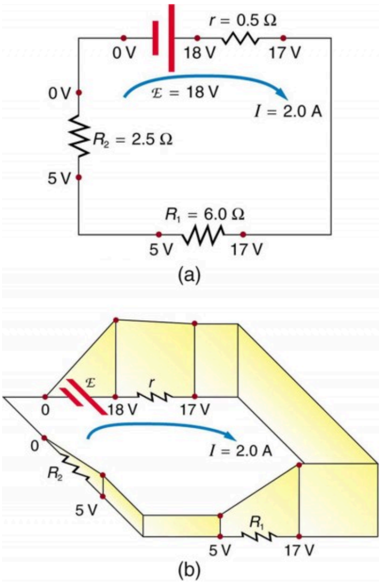


Figure 3. The loop rule. An example of Kirchhoff's second rule where the sum of the changes in potential around a closed loop must be zero. (a) In this standard schematic of a simple series circuit, the emf supplies 18 V, which is reduced to zero by the resistances, with 1 V across the internal resistance, and 12 V and 5 V across the two load resistances, for a total of 18 V. (b) This perspective view represents the potential as something like a roller coaster, where charge is raised in potential by the emf and lowered by the resistances. (Note that the script E stands for emf.)

Applying Kirchhoff's Rules

By applying Kirchhoff's rules, we generate equations that allow us to find the unknowns in circuits. The unknowns may be currents, emfs, or resistances. Each time a rule is applied, an equation is produced. If there are as many independent equations as unknowns, then the problem can be solved. There are two decisions you must make when applying Kirchhoff's rules. These decisions determine the signs of various quantities in the equations you obtain from applying the rules.

1. When applying Kirchhoff's first rule, the junction rule, you must label the current in each branch and decide in what direction it is going. For example, in Figure 1, Figure 2, and Figure 3, currents are labeled I_1, I_2, I_3, and I, and arrows indicate their directions. There is no risk here, for if you choose the wrong direction, the current will be of the correct magnitude but negative.
2. When applying Kirchhoff's second rule, the loop rule, you must identify a closed loop and decide in which direction to go around it, clockwise or counterclockwise. For example, in

Figure 3 the loop was traversed in the same direction as the current (clockwise). Again, there is no risk; going around the circuit in the opposite direction reverses the sign of every term in the equation, which is like multiplying both sides of the equation by −1.

Figure 4 and the following points will help you get the plus or minus signs right when applying the loop rule. Note that the resistors and emfs are traversed by going from a to b. In many circuits, it will be necessary to construct more than one loop. In traversing each loop, one needs to be consistent for the sign of the change in potential. (See Example 4.)

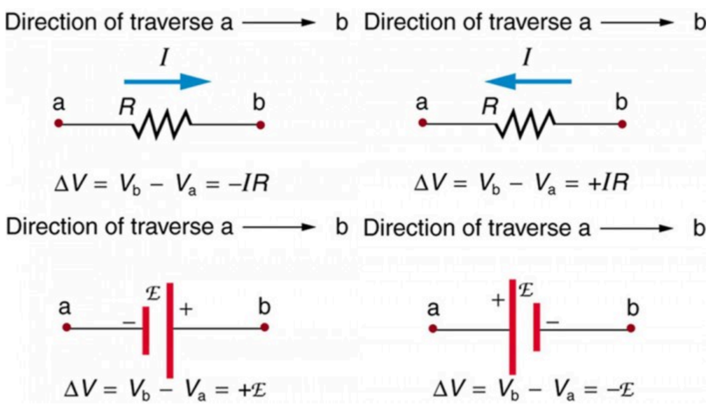


Figure 4. Each of these resistors and voltage sources is traversed from a to b. The potential changes are shown beneath each element and are explained in the text. (Note that the script E stands for emf.)

- When a resistor is traversed in the same direction as the current, the change in potential is $-IR$. (See Figure 4.)
- When a resistor is traversed in the direction opposite to the current, the change in potential is $+IR$. (See Figure 4.)
- When an emf is traversed from − to + (the same direction it moves positive charge), the change in potential is +emf. (See Figure 4.)

- When an emf is traversed from + to − (opposite to the direction it moves positive charge), the change in potential is − emf. (See Figure 4.)

Example 1. Calculating Current: Using Kirchhoff's Rules

Find the currents flowing in the circuit in Figure 5.

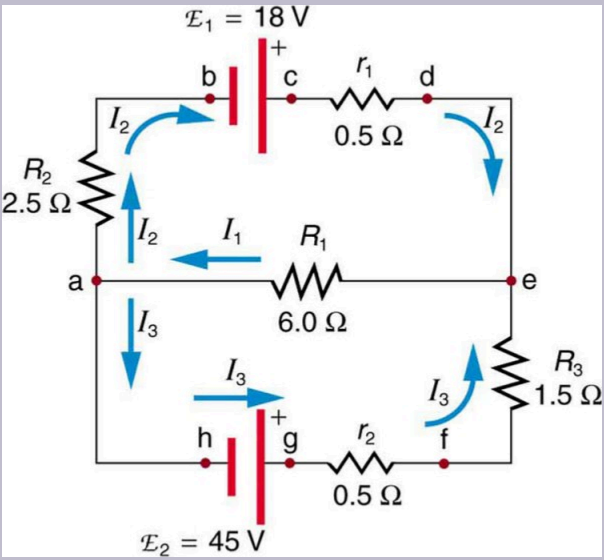


Figure 5. This circuit is similar to that in Figure 1, but the resistances and emfs are specified. (Each emf is denoted by script E.) The currents in each branch are labeled and assumed to move in the directions shown. This example uses Kirchhoff's rules to find the currents.

Strategy

This circuit is sufficiently complex that the currents cannot be found using Ohm's law and the series-parallel techniques—it is necessary to use Kirchhoff's rules. Currents have been labeled I_1, I_2, and I_3 in the figure and assumptions have been made about their directions. Locations on the diagram have been labeled with letters a through h. In the solution we will apply the junction and loop rules, seeking three independent equations to allow us to solve for the three unknown currents.

Solution

We begin by applying Kirchhoff's first or junction rule at point a. This gives

$$I_1 = I_2 + I_3,$$

since I_1 flows into the junction, while I_2 and I_3 flow out. Applying the junction rule at e produces exactly the same equation, so that no new information is obtained. This is a single equation with three unknowns—three independent equations are needed, and so the loop rule must be applied. Now we consider the loop abcdea. Going from a to b, we traverse R_2 in the same (assumed) direction of the current I_2, and so the change in potential is $-I_2R_2$. Then going from b to c, we go from − to +, so that the change in potential is $+\text{emf}_1$. Traversing the internal resistance r_1 from c to d gives $-I_2r_1$. Completing the loop by going from d to a again traverses a resistor in the same direction as its current, giving a change in potential of $-I_1R_1$. The loop rule states that the changes in potential sum to zero. Thus,

$$-I_2R_2 + \text{emf}_1 - I_2r_1 - I_1R_1 = -I_2(R_2 + r_1) + \text{emf}_1 - I_1R_1 = 0.$$

Substituting values from the circuit diagram for the resistances and emf, and canceling the ampere unit gives

$$-3I_2 + 18 - 6I_1 = 0.$$

Now applying the loop rule to aefgha (we could have chosen abcdefgha as well) similarly gives

$$+ I_1R_1 + I_3R_3 + I_3r_2 - \text{emf}_2 = + I_1R_1 + I_3(R_3 + r_2) - \text{emf}_2 = 0.$$

Note that the signs are reversed compared with the other loop, because elements are traversed in the opposite direction. With values entered, this becomes

$$+6I_1 + 2I_3 - 45 = 0.$$

These three equations are sufficient to solve for the three unknown currents. First, solve the second equation for I_2:

$$I_2 = 6 - 2I_1.$$

Now solve the third equation for I_3:

$$I_3 = 22.5 - 3I_1.$$

Substituting these two new equations into the first one allows us to find a value for I_1:

$$I_1 = I_2 + I_3 = (6-2I_1) + (22.5- 3I_1) = 28.5 - 5I_1.$$

Combining terms gives

$$6I_1 = 28.5, \text{ and}$$
$$I_1 = 4.75 \text{ A}.$$

Substituting this value for I_1 back into the fourth equation gives

$$I_2 = 6 - 2I_1 = 6 - 9.50$$
$$I_2 = -3.50 \text{ A}.$$

The minus sign means I_2 flows in the direction opposite to that assumed in Figure 5. Finally, substituting the value for I_1 into the fifth equation gives

$$I_3 = 22.5 - 3I_1 = 22.5 - 14.25$$
$$I_3 = 8.25 \text{ A}.$$

Discussion

Just as a check, we note that indeed $I_1 = I_2 + I_3$. The results could also have been checked by entering all of the values into the equation for the abcdefgha loop.

Problem-Solving Strategies for Kirchhoff's Rules

1. Make certain there is a clear circuit diagram on which you can label all known and unknown resistances, emfs, and currents. If a current is unknown, you must assign it a direction. This is necessary for determining the signs of potential changes. If you assign the direction incorrectly, the current will be found to have a negative value—no harm done.
2. Apply the junction rule to any junction in the

circuit. Each time the junction rule is applied, you should get an equation with a current that does not appear in a previous application—if not, then the equation is redundant.

3. Apply the loop rule to as many loops as needed to solve for the unknowns in the problem. (There must be as many independent equations as unknowns.) To apply the loop rule, you must choose a direction to go around the loop. Then carefully and consistently determine the signs of the potential changes for each element using the four bulleted points discussed above in conjunction with Figure 4.
4. Solve the simultaneous equations for the unknowns. This may involve many algebraic steps, requiring careful checking and rechecking.
5. Check to see whether the answers are reasonable and consistent. The numbers should be of the correct order of magnitude, neither exceedingly large nor vanishingly small. The signs should be reasonable—for example, no resistance should be negative. Check to see that the values obtained satisfy the various equations obtained from applying the rules. The currents should satisfy the junction rule, for example.

The material in this section is correct in theory. We should be able to verify it by making measurements of current and voltage. In fact, some of the devices used to make such measurements are straightforward applications of the principles covered so far and are explored in the next modules. As we shall see, a very basic, even profound, fact results—making a measurement alters the quantity being measured.

Check Your Understanding

Can Kirchhoff's rules be applied to simple series and parallel circuits or are they restricted for use in more complicated circuits that are not combinations of series and parallel?

Solution

Kirchhoff's rules can be applied to any circuit since they are applications to circuits of two conservation laws. Conservation laws are the most broadly applicable principles in physics. It is usually mathematically simpler to use the rules for series and parallel in simpler circuits so we emphasize Kirchhoff's rules for use in more complicated situations. But the rules for series and parallel can be derived from Kirchhoff's rules. Moreover, Kirchhoff's rules can be expanded to devices other than resistors and emfs, such as capacitors, and are one of the basic analysis devices in circuit analysis.

Section Summary

- Kirchhoff's rules can be used to analyze any circuit, simple or complex.

- Kirchhoff's first rule—the junction rule: The sum of all currents entering a junction must equal the sum of all currents leaving the junction.
- Kirchhoff's second rule—the loop rule: The algebraic sum of changes in potential around any closed circuit path (loop) must be zero.
- The two rules are based, respectively, on the laws of conservation of charge and energy.
- When calculating potential and current using Kirchhoff's rules, a set of conventions must be followed for determining the correct signs of various terms.
- The simpler series and parallel rules are special cases of Kirchhoff's rules.

Conceptual Questions

1. Can all of the currents going into the junction in Figure 6 be positive? Explain.

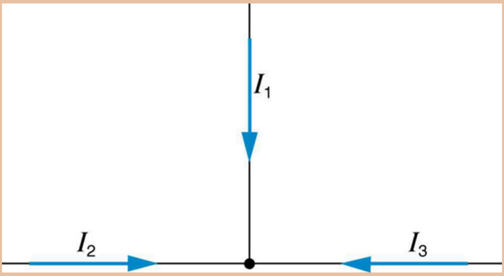


Figure 6.

2. Apply the junction rule to junction b in Figure 7. Is any

new information gained by applying the junction rule at e? (In the figure, each emf is represented by script E.)

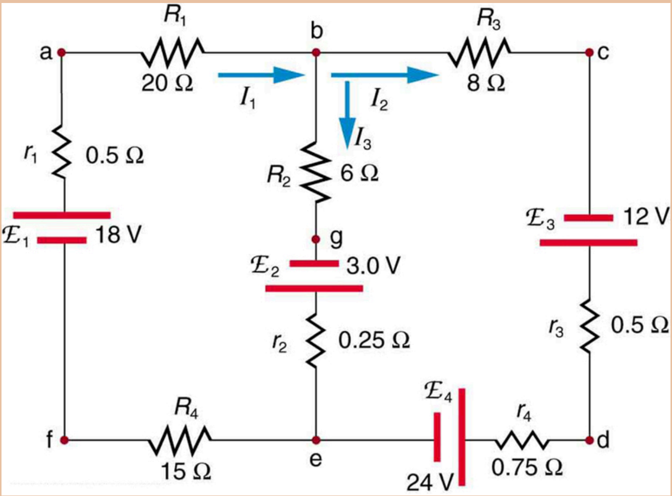


Figure 7.

3. (a) What is the potential difference going from point a to point b in Figure 7? (b) What is the potential difference going from c to b? (c) From e to g? (d) From e to d?

4. Apply the loop rule to loop afedcba in Figure 7.

5. Apply the loop rule to loops abgefa and cbgedc in Figure 7.

Problems & Exercises

1. Apply the loop rule to loop abcdefgha in Figure 5 (shown again below).

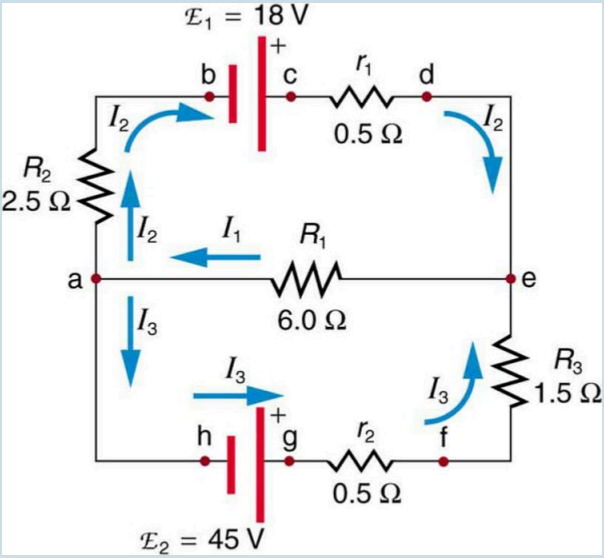


Figure 5. This circuit is similar to that in Figure 1, but the resistances and emfs are specified. (Each emf is denoted by script E.) The currents in each branch are labeled and assumed to move in the directions shown. This example uses Kirchhoff's rules to find the currents.

2. Apply the loop rule to loop aedcba in Figure 5.

3. Verify the second equation in *Example 1 Calculating Current: Using Kirchhoff's Rules* (in the text above) by substituting the values found for the currents I_1 and I_2.

4. Verify the third equation in *Example 1 Calculating Current: Using Kirchhoff's Rules* (in the text above) by substituting the values found for the currents I_1 and I_3.

5. Apply the junction rule at point a in Figure 8.

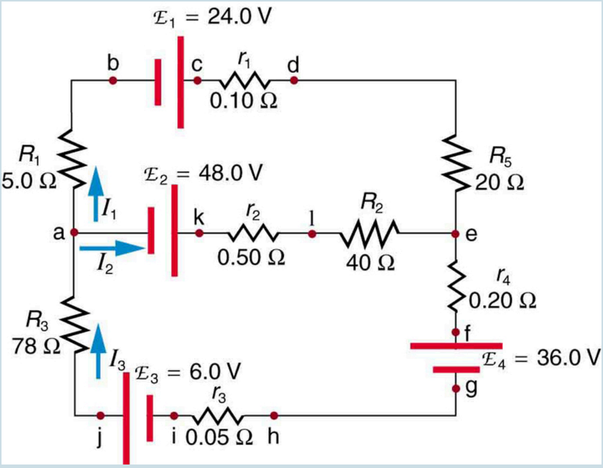


Figure 8.

6. Apply the loop rule to loop abcdefghija in Figure 8.

7. Apply the loop rule to loop akledcba in Figure 8.

8. Find the currents flowing in the circuit in Figure 8. Explicitly show how you follow the steps in the *Problem-Solving Strategies for Series and Parallel Resistors* above.

9. Solve *Example 1 Calculating Current: Using Kirchhoff's Rules* (in the text above), but use loop abcdefgha instead of loop akledcba. Explicitly show how you follow the steps in the *Problem-Solving Strategies for Series and Parallel Resistors.*

10. Find the currents flowing in the circuit in Figure 7 (shown again below).

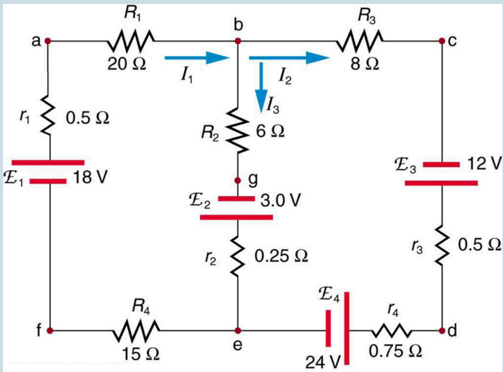


Figure 7.

11. **Unreasonable Results** Consider the circuit in Figure 9, and suppose that the emfs are unknown and the currents are given to be $I_1 = 5.00$ A, $I_2 = 3.0$ A, and $I_3 = -2.00$ A. (a) Could you find the emfs? (b) What is wrong with the assumptions?

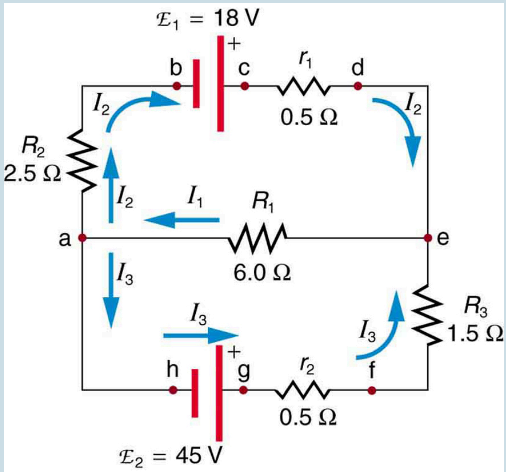


Figure 9.

Glossary

Kirchhoff's rules:
: a set of two rules, based on conservation of charge and energy, governing current and changes in potential in an electric circuit

junction rule:
: Kirchhoff's first rule, which applies the conservation of charge to a junction; current is the flow of charge; thus, whatever charge flows into the junction must flow out; the rule can be stated $I_1 = I_2 + I_3$

loop rule:
: Kirchhoff's second rule, which states that in a closed loop,

whatever energy is supplied by emf must be transferred into other forms by devices in the loop, since there are no other ways in which energy can be transferred into or out of the circuit. Thus, the emf equals the sum of the IR (voltage) drops in the loop and can be stated: $\mathrm{emf} = \mathrm{Ir} + \mathrm{IR}_1 + \mathrm{IR}_2$

conservation laws:

require that energy and charge be conserved in a system

Selected Solutions to Problems & Exercises

1.
$$-I_2R_2 + \mathrm{emf}_1 - \mathrm{I}_2r_1 + \mathrm{I}_3R_3 + \mathrm{I}_3r_2 - \mathrm{emf}_2 = 0$$

5. $I_3 = I_1 + I_2$

7.
$$\mathrm{emf}_2 - \mathrm{I}_2r_2 - \mathrm{I}_2R_2 + \mathrm{I}_1R_5 + I_1r_1 - \mathrm{emf}_1 + \mathrm{I}_1R_1 = 0$$

9.(a) $I_1 = 4.75$ A (b) $I_2 = -3.5$ A (c) $I_3 = 8.25$ A

11. (a) No, you would get inconsistent equations to solve. (b) $I_1 \neq I_2 + I_3$. The assumed currents violate the junction rule.

33. DC Voltmeters and Ammeters

Learning Objectives

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

- Explain why a voltmeter must be connected in parallel with the circuit.
- Draw a diagram showing an ammeter correctly connected in a circuit.
- Describe how a galvanometer can be used as either a voltmeter or an ammeter.
- Find the resistance that must be placed in series with a galvanometer to allow it to be used as a voltmeter with a given reading.
- Explain why measuring the voltage or current in a circuit can never be exact.

Voltmeters measure voltage, whereas *ammeters* measure current. Some of the meters in automobile dashboards, digital cameras, cell phones, and tuner-amplifiers are voltmeters or ammeters. (See Figure 1.) The internal construction of the simplest of these meters and how they are connected to the system they monitor give further insight into applications of series and parallel connections.

Figure 1. The fuel and temperature gauges (far right and far left, respectively) in this 1996 Volkswagen are voltmeters that register the voltage output of "sender" units, which are hopefully proportional to the amount of gasoline in the tank and the engine temperature. (credit: Christian Giersing)

Voltmeters are connected in parallel with whatever device's voltage is to be measured. A parallel connection is used because objects in parallel experience the same potential difference. (See Figure 2, where the voltmeter is represented by the symbol V.) Ammeters are connected in series with whatever device's current is to be measured. A series connection is used because objects in series have the same current passing through them. (See Figure 3, where the ammeter is represented by the symbol A.)

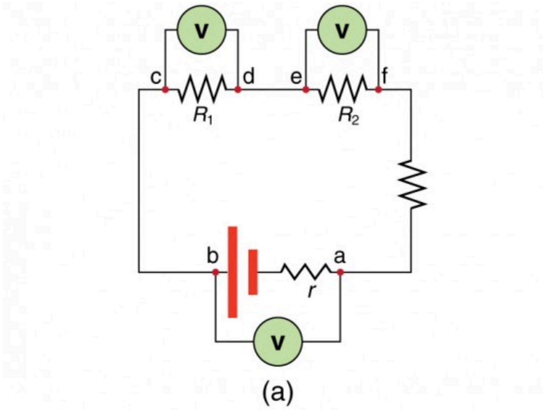


Figure 2. (a) To measure potential differences in this series circuit, the voltmeter (V) is placed in parallel with the voltage source or either of the resistors. Note that terminal voltage is measured between points a and b. It is not possible to connect the voltmeter directly across the emf without including its internal resistance, r. (b) A digital voltmeter in use. (credit: Messtechniker, Wikimedia Commons)

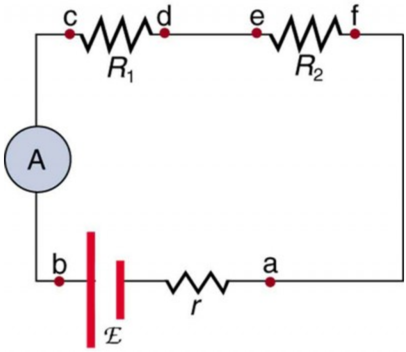


Figure 3. An ammeter (A) is placed in series to measure current. All of the current in this circuit flows through the meter. The ammeter would have the same reading if located between points d and e or between points f and a as it does in the position shown. (Note that the script capital E stands for emf, and r stands for the internal resistance of the source of potential difference.)

Analog Meters: Galvanometers

Analog meters have a needle that swivels to point at numbers on a scale, as opposed to *digital meters*, which have numerical readouts similar to a hand-held calculator. The heart of most analog meters is a device called a *galvanometer*, denoted by G. Current flow through a galvanometer, I_G, produces a proportional needle deflection. (This deflection is due to the force of a magnetic field upon a current-carrying wire.)

The two crucial characteristics of a given galvanometer are its resistance and current sensitivity. *Current sensitivity* is the current that gives a *full-scale deflection* of the galvanometer's needle, the maximum current that the instrument can measure. For example, a galvanometer with a current sensitivity of 50 μA has a maximum deflection of its needle when 50 μA flows through it, reads half-scale when 25 μA flows through it, and so on. If such a galvanometer has a 25-Ω resistance, then a voltage of only $V = IR = (50\ \mu A)(25\ \Omega) = 1.25\ mV$ produces a full-scale reading. By connecting resistors to this galvanometer in different ways, you can use it as either a voltmeter or ammeter that can measure a broad range of voltages or currents.

Galvanometer as Voltmeter

Figure 4 shows how a galvanometer can be used as a voltmeter by connecting it in series with a large resistance, R. The value of the resistance R is determined by the maximum voltage to be measured. Suppose you want 10 V to produce a full-scale deflection of a voltmeter containing a 25-Ω galvanometer with a 50-μA sensitivity. Then 10 V applied to the meter must produce a current of 50 μA. The total resistance must be

$$R_{\text{tot}} = R + r = \frac{V}{I} = \frac{10\ \text{V}}{50\ \mu\text{A}} = 200\ \text{k}\Omega \text{ or}$$
$$R = R_{\text{tot}} - r = 200\ \text{k}\Omega - 25\ \Omega \approx 200\ \text{k}\Omega.$$

(R is so large that the galvanometer resistance, r, is nearly negligible.) Note that 5 V applied to this voltmeter produces a half-scale deflection by producing a 25-μA current through the meter, and so the voltmeter's reading is proportional to voltage as desired. This voltmeter would not be useful for voltages less than about half a volt, because the meter deflection would be small and difficult to read accurately. For other voltage ranges, other resistances are placed in series with the galvanometer. Many meters have a choice of scales. That choice involves switching an appropriate resistance into series with the galvanometer.

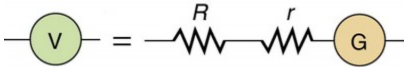


Figure 4. A large resistance R placed in series with a galvanometer G produces a voltmeter, the full-scale deflection of which depends on the choice of R. The larger the voltage to be measured, the larger R must be. (Note that r represents the internal resistance of the galvanometer.)

Galvanometer as Ammeter

The same galvanometer can also be made into an ammeter by placing it in parallel with a small resistance *R*, often called the *shunt resistance*, as shown in Figure 5. Since the shunt resistance is small, most of the current passes through it, allowing an ammeter to measure currents much greater than those producing a full-scale deflection of the galvanometer. Suppose, for example, an ammeter is needed that gives a full-scale deflection for 1.0 A, and contains the same 25-Ω galvanometer with its 50-μA sensitivity. Since *R* and *r* are in parallel, the voltage across them is the same. These *IR* drops are $IR = I_G r$ so that $IR = \frac{I_{\text{G}}}{I} = \frac{R}{r}$. Solving for R, and noting that I_G is 50 μA and I is 0.999950 A, we have

$$R = r\frac{I_{\text{G}}}{I} = (25\ \Omega)\frac{50\ \mu\text{A}}{0.999950\ \text{A}} = 1.25 \times 10^{-3}\ \Omega$$

.

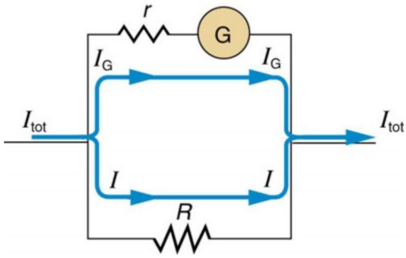


Figure 5. A small shunt resistance R placed in parallel with a galvanometer G produces an ammeter, the full-scale deflection of which depends on the choice of R. The larger the current to be measured, the smaller R must be. Most of the current (I) flowing through the meter is shunted through R to protect the galvanometer. (Note that r represents the internal resistance of the galvanometer.) Ammeters may also have multiple scales for greater flexibility in application. The various scales are achieved by switching various shunt resistances in parallel with the galvanometer—the greater the maximum current to be measured, the smaller the shunt resistance must be.

Taking Measurements Alters the Circuit

When you use a voltmeter or ammeter, you are connecting another resistor to an existing circuit and, thus, altering the circuit. Ideally, voltmeters and ammeters do not appreciably affect the circuit, but it is instructive to examine the circumstances under which they do or do not interfere. First, consider the voltmeter, which is always placed in parallel with the device being measured. Very little current flows through the voltmeter if its resistance is a few orders of magnitude greater than the device, and so the circuit is not appreciably affected. (See Figure 6(a).) (A large resistance in parallel with a small one has a combined resistance essentially equal to the small one.) If, however, the voltmeter's resistance is comparable to that of the device being measured, then the two in parallel have a smaller resistance, appreciably affecting the circuit. (See Figure 6(b).) The voltage across the device is not the same as when the voltmeter is out of the circuit.

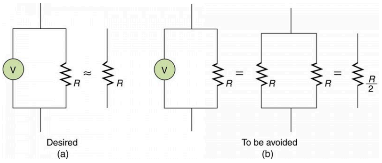


Figure 6. (a) A voltmeter having a resistance much larger than the device (RVoltmeter>>R) with which it is in parallel produces a parallel resistance essentially the same as the device and does not appreciably affect the circuit being measured. (b) Here the voltmeter has the same resistance as the device (RVoltmeter≅R), so that the parallel resistance is half of what it is when the voltmeter is not connected. This is an example of a significant alteration of the circuit and is to be avoided.

An ammeter is placed in series in the branch of the circuit being measured, so that its resistance adds to that branch. Normally, the ammeter's resistance is very small compared with the resistances of the devices in the circuit, and so the extra resistance is negligible. (See Figure 7(a).) However, if very small load resistances are involved, or if the ammeter is not as low in resistance as it should be, then the total series resistance is significantly greater, and the current in the branch being measured is reduced. (See Figure 7(b).) A practical problem can occur if the ammeter is connected incorrectly. If it was put in parallel with the resistor to measure the current in it, you could possibly damage the meter; the low resistance of the ammeter would allow most of the current in the circuit to go through the galvanometer, and this current would be larger since the effective resistance is smaller.

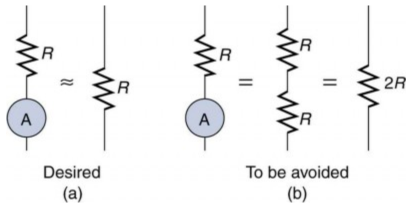


Figure 7. (a) An ammeter normally has such a small resistance that the total series resistance in the branch being measured is not appreciably increased. The circuit is essentially unaltered compared with when the ammeter is absent. (b) Here the ammeter's resistance is the same as that of the branch, so that the total resistance is doubled and the current is half what it is without the ammeter. This significant alteration of the circuit is to be avoided.

One solution to the problem of voltmeters and ammeters interfering with the circuits being measured is to use galvanometers with greater sensitivity. This allows construction of voltmeters with greater resistance and ammeters with smaller resistance than when less sensitive galvanometers are used. There are practical limits to galvanometer sensitivity, but it is possible to get analog meters that make measurements accurate to a few percent. Note that the inaccuracy comes from altering the circuit, not from a fault in the meter.

Connections: Limits to Knowledge

Making a measurement alters the system being measured in a manner that produces uncertainty in the measurement. For macroscopic systems, such as the circuits discussed in this module, the alteration can usually be made negligibly small, but it cannot be eliminated entirely. For submicroscopic systems, such as atoms, nuclei, and smaller particles, measurement alters the system in a manner that cannot be made arbitrarily small. This actually limits knowledge of the system—even limiting what nature can know about itself. We shall see profound implications of this when the Heisenberg uncertainty principle is discussed in the modules on quantum mechanics.

There is another measurement technique based on drawing no current at all and, hence, not altering the circuit at all. These are called null measurements and are the topic of Null Measurements. Digital meters that employ solid-state electronics and null measurements can attain accuracies of one part in 10^6.

Check Your Understanding

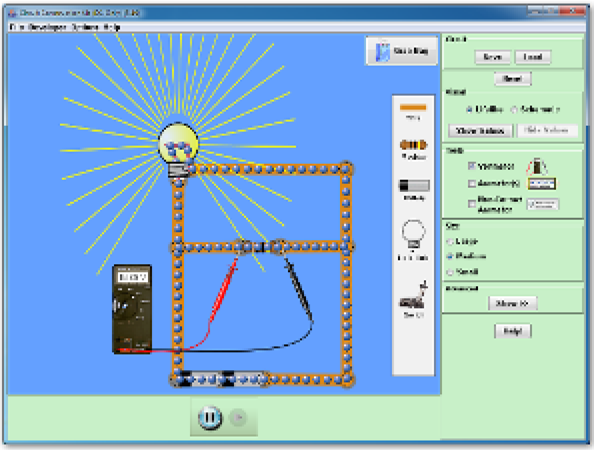
Digital meters are able to detect smaller currents than analog meters employing galvanometers. How does this explain their ability to measure voltage and current more accurately than analog meters?

Solution

Since digital meters require less current than analog meters, they alter the circuit less than analog meters. Their resistance as a voltmeter can be far greater than an analog meter, and their resistance as an ammeter can be far less than an analog meter. Consult Figure 2 and Figure 3 and their discussion in the text.

PhET Explorations: Circuit Construction Kit (DC Only), Virtual Lab

Stimulate a neuron and monitor what happens. Pause, rewind, and move forward in time in order to observe the ions as they move across the neuron membrane.



Click to download the simulation. Run using Java.

Section Summary

- Voltmeters measure voltage, and ammeters measure current.
- A voltmeter is placed in parallel with the voltage source to receive full voltage and must have a large resistance to limit its effect on the circuit.
- An ammeter is placed in series to get the full current flowing through a branch and must have a small resistance to limit its effect on the circuit.
- Both can be based on the combination of a resistor and a galvanometer, a device that gives an analog reading of current.
- Standard voltmeters and ammeters alter the circuit being

measured and are thus limited in accuracy.

Conceptual Questions

1. Why should you not connect an ammeter directly across a voltage source as shown in Figure 9? (Note that script E in the figure stands for emf.)

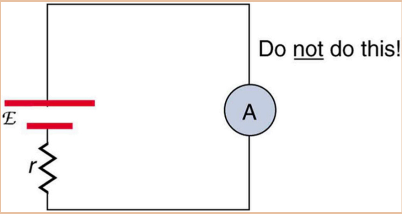


Figure 9.

2. Suppose you are using a multimeter (one designed to measure a range of voltages, currents, and resistances) to measure current in a circuit and you inadvertently leave it in a voltmeter mode. What effect will the meter have on the circuit? What would happen if you were measuring voltage but accidentally put the meter in the ammeter mode?

3. Specify the points to which you could connect a voltmeter to measure the following potential differences in Figure 10: (a) the potential difference of the voltage source; (b) the potential difference across R_1; (c) across R_2; (d) across R_3; (e) across R_2 and R_3. Note that there may be more than one answer to each part.

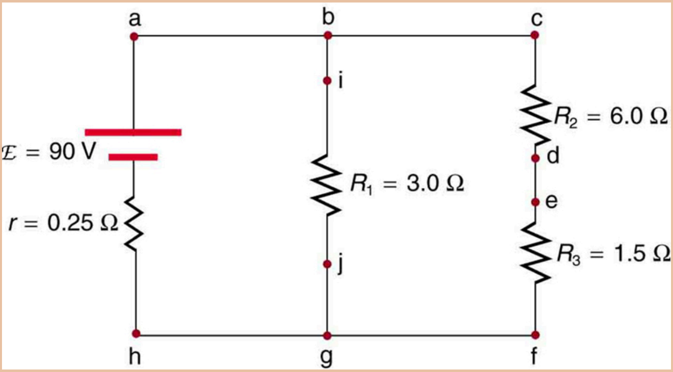


Figure 10.

4. To measure currents in Figure 10, you would replace a wire between two points with an ammeter. Specify the points between which you would place an ammeter to measure the following: (a) the total current; (b) the current flowing through R_1; (c) through R_2; (d) through R_3. Note that there may be more than one answer to each part.

Problems & Exercises

1. What is the sensitivity of the galvanometer (that is, what current gives a full-scale deflection) inside a voltmeter that has a 1.00-MΩ resistance on its 30.0-V scale?

2. What is the sensitivity of the galvanometer (that is,

what current gives a full-scale deflection) inside a voltmeter that has a 25.0-kΩ resistance on its 100-V scale?

3. Find the resistance that must be placed in series with a 25.0-Ω galvanometer having a 50.0-μA sensitivity (the same as the one discussed in the text) to allow it to be used as a voltmeter with a 0.100-V full-scale reading.

4. Find the resistance that must be placed in series with a 25.0-Ω galvanometer having a 50.0-μA sensitivity (the same as the one discussed in the text) to allow it to be used as a voltmeter with a 3000-V full-scale reading. Include a circuit diagram with your solution.

5. Find the resistance that must be placed in parallel with a 25.0-Ω galvanometer having a 50.0-μA sensitivity (the same as the one discussed in the text) to allow it to be used as an ammeter with a 10.0-A full-scale reading. Include a circuit diagram with your solution.

6. Find the resistance that must be placed in parallel with a 25.0-Ω galvanometer having a 50.0-μA sensitivity (the same as the one discussed in the text) to allow it to be used as an ammeter with a 300-mA full-scale reading.

7. Find the resistance that must be placed in series with a 10.0-Ω galvanometer having a 100-μA sensitivity to allow it to be used as a voltmeter with: (a) a 300-V full-scale reading, and (b) a 0.300-V full-scale reading.

8. Find the resistance that must be placed in parallel with a 10.0-Ω galvanometer having a 100-μA sensitivity to allow it to be used as an ammeter with: (a) a 20.0-A full-scale reading, and (b) a 100-mA full-scale reading.

9. Suppose you measure the terminal voltage of a 1.585-V alkaline cell having an internal resistance of 0.100Ω by placing a 1.00-kΩ voltmeter across its terminals. (See Figure 11.) (a) What current flows? (b) Find the terminal voltage. (c) To see how close the measured terminal voltage is to the emf, calculate their ratio.

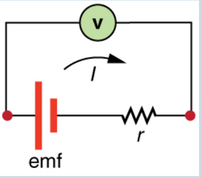


Figure 11.

10. Suppose you measure the terminal voltage of a 3.200-V lithium cell having an internal resistance of 5.00 Ω by placing a 1.00-kΩ voltmeter across its terminals. (a) What current flows? (b) Find the terminal voltage. (c) To see how close the measured terminal voltage is to the emf, calculate their ratio.

11. A certain ammeter has a resistance of $5.00 \times 10^{-5}\Omega$ on its 3.00-A scale and contains a 10.0-Ω galvanometer. What is the sensitivity of the galvanometer?

12. A 1.00-MΩ voltmeter is placed in parallel with a 75.0-kΩ resistor in a circuit. (a) Draw a circuit diagram of the connection. (b) What is the resistance of the combination? (c) If the voltage across the combination is kept the same as it was across the 75.0-kΩ resistor alone, what is the percent increase in current? (d) If the current

through the combination is kept the same as it was through the 75.0-kΩ resistor alone, what is the percentage decrease in voltage? (e) Are the changes found in parts (c) and (d) significant? Discuss.

13. A 0.0200-Ω ammeter is placed in series with a 10.00-Ω resistor in a circuit. (a) Draw a circuit diagram of the connection. (b) Calculate the resistance of the combination. (c) If the voltage is kept the same across the combination as it was through the 10.00-Ω resistor alone, what is the percent decrease in current? (d) If the current is kept the same through the combination as it was through the 10.00-Ω resistor alone, what is the percent increase in voltage? (e) Are the changes found in parts (c) and (d) significant? Discuss.

14. **Unreasonable Results** Suppose you have a40.0-Ω galvanometer with a 25.0-μA sensitivity. (a) What resistance would you put in series with it to allow it to be used as a voltmeter that has a full-scale deflection for 0.500 mV? (b) What is unreasonable about this result? (c) Which assumptions are responsible?

15. **Unreasonable Results** (a) What resistance would you put in parallel with a 40.0-Ω galvanometer having a 25.0-μA sensitivity to allow it to be used as an ammeter that has a full-scale deflection for 10.0-μA? (b) What is unreasonable about this result? (c) Which assumptions are responsible?

Glossary

voltmeter:
: an instrument that measures voltage

ammeter:
: an instrument that measures current

analog meter:
: a measuring instrument that gives a readout in the form of a needle movement over a marked gauge

digital meter:
: a measuring instrument that gives a readout in a digital form

galvanometer:
: an analog measuring device, denoted by G, that measures current flow using a needle deflection caused by a magnetic field force acting upon a current-carrying wire

current sensitivity:
: the maximum current that a galvanometer can read

full-scale deflection:
: the maximum deflection of a galvanometer needle, also known as current sensitivity; a galvanometer with a full-scale deflection of 50 μA has a maximum deflection of its needle when 50 μA flows through it

shunt resistance:
: a small resistance R placed in parallel with a galvanometer G to produce an ammeter; the larger the current to be measured, the smaller R must be; most of the current flowing through the meter is shunted through R to protect the galvanometer

Selected Solutions to Problems & Exercises

1. 30 μA

3. 1.98 kΩ

5. 1.25×10^{-4} Ω

7. (a) 3.00 MΩ (b) 2.99 kΩ

9. (a) 1.58 mA (b) 1.5848 V (need four digits to see the difference) (c) 0.99990 (need five digits to see the difference from unity)

11. 15.0 µA

12.

Figure 12.

(a)

(b) 10.02 Ω

(c) 0.9980, or a 2.0×10^{-1} percent decrease

(d) 1.002, or a 2.0×10^{-1} percent increase

(e) Not significant.

15. (a) −66.7 Ω (b) You can't have negative resistance. (c) It is unreasonable that I_G is greater than I_{tot} (see Figure 5). You cannot achieve a full-scale deflection using a current less than the sensitivity of the galvanometer.

34. Null Measurements

Learning Objectives

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

- Explain why a null measurement device is more accurate than a standard voltmeter or ammeter.
- Demonstrate how a Wheatstone bridge can be used to accurately calculate the resistance in a circuit.

Standard measurements of voltage and current alter the circuit being measured, introducing uncertainties in the measurements. Voltmeters draw some extra current, whereas ammeters reduce current flow. *Null measurements* balance voltages so that there is no current flowing through the measuring device and, therefore, no alteration of the circuit being measured. Null measurements are generally more accurate but are also more complex than the use of standard voltmeters and ammeters, and they still have limits to their precision. In this module, we shall consider a few specific types of null measurements, because they are common and interesting, and they further illuminate principles of electric circuits.

The Potentiometer

Suppose you wish to measure the emf of a battery. Consider what happens if you connect the battery directly to a standard voltmeter as shown in Figure 1. (Once we note the problems with this measurement, we will examine a null measurement that improves accuracy.) As discussed before, the actual quantity measured is the terminal voltage V, which is related to the emf of the battery by V = emf − *Ir*, where *I* is the current that flows and *r* is the internal resistance of the battery. The emf could be accurately calculated if *r* were very accurately known, but it is usually not. If the current *I* could be made zero, then V=emf, and so emf could be directly measured. However, standard voltmeters need a current to operate; thus, another technique is needed.

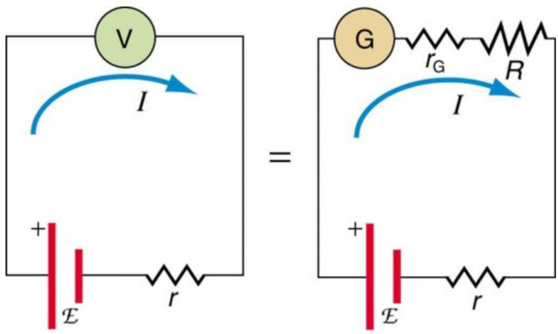


Figure 1. An analog voltmeter attached to a battery draws a small but nonzero current and measures a terminal voltage that differs from the emf of the battery. (Note that the script capital E symbolizes electromotive force, or emf.) Since the internal resistance of the battery is not known precisely, it is not possible to calculate the emf precisely.

A *potentiometer* is a null measurement device for measuring potentials (voltages). (See Figure 2.) A voltage source is connected to a resistor R, say, a long wire, and passes a constant current through it. There is a steady drop in potential (an *IR* drop) along the wire, so that a variable potential can be obtained by making contact at varying locations along the wire. Figure 2(b) shows an unknown emf_x (represented by script E_x in the figure) connected in series with a galvanometer. Note that emf_x opposes the other voltage source. The location of the contact point (see the arrow on the drawing) is adjusted until the galvanometer reads zero. When the galvanometer reads zero, $\text{emf}_x = IR_x$, where R_x is the resistance of the section of wire up to the contact point. Since no current flows through the galvanometer, none flows through the unknown emf, and so emf_x is directly sensed. Now, a very precisely known standard emf_s is substituted for emf_x, and the contact point is adjusted until the galvanometer again reads zero, so that $\text{emf}_s = IR_s$. In both cases, no current passes through the galvanometer, and so the current I through the long wire is the same. Upon taking the ratio $\frac{\text{emf}_x}{\text{emf}_s}$, I cancels, giving

$$\frac{\text{emf}_x}{\text{emf}_s} = \frac{IR_x}{IR_s} = \frac{R_x}{R_s}.$$

Solving for emf_x gives

$$\text{emf}_x = \text{emf}_s \frac{R_x}{R_s}.$$

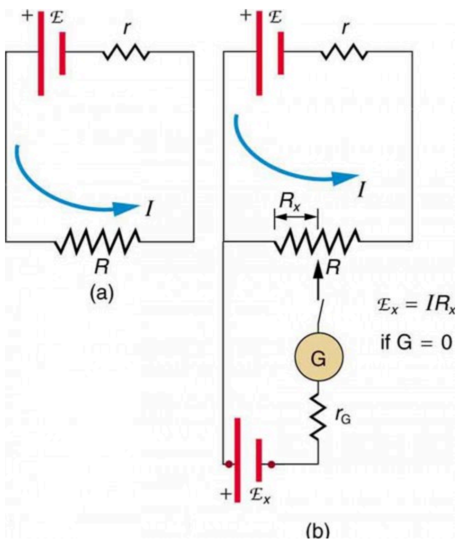


Figure 2. The potentiometer, a null measurement device. (a) A voltage source connected to a long wire resistor passes a constant current I through it. (b) An unknown emf (labeled script E_X in the figure) is connected as shown, and the point of contact along R is adjusted until the galvanometer reads zero. The segment of wire has a resistance R_X and script E_X, where I is unaffected by the connection since no current flows through the galvanometer. The unknown emf is thus proportional to the resistance of the wire segment.

Because a long uniform wire is used for R, the ratio of resistances R_X/R_S is the same as the ratio of the lengths of wire that zero the galvanometer for each emf. The three quantities on the right-hand side of the equation are now known or measured, and emf_X can be calculated. The uncertainty in this calculation can be considerably smaller than when using a voltmeter directly, but it is not zero. There is always some uncertainty in the ratio of resistances R_X/R_S and in the standard emf_S. Furthermore, it is not possible to tell when the galvanometer reads exactly zero, which introduces error into both R_X and R_S, and may also affect the current I.

Resistance Measurements and the Wheatstone Bridge

There is a variety of so-called *ohmmeters* that purport to measure resistance. What the most common ohmmeters actually do is to apply a voltage to a resistance, measure the current, and calculate the resistance using Ohm's law. Their readout is this calculated resistance. Two configurations for ohmmeters using standard voltmeters and ammeters are shown in Figure 3. Such configurations are limited in accuracy, because the meters alter both the voltage applied to the resistor and the current that flows through it.

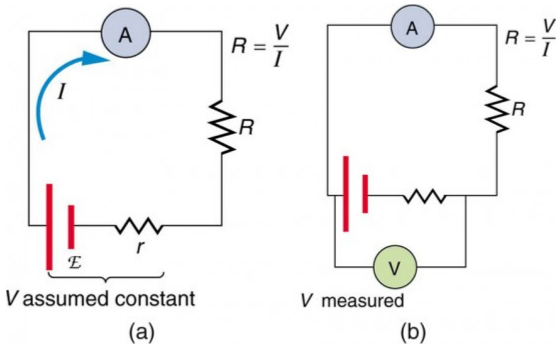


Figure 3. Two methods for measuring resistance with standard meters. (a) Assuming a known voltage for the source, an ammeter measures current, and resistance is calculated as [latex]R=\frac{V}{I}\\[/latex] . (b) Since the terminal voltage V *varies with current, it is better to measure it.* V *is most accurately known when* I *is small, but* I *itself is most accurately known when it is large.*

The *Wheatstone bridge* is a null measurement device for calculating resistance by balancing potential drops in a circuit. (See Figure 4.) The device is called a bridge because the galvanometer forms a bridge between two branches. A variety of *bridge devices* are used to make null measurements in circuits. Resistors R_1 and R_2 are precisely known, while the arrow through R_3 indicates that it is a variable resistance. The value of R_3 can be precisely read. With the unknown resistance R_x in the circuit, R_3 is adjusted until the galvanometer reads zero. The potential difference between points b and d is then zero, meaning that b and d are at the same potential. With no current running through the galvanometer, it has no effect on the rest of the circuit. So the branches abc and adc are in parallel, and each branch has the full voltage of the source. That is, the *IR* drops along abc and adc are the same. Since b and d are at the same potential, the *IR* drop along ad must equal the *IR* drop along ab.

Thus,

$$I_1R_1 = I_2R_3.$$

Again, since b and d are at the same potential, the *IR* drop along dc must equal the *IR* drop along bc. Thus,

$$I_1R_2 = I_2R_x.$$

Taking the ratio of these last two expressions gives

$$\frac{I_1R_1}{I_1R_2} = \frac{I_2R_3}{I_2R_x}.$$

Canceling the currents and solving for R_x yields

$$R_\text{x} = R_3\frac{R_2}{R_1}.$$

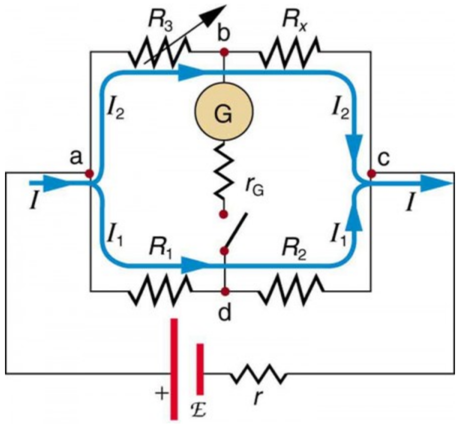


Figure 4. The Wheatstone bridge is used to calculate unknown resistances. The variable resistance R_3 is adjusted until the galvanometer reads zero with the switch closed. This simplifies the circuit, allowing R_x to be calculated based on the IR drops as discussed in the text.

This equation is used to calculate the unknown resistance when current through the galvanometer is zero. This method can be very accurate (often to four significant digits), but it is limited by two factors. First, it is not possible to get the current through the galvanometer to be exactly zero. Second, there are always uncertainties in R_1, R_2, and R_3, which contribute to the uncertainty in R_x.

Check Your Understanding

Identify other factors that might limit the accuracy of null measurements. Would the use of a digital device that is more sensitive than a galvanometer improve the accuracy of null measurements?

Solution

One factor would be resistance in the wires and connections in a null measurement. These are impossible to make zero, and they can change over time. Another factor would be temperature variations in resistance, which can be reduced but not completely eliminated by choice of material. Digital devices sensitive to smaller currents than analog devices do improve the accuracy of null measurements because they allow you to get the current closer to zero.

Section Summary

- Null measurement techniques achieve greater accuracy by balancing a circuit so that no current flows through the measuring device.
- One such device, for determining voltage, is a potentiometer.
- Another null measurement device, for determining resistance, is the Wheatstone bridge.
- Other physical quantities can also be measured with null measurement techniques.

Conceptual questions

1. Why can a null measurement be more accurate than one using standard voltmeters and ammeters? What factors limit the accuracy of null measurements?

2. If a potentiometer is used to measure cell emfs on the order of a few volts, why is it most accurate for the standard emf_s to be the same order of magnitude and the resistances to be in the range of a few ohms?

Problems & Exercises

1. What is the emf_x of a cell being measured in a potentiometer, if the standard cell's emf is 12.0 V and the potentiometer balances for $R_x = 5.000\ \Omega$ and $R_s = 2.500\ \Omega$?

2. Calculate the emf_x of a dry cell for which a potentiometer is balanced when $R_x = 1.200\ \Omega$, while an alkaline standard cell with an emf of 1.600 V requires $R_s = 1.247\ \Omega$ to balance the potentiometer.

3. When an unknown resistance R_x is placed in a Wheatstone bridge, it is possible to balance the bridge by adjusting R_3 to be 2500 Ω. What is R_x if $\frac{R_2}{R_1} = 0.625$?

4. To what value must you adjust R_3 to balance a

Wheatstone bridge, if the unknown resistance R_x is 100 Ω, R_1 is 50.0 Ω, and R_2 is 175 Ω?

5. (a) What is the unknown emf_x in a potentiometer that balances when R_x is 10.0 Ω, and balances when R_s is 15.0 Ω for a standard 3.000-V emf? (b) The same emf_x is placed in the same potentiometer, which now balances when R_s is 15.0 Ω for a standard emf of 3.100 V. At what resistance R_x will the potentiometer balance?

6. Suppose you want to measure resistances in the range from 10.0 Ω to 10.0 kΩ using a Wheatstone bridge that has $\frac{R_2}{R_1} = 2.000$. Over what range should R_3 be adjustable?

Glossary

null measurements:
: methods of measuring current and voltage more accurately by balancing the circuit so that no current flows through the measurement device

potentiometer:
: a null measurement device for measuring potentials (voltages)

ohmmeter:
: an instrument that applies a voltage to a resistance, measures the current, calculates the resistance using Ohm's law, and provides a readout of this calculated resistance

bridge device:
: a device that forms a bridge between two branches of a circuit; some bridge devices are used to make null measurements in

circuits

Wheatstone bridge:
a null measurement device for calculating resistance by balancing potential drops in a circuit

Selected Solutions to Problems & Exercises

1. 24.0 V

3. 1.56 kΩ

5. (a) 2.00 V (b)9.68 Ω

6. Range = 5.00 Ω t o5.00 kΩ

35. DC Circuits Containing Resistors and Capacitors

Learning Objectives

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

- Explain the importance of the time constant, τ, and calculate the time constant for a given resistance and capacitance.
- Explain why batteries in a flashlight gradually lose power and the light dims over time.
- Describe what happens to a graph of the voltage across a capacitor over time as it charges.
- Explain how a timing circuit works and list some applications.
- Calculate the necessary speed of a strobe flash needed to "stop" the movement of an object over a particular length.

When you use a flash camera, it takes a few seconds to charge the capacitor that powers the flash. The light flash discharges the capacitor in a tiny fraction of a second. Why does charging take longer than discharging? This question and a number of other phenomena that involve charging and discharging capacitors are discussed in this module.

RC Circuits

An ***RC circuit*** is one containing a *resistor* R and a *capacitor* C. The capacitor is an electrical component that stores electric charge.

Figure 1 shows a simple *RC* circuit that employs a DC (direct current) voltage source. The capacitor is initially uncharged. As soon as the switch is closed, current flows to and from the initially uncharged capacitor. As charge increases on the capacitor plates, there is increasing opposition to the flow of charge by the repulsion of like charges on each plate.

In terms of voltage, this is because voltage across the capacitor is given by $V_C = Q/C$, where Q is the amount of charge stored on each plate and C is the *capacitance*. This voltage opposes the battery, growing from zero to the maximum emf when fully charged. The current thus decreases from its initial value of $I_o = \frac{\text{emf}}{R}$ to zero as the voltage on the capacitor reaches the same value as the emf. When there is no current, there is no *IR* drop, and so the voltage on the capacitor must then equal the emf of the voltage source. This can also be explained with Kirchhoff's second rule (the loop rule), discussed in Kirchhoff's Rules, which says that the algebraic sum of changes in potential around any closed loop must be zero.

The initial current is $I_o = \frac{\text{emf}}{R}$, because all of the *IR* drop is in the resistance. Therefore, the smaller the resistance, the faster a given capacitor will be charged. Note that the internal resistance of the voltage source is included in R, as are the resistances of the capacitor and the connecting wires. In the flash camera scenario above, when the batteries powering the camera begin to wear out, their internal resistance rises, reducing the current and lengthening the time it takes to get ready for the next flash.

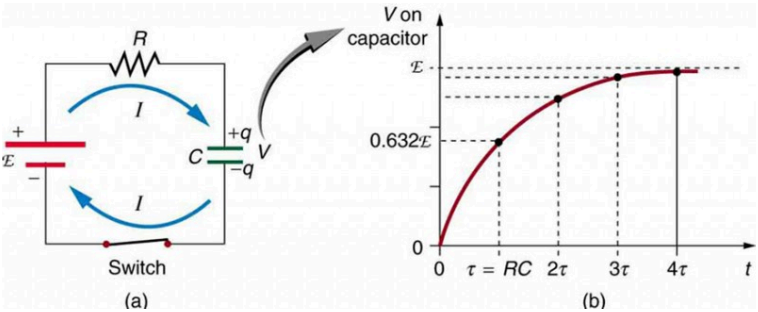


Figure 1. (a) An circuit with an initially uncharged capacitor. Current flows in the direction shown (opposite of electron flow) as soon as the switch is closed. Mutual repulsion of like charges in the capacitor progressively slows the flow as the capacitor is charged, stopping the current when the capacitor is fully charged and Q = C · *emf. (b)* A *graph of voltage across the capacitor versus time, with the switch closing at time* t = 0. *(Note that in the two parts of the figure, the capital script* E *stands for emf, q stands for the charge stored on the capacitor, and τ is the* RC *time constant.)*

Voltage on the capacitor is initially zero and rises rapidly at first, since the initial current is a maximum. Figure 1(b) shows a graph of capacitor voltage versus time (t) starting when the switch is closed at $t = 0$. The voltage approaches emf asymptotically, since the closer it gets to emf the less current flows. The equation for voltage versus time when charging a capacitor C through a resistor R, derived using calculus, is

$$V = \text{emf}(1 - e^{-t/RC}) \text{ (charging)},$$

where V is the voltage across the capacitor, emf is equal to the emf of the DC voltage source, and the exponential e = 2.718 … is the base of the natural logarithm. Note that the units of RC are seconds. We define

$$\tau = RC$$

where τ (the Greek letter tau) is called the time constant for an RC circuit. As noted before, a small resistance R allows the capacitor to charge faster. This is reasonable, since a larger current flows through a smaller resistance. It is also reasonable that the smaller the capacitor C, the less time needed to charge it. Both factors are contained in τ = RC. More quantitatively, consider what happens when t = τ = RC. Then the voltage on the capacitor is

$$V = \text{emf}\,(1 - e^{-1}) = \text{emf}\,(1 - 0.368) = 0.632 \cdot \text{emf}.$$

This means that in the time τ = RC, the voltage rises to 0.632 of its final value. The voltage will rise 0.632 of the remainder in the next time τ. It is a characteristic of the exponential function that the final value is never reached, but 0.632 of the remainder to that value is achieved in every time, τ. In just a few multiples of the time constant τ, then, the final value is very nearly achieved, as the graph in Figure 1(b) illustrates.

Discharging a Capacitor

Discharging a capacitor through a resistor proceeds in a similar fashion, as Figure 2 illustrates. Initially, the current is $I_0 = \dfrac{V_0}{R}$, driven by the initial voltage V_0 on the capacitor. As the voltage decreases, the current and hence the rate of discharge decreases, implying another exponential formula for V. Using calculus, the voltage V on a capacitor C being discharged through a resistor R is found to be

$$V = V_0 e^{-t/RC} \text{(discharging)}.$$

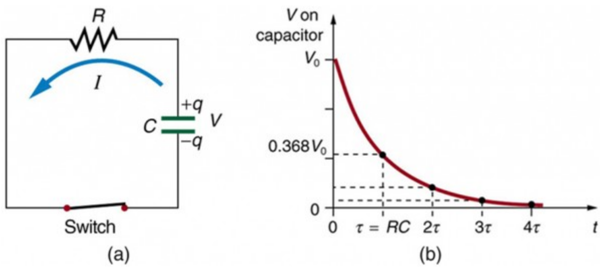


Figure 2. (a) Closing the switch discharges the capacitor C through the resistor R. Mutual repulsion of like charges on each plate drives the current. (b) A graph of voltage across the capacitor versus time, with $V = V_0$ at $t = 0$. The voltage decreases exponentially, falling a fixed fraction of the way to zero in each subsequent time constant τ.

The graph in Figure 2(b) is an example of this exponential decay. Again, the time constant is $\tau = RC$. A small resistance R allows the capacitor to discharge in a small time, since the current is larger. Similarly, a small capacitance requires less time to discharge, since less charge is stored. In the first time interval $\tau = RC$ after the switch is closed, the voltage falls to 0.368 of its initial value, since $V = V_0 \cdot e^{-1} = 0.368V_0$.

During each successive time τ, the voltage falls to 0.368 of its preceding value. In a few multiples of τ, the voltage becomes very close to zero, as indicated by the graph in Figure 2(b). Now we can explain why the flash camera in our scenario takes so much longer to charge than discharge; the resistance while charging is significantly greater than while discharging. The internal resistance of the battery accounts for most of the resistance while charging. As the battery ages, the increasing internal resistance makes the charging process even slower. (You may have noticed this.)

The flash discharge is through a low-resistance ionized gas in the flash tube and proceeds very rapidly. Flash photographs, such as in Figure 3, can capture a brief instant of a rapid motion because the flash can be less than a microsecond in duration. Such flashes can be made extremely intense. During World War II, nighttime reconnaissance photographs were made from the air with a single flash illuminating more than a square kilometer of enemy territory. The brevity of the flash eliminated blurring due to the surveillance aircraft's motion. Today, an important use of intense flash lamps is to pump energy into a laser. The short intense flash can rapidly energize a laser and allow it to reemit the energy in another form.

Figure 3. This stop-motion photograph of a rufous hummingbird (Selasphorus rufus) feeding on a flower was obtained with an extremely brief and intense flash of light powered by the discharge of a capacitor through a gas. (credit: Dean E. Biggins, U.S. Fish and Wildlife Service)

Example 1. Integrated Concept Problem: Calculating Capacitor Size—Strobe Lights

High-speed flash photography was pioneered by Doc Edgerton in the 1930s, while he was a professor of electrical engineering at MIT. You might have seen examples of his work in the amazing shots of hummingbirds in motion, a drop of milk splattering on a table, or a bullet penetrating an apple (see Figure 3). To stop the motion and capture these pictures, one needs a high-intensity, very short pulsed flash, as mentioned earlier in this module.

Suppose one wished to capture the picture of a bullet (moving at 5.0×10^2 m/s) that was passing through an apple. The duration of the flash is related to the RC time constant, τ. What size capacitor would one need in the RC circuit to succeed, if the resistance of the flash tube was 10.0 Ω? Assume the apple is a sphere with a diameter of 8.0×10^{-2}m.

Strategy

We begin by identifying the physical principles involved. This example deals with the strobe light, as discussed above. Figure 2 shows the circuit for this probe. The characteristic time τ of the strobe is given as $\tau = RC$.

Solution

We wish to find C, but we don't know τ. We want the flash to be on only while the bullet traverses the apple. So we need to use the kinematic equations that describe the relationship between distance x, velocity v, and time t:

$$x = vt \text{ or } t = \frac{x}{v}.$$

The bullet's velocity is given as 5.0×10^2 m/s, and the distance x is 8.0×10^{-2} m The traverse time, then, is

$$t=\frac{x}{v}=\frac{8.0\times10^{-2}\text{ m}}{5.0\times10^{2}\text{ m/s}}=1.6\times10^{-4}\text{ s.}$$

We set this value for the crossing time t equal to τ. Therefore,

$$C=\frac{t}{R}=\frac{1.6\times10^{-4}\text{ s}}{10.0\ \Omega}=16\ \mu\text{F.}$$

(Note: Capacitance C is typically measured in farads, F, defined as Coulombs per volt. From the equation, we see that C can also be stated in units of seconds per ohm.)

Discussion

The flash interval of 160 μs (the traverse time of the bullet) is relatively easy to obtain today. Strobe lights have opened up new worlds from science to entertainment. The information from the picture of the apple and bullet was used in the Warren Commission Report on the assassination of President John F. Kennedy in 1963 to confirm that only one bullet was fired.

RC Circuits for Timing

RC circuits are commonly used for timing purposes. A mundane example of this is found in the ubiquitous intermittent wiper systems of modern cars. The time between wipes is varied by adjusting the resistance in an RC circuit. Another example of an *RC* circuit is found in novelty jewelry, Halloween costumes, and various toys that have battery-powered flashing lights. (See Figure 4 for a timing circuit.)

A more crucial use of *RC* circuits for timing purposes is in the artificial pacemaker, used to control heart rate. The heart rate is normally controlled by electrical signals generated by the sino-atrial (SA) node, which is on the wall of the right atrium chamber. This causes the muscles to contract and pump blood. Sometimes the heart rhythm is abnormal and the heartbeat is too high or too low. The artificial pacemaker is inserted near the heart to provide electrical signals to the heart when needed with the appropriate time constant. Pacemakers have sensors that detect body motion and breathing to increase the heart rate during exercise to meet the body's increased needs for blood and oxygen.

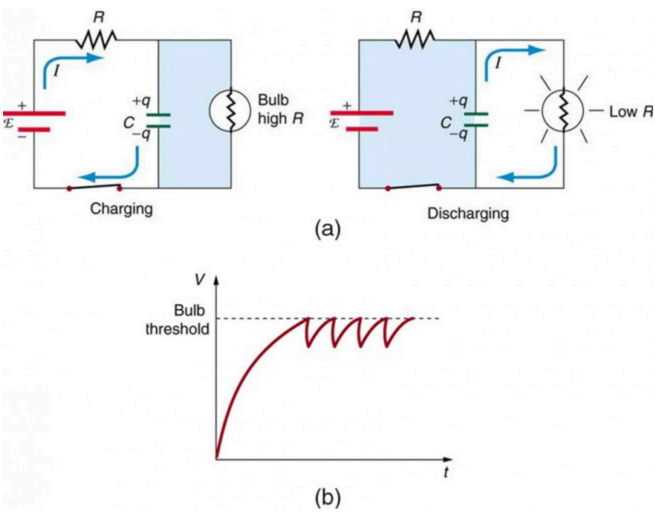


Figure 4. (a) The lamp in this RC circuit ordinarily has a very high resistance, so that the battery charges the capacitor as if the lamp were not there. When the voltage reaches a threshold value, a current flows through the lamp that dramatically reduces its resistance, and the capacitor discharges through the lamp as if the battery and charging resistor were not there. Once discharged, the process starts again, with the flash period determined by the RC constant τ. (b) A graph of voltage versus time for this circuit.

Example 2. Calculating Time: RC Circuit in a Heart Defibrillator

A heart defibrillator is used to resuscitate an accident victim by discharging a capacitor through the trunk of her body. A simplified version of the circuit is seen in Figure 2. (a) What is the time constant if an 8.00-μF capacitor is used and the path resistance through her body is $1.00 \times 10^3\ \Omega$? (b) If the initial voltage is 10.0 kV, how long does it take to decline to 5.00×10^2 V?

Strategy

Since the resistance and capacitance are given, it is straightforward to multiply them to give the time constant asked for in part (a). To find the time for the voltage to decline to 5.00×10^2 V, we repeatedly multiply the initial voltage by 0.368 until a voltage less than or equal to 5.00×10^2 V is obtained. Each multiplication corresponds to a time of τ seconds.

Solution for (a)

The time constant τ is given by the equation $\tau = RC$. Entering the given values for resistance and capacitance (and remembering that units for a farad can be expressed as s/Ω) gives

$$\tau = RC = (1.00 \times 10^3\ \Omega)\ (8.00\ \mu F) = 8.00\ \text{ms}.$$

Solution for (b)

In the first 8.00 ms, the voltage (10.0 kV) declines to 0.368 of its initial value. That is:

$$V = 0.368\ V_0 = 3.680 \times 10^3\ \text{V at } t = 8.00\ \text{ms}.$$

(Notice that we carry an extra digit for each intermediate calculation.) After another 8.00 ms, we multiply by 0.368 again, and the voltage is

$$\begin{aligned} V\prime &= 0.368\ \text{V} \\ &= (0.368)\left(3.680 \times 10^3\ \text{V}\right) \\ &= 1.354 \times 10^3\ \text{Vat}\ t = 16.0\ \text{ms} \end{aligned}$$

Similarly, after another 8.00 ms, the voltage is

$$\begin{aligned} V'' &= 0.368\ V' = (0.368)\left(1.354 \times 10^3\ \text{V}\right) \\ &= 498\ \text{V at}\ t = 24.0\ \text{ms} \end{aligned}$$

.

Discussion

So after only 24.0 ms, the voltage is down to 498 V, or 4.98% of its original value.Such brief times are useful in heart defibrillation, because the brief but intense current causes a brief but effective contraction of the heart. The actual circuit in a heart defibrillator is slightly more complex than the one in Figure 2, to compensate for magnetic and AC effects that will be covered in Magnetism.

Check Your Understanding

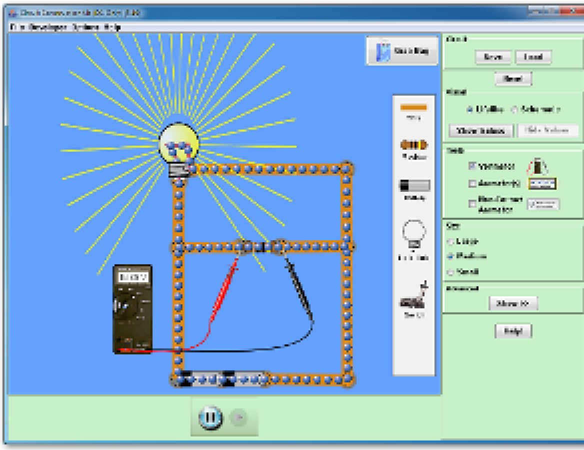
When is the potential difference across a capacitor an emf?

Solution

Only when the current being drawn from or put into the capacitor is zero. Capacitors, like batteries, have internal resistance, so their output voltage is not an emf unless current is zero. This is difficult to measure in practice so we refer to a capacitor's voltage rather than its emf. But the source of potential difference in a capacitor is fundamental and it is an emf.

PhET Explorations: Circuit Construction Kit (DC only)

An electronics kit in your computer! Build circuits with resistors, light bulbs, batteries, and switches. Take measurements with the realistic ammeter and voltmeter. View the circuit as a schematic diagram, or switch to a life-like view.



Click to download the simulation. Run using Java.

Section Summary

- An RC circuit is one that has both a resistor and a capacitor.
- The time constant τ for an RC circuit is $\tau = RC$.
- When an initially uncharged ($V_0 = 0$ at $t = 0$) capacitor in series with a resistor is charged by a DC voltage source, the voltage rises, asymptotically approaching the emf of the voltage source; as a function of time,

$$V = \text{emf}(1 - e^{-t/RC}) \text{ (charging)},$$

- Within the span of each time constant τ, the voltage rises by 0.632 of the remaining value, approaching the final voltage asymptotically.

- If a capacitor with an initial voltage V_0 is discharged through a resistor starting at $t = 0$, then its voltage decreases exponentially as given by

$$V = V_0 e^{-t/RC} \text{(discharging)}.$$

- In each time constant τ, the voltage falls by 0.368 of its remaining initial value, approaching zero asymptotically.

Conceptual questions

1. Regarding the units involved in the relationship $\tau = RC$, verify that the units of resistance times capacitance are time, that is, $\Omega \cdot \text{F} = \text{s}$.

2. The *RC* time constant in heart defibrillation is crucial to limiting the time the current flows. If the capacitance in the defibrillation unit is fixed, how would you manipulate resistance in the circuit to adjust the *RC* constant τ? Would an adjustment of the applied voltage also be needed to ensure that the current delivered has an appropriate value?

3. When making an ECG measurement, it is important to measure voltage variations over small time intervals. The time is limited by the *RC* constant of the circuit—it is not possible to measure time variations shorter than *RC*. How would you manipulate *R* and *C* in the circuit to allow the necessary measurements?

4. Draw two graphs of charge versus time on a capacitor. Draw one for charging an initially uncharged capacitor in

series with a resistor, as in the circuit in Figure 1 (above), starting from $t = 0$. Draw the other for discharging a capacitor through a resistor, as in the circuit in Figure 2 (above), starting at $t = 0$, with an initial charge Q_0. Show at least two intervals of τ.

5. When charging a capacitor, as discussed in conjunction with Figure 2, how long does it take for the voltage on the capacitor to reach emf? Is this a problem?

6. When discharging a capacitor, as discussed in conjunction with Figure 2, how long does it take for the voltage on the capacitor to reach zero? Is this a problem?

7. Referring to Figure 1, draw a graph of potential difference across the resistor versus time, showing at least two intervals of τ. Also draw a graph of current versus time for this situation.

8. A long, inexpensive extension cord is connected from inside the house to a refrigerator outside. The refrigerator doesn't run as it should. What might be the problem?

9. In Figure 4 (above), does the graph indicate the time constant is shorter for discharging than for charging? Would you expect ionized gas to have low resistance? How would you adjust R to get a longer time between flashes? Would adjusting R affect the discharge time?

10. An electronic apparatus may have large capacitors at high voltage in the power supply section, presenting a shock hazard even when the apparatus is switched off. A "bleeder resistor" is therefore placed across such a

capacitor, as shown schematically in Figure 6, to bleed the charge from it after the apparatus is off. Why must the bleeder resistance be much greater than the effective resistance of the rest of the circuit? How does this affect the time constant for discharging the capacitor?

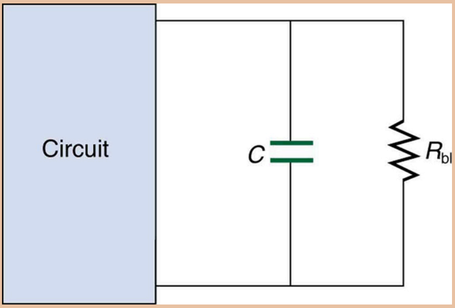


Figure 6. A bleeder resistor R_{bl} discharges the capacitor in this electronic device once it is switched off.

Problems & Exercises

1. The timing device in an automobile's intermittent wiper system is based on an RC time constant and utilizes a 0.500-μF capacitor and a variable resistor. Over what range must R be made to vary to achieve time constants from 2.00 to 15.0 s?

2. A heart pacemaker fires 72 times a minute, each time a

25.0-nF capacitor is charged (by a battery in series with a resistor) to 0.632 of its full voltage. What is the value of the resistance?

3. The duration of a photographic flash is related to an *RC* time constant, which is 0.100 μs for a certain camera. (a) If the resistance of the flash lamp is 0.0400 Ω during discharge, what is the size of the capacitor supplying its energy? (b) What is the time constant for charging the capacitor, if the charging resistance is 800 kΩ?

4. A 2.00- and a 7.50-μF capacitor can be connected in series or parallel, as can a 25.0- and a 100-kΩ resistor. Calculate the four *RC* time constants possible from connecting the resulting capacitance and resistance in series.

5. After two time constants, what percentage of the final voltage, emf, is on an initially uncharged capacitor *C*, charged through a resistance *R*?

6. A 500-Ω resistor, an uncharged 1.50-μF capacitor, and a 6.16-V emf are connected in series. (a) What is the initial current? (b) What is the *RC* time constant? (c) What is the current after one time constant? (d) What is the voltage on the capacitor after one time constant?

7. A heart defibrillator being used on a patient has an *RC* time constant of 10.0 ms due to the resistance of the patient and the capacitance of the defibrillator. (a) If the defibrillator has an 8.00-μF capacitance, what is the resistance of the path through the patient? (You may neglect the capacitance of the patient and the resistance of the defibrillator.) (b) If the initial voltage is 12.0 kV, how long does it take to decline to 6.00×10^2 V?

8. An ECG monitor must have an *RC* time constant less

than 1.00×10^2 μs to be able to measure variations in voltage over small time intervals. (a) If the resistance of the circuit (due mostly to that of the patient's chest) is 1.00 kΩ, what is the maximum capacitance of the circuit? (b) Would it be difficult in practice to limit the capacitance to less than the value found in (a)?

9. Figure 7 shows how a bleeder resistor is used to discharge a capacitor after an electronic device is shut off, allowing a person to work on the electronics with less risk of shock. (a) What is the time constant? (b) How long will it take to reduce the voltage on the capacitor to 0.250% (5% of 5%) of its full value once discharge begins? (c) If the capacitor is charged to a voltage V_0 through a 100-Ω resistance, calculate the time it takes to rise to 0.865 V_0 (This is about two time constants.)

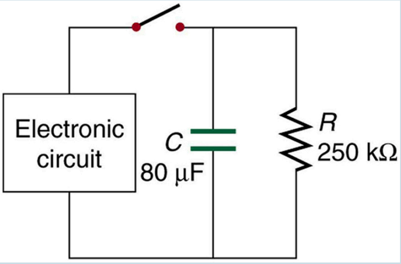


Figure 7.

10. Using the exact exponential treatment, find how much time is required to discharge a 250-μF capacitor through a 500-Ω resistor down to 1.00% of its original voltage.

11. Using the exact exponential treatment, find how much

time is required to charge an initially uncharged 100-pF capacitor through a 75.0-MΩ resistor to 90.0% of its final voltage.

12. **Integrated Concepts** If you wish to take a picture of a bullet traveling at 500 m/s, then a very brief flash of light produced by an *RC* discharge through a flash tube can limit blurring. Assuming 1.00 mm of motion during one *RC* constant is acceptable, and given that the flash is driven by a 600-μF capacitor, what is the resistance in the flash tube?

13. **Integrated Concepts** A flashing lamp in a Christmas earring is based on an *RC* discharge of a capacitor through its resistance. The effective duration of the flash is 0.250 s, during which it produces an average 0.500 W from an average 3.00 V. (a) What energy does it dissipate? (b) How much charge moves through the lamp? (c) Find the capacitance. (d) What is the resistance of the lamp?

14. **Integrated Concepts** A 160-μF capacitor charged to 450 V is discharged through a 31.2-kΩ resistor. (a) Find the time constant. (b) Calculate the temperature increase of the resistor, given that its mass is 2.50 g and its specific heat is $1.67 \frac{\text{kJ}}{\text{kg}\cdot^{\text{o}}\text{C}}$, noting that most of the thermal energy is retained in the short time of the discharge. (c) Calculate the new resistance, assuming it is pure carbon. (d) Does this change in resistance seem significant?

15. **Unreasonable Results** (a) Calculate the capacitance needed to get an *RC* time constant of 1.00×10^3 with a 0.100-Ω resistor. (b) What is unreasonable about this result? (c) Which assumptions are responsible?

16. **Construct Your Own Problem** Consider a camera's

flash unit. Construct a problem in which you calculate the size of the capacitor that stores energy for the flash lamp. Among the things to be considered are the voltage applied to the capacitor, the energy needed in the flash and the associated charge needed on the capacitor, the resistance of the flash lamp during discharge, and the desired *RC* time constant.

17. **Construct Your Own Problem** Consider a rechargeable lithium cell that is to be used to power a camcorder. Construct a problem in which you calculate the internal resistance of the cell during normal operation. Also, calculate the minimum voltage output of a battery charger to be used to recharge your lithium cell. Among the things to be considered are the emf and useful terminal voltage of a lithium cell and the current it should be able to supply to a camcorder.

Glossary

RC circuit:
: a circuit that contains both a resistor and a capacitor

capacitor:
: an electrical component used to store energy by separating electric charge on two opposing plates

capacitance:
: the maximum amount of electric potential energy that can be stored (or separated) for a given electric potential

Selected Solutions to Problems & Exercises

1. range 4.00 to 30.0 MΩ

3. (a) 2.50 μF (b) 2.00 s

5. 86.5%

7. (a) 1.25 kΩ (b) 30.0 ms

9. (a) 20.0 s (b) 120 s (c) 16.0 ms

11. 1.73×10^{-2} s

12. 3.33×10^{-3} Ω

14. (a) 4.99 s (b) 3.87ºC (c) 31.1 kΩ (d) No

36. Video: DC Circuits

Watch the following Physics Concept Trailer to see the important role that circuits play in operating electric cars.

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://library.achievingthedream.org/austinccphysics2/?p=57#oembed-1

PART V

ELECTROMAGNETIC INDUCTION, AC CIRCUITS, AND ELECTRICAL TECHNOLOGIES, PART I

37. Inductance

Learning Objectives

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

- Calculate the inductance of an inductor.
- Calculate the energy stored in an inductor.
- Calculate the emf generated in an inductor.

Inductors

Induction is the process in which an emf is induced by changing magnetic flux. Many examples have been discussed so far, some more effective than others. Transformers, for example, are designed to be particularly effective at inducing a desired voltage and current with very little loss of energy to other forms. Is there a useful physical quantity related to how "effective" a given device is? The answer is yes, and that physical quantity is called *inductance*. *Mutual inductance* is the effect of Faraday's law of induction for one device upon another, such as the primary coil in transmitting energy to the secondary in a transformer. See Figure 1, where simple coils induce emfs in one another.

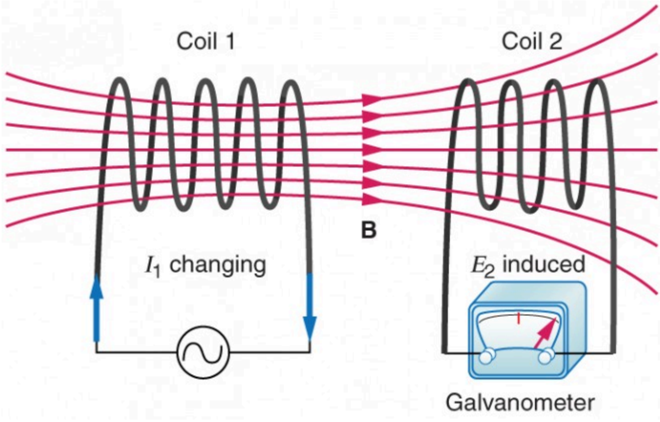


Figure 1. These coils can induce emfs in one another like an inefficient transformer. Their mutual inductance M indicates the effectiveness of the coupling between them. Here a change in current in coil 1 is seen to induce an emf in coil 2. (Note that " E_2 *induced" represents the induced emf in coil 2.)*

In the many cases where the geometry of the devices is fixed, flux is changed by varying current. We therefore concentrate on the rate of change of current, $\Delta I/\Delta t$, as the cause of induction. A change in the current I_1 in one device, coil 1 in the figure, induces an emf_2 in the other. We express this in equation form as

$$\text{emf}_2 = -M\frac{\Delta I_1}{\Delta t},$$

where M is defined to be the mutual inductance between the two devices. The minus sign is an expression of Lenz's law. The larger the mutual inductance M, the more effective the coupling. For example, the coils in Figure 1 have a small M compared with the transformer coils in Figure 3 from Transformers. Units for M are $(\text{V}\cdot\text{s})/\text{A} = \Omega\cdot\text{s}$, which is named a *henry* (H), after Joseph Henry. That is, $1\ \text{H} = 1\ \Omega\cdot\text{s}$. Nature is symmetric here. If we change the current I_2 in coil 2, we induce an emf_1 in coil 1, which is given by

$$\text{emf}_1 = -M\frac{\Delta I_2}{\Delta t},$$

where M is the same as for the reverse process. Transformers run backward with the same effectiveness, or mutual inductance M. A large mutual inductance M may or may not be desirable. We want a transformer to have a large mutual inductance. But an appliance, such as an electric clothes dryer, can induce a dangerous emf on its case if the mutual inductance between its coils and the case is large. One way to reduce mutual inductance M is to counterwind coils to cancel the magnetic field produced. (See Figure 2.)

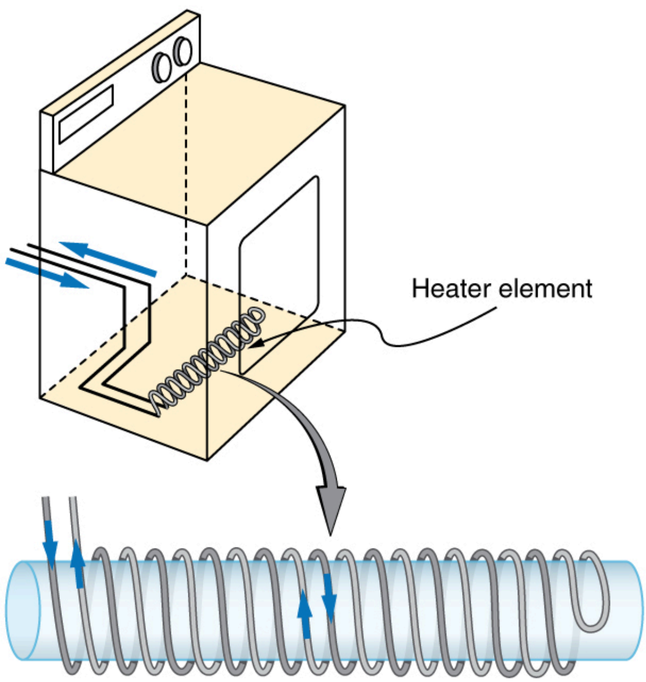


Figure 2. The heating coils of an electric clothes dryer can be counter-wound so that their magnetic fields cancel one another, greatly reducing the mutual inductance with the case of the dryer.

Self-inductance, the effect of Faraday's law of induction of a device on itself, also exists. When, for example, current through a coil is increased, the magnetic field and flux also increase, inducing a counter emf, as required by Lenz's law. Conversely, if the current is decreased, an emf is induced that opposes the decrease. Most devices have a fixed geometry, and so the change in flux is due entirely to the change in current ΔI through the device. The induced emf is related to the physical geometry of the device and the rate of change of current. It is given by

$$\mathrm{emf} = -L\frac{\Delta I}{\Delta t},$$

where L is the self-inductance of the device. A device that exhibits significant self-inductance is called an *inductor*, and given the symbol in Figure 3.

Figure 3.

The minus sign is an expression of Lenz's law, indicating that emf opposes the change in current. Units of self-inductance are henries (H) just as for mutual inductance. The larger the self-inductance L of a device, the greater its opposition to any change in current through it. For example, a large coil with many turns and an iron core has a large L and will not allow current to change quickly. To avoid this effect, a small L must be achieved, such as by counterwinding coils as in Figure 2. A 1 H inductor is a large inductor. To illustrate this, consider a device with L = 1.0 H that has a 10 A current flowing through it. What happens if we try to shut off the current rapidly, perhaps in only 1.0 ms? An emf, given by emf = −L(ΔI/Δt), will oppose the change. Thus an emf will be induced given by emf = −L(ΔI/Δt) = (1.0 H)[(10 A)/(1.0 ms)] = 10,000 V. The positive sign means this large voltage is in the same direction as the current, opposing its decrease. Such large emfs can cause arcs, damaging switching equipment, and so it may be necessary to change current more slowly. There are uses for such a large induced voltage. Camera flashes use a battery, two inductors that function as a transformer, and a switching system or oscillator to induce large voltages. (Remember that we need a changing magnetic field, brought about by a changing current, to induce a voltage in another coil.) The oscillator system will do this many times as the battery voltage is boosted to over one thousand volts. (You may hear the high pitched whine from the transformer as the capacitor is being charged.) A capacitor stores the high voltage for later use in powering the flash. (See Figure 4.)

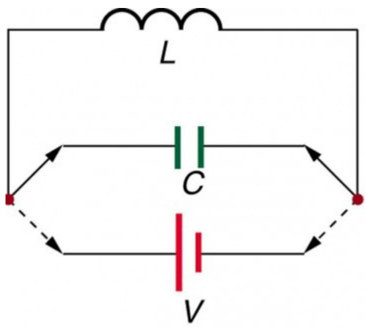


Figure 4. Through rapid switching of an inductor, 1.5 V batteries can be used to induce emfs of several thousand volts. This voltage can be used to store charge in a capacitor for later use, such as in a camera flash attachment.

It is possible to calculate L for an inductor given its geometry (size and shape) and knowing the magnetic field that it produces. This is difficult in most cases, because of the complexity of the field created. So in this text the inductance L is usually a given quantity. One exception is the solenoid, because it has a very uniform field inside, a nearly zero field outside, and a simple shape. It is instructive to derive an equation for its inductance. We start by noting that the induced emf is given by Faraday's law of induction as emf = $-N(\Delta\Phi/\Delta t)$ and, by the definition of self-inductance, as emf = $-L(\Delta I/\Delta t)$. Equating these yields

$$\text{emf} = -N\frac{\Delta\Phi}{\Delta t} = -L\frac{\Delta I}{\Delta t}$$

Solving for L gives

$$L = N\frac{\Delta\Phi}{\Delta I}$$

This equation for the self-inductance L of a device is always valid. It means that self-inductance L depends on how effective the current is in creating flux; the more effective, the greater $\Delta\Phi/\Delta I$ is. Let us use this last equation to find an expression for the inductance of a solenoid. Since the area A of a solenoid is fixed, the change in flux is Δ Φ = Δ(BA) = A Δ B. To find ΔB , we note that the magnetic field of a solenoid is given by $B=\mu_0 nI=\mu_0\frac{NI}{\ell}$. (Here n = N/l, where N is the number of coils and l is the solenoid's length.) Only the current changes, so that $\Delta\Phi=A\Delta B=\mu_0\text{NA}\frac{\Delta I}{\ell}$. Substituting Δ Φ into $L=N\frac{\Delta\Phi}{\Delta I}$ gives

$$L=N\frac{\Delta\Phi}{\Delta I}=N\frac{\mu_0\text{NA}\frac{\Delta I}{\ell}}{\Delta I}.$$

This simplifies to

$$L=\frac{\mu_0 N^2 A}{\ell}(\text{solenoid}).$$

This is the self-inductance of a solenoid of cross-sectional area A and length l. Note that the inductance depends only on the physical characteristics of the solenoid, consistent with its definition.

Example 1. Calculating the Self-inductance of a Moderate Size Solenoid

Calculate the self-inductance of a 10.0 cm long, 4.00 cm diameter solenoid that has 200 coils.

Strategy

This is a straightforward application of $L=\frac{\mu_0 N^2 A}{\ell}$, since all quantities in the equation except L are known.

Solution

Use the following expression for the self-inductance of a solenoid:

$$L=\frac{\mu_0 N^2 A}{\ell}$$

The cross-sectional area in this example is $A = \pi r^2 =$ $(3.14\ldots)(0.0200 \text{ m})^2 = 1.26 \times 10^{-3} \text{ m}^2$, N is given to be 200, and the length ℓ is 0.100 m. We know the permeability of free space is $\mu_0 = 4\pi \times 10^{-7} \text{ T} \cdot \text{m/A}$. Substituting these into the expression for L gives

$$\begin{aligned} L &= \frac{\left(4\pi\times10^{-7} \text{ T}\cdot \text{m/A}\right)(200)^2\left(1.26\times10^{-3} \text{ m}^2\right)}{0.100 \text{ m}}. \\ &= 0.632 \text{ mH} \end{aligned}$$

Discussion

This solenoid is moderate in size. Its inductance of nearly a millihenry is also considered moderate.

One common application of inductance is used in traffic lights that can tell when vehicles are waiting at the intersection. An electrical circuit with an inductor is placed in the road under the place a waiting car will stop over. The body of the car increases the inductance and the circuit changes sending a signal to the traffic lights to change colors. Similarly, metal detectors used for airport security employ the same technique. A coil or inductor in the metal detector frame acts as both a transmitter and a receiver. The pulsed signal in the transmitter coil induces a signal in the receiver. The self-inductance of the circuit is affected by any metal object in the path. Such detectors can be adjusted for sensitivity and also can indicate the approximate location of metal found on a person. (But they will not be able to detect any plastic explosive such as that found on the "underwear bomber.") See Figure 5.

Figure 5. The familiar security gate at an airport can not only detect metals but also indicate their approximate height above the floor. (credit: Alexbuirds, Wikimedia Commons)

Energy Stored in an Inductor

We know from Lenz's law that inductances oppose changes in current. There is an alternative way to look at this opposition that is based on energy. Energy is stored in a magnetic field. It takes time to build up energy, and it also takes time to deplete energy; hence, there is an opposition to rapid change. In an inductor, the magnetic field is directly proportional to current and to the inductance of the device. It can be shown that the *energy stored in an inductor* E_{ind} is given by

$$E_{\text{ind}} = \frac{1}{2}LI^2.$$

This expression is similar to that for the energy stored in a capacitor.

Example 2 Calculating the Energy Stored in the Field of a Solenoid

How much energy is stored in the 0.632 mH inductor of the preceding example when a 30.0 A current flows through it?

Strategy

The energy is given by the equation $E_{\text{ind}} = \frac{1}{2}LI^2$, and all quantities except E_{ind} are known.

Solution

Substituting the value for L found in the previous example and the given current into $E_{\text{ind}} = \dfrac{1}{2}LI^2$ gives

$$\begin{aligned} E_{\text{ind}} &= \tfrac{1}{2}LI^2 \\ &= 0.5\left(0.632 \times 10^{-3}\ \text{H}\right)\left(30.0\ \text{A}^2\right) = 0.284\ \text{J} \end{aligned}$$

Discussion

This amount of energy is certainly enough to cause a spark if the current is suddenly switched off. It cannot be built up instantaneously unless the power input is infinite.

Section Summary

- Inductance is the property of a device that tells how effectively it induces an emf in another device.
- Mutual inductance is the effect of two devices in inducing emfs in each other.
- A change in current $\Delta I_1/\Delta t$ in one induces an emf_2 in the second:

$$\text{emf}_2 = -M\frac{\Delta I_1}{\Delta t},$$

  where M is defined to be the mutual inductance between the two devices, and the minus sign is due to Lenz's law.
- Symmetrically, a change in current $\Delta I_2/\Delta t$ through the second device induces an emf_1 in the first:

$$\text{emf}_1 = -M\frac{\Delta I_2}{\Delta t},$$

where M is the same mutual inductance as in the reverse process.

- Current changes in a device induce an emf in the device itself.
- Self-inductance is the effect of the device inducing emf in itself.
- The device is called an inductor, and the emf induced in it by a change in current through it is
$$\text{emf} = -L\frac{\Delta I}{\Delta t},$$
where L is the self-inductance of the inductor, and $\Delta I/\Delta t$ is the rate of change of current through it. The minus sign indicates that emf opposes the change in current, as required by Lenz's law.
- The unit of self- and mutual inductance is the henry (H), where 1 H = 1 Ω·s.
- The self-inductance L of an inductor is proportional to how much flux changes with current. For an N-turn inductor,
$$L = N\frac{\Delta\Phi}{\Delta I}$$
- The self-inductance of a solenoid is
$$L = \frac{\mu_0 N^2 A}{\ell}\text{(solenoid)}\cdot$$
where N is its number of turns in the solenoid, A is its cross-sectional area, l is its length, and $\mu_0 = 4\pi \times 10^{-7}\ \text{T} \cdot \text{m/A}$ is the permeability of free space.
- The energy stored in an inductor E_{ind} is
$$E_{\text{ind}} = \frac{1}{2}LI^2.$$

Conceptual Questions

1. How would you place two identical flat coils in contact so that they had the greatest mutual inductance? The least?
2. How would you shape a given length of wire to give it the greatest self-inductance? The least?
3. Verify, as was concluded without proof in Example 1 (above), that units of $T \cdot m^2/A = \Omega \cdot s = H$.

Problems & Exercises

1. Two coils are placed close together in a physics lab to demonstrate Faraday's law of induction. A current of 5.00 A in one is switched off in 1.00 ms, inducing a 9.00 V emf in the other. What is their mutual inductance?

2. If two coils placed next to one another have a mutual inductance of 5.00 mH, what voltage is induced in one when the 2.00 A current in the other is switched off in 30.0 ms?

3. The 4.00 A current through a 7.50 mH inductor is switched off in 8.33 ms. What is the emf induced opposing this?

4. A device is turned on and 3.00 A flows through it 0.100

ms later. What is the self-inductance of the device if an induced 150 V emf opposes this?

5. Starting with $\text{emf}_2 = -M\dfrac{\Delta I_1}{\Delta t}$, show that the units of inductance are (V · s)/A = Ω · s.

6. Camera flashes charge a capacitor to high voltage by switching the current through an inductor on and off rapidly. In what time must the 0.100 A current through a 2.00 mH inductor be switched on or off to induce a 500 V emf?

7. A large research solenoid has a self-inductance of 25.0 H. (a) What induced emf opposes shutting it off when 100 A of current through it is switched off in 80.0 ms? (b) How much energy is stored in the inductor at full current? (c) At what rate in watts must energy be dissipated to switch the current off in 80.0 ms? (d) In view of the answer to the last part, is it surprising that shutting it down this quickly is difficult?

8. (a) Calculate the self-inductance of a 50.0 cm long, 10.0 cm diameter solenoid having 1000 loops. (b) How much energy is stored in this inductor when 20.0 A of current flows through it? (c) How fast can it be turned off if the induced emf cannot exceed 3.00 V?

9. A precision laboratory resistor is made of a coil of wire 1.50 cm in diameter and 4.00 cm long, and it has 500 turns. (a) What is its self-inductance? (b) What average emf is induced if the 12.0 A current through it is turned on in 5.00 ms (one-fourth of a cycle for 50 Hz AC)? (c) What is its

inductance if it is shortened to half its length and counter-wound (two layers of 250 turns in opposite directions)?

10. The heating coils in a hair dryer are 0.800 cm in diameter, have a combined length of 1.00 m, and a total of 400 turns. (a) What is their total self-inductance assuming they act like a single solenoid? (b) How much energy is stored in them when 6.00 A flows? (c) What average emf opposes shutting them off if this is done in 5.00 ms (one-fourth of a cycle for 50 Hz AC)?

11. When the 20.0 A current through an inductor is turned off in 1.50 ms, an 800 V emf is induced, opposing the change. What is the value of the self-inductance?

12. How fast can the 150 A current through a 0.250 H inductor be shut off if the induced emf cannot exceed 75.0 V?

13. **Integrated Concepts** A very large, superconducting solenoid such as one used in MRI scans, stores 1.00 MJ of energy in its magnetic field when 100 A flows. (a) Find its self-inductance. (b) If the coils "go normal," they gain resistance and start to dissipate thermal energy. What temperature increase is produced if all the stored energy goes into heating the 1000 kg magnet, given its average specific heat is 200 J/kg · ºC?

14. **Unreasonable Results** A 25.0 H inductor has 100 A of current turned off in 1.00 ms. (a) What voltage is induced to oppose this? (b) What is unreasonable about this result? (c) Which assumption or premise is responsible?

Glossary

inductance
: a property of a device describing how efficient it is at inducing emf in another device

mutual inductance
: how effective a pair of devices are at inducing emfs in each other

henry
: the unit of inductance; 1 H = 1 Ω·s

self-inductance
: how effective a device is at inducing emf in itself

inductor
: a device that exhibits significant self-inductance

energy stored in an inductor
: self-explanatory; calculated by $E_{\text{ind}} = \frac{1}{2}LI^2$

Selected Solutions to Problems & Exercises

1. 1.80 mH

3. 3.60 V

7. (a) 31.3 kV (b) 125 kJ (c) 1.56 MW (d) No, it is not surprising since this power is very high.

9. (a) 1.39 mH (b) 3.33 V (c) Zero

11. 60.0 mH

13. (a) 200 H (b) 500 °C

38. RL Circuits

Learning Objectives

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

- Calculate the current in an RL circuit after a specified number of characteristic time steps.
- Calculate the characteristic time of an RL circuit.
- Sketch the current in an RL circuit over time.

We know that the current through an inductor *L* cannot be turned on or off instantaneously. The change in current changes flux, inducing an emf opposing the change (Lenz's law). How long does the opposition last? Current *will* flow and *can* be turned off, but how long does it take? Figure 1 shows a switching circuit that can be used to examine current through an inductor as a function of time.

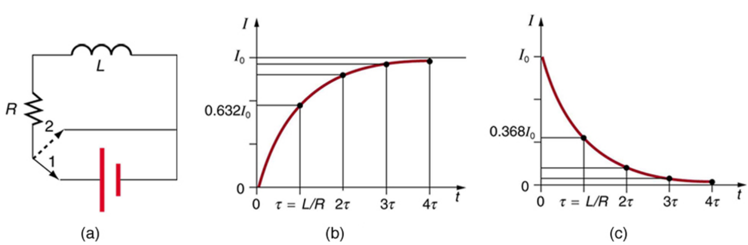


Figure 1. (a) An RL circuit with a switch to turn current on and off. When in position 1, the battery, resistor, and inductor are in series and a current is established. In position 2, the battery is removed and the current eventually stops because of energy loss in the resistor. (b) A graph of current growth versus time when the switch is moved to position 1. (c) A graph of current decay when the switch is moved to position 2.

When the switch is first moved to position 1 (at $t = 0$), the current is zero and it eventually rises to $I_0 = V/R$, where R is the total resistance of the circuit. The opposition of the inductor L is greatest at the beginning, because the amount of change is greatest. The opposition it poses is in the form of an induced emf, which decreases to zero as the current approaches its final value. The opposing emf is proportional to the amount of change left. This is the hallmark of an exponential behavior, and it can be shown with calculus that

$$I = I_0 \left(1 - e^{-t/\tau}\right) \qquad \text{(turning on)},$$

is the current in an *RL* circuit when switched on (Note the similarity to the exponential behavior of the voltage on a charging capacitor). The initial current is zero and approaches $I_0 = V/R$ with a *characteristic time constant* τ for an *RL* circuit, given by

$$\tau = \frac{L}{R},$$

where τ has units of seconds, since 1 H = 1 Ω·s. In the first period of time τ, the current rises from zero to $0.632\ I_0$, since $I = I_0(1 - e^{-1}) = I_0\ (1 - 0.368) = 0.632\ I_0$. The current will go 0.632 of the remainder in the next time τ. A well-known property of the exponential is that the final value is never exactly reached, but 0.632 of the remainder to that value is achieved in every characteristic time τ. In just a few multiples of the time τ, the final value is very nearly achieved, as the graph in Figure 1(b) illustrates.

The characteristic time τ depends on only two factors, the inductance L and the resistance R. The greater the inductance L, the greater τ is, which makes sense since a large inductance is very effective in opposing change. The smaller the resistance R, the greater τ is. Again this makes sense, since a small resistance means a large final current and a greater change to get there. In both cases—large L and small R —more energy is stored in the inductor and more time is required to get it in and out.

When the switch in Figure 1(a) is moved to position 2 and cuts the battery out of the circuit, the current drops because of energy dissipation by the resistor. But this is also not instantaneous, since the inductor opposes the decrease in current by inducing an emf in the same direction as the battery that drove the current. Furthermore, there is a certain amount of energy, $(1/2)LI_0{}^2$, stored in the inductor, and it is dissipated at a finite rate. As the current approaches zero, the rate of decrease slows, since the energy dissipation rate is I^2R. Once again the behavior is exponential, and I is found to be

$$I = I_0 e^{-t/\tau} \qquad \text{(turning off)}.$$

(See Figure 1(c).) In the first period of time $\tau = L/R$ after the switch is closed, the current falls to 0.368 of its initial value, since $I = I_0 e^{-1} = 0.368\, I_0$. In each successive time τ, the current falls to 0.368 of the preceding value, and in a few multiples of τ, the current becomes very close to zero, as seen in the graph in Figure 1(c).

Example 1. Calculating Characteristic Time and Current in an RL Circuit

(a) What is the characteristic time constant for a 7.50 mH inductor in series with a 3.00 Ω resistor? (b) Find the current 5.00 ms after the switch is moved to position 2 to disconnect the battery, if it is initially 10.0 A.

Strategy for (a)

The time constant for an RL circuit is defined by $\tau = L/R$.

Solution for (a)

Entering known values into the expression for τ given in $\tau = L/R$ yields

$$\tau = \frac{L}{R} = \frac{7.50\ \text{mH}}{3.00\ \Omega} = 2.50\ \text{ms}.$$

Discussion for (a)

This is a small but definitely finite time. The coil will be very close to its full current in about ten time constants, or about 25 ms.

Strategy for (b)

We can find the current by using $I = I_0 e^{-t/\tau}$, or by considering the decline in steps. Since the time is twice the characteristic time, we consider the process in steps.

Solution for (b)

In the first 2.50 ms, the current declines to 0.368 of its initial value, which is

$$\begin{aligned} I &= 0.368\ I_0 = (0.368)\,(10.0\ \text{A}) \\ &= 3.68\ \text{A at } t = 2.50\ \text{ms} \end{aligned}.$$

After another 2.50 ms, or a total of 5.00 ms, the current declines to 0.368 of the value just found. That is,

$$\begin{aligned} I' &= 0.368\ I = (0.368)\,(3.68\ \text{A}) \\ &= 1.35\ \text{A at } t = 5.00\ \text{ms} \end{aligned}.$$

Discussion for (b)

After another 5.00 ms has passed, the current will be 0.183 A (see Exercise 1 in the Problems & Exercises section); so, although it does die out, the current certainly does not go to zero instantaneously.

In summary, when the voltage applied to an inductor is changed, the current also changes, *but the change in current lags the change in voltage in an RL circuit*. In Reactance, Inductive and Capacitive, we explore how an *RL* circuit behaves when a sinusoidal AC voltage is applied.

Section Summary

- When a series connection of a resistor and an inductor—an *RL* circuit—is connected to a voltage source, the time variation of the current is
$$I = I_0\left(1 - e^{-t/\tau}\right) \qquad \text{(turning on)},$$
where $I_0 = V/R$ is the final current.
- The characteristic time constant τ is $\tau = \frac{L}{R}$, where L is the inductance and R is the resistance.
- In the first time constant τ, the current rises from zero to $0.632I_0$, and 0.632 of the remainder in every subsequent time interval τ.
- When the inductor is shorted through a resistor, current decreases as

$$I = I_0 e^{-t/\tau} \qquad \text{(turning off)}.$$

Here I_0 is the initial current.

- Current falls to $0.368I_0$ in the first time interval τ, and 0.368 of the remainder toward zero in each subsequent time τ.

Problems & Exercises

1. If you want a characteristic *RL* time constant of 1.00 s, and you have a 500 Ω resistor, what value of self-inductance is needed?

2. Your *RL* circuit has a characteristic time constant of 20.0 ns, and a resistance of 5.00 MΩ. (a) What is the inductance of the circuit? (b) What resistance would give you a 1.00 ns time constant, perhaps needed for quick response in an oscilloscope?

3. A large superconducting magnet, used for magnetic resonance imaging, has a 50.0 H inductance. If you want current through it to be adjustable with a 1.00 s characteristic time constant, what is the minimum resistance of system?

4. Verify that after a time of 10.0 ms, the current for the situation considered in Example 1 (above) will be 0.183 A as stated.

5. Suppose you have a supply of inductors ranging from 1.00 nH to 10.0 H, and resistors ranging from 0.100 Ω to 1.00 MΩ. What is the range of characteristic *RL* time constants

you can produce by connecting a single resistor to a single inductor?

6. (a) What is the characteristic time constant of a 25.0 mH inductor that has a resistance of 4.00 Ω? (b) If it is connected to a 12.0 V battery, what is the current after 12.5 ms?

7. What percentage of the final current I_0 flows through an inductor L in series with a resistor R, three time constants after the circuit is completed?

8. The 5.00 A current through a 1.50 H inductor is dissipated by a 2.00 Ω resistor in a circuit like that in Figure 1 (above) with the switch in position 2. (a) What is the initial energy in the inductor? (b) How long will it take the current to decline to 5.00% of its initial value? (c) Calculate the average power dissipated, and compare it with the initial power dissipated by the resistor.

9. (a) Use the exact exponential treatment to find how much time is required to bring the current through an 80.0 mH inductor in series with a 15.0 Ω resistor to 99.0% of its final value, starting from zero. (b) Compare your answer to the approximate treatment using integral numbers of τ. (c) Discuss how significant the difference is.

10. (a) Using the exact exponential treatment, find the time required for the current through a 2.00 H inductor in series with a 0.500 Ω resistor to be reduced to 0.100% of its original value. (b) Compare your answer to the approximate treatment using integral numbers of τ. (c) Discuss how significant the difference is.

Glossary

characteristic time constant:
denoted by τ, of a particular series *RL* circuit is calculated by $\tau = \frac{L}{R}$, where *L* is the inductance and *R* is the resistance

Selected Solutions to Problems & Exercises

1. 500 H

3. 50.0 Ω

5. 1.00×10^{-18} s to 0.100 s

7. 95.0%

9. (a) 24.6 ms (b) 26.7 ms (c) 9% difference, which is greater than the inherent uncertainty in the given parameters.

39. Reactance, Inductive and Capacitive

Learning Objectives

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

- Sketch voltage and current versus time in simple inductive, capacitive, and resistive circuits.
- Calculate inductive and capacitive reactance.
- Calculate current and/or voltage in simple inductive, capacitive, and resistive circuits.

Many circuits also contain capacitors and inductors, in addition to resistors and an AC voltage source. We have seen how capacitors and inductors respond to DC voltage when it is switched on and off. We will now explore how inductors and capacitors react to sinusoidal AC voltage.

Inductors and Inductive Reactance

Suppose an inductor is connected directly to an AC voltage source, as shown in Figure 1. It is reasonable to assume negligible resistance, since in practice we can make the resistance of an inductor so small that it has a negligible effect on the circuit. Also shown is a graph of voltage and current as functions of time.

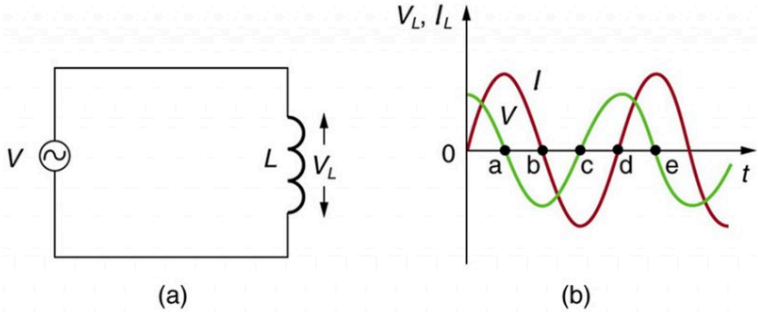


Figure 1. (a) An AC voltage source in series with an inductor having negligible resistance. (b) Graph of current and voltage across the inductor as functions of time.

The graph in Figure 1(b) starts with voltage at a maximum. Note that the current starts at zero and rises to its peak *after* the voltage that drives it, just as was the case when DC voltage was switched on in the preceding section. When the voltage becomes negative at point a, the current begins to decrease; it becomes zero at point b, where voltage is its most negative. The current then becomes negative, again following the voltage. The voltage becomes positive at point c and begins to make the current less negative. At point d, the current goes through zero just as the voltage reaches its positive peak to start another cycle. This behavior is summarized as follows:

AC Voltage in an Inductor

When a sinusoidal voltage is applied to an inductor, the voltage leads the current by one-fourth of a cycle, or by a 90° phase angle.

Current lags behind voltage, since inductors oppose change in current. Changing current induces a back emf $V = -L(\Delta I/\Delta t)$. This is considered to be an effective resistance of the inductor to AC. The rms current I through an inductor L is given by a version of Ohm's law:

$$I = \frac{V}{X_L},$$

where V is the rms voltage across the inductor and X_L is defined to be

$$X_L = 2\pi f L,$$

with f the frequency of the AC voltage source in hertz (An analysis of the circuit using Kirchhoff's loop rule and calculus actually produces this expression). X_L is called the *inductive reactance*, because the inductor reacts to impede the current. X_L has units of ohms (1 H = 1 Ω · s, so that frequency times inductance has units of (cycles/s)(Ω · s)=Ω)), consistent with its role as an effective resistance. It makes sense that X_L is proportional to L, since the greater the induction the greater its resistance to change. It is also reasonable that X_L is proportional to frequency f, since greater frequency means greater change in current. That is, $\Delta I/\Delta t$ is large for large frequencies (large f, small Δt). The greater the change, the greater the opposition of an inductor.

Example 1. Calculating Inductive Reactance and then Current

(a) Calculate the inductive reactance of a 3.00 mH inductor when 60.0 Hz and 10.0 kHz AC voltages are applied. (b) What is the rms current at each frequency if the applied rms voltage is 120 V?

Strategy

The inductive reactance is found directly from the expression $X_L = 2\pi fL$. Once X_L has been found at each frequency, Ohm's law as stated in the equation $I = V/X_L$ can be used to find the current at each frequency.

Solution for (a)

Entering the frequency and inductance into the equation $X_L = 2\pi fL$ gives

$$X_L = 2\pi fL = 6.28(60.0/\text{s})(3.00\text{ mH}) = 1.13\ \Omega \text{ at } 60\text{ Hz}.$$

Similarly, at 10 kHz,

$$X_L = 2\pi fL = 6.28(1.00 \times 10^4/\text{s})(3.00\text{ mH}) = 188\ \Omega \text{ at } 10\text{ kHz}.$$

Solution for (b)

The rms current is now found using the version of Ohm's law in Equation $I = V/X_L$, given the applied rms voltage is 120 V. For the first frequency, this yields

$$I = \frac{V}{X_L} = \frac{120\ \text{V}}{1.13\ \Omega} = 106\ \text{A at 60 Hz}.$$

Similarly, at 10 kHz,

$$I = \frac{V}{X_L} = \frac{120\ \text{V}}{188\ \Omega} = 0.637\ \text{A at 10 kHz}.$$

Discussion

The inductor reacts very differently at the two different frequencies. At the higher frequency, its reactance is large and the current is small, consistent with how an inductor impedes rapid change. Thus high frequencies are impeded the most. Inductors can be used to filter out high frequencies; for example, a large inductor can be put in series with a sound reproduction system or in series with your home computer to reduce high-frequency sound output from your speakers or high-frequency power spikes into your computer.

Note that although the resistance in the circuit considered is negligible, the AC current is not extremely large because inductive reactance impedes its flow. With AC, there is no time for the current to become extremely large.

Capacitors and Capacitive Reactance

Consider the capacitor connected directly to an AC voltage source as shown in Figure 2. The resistance of a circuit like this can be made so small that it has a negligible effect compared with the capacitor, and so we can assume negligible resistance. Voltage across the capacitor and current are graphed as functions of time in the figure.

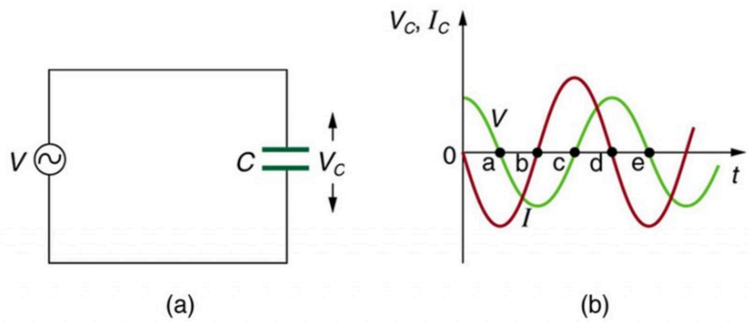


Figure 2. (a) An AC voltage source in series with a capacitor C having negligible resistance. (b) Graph of current and voltage across the capacitor as functions of time.

The graph in Figure 2 starts with voltage across the capacitor at a maximum. The current is zero at this point, because the capacitor is fully charged and halts the flow. Then voltage drops and the current becomes negative as the capacitor discharges. At point a, the capacitor has fully discharged (Q = 0 on it) and the voltage across it is zero. The current remains negative between points a and b, causing the voltage on the capacitor to reverse. This is complete at point b, where the current is zero and the voltage has its most negative value. The current becomes positive after point b, neutralizing the charge on the capacitor and bringing the voltage to zero at point c, which allows the current to reach its maximum. Between points c and d, the current drops to zero as the voltage rises to its peak, and the process starts to repeat. Throughout the cycle, the voltage follows what the current is doing by one-fourth of a cycle:

AC Voltage in a Capacitor

When a sinusoidal voltage is applied to a capacitor, the voltage follows the current by one-fourth of a cycle, or by a 90° phase angle.

The capacitor is affecting the current, having the ability to stop it altogether when fully charged. Since an AC voltage is applied, there is an rms current, but it is limited by the capacitor. This is considered to be an effective resistance of the capacitor to AC, and so the rms current I in the circuit containing only a capacitor C is given by another version of Ohm's law to be

$$I = \frac{V}{X_C},$$

where V is the rms voltage and X_C is defined (As with X_L, this expression for X_C results from an analysis of the circuit using Kirchhoff's rules and calculus) to be

$$X_C = \frac{1}{2\pi f C},$$

where X_C is called the *capacitive reactance*, because the capacitor reacts to impede the current. X_C has units of ohms (verification left as an exercise for the reader). X_C is inversely proportional to the capacitance C; the larger the capacitor, the greater the charge it can store and the greater the current that can flow. It is also inversely proportional to the frequency *f*; the greater the frequency, the less time there is to fully charge the capacitor, and so it impedes current less.

Example 2. Calculating Capacitive Reactance and then Current

(a) Calculate the capacitive reactance of a 5.00 mF capacitor when 60.0 Hz and 10.0 kHz AC voltages are applied. (b) What is the rms current if the applied rms voltage is 120 V?

Strategy

The capacitive reactance is found directly from the expression in $X_C = \dfrac{1}{2\pi fC}$. Once X_C has been found at each frequency, Ohm's law stated as $I = V/X_C$ can be used to find the current at each frequency.

Solution for (a)

Entering the frequency and capacitance into $X_C = \dfrac{1}{2\pi fC}$, gives

$$\begin{aligned} X_C &= \frac{1}{2\pi fC} \\ &= \frac{1}{6.28(60.0/\text{s})(5.00\ \mu\text{F})} = 531\ \Omega \text{ at } 60\ \text{Hz} \end{aligned}$$

.

Similarly, at 10 kHz,

$$\begin{aligned} X_C &= \frac{1}{2\pi fC} = \frac{1}{6.28\left(1.00\times10^4/\text{s}\right)(5.00\mu\text{F})}. \\ &= 3.18\ \Omega \text{ at } 10\text{ kHz} \end{aligned}$$

Solution for (b)

The rms current is now found using the version of Ohm's law in I = V/X_C, given the applied rms voltage is 120 V. For the first frequency, this yields

$$I = \frac{V}{X_C} = \frac{120\text{ V}}{531\ \Omega} = 0.226\text{ A at }60\text{ Hz}.$$

Similarly, at 10 kHz,

$$I = \frac{V}{X_C} = \frac{120\text{ V}}{3.18\ \Omega} = 3.37\text{ A at }10\text{ Hz}.$$

Discussion

The capacitor reacts very differently at the two different frequencies, and in exactly the opposite way an inductor reacts. At the higher frequency, its reactance is small and the current is large. Capacitors favor change, whereas inductors oppose change. Capacitors impede low frequencies the most, since low frequency allows them time to become charged and stop the current. Capacitors can be used to filter out low frequencies. For example, a capacitor in series with a sound reproduction system rids it of the 60 Hz hum.

Although a capacitor is basically an open circuit, there is an rms current in a circuit with an AC voltage applied to a capacitor. This is because the voltage is continually reversing, charging and discharging the capacitor. If the frequency goes to zero (DC), X_C tends to infinity, and the current is zero once the capacitor is charged. At very high frequencies, the capacitor's reactance tends to zero—it has a negligible reactance and does not impede the current (it acts like a simple wire). *Capacitors have the opposite effect on AC circuits that inductors have.*

Resistors in an AC Circuit

Just as a reminder, consider Figure 3, which shows an AC voltage applied to a resistor and a graph of voltage and current versus time. The voltage and current are exactly *in phase* in a resistor. There is no frequency dependence to the behavior of plain resistance in a circuit:

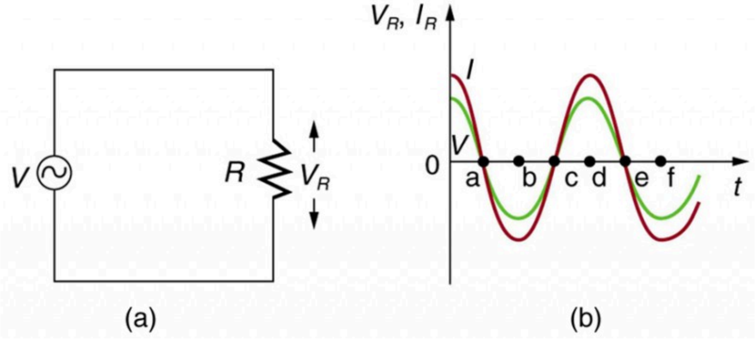


Figure 3. (a) An AC voltage source in series with a resistor. (b) Graph of current and voltage across the resistor as functions of time, showing them to be exactly in phase.

AC Voltage in a Resistor

When a sinusoidal voltage is applied to a resistor, the voltage is exactly in phase with the current—they have a 0° phase angle.

Section Summary

- For inductors in AC circuits, we find that when a sinusoidal voltage is applied to an inductor, the voltage leads the current by one-fourth of a cycle, or by a 90° phase angle.
- The opposition of an inductor to a change in current is expressed as a type of AC resistance.
- Ohm's law for an inductor is
$$I = \frac{V}{X_L},$$
where V is the rms voltage across the inductor.
- X_L is defined to be the inductive reactance, given by
$$X_L = 2\pi f L,$$
with *f* the frequency of the AC voltage source in hertz.
- Inductive reactance X_L has units of ohms and is greatest at high frequencies.
- For capacitors, we find that when a sinusoidal voltage is applied to a capacitor, the voltage follows the current by one-fourth of a cycle, or by a 90° phase angle.

- Since a capacitor can stop current when fully charged, it limits current and offers another form of AC resistance; Ohm's law for a capacitor is

$$I = \frac{V}{X_C},$$

where V is the rms voltage across the capacitor.
- X_C is defined to be the capacitive reactance, given by

$$X_C = \frac{1}{2\pi fC}.$$

- X_C has units of ohms and is greatest at low frequencies.

Conceptual Questions

1. Presbycusis is a hearing loss due to age that progressively affects higher frequencies. A hearing aid amplifier is designed to amplify all frequencies equally. To adjust its output for presbycusis, would you put a capacitor in series or parallel with the hearing aid's speaker? Explain.

2. Would you use a large inductance or a large capacitance in series with a system to filter out low frequencies, such as the 100 Hz hum in a sound system? Explain.

3. High-frequency noise in AC power can damage computers. Does the plug-in unit designed to prevent this damage use a large inductance or a large capacitance (in series with the computer) to filter out such high frequencies? Explain.

4. Does inductance depend on current, frequency, or both? What about inductive reactance?

5. Explain why the capacitor in Figure 4(a) acts as a low-frequency filter between the two circuits, whereas that in Figure 4(b) acts as a high-frequency filter.

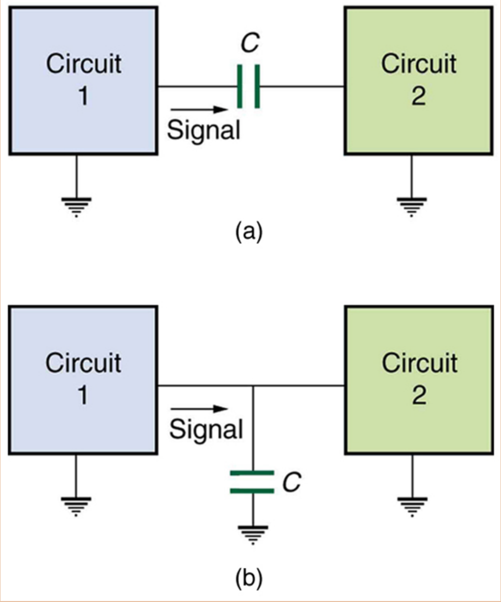


Figure 4. Capacitors and inductors. Capacitor with high frequency and low frequency.

6. If the capacitors in Figure 4 are replaced by inductors, which acts as a low-frequency filter and which as a high-frequency filter?

Problems & Exercises

1. At what frequency will a 30.0 mH inductor have a reactance of 100 Ω?

2. What value of inductance should be used if a 20.0 kΩ reactance is needed at a frequency of 500 Hz?

3. What capacitance should be used to produce a 2.00 MΩ reactance at 60.0 Hz?

4. At what frequency will an 80.0 mF capacitor have a reactance of 0.250 Ω?

5. (a) Find the current through a 0.500 H inductor connected to a 60.0 Hz, 480 V AC source. (b) What would the current be at 100 kHz?

6. (a) What current flows when a 60.0 Hz, 480 V AC source is connected to a 0.250 μF capacitor? (b) What would the current be at 25.0 kHz?

7. A 20.0 kHz, 16.0 V source connected to an inductor produces a 2.00 A current. What is the inductance?

8. A 20.0 Hz, 16.0 V source produces a 2.00 mA current when connected to a capacitor. What is the capacitance?

9. (a) An inductor designed to filter high-frequency noise from power supplied to a personal computer is placed in series with the computer. What minimum inductance should it have to produce a 2.00 kΩ reactance for 15.0 kHz noise? (b) What is its reactance at 60.0 Hz?

10. The capacitor in Figure 4(a) is designed to filter low-frequency signals, impeding their transmission between circuits. (a) What capacitance is needed to produce a 100 kΩ reactance at a frequency of 120 Hz? (b) What would its reactance be at 1.00 MHz? (c) Discuss the implications of your answers to (a) and (b).

11. The capacitor in Figure 4(b) will filter high-frequency signals by shorting them to earth/ground. (a) What capacitance is needed to produce a reactance of $\text{10.0 m}\Omega$ for a 5.00 kHz signal? (b) What would its reactance be at 3.00 Hz? (c) Discuss the implications of your answers to (a) and (b).

12. **Unreasonable Results** In a recording of voltages due to brain activity (an EEG), a 10.0 mV signal with a 0.500 Hz frequency is applied to a capacitor, producing a current of 100 mA. Resistance is negligible. (a) What is the capacitance? (b) What is unreasonable about this result? (c) Which assumption or premise is responsible?

13. **Construct Your Own Problem** Consider the use of an inductor in series with a computer operating on 60 Hz electricity. Construct a problem in which you calculate the relative reduction in voltage of incoming high frequency noise compared to 60 Hz voltage. Among the things to consider are the acceptable series reactance of the inductor for 60 Hz power and the likely frequencies of noise coming through the power lines.

Glossary

inductive reactance:
: the opposition of an inductor to a change in current; calculated by $X_L = 2\pi fL$

capacitive reactance:
: the opposition of a capacitor to a change in current; calculated by $X_C = \dfrac{1}{2\pi fC}$

Selected Solutions to Problems & Exercises

1. 531 Hz

3. 1.33 nF

5. (a) 2.55 A (b) 1.53 mA

7. 63.7 μH

9. (a) 21.2 mH (b) 8.00 Ω

40. RLC Series AC Circuits

Learning Objectives

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

- Calculate the impedance, phase angle, resonant frequency, power, power factor, voltage, and/or current in a RLC series circuit.
- Draw the circuit diagram for an RLC series circuit.
- Explain the significance of the resonant frequency.

Impedance

When alone in an AC circuit, inductors, capacitors, and resistors all impede current. How do they behave when all three occur together? Interestingly, their individual resistances in ohms do not simply add. Because inductors and capacitors behave in opposite ways, they partially to totally cancel each other's effect. Figure 1 shows an *RLC* series circuit with an AC voltage source, the behavior of which is the subject of this section. The crux of the analysis of an *RLC* circuit is the frequency dependence of X_L and X_C, and the effect they have on the phase of voltage versus current (established in the preceding section). These give rise to the frequency dependence of the circuit, with important "resonance" features that are the basis of many applications, such as radio tuners.

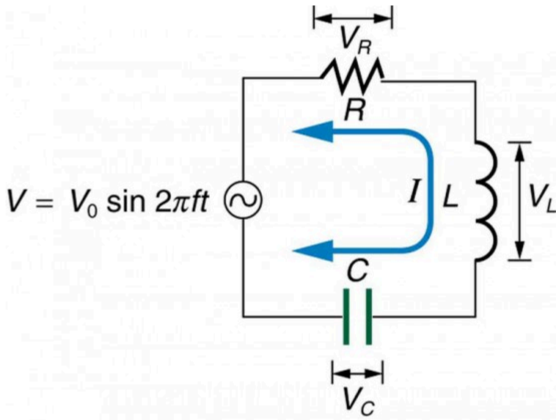


Figure 1. An RLC series circuit with an AC voltage source.

The combined effect of resistance R, inductive reactance X_L, and capacitive reactance X_C is defined to be *impedance*, an AC analogue to resistance in a DC circuit. Current, voltage, and impedance in an RLC circuit are related by an AC version of Ohm's law:

$$I_0 = \frac{V_0}{Z} \text{ or } I_{\text{rms}} = \frac{V_{\text{rms}}}{Z}.$$

Here I_0 is the peak current, V_0 the peak source voltage, and Z is the impedance of the circuit. The units of impedance are ohms, and its effect on the circuit is as you might expect: the greater the impedance, the smaller the current. To get an expression for Z in terms of R , X_L, and X_C, we will now examine how the voltages across the various components are related to the source voltage. Those voltages are labeled V_R, V_L, and V_C in Figure 1. Conservation of charge requires current to be the same in each part of the circuit at all times, so that we can say the currents in R, L, and C are equal and in phase. But we know from the preceding section that the voltage across the inductor V_L leads the current by one-fourth of a cycle, the voltage across the capacitor V_C follows the current by one-fourth of a cycle, and the voltage across the resistor V_R is exactly in phase with the current. Figure 2 shows these relationships in one graph, as well as showing the total voltage around the circuit $V = V_R + V_L + V_C$, where all four voltages are the instantaneous values. According to Kirchhoff's loop rule, the total voltage around the circuit V is also the voltage of the source. You can see from Figure 2 that while V_R is in phase with the current, V_L leads by 90°, and V_C follows by 90°. Thus V_L and V_C are 180° out of phase (crest to trough) and tend to cancel, although not completely unless they have the same magnitude. Since the peak voltages are not aligned (not in phase), the peak voltage V_0 of the source does *not* equal the sum of the peak voltages across R, L, and C. The actual relationship is

$$V_0 = \sqrt{{V_{0R}}^2 + (V_{0L} - V_{0C})^2},$$

where V_{0R}, V_{0L}, and V_{0C} are the peak voltages across R, L, and C, respectively. Now, using Ohm's law and definitions from [Reactance, Inductive and Capacitive](), we substitute $V_0 = I_0 Z$ into the above, as well as $V_{0R} = I_0 R$, $V_{0L} = I_0 X_L$, and $V_{0C} = I_0 X_C$, yielding

$$I_0 Z = \sqrt{{I_0}^2 R^2 + (I_0 X_L - I_0 X_C)^2} = I_0 \sqrt{R^2 + (X_L - X_C)^2}$$

I_0 cancels to yield an expression for Z:

$$Z = \sqrt{R^2 + (X_L - X_C)^2},$$

which is the impedance of an RLC series AC circuit. For circuits without a resistor, take R = 0; for those without an inductor, take $X_L = 0$; and for those without a capacitor, take $X_C = 0$.

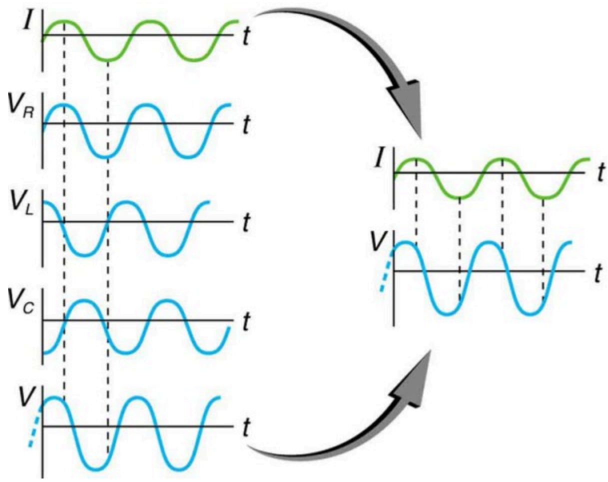


Figure 2. This graph shows the relationships of the voltages in an RLC circuit to the current. The voltages across the circuit elements add to equal the voltage of the source, which is seen to be out of phase with the current.

Example 1. Calculating Impedance and Current

An *RLC* series circuit has a 40.0 Ω resistor, a 3.00 mH inductor, and a 5.00 μF capacitor. (a) Find the circuit's impedance at 60.0 Hz and 10.0 kHz, noting that these frequencies and the values for L and C are the same as in Example 1 and Example 2 from Reactance, Inductive, and Capacitive. (b) If the voltage source has V_{rms} = 120 V, what is I_{rms} at each frequency?

Strategy

For each frequency, we use $Z = \sqrt{R^2 + (X_L - X_C)^2}$ to find the impedance and then Ohm's law to find current. We can take advantage of the results of the previous two examples rather than calculate the reactances again.

Solution for (a)

At 60.0 Hz, the values of the reactances were found in Example 1 from Reactance, Inductive, and Capacitive to be X_L = 1.13 Ω and in Example 2 from Reactance, Inductive, and Capacitive to be X_C = 531 Ω. Entering these and the given 40.0 Ω for resistance into $Z = \sqrt{R^2 + (X_L - X_C)^2}$ yields

$$\begin{aligned} Z &= \sqrt{R^2 + (X_L - X_C)^2} \\ &= \sqrt{(40.0\ \Omega)^2 + (1.13\ \Omega - 531\ \Omega)^2} \\ &= 531\ \Omega \text{ at } 60.0\ \text{Hz} \end{aligned}$$

Similarly, at 10.0 kHz, X_L = 188 Ω and X_C = 3.18 Ω, so that

$$\begin{aligned} Z &= \sqrt{(40.0\ \Omega)^2 + (188\ \Omega - 3.18\ \Omega)^2} \\ &= 190\ \Omega \text{ at } 10.0\ \text{kHz} \end{aligned}$$

Discussion for (a)

In both cases, the result is nearly the same as the largest value, and the impedance is definitely not the sum of the individual values. It is clear that X_L dominates at high frequency and X_C dominates at low frequency.

Solution for (b)

The current I_{rms} can be found using the AC version of Ohm's law in the equation $I_{rms}=V_{rms}/Z$:

$$I_{\text{rms}} = \frac{V_{\text{rms}}}{Z} = \frac{120\text{V}}{531\ \Omega} = 0.226\ \text{A} \text{ at 60.0 Hz}$$

Finally, at 10.0 kHz, we find

$$I_{\text{rms}} = \frac{V_{\text{rms}}}{Z} = \frac{120\ \text{V}}{190\ \Omega} = 0.633\ \text{A} \text{ at 10.0 kHz}$$

Discussion for (a)

The current at 60.0 Hz is the same (to three digits) as found for the capacitor alone in Example 2 from Reactance, Inductive, and Capacitive. The capacitor dominates at low frequency. The current at 10.0 kHz is only slightly different from that found for the inductor alone in Example 1 from Reactance, Inductive, and Capacitive. The inductor dominates at high frequency.

Resonance in *RLC* Series AC Circuits

How does an RLC circuit behave as a function of the frequency of the driving voltage source? Combining Ohm's law, $I_{rms} = V_{rms}/Z$, and the expression for impedance Z from $Z = \sqrt{R^2 + (X_L - X_C)^2}$ gives

$$I_{\text{rms}} = \frac{V_{\text{rms}}}{\sqrt{R^2 + (X_L - X_C)^2}}$$

The reactances vary with frequency, with X_L large at high frequencies and X_C large at low frequencies, as we have seen in three previous examples. At some intermediate frequency f_0, the reactances will be equal and cancel, giving Z = R — this is a minimum value for impedance, and a maximum value for I_{rms} results. We can get an expression for f_0 by taking

$$X_L = X_C.$$

Substituting the definitions of X_L and X_C,

$$2\pi f_0 L = \frac{1}{2\pi f_0 C}.$$

Solving this expression for f_0 yields

$$f_0 = \frac{1}{2\pi\sqrt{LC}},$$

where f_0 is the *resonant frequency* of an RLC series circuit. This is also the *natural frequency* at which the circuit would oscillate if not driven by the voltage source. At f_0, the effects of the inductor and capacitor cancel, so that Z = R, and I_{rms} is a maximum.

Resonance in AC circuits is analogous to mechanical resonance, where resonance is defined to be a forced oscillation—in this case, forced by the voltage source—at the natural frequency of the system. The receiver in a radio is an RLC circuit that oscillates best at its f_0. A variable capacitor is often used to adjust f_0 to receive a desired frequency and to reject others. Figure 3 is a graph of current as a function of frequency, illustrating a resonant peak in I_{rms} at f_0. The two curves are for two different circuits, which differ only in the amount of resistance in them. The peak is lower and broader for the higher-resistance circuit. Thus the higher-resistance circuit does not resonate as strongly and would not be as selective in a radio receiver, for example.

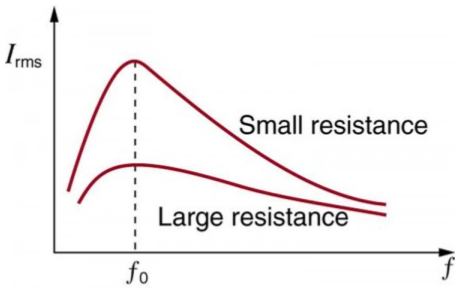


Figure 3. A graph of current versus frequency for two RLC series circuits differing only in the amount of resistance. Both have a resonance at f_0, but that for the higher resistance is lower and broader. The driving AC voltage source has a fixed amplitude V_0.

Example 2. Calculating Resonant Frequency and Current

For the same RLC series circuit having a 40.0 Ω resistor, a 3.00 mH inductor, and a 5.00 μF capacitor: (a) Find the resonant frequency. (b) Calculate I_{rms} at resonance if V_{rms} is 120 V.

Strategy

The resonant frequency is found by using the expression in $f_0 = \dfrac{1}{2\pi\sqrt{LC}}$. The current at that frequency is the same as if the resistor alone were in the circuit.

Solution for (a)

Entering the given values for L and C into the expression given for f_0 in $f_0 = \dfrac{1}{2\pi\sqrt{LC}}$ yields

$$\begin{aligned} f_0 &= \frac{1}{2\pi\sqrt{LC}} \\ &= \frac{1}{2\pi\sqrt{\left(3.00\times10^{-3}\text{ H}\right)\left(5.00\times10^{-6}\text{ F}\right)}} = 1.30\text{ kHz} \end{aligned}$$

Discussion for (a)

We see that the resonant frequency is between 60.0 Hz and 10.0 kHz, the two frequencies chosen in earlier examples. This was to be expected, since the capacitor dominated at the low frequency and the inductor dominated at the high frequency. Their effects are the same at this intermediate frequency.

Solution for (b)

The current is given by Ohm's law. At resonance, the two reactances are equal and cancel, so that the impedance equals the resistance alone. Thus,

$$I_{\text{rms}} = \frac{V_{\text{rms}}}{Z} = \frac{120\text{ V}}{40.0\ \Omega} = 3.00\text{ A}.$$

Discussion for (b)

At resonance, the current is greater than at the higher and lower frequencies considered for the same circuit in the preceding example.

Power in *RLC* Series AC Circuits

If current varies with frequency in an *RLC* circuit, then the power delivered to it also varies with frequency. But the average power is not simply current times voltage, as it is in purely resistive circuits. As was seen in Figure 2, voltage and current are out of phase in an *RLC* circuit. There is a *phase angle* φ between the source voltage V and the current I, which can be found from

$$\cos\varphi = \frac{R}{Z}$$

For example, at the resonant frequency or in a purely resistive circuit Z = R, so that $\cos\varphi = 1$. This implies that $\varphi = 0°$ and that voltage and current are in phase, as expected for resistors. At other frequencies, average power is less than at resonance. This is both because voltage and current are out of phase and because I_{rms} is lower. The fact that source voltage and current are out of phase affects the power delivered to the circuit. It can be shown that the *average power* is

$$P_{\text{ave}} = I_{\text{rms}}V_{\text{rms}}\cos\varphi,$$

Thus cos φ is called the *power factor*, which can range from 0 to 1. Power factors near 1 are desirable when designing an efficient motor, for example. At the resonant frequency, cos φ = 1.

Example 3. Calculating the Power Factor and Power

For the same *RLC* series circuit having a 40.0 Ω resistor, a 3.00 mH inductor, a 5.00 μF capacitor, and a voltage source with a V_{rms} of 120 V: (a) Calculate the power factor and phase angle for *f* = 60.0Hz. (b) What is the average power at 50.0 Hz? (c) Find the average power at the circuit's resonant frequency.

Strategy and Solution for (a)

The power factor at 60.0 Hz is found from

$$\cos\varphi = \frac{R}{Z}.$$

We know Z = 531 Ω from *Example 1: Calculating Impedance and Current*, so that

$$\cos\varphi = \frac{40.0\ \Omega}{531\ \Omega} = 0.0753\text{ at }60.0\text{ Hz}.$$

This small value indicates the voltage and current are significantly out of phase. In fact, the phase angle is

$$\varphi = \cos^{-1}0.0753 = 85.7^{\text{o}}\ \text{ at }60.0\text{ Hz}.$$

Discussion for (a)

The phase angle is close to 90°, consistent with the fact that the capacitor dominates the circuit at this low frequency (a pure *RC* circuit has its voltage and current 90° out of phase).

Strategy and Solution for (b)

The average power at 60.0 Hz is

$$P_{ave} = I_{rms}V_{rms}\cos\varphi.$$

I_{rms} was found to be 0.226 A in *Example 1: Calculating Impedance and Current*. Entering the known values gives

$$P_{ave} = (0.226\text{ A})(120\text{ V})(0.0753) = 2.04\text{ W at }60.0\text{ Hz}.$$

Strategy and Solution for (c)

At the resonant frequency, we know $\cos\varphi = 1$, and I_{rms} was found to be 6.00 A in *Example 3: Calculating Resonant Frequency and Current.*
Thus, $P_{ave} = (3.00\text{ A})(120\text{ V})(1) = 360\text{ W}$ at resonance (1.30 kHz)

Discussion

Both the current and the power factor are greater at resonance, producing significantly greater power than at higher and lower frequencies.

Power delivered to an *RLC* series AC circuit is dissipated by the resistance alone. The inductor and capacitor have energy input and output but do not dissipate it out of the circuit. Rather they transfer energy back and forth to one another, with the resistor dissipating exactly what the voltage source puts into the circuit. This assumes no significant electromagnetic radiation from the inductor and capacitor, such as radio waves. Such radiation can happen and may even be desired, as we will see in the next chapter on electromagnetic radiation, but it can also be suppressed as is the case in this chapter. The circuit is analogous to the wheel of a car driven over a corrugated road as shown in Figure 4. The regularly spaced bumps in the road are analogous to the voltage source, driving the wheel up and down. The shock absorber is analogous to the resistance damping and limiting the amplitude of the oscillation. Energy within the system goes back and forth between kinetic (analogous to maximum current, and energy stored in an inductor) and potential energy stored in the car spring (analogous to no current, and energy stored in the electric field of a capacitor). The amplitude of the wheels' motion is a maximum if the bumps in the road are hit at the resonant frequency.

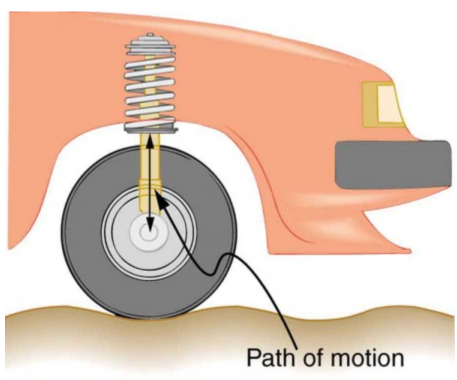


Figure 4. The forced but damped motion of the wheel on the car spring is analogous to an RLC series AC circuit. The shock absorber damps the motion and dissipates energy, analogous to the resistance in an RLC circuit. The mass and spring determine the resonant frequency.

A pure *LC* circuit with negligible resistance oscillates at f_0, the same resonant frequency as an *RLC* circuit. It can serve as a frequency standard or clock circuit—for example, in a digital wristwatch. With a very small resistance, only a very small energy input is necessary to maintain the oscillations. The circuit is analogous to a car with no shock absorbers. Once it starts oscillating, it continues at its natural frequency for some time. Figure 5 shows the analogy between an *LC* circuit and a mass on a spring.

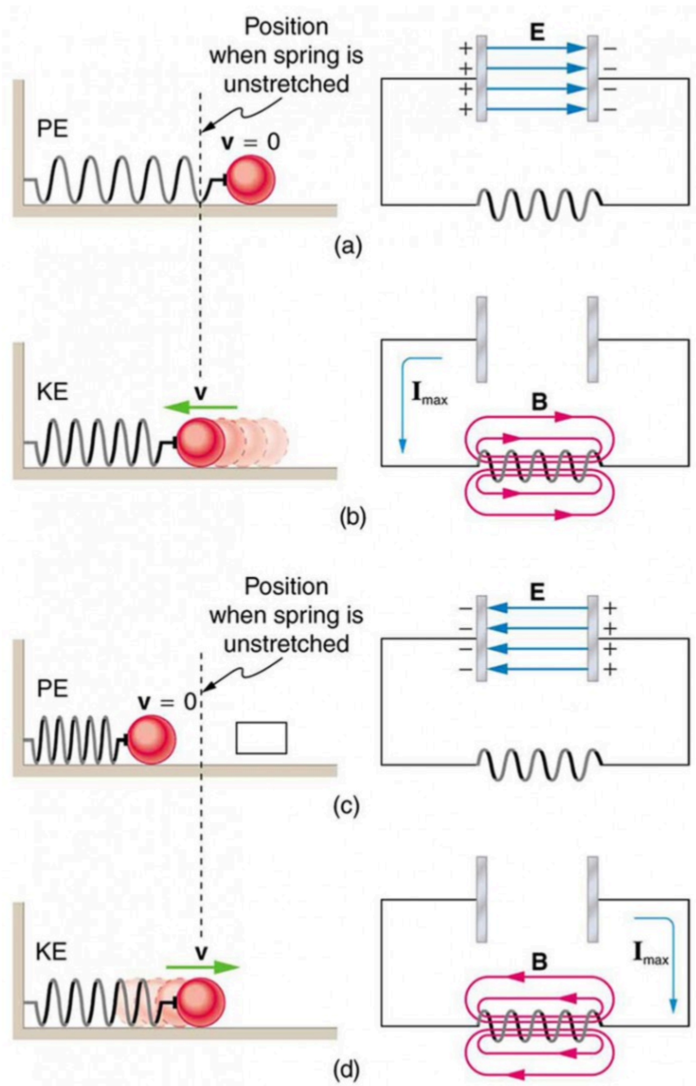
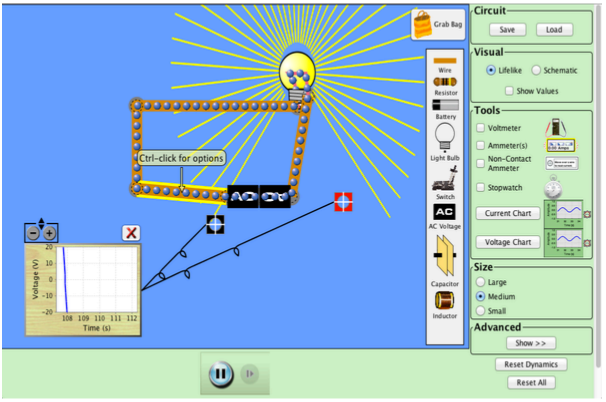


Figure 5. An LC circuit is analogous to a mass oscillating on a spring with no friction and no driving force. Energy moves back and forth between the inductor and capacitor, just as it moves from kinetic to potential in the mass-spring system.

PhET Explorations: Circuit Construction Kit (AC+DC), Virtual Lab

Build circuits with capacitors, inductors, resistors and AC or DC voltage sources, and inspect them using lab instruments such as voltmeters and ammeters.



Click to download the simulation. Run using Java.

Section Summary

- The AC analogy to resistance is impedance Z, the combined effect of resistors, inductors, and capacitors, defined by the AC version of Ohm's law:

$$I_0 = \frac{V_0}{Z} \text{ or } I_{\text{rms}} = \frac{V_{\text{rms}}}{Z},$$

where I_0 is the peak current and V_0 is the peak source voltage.

- Impedance has units of ohms and is given by $Z = \sqrt{R^2 + (X_L - X_C)^2}$.
- The resonant frequency f_0, at which $X_L = X_C$, is
$$f_0 = \frac{1}{2\pi\sqrt{LC}}$$
- In an AC circuit, there is a phase angle φ between source voltage V and the current I, which can be found from
$$\cos\varphi = \frac{R}{Z},$$
- $\varphi = 0°$ for a purely resistive circuit or an *RLC* circuit at resonance.
- The average power delivered to an *RLC* circuit is affected by the phase angle and is given by
$$P_{\text{ave}} = I_{\text{rms}} V_{\text{rms}} \cos\varphi,$$

cos φ is called the power factor, which ranges from 0 to 1.

Conceptual Questions

1. Does the resonant frequency of an AC circuit depend on the peak voltage of the AC source? Explain why or why not.

2. Suppose you have a motor with a power factor significantly less than 1. Explain why it would be better to improve the power factor as a method of improving the motor's output, rather than to increase the voltage input.

Problems & Exercises

1. An *RL* circuit consists of a 40.0 Ω resistor and a 3.00 mH inductor. (a) Find its impedance Z at 60.0 Hz and 10.0 kHz. (b) Compare these values of Z with those found in *Example 1: Calculating Impedance and Current* in which there was also a capacitor.

2. An *RC* circuit consists of a 40.0 Ω resistor and a 5.00 μF capacitor. (a) Find its impedance at 60.0 Hz and 10.0 kHz. (b) Compare these values of Z with those found in *Example 1: Calculating Impedance and Current*, in which there was also an inductor.

3. An *LC* circuit consists of a 3.00mH inductor and a 5.00 μF capacitor. (a) Find its impedance at 60.0 Hz and 10.0 kHz. (b) Compare these values of Z with those found in *Example 1: Calculating Impedance and Current* in which there was also a resistor.

4. What is the resonant frequency of a 0.500 mH inductor connected to a 40.0 μF capacitor?

5. To receive AM radio, you want an *RLC* circuit that can be made to resonate at any frequency between 500 and 1650 kHz. This is accomplished with a fixed 1.00 μH inductor connected to a variable capacitor. What range of capacitance is needed?

6. Suppose you have a supply of inductors ranging from 1.00 nH to 10.0 H, and capacitors ranging from 1.00 pF to 0.100 F. What is the range of resonant frequencies that can

be achieved from combinations of a single inductor and a single capacitor?

7. What capacitance do you need to produce a resonant frequency of 1.00 GHz, when using an 8.00 nH inductor?

8. What inductance do you need to produce a resonant frequency of 60.0 Hz, when using a 2.00 μF capacitor?

9. The lowest frequency in the FM radio band is 88.0 MHz. (a) What inductance is needed to produce this resonant frequency if it is connected to a 2.50 pF capacitor? (b) The capacitor is variable, to allow the resonant frequency to be adjusted to as high as 108 MHz. What must the capacitance be at this frequency?

10. An *RLC* series circuit has a 2.50 Ω resistor, a 100 μH inductor, and an 80.0 μF capacitor.(a) Find the circuit's impedance at 120 Hz. (b) Find the circuit's impedance at 5.00 kHz. (c) If the voltage source has V_{rms} = 5.60 V, what is I_{rms} at each frequency? (d) What is the resonant frequency of the circuit? (e) What is I_{rms} at resonance?

11. An *RLC* series circuit has a 1.00 kΩ resistor, a 150 μH inductor, and a 25.0 nF capacitor. (a) Find the circuit's impedance at 500 Hz. (b) Find the circuit's impedance at 7.50 kHz. (c) If the voltage source has V_{rms} = 408 V, what is I_{rms} at each frequency? (d) What is the resonant frequency of the circuit? (e) What is I_{rms} at resonance?

12. An *RLC* series circuit has a 2.50 Ω resistor, a 100 μH inductor, and an 80.0 μF capacitor. (a) Find the power factor at f = 120 Hz. (b) What is the phase angle at 120 Hz? (c) What

is the average power at 120 Hz? (d) Find the average power at the circuit's resonant frequency.

13. An *RLC* series circuit has a 1.00 kΩ resistor, a 150 μH inductor, and a 25.0 nF capacitor. (a) Find the power factor at f = 7.50 Hz. (b) What is the phase angle at this frequency? (c) What is the average power at this frequency? (d) Find the average power at the circuit's resonant frequency.

14. An *RLC* series circuit has a 200 Ω resistor and a 25.0 mH inductor. At 8000 Hz, the phase angle is 45.0°. (a) What is the impedance? (b) Find the circuit's capacitance. (c) If V_{rms} = 408 V is applied, what is the average power supplied?

15. Referring to Example 3. Calculating the Power Factor and Power, find the average power at 10.0 kHz.

Glossary

impedance:
: the AC analogue to resistance in a DC circuit; it is the combined effect of resistance, inductive reactance, and capacitive reactance in the form

$$Z = \sqrt{R^2 + (X_L - X_C)^2}$$

resonant frequency:
: the frequency at which the impedance in a circuit is at a minimum, and also the frequency at which the circuit would oscillate if not driven by a voltage source; calculated by

$$f_0 = \frac{1}{2\pi\sqrt{LC}}$$

phase angle:

denoted by φ, the amount by which the voltage and current are out of phase with each other in a circuit

power factor:

the amount by which the power delivered in the circuit is less than the theoretical maximum of the circuit due to voltage and current being out of phase; calculated by $\cos\varphi$

Selected Solutions to Problems & Exercises

1. (a) 40.02 Ω at 60.0 Hz, 193 Ω at 10.0 kHz (b) At 60 Hz, with a capacitor, Z = 531 Ω, over 13 times as high as without the capacitor. The capacitor makes a large difference at low frequencies. At 10 kHz, with a capacitor Z = 190 Ω, about the same as without the capacitor. The capacitor has a smaller effect at high frequencies.

3. (a) 529 Ω at 60.0 Hz, 185 Ω at 10.0 kHz (b) These values are close to those obtained in *Example 1: Calculating Impedance and Current* because at low frequency the capacitor dominates and at high frequency the inductor dominates. So in both cases the resistor makes little contribution to the total impedance.

5. 9.30 nF to 101 nF

7. 3.17 pF

9. (a) 1.31 μH (b) 1.66 pF

11. (a) 12.8 kΩ (b) 1.31 kΩ (c) 31.9 mA at 500 Hz, 312 mA at 7.50 kHz (d) 82.2 kHz (e) 0.408 A

13. (a) 0.159 (b) 80.9º (c) 26.4 W (d) 166 W

15. 16.0 W

PART VI
MAGNETISM

41. Introduction to Magnetism

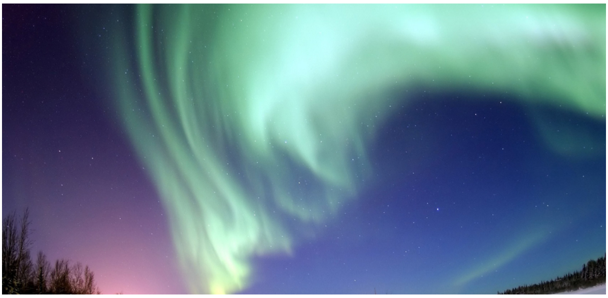


Figure 1. The magnificent spectacle of the Aurora Borealis, or northern lights, glows in the northern sky above Bear Lake near Eielson Air Force Base, Alaska. Shaped by the Earth's magnetic field, this light is produced by radiation spewed from solar storms. (credit: Senior Airman Joshua Strang, via Flickr)

One evening, an Alaskan sticks a note to his refrigerator with a small magnet. Through the kitchen window, the Aurora Borealis glows in the night sky. This grand spectacle is shaped by the same force that holds the note to the refrigerator.

People have been aware of magnets and magnetism for thousands of years. The earliest records date to well before the time of Christ, particularly in a region of Asia Minor called Magnesia (the name of this region is the source of words like *magnetic*). Magnetic rocks found in Magnesia, which is now part of western Turkey, stimulated interest during ancient times. A practical application for magnets was found later, when they were employed as navigational compasses. The use of magnets in compasses resulted not only in improved long-distance sailing, but also in the names of "north" and "south" being given to the two types of magnetic poles.

Today magnetism plays many important roles in our lives. Physicists' understanding of magnetism has enabled the development of technologies that affect our everyday lives. The iPod in your purse or backpack, for example, wouldn't have been possible without the applications of magnetism and electricity on a small scale.

The discovery that weak changes in a magnetic field in a thin film of iron and chromium could bring about much larger changes in electrical resistance was one of the first large successes of nanotechnology. The 2007 Nobel Prize in Physics went to Albert Fert from France and Peter Grunberg from Germany for this discovery of *giant magnetoresistance* and its applications to computer memory.

All electric motors, with uses as diverse as powering refrigerators, starting cars, and moving elevators, contain magnets. Generators, whether producing hydroelectric power or running bicycle lights, use magnetic fields. Recycling facilities employ magnets to separate iron from other refuse. Hundreds of millions of dollars are spent annually on magnetic containment of fusion as a future energy source. Magnetic resonance imaging (MRI) has become an important diagnostic tool in the field of medicine, and the use of magnetism to explore brain activity is a subject of contemporary research and development. The list of applications also includes computer hard drives, tape recording, detection of inhaled asbestos, and levitation of high-speed trains. Magnetism is used to explain atomic energy levels, cosmic rays, and charged particles trapped in the Van Allen belts. Once again, we will find all these disparate phenomena are linked by a small number of underlying physical principles.

Figure 2. Engineering of technology like iPods would not be possible without a deep understanding magnetism. (credit: Jesse! S?, Flickr)

42. 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.

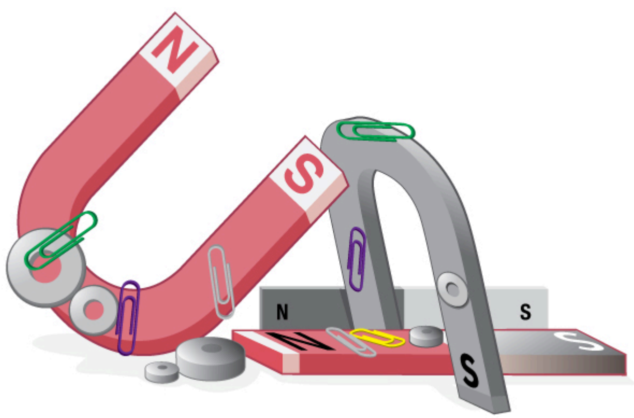


Figure 1. Magnets come in various shapes, sizes, and strengths. All have both a north pole and a south pole. There is never an isolated pole (a monopole).

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 Magnetic 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.

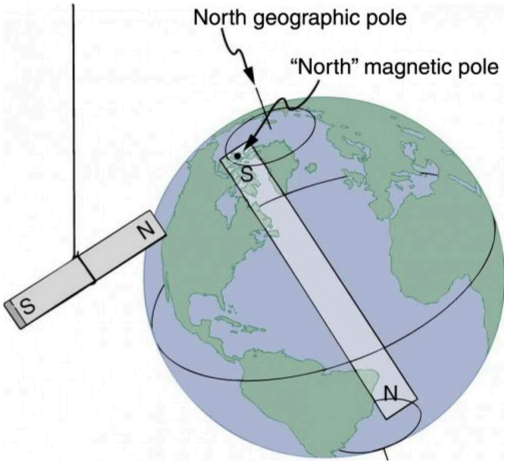


Figure 2. One end of a bar magnet is suspended from a thread that points toward north. The magnet's two poles are labeled N and S for north-seeking and south-seeking poles, respectively.

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.

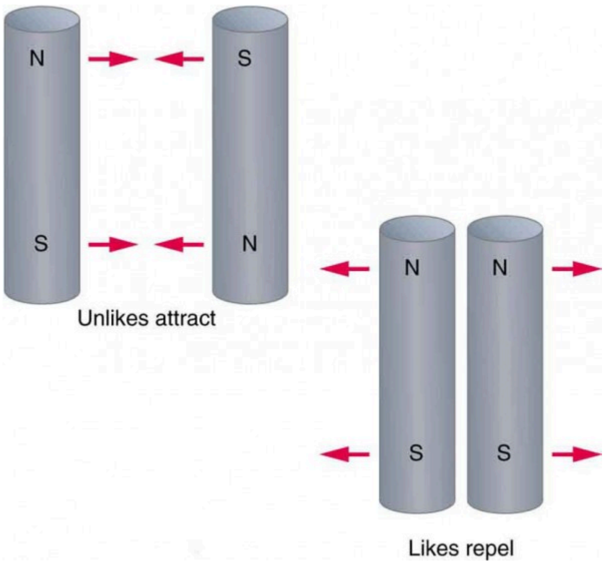


Figure 3. Unlike poles attract, whereas like poles repel.

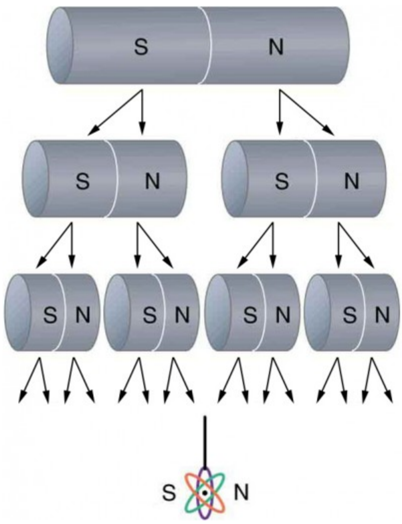


Figure 4. North and south poles always occur in pairs. Attempts to separate them result in more pairs of poles. If we continue to split the magnet, we will eventually get down to an iron atom with a north pole and a south pole—these, too, cannot be separated.

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?

Section 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.

Conceptual Questions

1. Volcanic and other such activity at the mid-Atlantic ridge extrudes material to fill the gap between separating tectonic plates associated with continental drift. The magnetization of rocks is found to reverse in a coordinated manner with distance from the ridge. What does this imply about the Earth's magnetic field and how could the knowledge of the spreading rate be used to give its historical record?

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

43. 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.

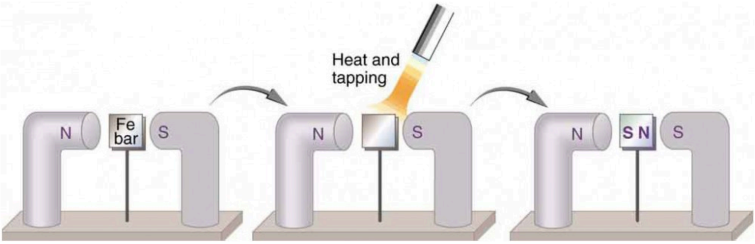


Figure 1. An unmagnetized piece of iron is placed between two magnets, heated, and then cooled, or simply tapped when cold. The iron becomes a permanent magnet with the poles aligned as shown: its south pole is adjacent to the north pole of the original magnet, and its north pole is adjacent to the south 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 1. (This results in the attraction of the previously unmagnetized material to the magnet.) What happens on a microscopic scale is illustrated in Figure 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 2(b). This induced magnetization can be made permanent if the material is heated and then cooled, or simply tapped in the presence of other magnets.

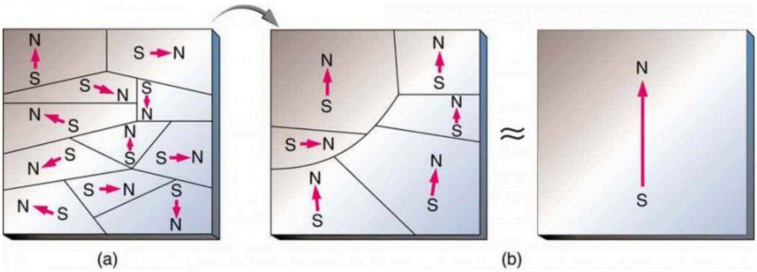


Figure 2. (a) An unmagnetized piece of iron (or other ferromagnetic material) has randomly oriented domains. (b) When magnetized by an external field, the domains show greater alignment, and some grow at the expense of others. Individual atoms are aligned within domains; each atom acts like a tiny bar 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°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 (See Figure 3).

Figure 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 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.

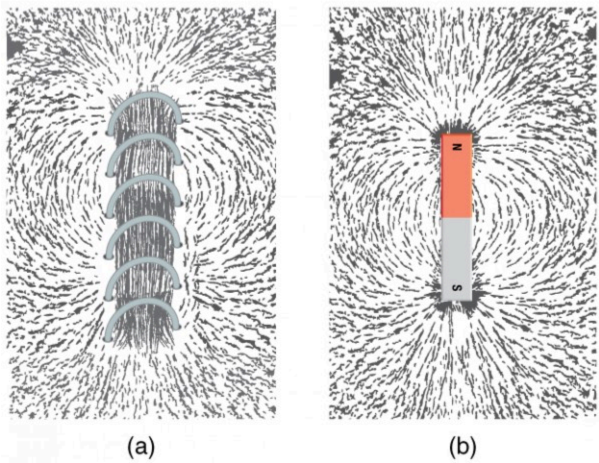


Figure 4. Iron filings near (a) a current-carrying coil and (b) a magnet act like tiny compass needles, showing the shape of their fields. Their response to a current-carrying coil and a permanent magnet is seen to be very similar, especially near the ends of the coil and the magnet.

Combining a ferromagnet with an electromagnet can produce particularly strong magnetic effects. (See Figure 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.

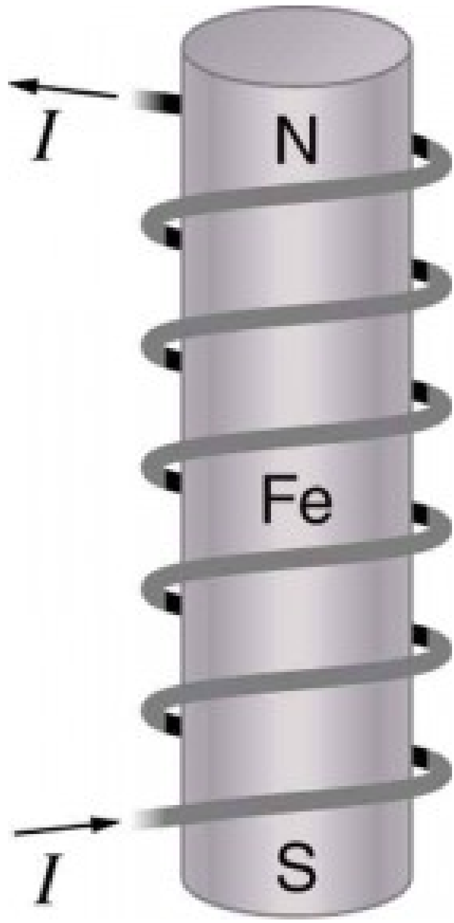


Figure 5. An electromagnet with a ferromagnetic core can produce very strong magnetic effects. Alignment of domains in the core produces a magnet, the poles of which are aligned with the electromagnet.

Figure 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.

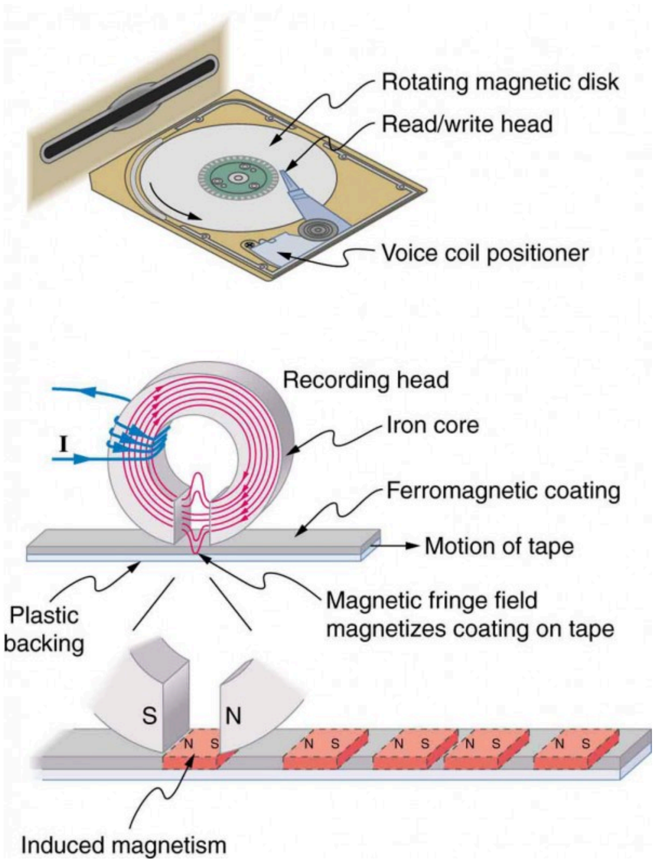


Figure 6. An electromagnet induces regions of permanent magnetism on a floppy disk coated with a ferromagnetic material. The information stored here is digital (a region is either magnetic or not); in other applications, it can be analog (with a varying strength), such as on audiotapes.

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 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.

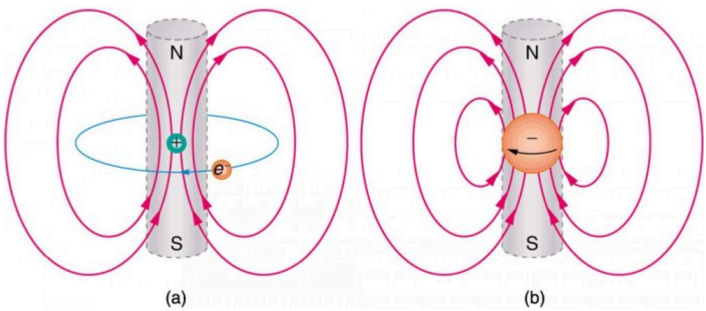
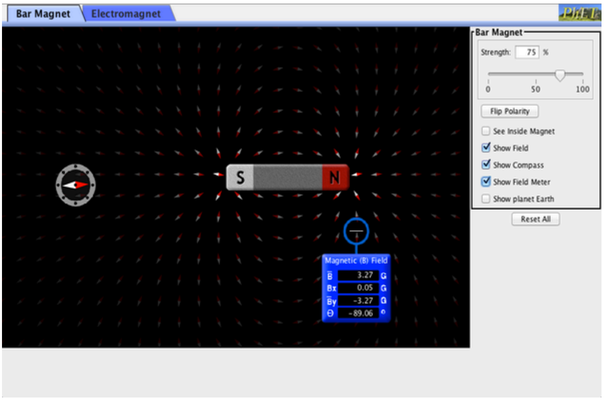


Figure 7. (a) In the planetary model of the atom, an electron orbits a nucleus, forming a closed-current loop and producing a magnetic field with a north pole and a south pole. (b) Electrons have spin and can be crudely pictured as rotating charge, forming a current that produces a magnetic field with 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?



Click to download the simulation. Run using Java.

Section 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)

44. Magnetic Fields and Magnetic Field Lines

Learning Objective

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

- Define magnetic field and describe the magnetic field lines of various magnetic fields.

Einstein is said to have been fascinated by a compass as a child, perhaps musing on how the needle felt a force without direct physical contact. His ability to think deeply and clearly about action at a distance, particularly for gravitational, electric, and magnetic forces, later enabled him to create his revolutionary theory of relativity. Since magnetic forces act at a distance, we define a *magnetic field* to represent magnetic forces. The pictorial representation of *magnetic field lines* is very useful in visualizing the strength and direction of the magnetic field. As shown in Figure 1, the *direction of magnetic field lines* is defined to be the direction in which the north end of a compass needle points. The magnetic field is traditionally called the *B-field*.

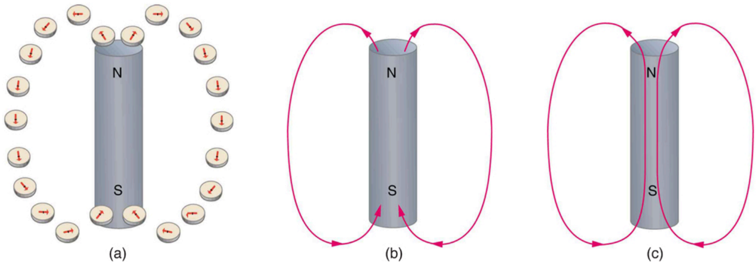


Figure 1. Magnetic field lines are defined to have the direction that a small compass points when placed at a location. (a) If small compasses are used to map the magnetic field around a bar magnet, they will point in the directions shown: away from the north pole of the magnet, toward the south pole of the magnet. (Recall that the Earth's north magnetic pole is really a south pole in terms of definitions of poles on a bar magnet.) (b) Connecting the arrows gives continuous magnetic field lines. The strength of the field is proportional to the closeness (or density) of the lines. (c) If the interior of the magnet could be probed, the field lines would be found to form continuous closed loops.

Small compasses used to test a magnetic field will not disturb it. (This is analogous to the way we tested electric fields with a small test charge. In both cases, the fields represent only the object creating them and not the probe testing them.) Figure 2 shows how the magnetic field appears for a current loop and a long straight wire, as could be explored with small compasses. A small compass placed in these fields will align itself parallel to the field line at its location, with its north pole pointing in the direction of *B*. Note the symbols used for field into and out of the paper.

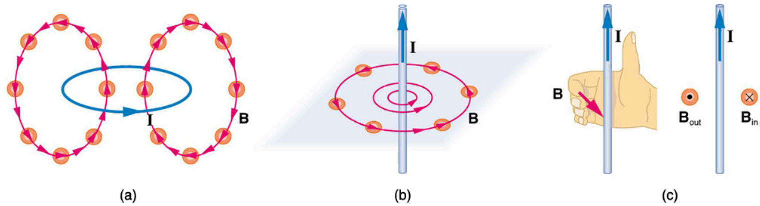


Figure 2. Small compasses could be used to map the fields shown here. (a) The magnetic field of a circular current loop is similar to that of a bar magnet. (b) A long and straight wire creates a field with magnetic field lines forming circular loops. (c) When the wire is in the plane of the paper, the field is perpendicular to the paper. Note that the symbols used for the field pointing inward (like the tail of an arrow) and the field pointing outward (like the tip of an arrow).

Making Connections: Concept of a Field

A field is a way of mapping forces surrounding any object that can act on another object at a distance without apparent physical connection. The field represents the object generating it. Gravitational fields map gravitational forces, electric fields map electrical forces, and magnetic fields map magnetic forces.

Extensive exploration of magnetic fields has revealed a number of hard-and-fast rules. We use magnetic field lines to represent the field (the lines are a pictorial tool, not a physical entity in and of themselves). The properties of magnetic field lines can be summarized by these rules:

1. The direction of the magnetic field is tangent to the field line at any point in space. A small compass will point in the direction of the field line.

2. The strength of the field is proportional to the closeness of the lines. It is exactly proportional to the number of lines per unit area perpendicular to the lines (called the areal density).
3. Magnetic field lines can never cross, meaning that the field is unique at any point in space.
4. Magnetic field lines are continuous, forming closed loops without beginning or end. They go from the north pole to the south pole.

The last property is related to the fact that the north and south poles cannot be separated. It is a distinct difference from electric field lines, which begin and end on the positive and negative charges. If magnetic monopoles existed, then magnetic field lines would begin and end on them.

Section Summary

- Magnetic fields can be pictorially represented by magnetic field lines, the properties of which are as follows:
  - The field is tangent to the magnetic field line.
  - Field strength is proportional to the line density.
  - Field lines cannot cross.
  - Field lines are continuous loops.

Conceptual Questions

1. Explain why the magnetic field would not be unique (that is, not have a single value) at a point in space where magnetic field lines might cross. (Consider the

direction of the field at such a point.)

2. List the ways in which magnetic field lines and electric field lines are similar. For example, the field direction is tangent to the line at any point in space. Also list the ways in which they differ. For example, electric force is parallel to electric field lines, whereas magnetic force on moving charges is perpendicular to magnetic field lines.
3. Noting that the magnetic field lines of a bar magnet resemble the electric field lines of a pair of equal and opposite charges, do you expect the magnetic field to rapidly decrease in strength with distance from the magnet? Is this consistent with your experience with magnets?
4. Is the Earth's magnetic field parallel to the ground at all locations? If not, where is it parallel to the surface? Is its strength the same at all locations? If not, where is it greatest?

Glossary

magnetic field:
the representation of magnetic forces

B-field:
another term for magnetic field

magnetic field lines:
the pictorial representation of the strength and the direction of a magnetic field

direction of magnetic field lines:
the direction that the north end of a compass needle points

45. Video: Magnetic Forces

Watch the following Physics Concept Trailer to see an explanation of how magnetic forces come into play to create the spectacular lights and colors we see in auroras.

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://library.achievingthedream.org/austinccphysics2/?p=68#oembed-1

46. Magnetic Field Strength: Force on a Moving Charge in a Magnetic Field

Learning Objectives

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

- Describe the effects of magnetic fields on moving charges.
- Use the right hand rule 1 to determine the velocity of a charge, the direction of the magnetic field, and the direction of the magnetic force on a moving charge.
- Calculate the magnetic force on a moving charge.

What is the mechanism by which one magnet exerts a force on another? The answer is related to the fact that all magnetism is caused by current, the flow of charge. *Magnetic fields exert forces on moving charges*, and so they exert forces on other magnets, all of which have moving charges.

Right Hand Rule 1

The magnetic force on a moving charge is one of the most fundamental known. Magnetic force is as important as the electrostatic or Coulomb force. Yet the magnetic force is more complex, in both the number of factors that affects it and in its direction, than the relatively simple Coulomb force. The magnitude of the *magnetic force* F on a charge q moving at a speed v in a magnetic field of strength B is given by

$$F = qvB \sin \theta,$$

where θ is the angle between the directions of v and B. This force is often called the *Lorentz force*. In fact, this is how we define the magnetic field strength B—in terms of the force on a charged particle moving in a magnetic field. The SI unit for magnetic field strength B is called the *tesla* (T) after the eccentric but brilliant inventor Nikola Tesla (1856–1943). To determine how the tesla relates to other SI units, we solve $F = qvB \sin \theta$ for B.

$$B = \frac{F}{qv \sin \theta}$$

Because $\sin \theta$ is unitless, the tesla is

$$1\ \mathrm{T} = \frac{1\ \mathrm{N}}{\mathrm{C} \cdot \ \mathrm{m/s}} = \frac{1\ \mathrm{N}}{\mathrm{A} \cdot \ \mathrm{m}}$$

(note that C/s = A). Another smaller unit, called the *gauss* (G), where $1\ \mathrm{G} = 10^{-4}\ \mathrm{T}$, is sometimes used. The strongest permanent magnets have fields near 2 T; superconducting electromagnets may attain 10 T or more. The Earth's magnetic field on its surface is only about 5×10^{-5} T, or 0.5 G.

The *direction* of the magnetic force **F** is perpendicular to the plane formed by **v** and **B**, as determined by the *right hand rule* 1 (or **RHR-1**), which is illustrated in Figure 1. RHR-1 states that, to determine the direction of the magnetic force on a positive moving charge, you point the thumb of the right hand in the direction of **v**, the fingers in the direction of **B**, and a perpendicular to the palm points in the direction of **F**. One way to remember this is that there is one velocity, and so the thumb represents it. There are many field lines, and so the fingers represent them. The force is in the direction you would push with your palm. The force on a negative charge is in exactly the opposite direction to that on a positive charge.

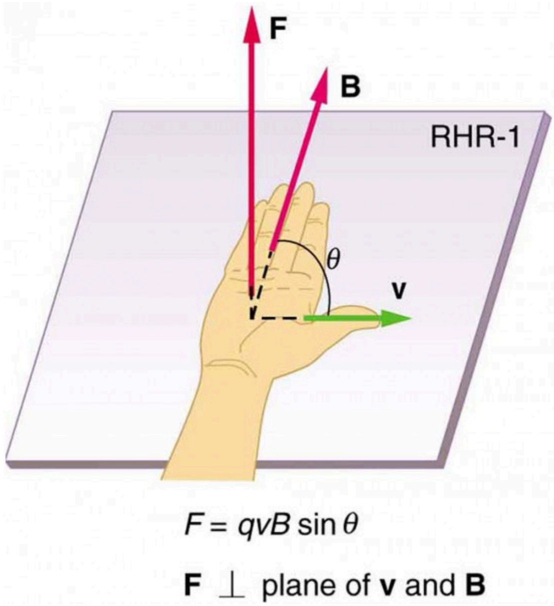


Figure 1. Magnetic fields exert forces on moving charges. This force is one of the most basic known. The direction of the magnetic force on a moving charge is perpendicular to the plane formed by v and B and follows right hand rule–1 (RHR-1) as shown. The magnitude of the force is proportional to q, v, B, and the sine of the angle between v and B.

Making Connections: Charges and Magnets

There is no magnetic force on static charges. However, there is a magnetic force on moving charges. When charges are stationary, their electric fields do not affect magnets. But, when charges move, they produce magnetic fields that exert forces on other magnets. When there is relative motion, a connection between electric and magnetic fields emerges—each affects the other.

Example 1. Calculating Magnetic Force: Earth's Magnetic Field on a Charged Glass Rod

With the exception of compasses, you seldom see or personally experience forces due to the Earth's small magnetic field. To illustrate this, suppose that in a physics lab you rub a glass rod with silk, placing a 20-nC positive charge on it. Calculate the force on the rod due to the Earth's magnetic field, if you throw it with a horizontal velocity of 10 m/s due west in a place where the Earth's field is due north parallel to the ground. (The direction of the force is determined with right hand rule 1 as shown in Figure 2.)

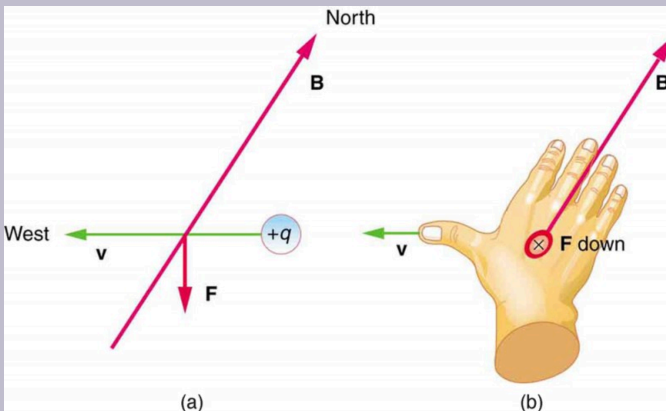


Figure 2. A positively charged object moving due west in a region where the Earth's magnetic field is due north experiences a force that is straight down as shown. A negative charge moving in the same direction would feel a force straight up.

Strategy

We are given the charge, its velocity, and the magnetic field strength and direction. We can thus use the equation $F = qvB \sin \theta$ to find the force.

Solution

The magnetic force is

$$F = qvB \sin \theta$$

We see that $\sin \theta = 1$, since the angle between the velocity and the direction of the field is 90°. Entering the other given quantities yields

$$
\begin{aligned}
F &= \left(20 \times 10^{-9}\ \mathrm{C}\right)(10\ \mathrm{m/s})\left(5 \times 10^{-5}\ \mathrm{T}\right) \\
&= 1 \times 10^{-11}\ (\mathrm{C} \cdot\ \mathrm{m/s})\left(\frac{N}{\mathrm{C} \cdot \mathrm{m/s}}\right) = 1 \times 10^{-11}\ \mathrm{N}
\end{aligned}
$$

.

Discussion

This force is completely negligible on any macroscopic object, consistent with experience. (It is calculated to only one digit, since the Earth's field varies with location and is given to only one digit.) The Earth's magnetic field, however, does produce very important effects, particularly on submicroscopic particles. Some of these are explored in Force on a Moving Charge in a Magnetic Field: Examples and Applications.

Section Summary

- Magnetic fields exert a force on a moving charge q, the magnitude of which is

  $$F = qvB \sin \theta,$$

  where θ is the angle between the directions of v and B.
- The SI unit for magnetic field strength B is the tesla (T), which is related to other units by

  $$1\ \mathrm{T} = \frac{1\ \mathrm{N}}{\mathrm{C} \cdot\ \mathrm{m/s}} = \frac{1\ \mathrm{N}}{\mathrm{A} \cdot\ \mathrm{m}}$$
- The *direction* of the force on a moving charge is given by right hand rule 1 (RHR-1): Point the thumb of the right hand in the

direction of v, the fingers in the direction of B, and a perpendicular to the palm points in the direction of F.
- The force is perpendicular to the plane formed by $\boldsymbol{v}$ and $\mathbf{B}$. Since the force is zero if $\boldsymbol{v}$ is parallel to $\mathbf{B}$, charged particles often follow magnetic field lines rather than cross them.

Conceptual Questions

1. If a charged particle moves in a straight line through some region of space, can you say that the magnetic field in that region is necessarily zero?

Problems & Exercises

1. What is the direction of the magnetic force on a positive charge that moves as shown in each of the six cases shown in Figure 3?

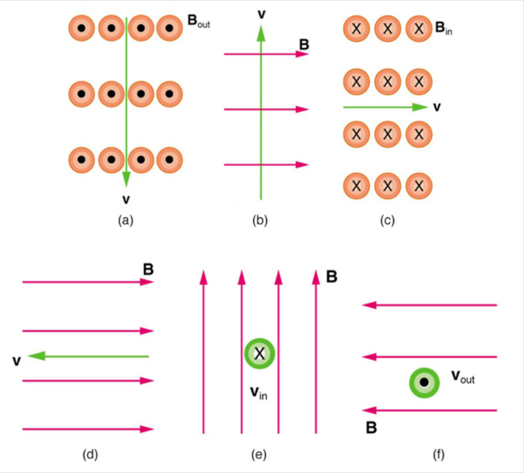


Figure 3.

2. Repeat Exercise 1 for a negative charge.

3. What is the direction of the velocity of a negative charge that experiences the magnetic force shown in each of the three cases in Figure 4, assuming it moves perpendicular to **B**?

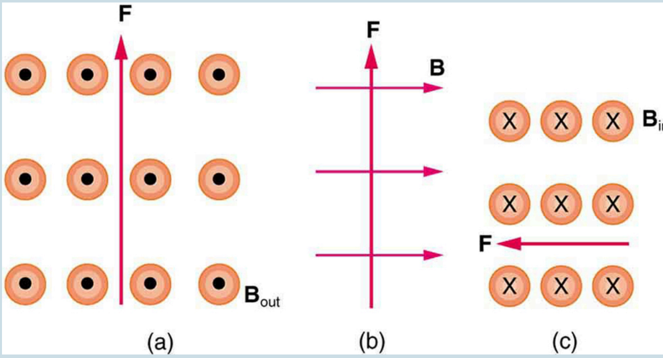


Figure 4.

4. Repeat Figure 4 for a positive charge.

5. What is the direction of the magnetic field that produces the magnetic force on a positive charge as shown in each of the three cases in the figure below, assuming **B** is perpendicular to **v**?

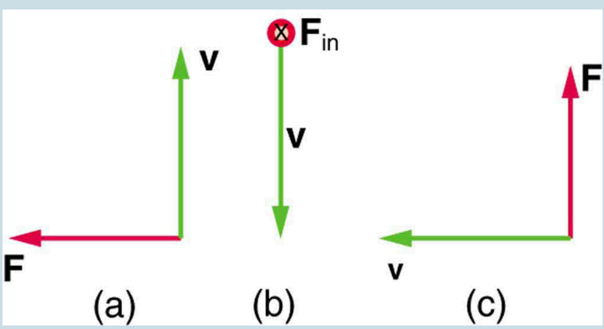


Figure 5.

6. Repeat Exercise 5 for a negative charge.

7. What is the maximum force on an aluminum rod with a 0.100-μC charge that you pass between the poles of a 1.50-T permanent magnet at a speed of 5.00 m/s? In what direction is the force?

8. (a) Aircraft sometimes acquire small static charges. Suppose a supersonic jet has a0.500-μC charge and flies due west at a speed of 660 m/s over the Earth's south magnetic pole, where the 8.00×10^{-5}-T magnetic field points straight up. What are the direction and the magnitude of the magnetic force on the plane? (b) Discuss whether the value obtained in part (a) implies this is a significant or negligible effect.

9. (a) A cosmic ray proton moving toward the Earth at 5.00×10^{7} experiences a magnetic force of 1.70×10^{-16} N. What is the strength of the magnetic field if there is a 45º angle between it and the proton's velocity? (b) Is the value obtained in part (a) consistent with the known strength of the Earth's magnetic field on its surface? Discuss.

10. An electron moving at 4.00×10^{3} m/s in a 1.25-T magnetic field experiences a magnetic force of 1.40×10^{-16} N. What angle does the velocity of the electron make with the magnetic field? There are two answers.

11. (a) A physicist performing a sensitive measurement wants to limit the magnetic force on a moving charge in her equipment to less than 1.00×10^{-12} N. What is the greatest the charge can be if it moves at a maximum speed of 30.0

m/s in the Earth's field? (b) Discuss whether it would be difficult to limit the charge to less than the value found in (a) by comparing it with typical static electricity and noting that static is often absent.

Glossary

right hand rule 1 (RHR-1):
: the rule to determine the direction of the magnetic force on a positive moving charge: when the thumb of the right hand points in the direction of the charge's velocity **v** and the fingers point in the direction of the magnetic field **B**, then the force on the charge is perpendicular and away from the palm; the force on a negative charge is perpendicular and into the palm

Lorentz force:
: the force on a charge moving in a magnetic field

tesla:
: T, the SI unit of the magnetic field strength;

$$1\text{ T} = \frac{1\text{ N}}{\text{A}\cdot\text{ m}}$$

magnetic force:
: the force on a charge produced by its motion through a magnetic field; the Lorentz force

gauss:
: G, the unit of the magnetic field strength; 1 G = 10^{-4}T

Selected Solutions to Problems & Exercises

1. (a) Left (West) (b) Into the page (c) Up (North) (d) No force (e) Right (East) (f) Down (South)

3. (a) East (right) (b) Into page (c) South (down)

5. (a) Into page (b) West (left) (c) Out of page

7. 7.50×10^{-7} N perpendicular to both the magnetic field lines and the velocity

9. (a) 3.01×10^{-5} T (b) This is slightly less then the magnetic field strength of 5×10^{-5} T at the surface of the Earth, so it is consistent.

10. (a) 6.67×10^{-10} C (taking the Earth's field to be 5.00×10^{-5} T) (b) Less than typical static, therefore difficult

47. Force on a Moving Charge in a Magnetic Field: Examples and Applications

Learning Objectives

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

- Describe the effects of a magnetic field on a moving charge.
- Calculate the radius of curvature of the path of a charge that is moving in a magnetic field.

Magnetic force can cause a charged particle to move in a circular or spiral path. Cosmic rays are energetic charged particles in outer space, some of which approach the Earth. They can be forced into spiral paths by the Earth's magnetic field. Protons in giant accelerators are kept in a circular path by magnetic force. The bubble chamber photograph in Figure 1 shows charged particles moving in such curved paths. The curved paths of charged particles in magnetic fields are the basis of a number of phenomena and can even be used analytically, such as in a mass spectrometer.

Figure 1. Trails of bubbles are produced by high-energy charged particles moving through the superheated liquid hydrogen in this artist's rendition of a bubble chamber. There is a strong magnetic field perpendicular to the page that causes the curved paths of the particles. The radius of the path can be used to find the mass, charge, and energy of the particle.

So does the magnetic force cause circular motion? Magnetic force is always perpendicular to velocity, so that it does no work on the charged particle. The particle's kinetic energy and speed thus remain constant. The direction of motion is affected, but not the speed. This is typical of uniform circular motion. The simplest case occurs when a charged particle moves perpendicular to a uniform B-field, such as shown in Figure 2. (If this takes place in a vacuum, the magnetic field is the dominant factor determining the motion.) Here, the magnetic force supplies the centripetal force $F_c = mv^2/r$. Noting that $\sin\theta = 1$, we see that $F = qvB$.

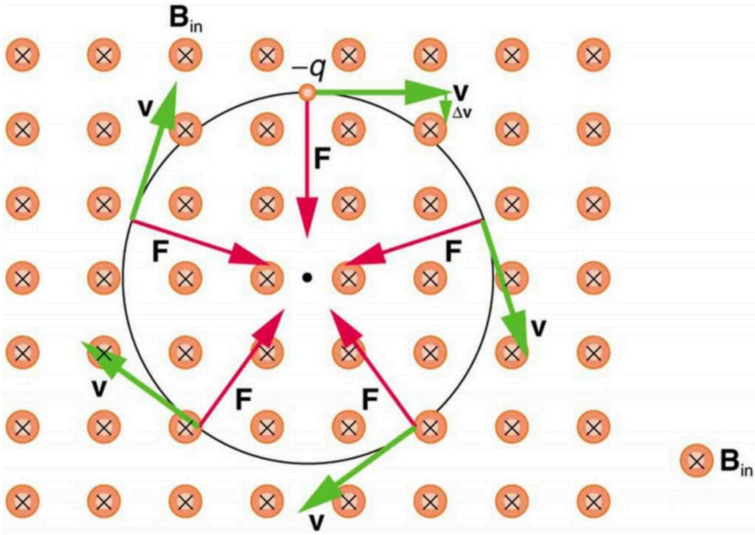


Figure 2. A negatively charged particle moves in the plane of the page in a region where the magnetic field is perpendicular into the page (represented by the small circles with x's—like the tails of arrows). The magnetic force is perpendicular to the velocity, and so velocity changes in direction but not magnitude. Uniform circular motion results.

Because the magnetic force F supplies the centripetal force F_c, we have

$$qvB = \frac{mv^2}{r}.$$

Solving for r yields

$$r = \frac{mv}{qB}.$$

Here, r is the radius of curvature of the path of a charged particle with mass m and charge q, moving at a speed v perpendicular to a magnetic field of strength B. If the velocity is not perpendicular to the magnetic field, then v is the component of the velocity perpendicular to the field. The component of the velocity parallel to the field is unaffected, since the magnetic force is zero for motion parallel to the field. This produces a spiral motion rather than a circular one.

Example 1. Calculating the Curvature of the Path of an Electron Moving in a Magnetic Field: A Magnet on a TV Screen

A magnet brought near an old-fashioned TV screen such as in Figure 1 (TV sets with cathode ray tubes instead of LCD screens) severely distorts its picture by altering the path of the electrons that make its phosphors glow. ***(Don't try this at home, as it will permanently magnetize and ruin the TV.)*** To illustrate this, calculate the radius of curvature of the path of an electron having a velocity of 6.00×10^7 m/s (corresponding to the accelerating voltage of about 10.0 kV used in some TVs) perpendicular to a magnetic field of strength $B = 0.500$ T (obtainable with permanent magnets).

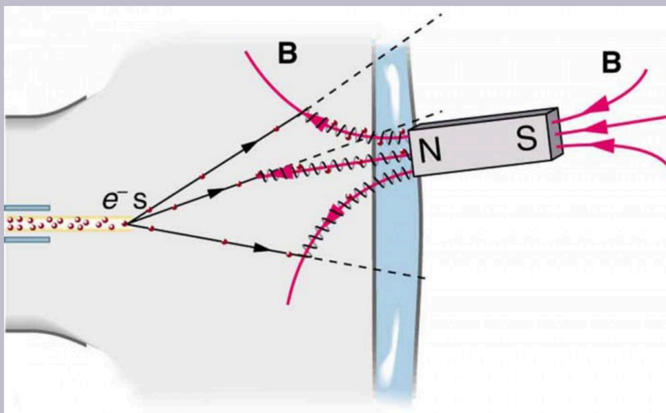


Figure 1. Side view showing what happens when a magnet comes in contact with a computer monitor or TV screen. Electrons moving toward the screen spiral about magnetic field lines, maintaining the component of their velocity parallel to the field lines. This distorts the image on the screen.

Strategy

We can find the radius of curvature *r* directly from the equation $r = \frac{mv}{qB}$, since all other quantities in it are given or known.

Solution

Using known values for the mass and charge of an electron, along with the given values of *v* and *B* gives us

$$
\begin{aligned}
r = \frac{mv}{qB} \quad &= \quad \frac{\left(9.11\times10^{-31}\text{ kg}\right)\left(6.00\times10^{7}\text{ m/s}\right)}{\left(1.60\times10^{-19}\text{ C}\right)\left(0.500\text{ T}\right)} \\
&= \quad 6.83\times10^{-4}\text{ m}
\end{aligned}
$$

or

r = 0.683 mm.

Discussion

The small radius indicates a large effect. The electrons in the TV picture tube are made to move in very tight circles, greatly altering their paths and distorting the image.

Figure 2 shows how electrons not moving perpendicular to magnetic field lines follow the field lines. The component of velocity parallel to the lines is unaffected, and so the charges spiral along the field lines. If field strength increases in the direction of motion, the field will exert a force to slow the charges, forming a kind of magnetic mirror, as shown below.

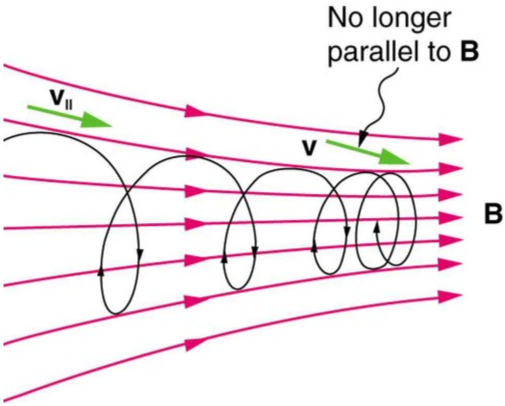


Figure 2. When a charged particle moves along a magnetic field line into a region where the field becomes stronger, the particle experiences a force that reduces the component of velocity parallel to the field. This force slows the motion along the field line and here reverses it, forming a "magnetic mirror."

The properties of charged particles in magnetic fields are related to such different things as the Aurora Australis or Aurora Borealis and particle accelerators. *Charged particles approaching magnetic field lines may get trapped in spiral orbits about the lines rather than crossing them*, as seen above. Some cosmic rays, for example, follow the Earth's magnetic field lines, entering the atmosphere near the magnetic poles and causing the southern or northern lights through their ionization of molecules in the atmosphere. This glow of energized atoms and molecules is seen in Figure 1 on page. Those particles that approach middle latitudes must cross magnetic field lines, and many are prevented from penetrating the atmosphere. Cosmic rays are a component of background radiation; consequently, they give a higher radiation dose at the poles than at the equator.

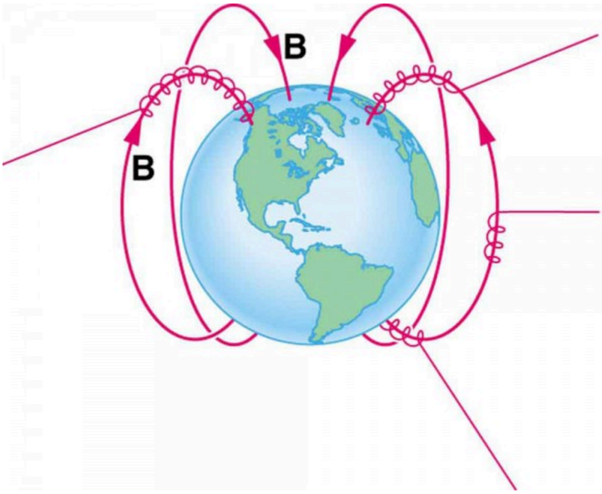


Figure 3. Energetic electrons and protons, components of cosmic rays, from the Sun and deep outer space often follow the Earth's magnetic field lines rather than cross them. (Recall that the Earth's north magnetic pole is really a south pole in terms of a bar magnet.)

Some incoming charged particles become trapped in the Earth's magnetic field, forming two belts above the atmosphere known as the Van Allen radiation belts after the discoverer James A. Van Allen, an American astrophysicist. (See Figure 4.) Particles trapped in these belts form radiation fields (similar to nuclear radiation) so intense that manned space flights avoid them and satellites with sensitive electronics are kept out of them. In the few minutes it took lunar missions to cross the Van Allen radiation belts, astronauts received radiation doses more than twice the allowed annual exposure for radiation workers. Other planets have similar belts, especially those having strong magnetic fields like Jupiter.

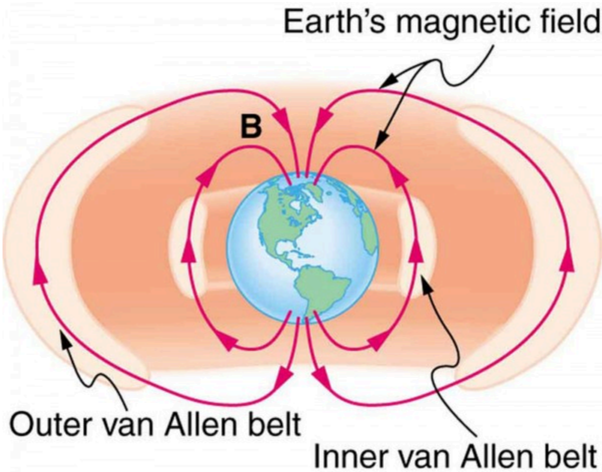


Figure 4. The Van Allen radiation belts are two regions in which energetic charged particles are trapped in the Earth's magnetic field. One belt lies about 300 km above the Earth's surface, the other about 16,000 km. Charged particles in these belts migrate along magnetic field lines and are partially reflected away from the poles by the stronger fields there. The charged particles that enter the atmosphere are replenished by the Sun and sources in deep outer space.

Back on Earth, we have devices that employ magnetic fields to contain charged particles. Among them are the giant particle accelerators that have been used to explore the substructure of matter. (See Figure 5.) Magnetic fields not only control the direction of the charged particles, they also are used to focus particles into beams and overcome the repulsion of like charges in these beams.

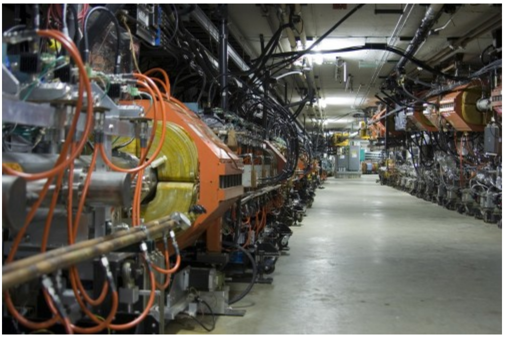


Figure 5. The Fermilab facility in Illinois has a large particle accelerator (the most powerful in the world until 2008) that employs magnetic fields (magnets seen here in orange) to contain and direct its beam. This and other accelerators have been in use for several decades and have allowed us to discover some of the laws underlying all matter. (credit: ammcrim, Flickr)

Thermonuclear fusion (like that occurring in the Sun) is a hope for a future clean energy source. One of the most promising devices is the *tokamak*, which uses magnetic fields to contain (or trap) and direct the reactive charged particles. (See Figure 6.) Less exotic, but more immediately practical, amplifiers in microwave ovens use a magnetic field to contain oscillating electrons. These oscillating electrons generate the microwaves sent into the oven.

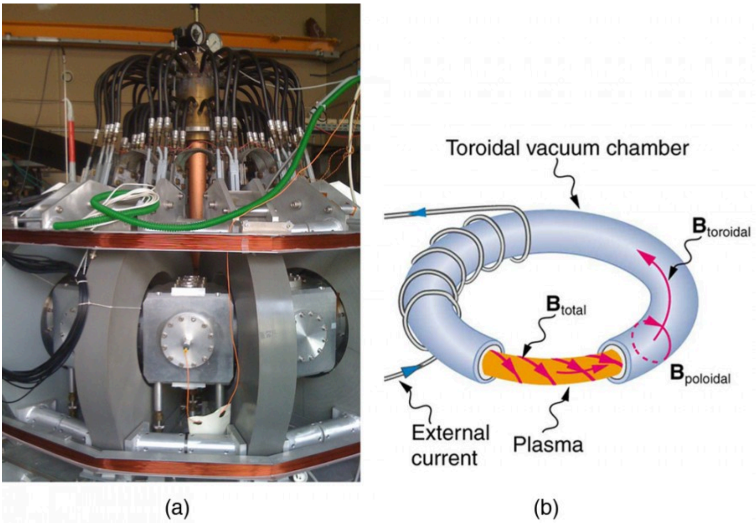


Figure 6. Tokamaks such as the one shown in the figure are being studied with the goal of economical production of energy by nuclear fusion. Magnetic fields in the doughnut-shaped device contain and direct the reactive charged particles. (credit: David Mellis, Flickr)

Mass spectrometers have a variety of designs, and many use magnetic fields to measure mass. The curvature of a charged particle's path in the field is related to its mass and is measured to obtain mass information. (See More Applications of Magnetism.) Historically, such techniques were employed in the first direct observations of electron charge and mass. Today, mass spectrometers (sometimes coupled with gas chromatographs) are used to determine the make-up and sequencing of large biological molecules.

Section Summary

- Magnetic force can supply centripetal force and cause a charged particle to move in a circular path of radius

$$r = \frac{mv}{qB}$$

where v is the component of the velocity perpendicular to B for a charged particle with mass m and charge q.

Conceptual Questions

1. How can the motion of a charged particle be used to distinguish between a magnetic and an electric field?

2. High-velocity charged particles can damage biological cells and are a component of radiation exposure in a variety of locations ranging from research facilities to natural background. Describe how you could use a magnetic field to shield yourself.

3. If a cosmic ray proton approaches the Earth from outer space along a line toward the center of the Earth that lies in the plane of the equator, in what direction will it be deflected by the Earth's magnetic field? What about an electron? A neutron?

4. What are the signs of the charges on the particles in Figure 9?

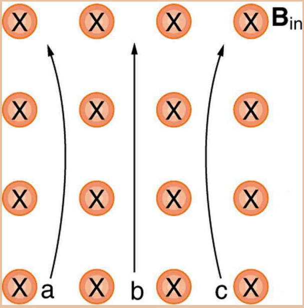


Figure 9.

5. Which of the particles in Figure 10 has the greatest velocity, assuming they have identical charges and masses?

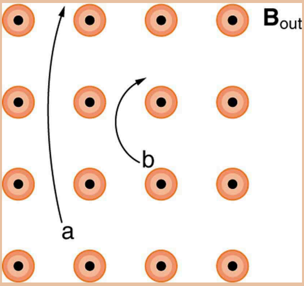


Figure 10.

6. Which of the particles in Figure 10 has the greatest mass, assuming all have identical charges and velocities?

7. While operating, a high-precision TV monitor is placed on its side during maintenance. The image on the monitor

changes color and blurs slightly. Discuss the possible relation of these effects to the Earth's magnetic field.

Problems & Exercises

If you need additional support for these problems, see More Applications of Magnetism.

1. A cosmic ray electron moves at 7.50×10^{6} m/s perpendicular to the Earth's magnetic field at an altitude where field strength is 1.00×10^{-5} T. What is the radius of the circular path the electron follows?

2. A proton moves at 7.50×10^{7} perpendicular to a magnetic field. The field causes the proton to travel in a circular path of radius 0.800 m. What is the field strength?

3. (a) Viewers of *Star Trek* hear of an antimatter drive on the Starship *Enterprise*. One possibility for such a futuristic energy source is to store antimatter charged particles in a vacuum chamber, circulating in a magnetic field, and then extract them as needed. Antimatter annihilates with normal matter, producing pure energy. What strength magnetic field is needed to hold antiprotons, moving at 5.00×10^{7} m/s in a circular path 2.00 m in radius? Antiprotons have the same mass as protons but the opposite (negative) charge. (b) Is this field strength obtainable with today's technology or is it a futuristic possibility?

4. (a) An oxygen-16 ion with a mass of 2.66×10^{-26} kg

travels at 5.00×10^6 m/s perpendicular to a 1.20-T magnetic field, which makes it move in a circular arc with a 0.231-m radius. What positive charge is on the ion? (b) What is the ratio of this charge to the charge of an electron? (c) Discuss why the ratio found in (b) should be an integer.

5. What radius circular path does an electron travel if it moves at the same speed and in the same magnetic field as the proton in number 2?

6. A velocity selector in a mass spectrometer uses a 0.100-T magnetic field. (a) What electric field strength is needed to select a speed of 4.00×10^6 m/s? (b) What is the voltage between the plates if they are separated by 1.00 cm?

7. An electron in a TV CRT moves with a speed of 6.00×10^7 m/s, in a direction perpendicular to the Earth's field, which has a strength of 5.00×10^{-5} T. (a) What strength electric field must be applied perpendicular to the Earth's field to make the electron moves in a straight line? (b) If this is done between plates separated by 1.00 cm, what is the voltage applied? (Note that TVs are usually surrounded by a ferromagnetic material to shield against external magnetic fields and avoid the need for such a correction.)

8. (a) At what speed will a proton move in a circular path of the same radius as the electron in question 2? (b) What would the radius of the path be if the proton had the same speed as the electron? (c) What would the radius be if the proton had the same kinetic energy as the electron? (d) The same momentum?

9. A mass spectrometer is being used to separate common oxygen-16 from the much rarer oxygen-18, taken from a sample of old glacial ice. (The relative abundance of these oxygen isotopes is related to climatic temperature at

the time the ice was deposited.) The ratio of the masses of these two ions is 16 to 18, the mass of oxygen-16 is 2.66×10^{-26} kg, and they are singly charged and travel at 5.00×10^{6} m/s in a 1.20-T magnetic field. What is the separation between their paths when they hit a target after traversing a semicircle?

10. (a) Triply charged uranium-235 and uranium-238 ions are being separated in a mass spectrometer. (The much rarer uranium-235 is used as reactor fuel.) The masses of the ions are 3.90×10^{-25} kg and 3.95×10^{-25} kg, respectively, and they travel at 3.00×10^{5} m/s in a 0.250-T field. What is the separation between their paths when they hit a target after traversing a semicircle? (b) Discuss whether this distance between their paths seems to be big enough to be practical in the separation of uranium-235 from uranium-238.

Selected Solutions to Problems & Exercises

1. 4.27 m

3. (a) 0.261 T (b) This strength is definitely obtainable with today’s technology. Magnetic field strengths of 0.500 T are obtainable with permanent magnets.

5. 4.36×10^{-4} m

7. (a) 3.00 kV/m (b) 30.0 V

9. 0.173 m

48. The Hall Effect

Learning Objectives

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

- Describe the Hall effect.
- Calculate the Hall emf across a current-carrying conductor.

We have seen effects of a magnetic field on free-moving charges. The magnetic field also affects charges moving in a conductor. One result is the Hall effect, which has important implications and applications. Figure 1 shows what happens to charges moving through a conductor in a magnetic field. The field is perpendicular to the electron drift velocity and to the width of the conductor. Note that conventional current is to the right in both parts of the figure. In part (a), electrons carry the current and move to the left. In part (b), positive charges carry the current and move to the right. Moving electrons feel a magnetic force toward one side of the conductor, leaving a net positive charge on the other side. This separation of charge *creates a voltage* ε, known as the *Hall emf*, *across* the conductor. The creation of a voltage *across* a current-carrying conductor by a magnetic field is known as the *Hall effect*, after Edwin Hall, the American physicist who discovered it in 1879.

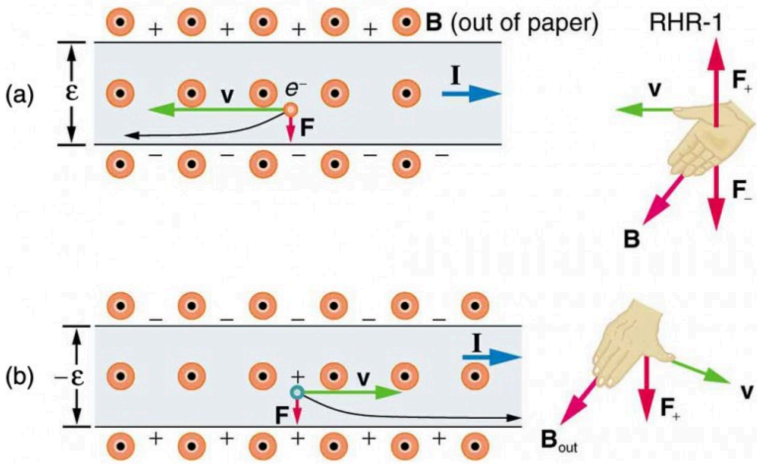


Figure 1. The Hall effect. (a) Electrons move to the left in this flat conductor (conventional current to the right). The magnetic field is directly out of the page, represented by circled dots; it exerts a force on the moving charges, causing a voltage ε, the Hall emf, across the conductor. (b) Positive charges moving to the right (conventional current also to the right) are moved to the side, producing a Hall emf of the opposite sign, −ε. Thus, if the direction of the field and current are known, the sign of the charge carriers can be determined from the Hall effect.

One very important use of the Hall effect is to determine whether positive or negative charges carries the current. Note that in Figure 1(b), where positive charges carry the current, the Hall emf has the sign opposite to when negative charges carry the current. Historically, the Hall effect was used to show that electrons carry current in metals and it also shows that positive charges carry current in some semiconductors. The Hall effect is used today as a research tool to probe the movement of charges, their drift velocities and densities, and so on, in materials. In 1980, it was discovered that the Hall effect is quantized, an example of quantum behavior in a macroscopic object.

The Hall effect has other uses that range from the determination of blood flow rate to precision measurement of magnetic field strength. To examine these quantitatively, we need an expression for the Hall emf, ε, across a conductor. Consider the balance of forces on a moving charge in a situation where B, v, and l are mutually perpendicular, such as shown in Figure 2. Although the magnetic force moves negative charges to one side, they cannot build up without limit. The electric field caused by their separation opposes the magnetic force, $F = qvB$, and the electric force, $F_e = qE$, eventually grows to equal it. That is,

$$qE = qvB$$

or

$$E = vB.$$

Note that the electric field E is uniform across the conductor because the magnetic field B is uniform, as is the conductor. For a uniform electric field, the relationship between electric field and voltage is $E = \varepsilon/l$, where l is the width of the conductor and ε is the Hall emf. Entering this into the last expression gives

$$\frac{\epsilon}{l} = vB.$$

Solving this for the Hall emf yields

$$\varepsilon = Blv \text{ (B, v, and l, mutually perpendicular)},$$

where ε is the Hall effect voltage across a conductor of width l through which charges move at a speed v.

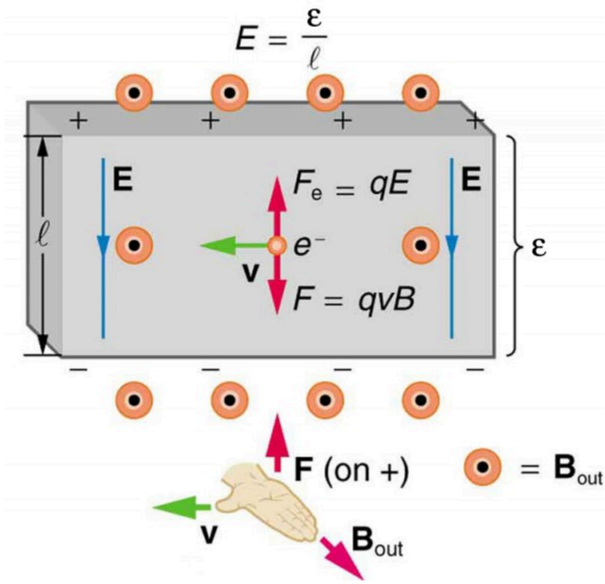


Figure 2. The Hall emf ε produces an electric force that balances the magnetic force on the moving charges. The magnetic force produces charge separation, which builds up until it is balanced by the electric force, an equilibrium that is quickly reached.

One of the most common uses of the Hall effect is in the measurement of magnetic field strength B. Such devices, called *Hall probes*, can be made very small, allowing fine position mapping. Hall probes can also be made very accurate, usually accomplished by careful calibration. Another application of the Hall effect is to measure fluid flow in any fluid that has free charges (most do). (See Figure 3.) A magnetic field applied perpendicular to the flow direction produces a Hall emf ε as shown. Note that the sign of ε depends not on the sign of the charges, but only on the directions of B and v. The magnitude of the Hall emf is $\varepsilon = Blv$, where l is the pipe diameter, so that the average velocity v can be determined from ε providing the other factors are known.

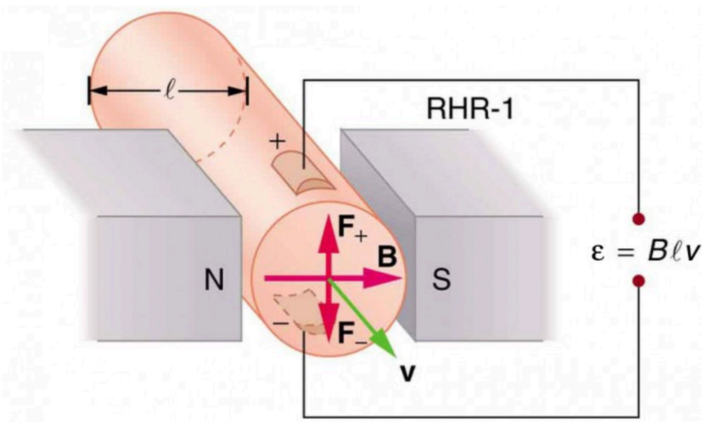


Figure 3. The Hall effect can be used to measure fluid flow in any fluid having free charges, such as blood. The Hall emf ε is measured across the tube perpendicular to the applied magnetic field and is proportional to the average velocity v.

Example 1. Calculating the Hall emf: Hall Effect for Blood Flow

A Hall effect flow probe is placed on an artery, applying a 0.100-T magnetic field across it, in a setup similar to that in Figure 3. What is the Hall emf, given the vessel's inside diameter is 4.00 mm and the average blood velocity is 20.0 cm/s?

Strategy

Because B, v, and l are mutually perpendicular, the equation $\varepsilon = Blv$ can be used to find ε.

Solution

Entering the given values for B, v, and l gives

$$\begin{aligned}\epsilon &= Blv = (0.100\ \text{T})\left(4.00 \times 10^{-3}\ \text{m}\right)(0.200\ \text{m/s}) \\ &= 80.0\ \mu\text{V}\end{aligned}$$

Discussion

This is the average voltage output. Instantaneous voltage varies with pulsating blood flow. The voltage is small in this type of measurement. ε is particularly difficult to measure, because there are voltages associated with heart action (ECG voltages) that are on the order of millivolts. In practice, this difficulty is overcome by applying an AC magnetic field, so that the Hall emf is AC with the same frequency. An amplifier can be very selective in picking out only the appropriate frequency, eliminating signals and noise at other frequencies.

Section Summary

- The Hall effect is the creation of voltage ε, known as the Hall emf, across a current-carrying conductor by a magnetic field.
- The Hall emf is given by

$$\varepsilon = Blv \text{ (B, } v, \textit{ and } l, \textit{ mutually perpendicular)}$$

for a conductor of width l through which charges move at a speed v.

Conceptual Questions

1. Discuss how the Hall effect could be used to obtain information on free charge density in a conductor. (Hint: Consider how drift velocity and current are related.)

Problems & Exercises

1. A large water main is 2.50 m in diameter and the average water velocity is 6.00 m/s. Find the Hall voltage produced if the pipe runs perpendicular to the Earth's 5.00×10^{-5}-T field.

2. What Hall voltage is produced by a 0.200-T field applied across a 2.60-cm-diameter aorta when blood velocity is 60.0 cm/s?

3. (a) What is the speed of a supersonic aircraft with a 17.0-m wingspan, if it experiences a 1.60-V Hall voltage between its wing tips when in level flight over the north magnetic pole, where the Earth's field strength is 8.00×10^{-5} T? (b) Explain why very little current flows as a result of this Hall voltage.

4. A nonmechanical water meter could utilize the Hall effect by applying a magnetic field across a metal pipe and measuring the Hall voltage produced. What is the average fluid velocity in a 3.00-cm-diameter pipe, if a 0.500-T field across it creates a 60.0-mV Hall voltage?

5. Calculate the Hall voltage induced on a patient's heart while being scanned by an MRI unit. Approximate the conducting path on the heart wall by a wire 7.50 cm long that moves at 10.0 cm/s perpendicular to a 1.50-T magnetic field.

6. A Hall probe calibrated to read 1.00 μV when placed in a 2.00-T field is placed in a 0.150-T field. What is its output voltage?

7. Using information in Example 2: Calculating Resistance: Hot-Filament Resistance, what would the Hall voltage be if a 2.00-T field is applied across a 10-gauge copper wire (2.588 mm in diameter) carrying a 20.0-A current?

8. Show that the Hall voltage across wires made of the same material, carrying identical currents, and subjected to the same magnetic field is inversely proportional to their diameters. (Hint: Consider how drift velocity depends on wire diameter.)

9. A patient with a pacemaker is mistakenly being scanned for an MRI image. A 10.0-cm-long section of pacemaker wire moves at a speed of 10.0 cm/s perpendicular to the MRI unit's magnetic field and a 20.0-mV Hall voltage is induced. What is the magnetic field strength?

Glossary

Hall effect:
: the creation of voltage across a current-carrying conductor by a magnetic field

Hall emf:
: the electromotive force created by a current-carrying conductor by a magnetic field, $\varepsilon = Blv$

Selected Solutions to Problems & Exercises

1. 7.50×10^{-4} V

3. (a) 1.18×10^{3} m/s (b) Once established, the Hall emf pushes charges one direction and the magnetic force acts in the opposite direction resulting in no net force on the charges. Therefore, no current flows in the direction of the Hall emf. This is the same as in a current-carrying conductor—current does not flow in the direction of the Hall emf.

5. 11.3 mV

7. 1.16 μV

9. 2.00 T

49. Magnetic Force on a Current-Carrying Conductor

Learning Objectives

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

- Describe the effects of a magnetic force on a current-carrying conductor.
- Calculate the magnetic force on a current-carrying conductor.

Because charges ordinarily cannot escape a conductor, the magnetic force on charges moving in a conductor is transmitted to the conductor itself.

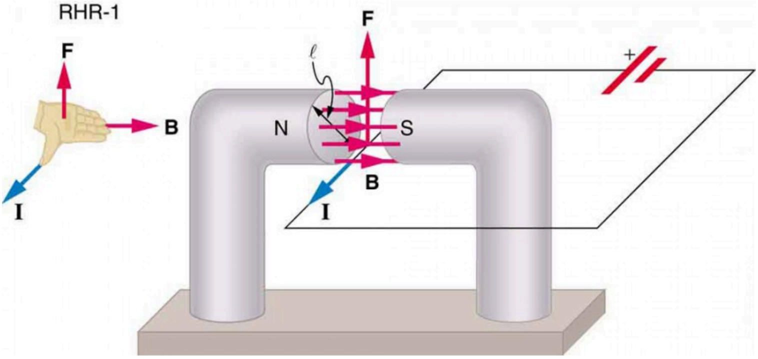


Figure 1. The magnetic field exerts a force on a current-carrying wire in a direction given by the right hand rule 1 (the same direction as that on the individual moving charges). This force can easily be large enough to move the wire, since typical currents consist of very large numbers of moving charges.

We can derive an expression for the magnetic force on a current by taking a sum of the magnetic forces on individual charges. (The forces add because they are in the same direction.) The force on an individual charge moving at the drift velocity v_d is given by F = qv_dB sin θ. Taking B to be uniform over a length of wire l and zero elsewhere, the total magnetic force on the wire is then F = (qv_dB sin θ)(N), where N is the number of charge carriers in the section of wire of length l. Now, N = nV, where n is the number of charge carriers per unit volume and V is the volume of wire in the field. Noting that V = Al, where A is the cross-sectional area of the wire, then the force on the wire is F = (qv_dB sin θ) (nAl). Gathering terms,

$$F = (nqAv_{\mathrm{d}})lB\sin\theta.$$

Because $nqAv_d$ = I (see Current),

$$F = IlB\sin\theta$$

is the equation for *magnetic force on a length l of wire carrying a current I in a uniform magnetic field* B, as shown in Figure 2. If we divide both sides of this expression by l, we find that the magnetic force per unit length of wire in a uniform field is $\frac{F}{l} = IB\sin\theta$. The direction of this force is given by RHR-1, with the thumb in the direction of the current I. Then, with the fingers in the direction of B, a perpendicular to the palm points in the direction of F, as in Figure 2.

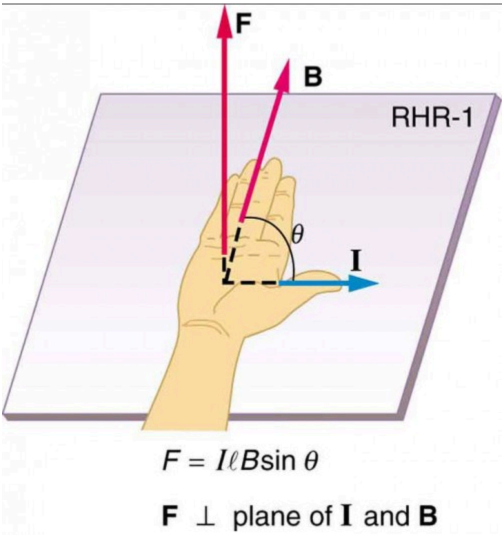


Figure 2. The force on a current-carrying wire in a magnetic field is F = IlB sin θ. Its direction is given by RHR-1.

Example 1. Calculating Magnetic Force on a Current-Carrying Wire: A Strong Magnetic Field

Calculate the force on the wire shown in Figure 1, given B = 1.50 T, l = 5.00 cm, and I = 20.0 A.

Strategy

The force can be found with the given information by using $F = IlB \sin \theta$ and noting that the angle θ between I and B is 90°, so that sin θ = 1.

Solution

Entering the given values into $F = IlB \sin \theta$ yields

$$F = IlB \sin \theta = (20.0 \text{ A})(0.0500 \text{ m})(1.50 \text{ T})(1).$$

The units for tesla are $1 \text{ T} = \frac{\text{N}}{\text{A} \cdot \text{m}}$; thus,

$$F = 1.50 \text{ N}.$$

Discussion

This large magnetic field creates a significant force on a small length of wire.

Magnetic force on current-carrying conductors is used to convert electric energy to work. (Motors are a prime example—they employ loops of wire and are considered in the next section.) Magnetohydrodynamics (MHD) is the technical name given to a clever application where magnetic force pumps fluids without moving mechanical parts. (See Figure 3.)

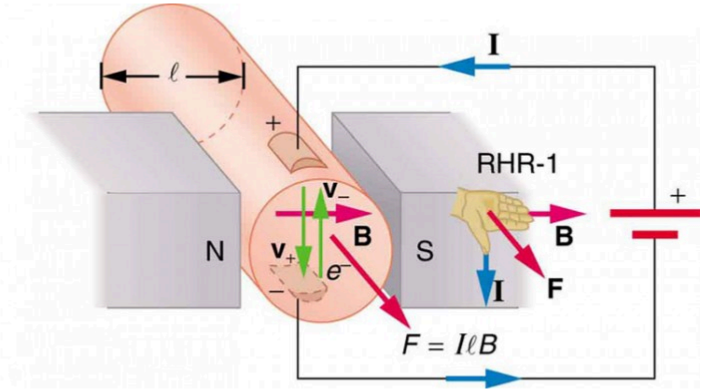


Figure 3. Magnetohydrodynamics. The magnetic force on the current passed through this fluid can be used as a nonmechanical pump.

A strong magnetic field is applied across a tube and a current is passed through the fluid at right angles to the field, resulting in a force on the fluid parallel to the tube axis as shown. The absence of moving parts makes this attractive for moving a hot, chemically active substance, such as the liquid sodium employed in some nuclear reactors. Experimental artificial hearts are testing with this technique for pumping blood, perhaps circumventing the adverse effects of mechanical pumps. (Cell membranes, however, are affected by the large fields needed in MHD, delaying its practical application in humans.) MHD propulsion for nuclear submarines has been proposed, because it could be considerably quieter than conventional propeller drives. The deterrent value of nuclear submarines is based on their ability to hide and survive a first or second nuclear strike. As we slowly disassemble our nuclear weapons arsenals, the submarine branch will be the last to be decommissioned because of this ability (See Figure 4.) Existing MHD drives are heavy and inefficient—much development work is needed.

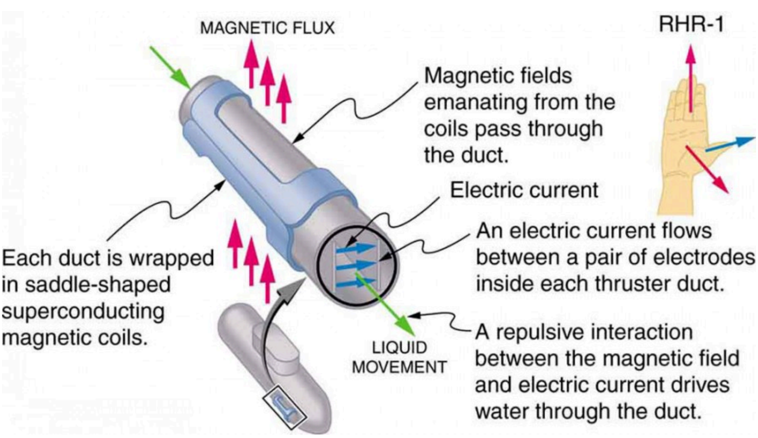


Figure 4. An MHD propulsion system in a nuclear submarine could produce significantly less turbulence than propellers and allow it to run more silently. The development of a silent drive submarine was dramatized in the book and the film The Hunt for Red October.

Section Summary

- The magnetic force on current-carrying conductors is given by
$$F = IlB\sin\theta$$
where I is the current, l is the length of a straight conductor in a uniform magnetic field B, and θ is the angle between I and B. The force follows RHR-1 with the thumb in the direction of I.

Conceptual Questions

1. Draw a sketch of the situation in Figure 1 showing the direction of electrons carrying the current, and use RHR-1 to verify the direction of the force on the wire.
2. Verify that the direction of the force in an MHD drive, such as that in Figure 3, does not depend on the sign of the charges carrying the current across the fluid.
3. Why would a magnetohydrodynamic drive work better in ocean water than in fresh water? Also, why would superconducting magnets be desirable?
4. Which is more likely to interfere with compass readings, AC current in your refrigerator or DC current when you start your car? Explain.

Problems & Exercises

1. What is the direction of the magnetic force on the current in each of the six cases in Figure 5?

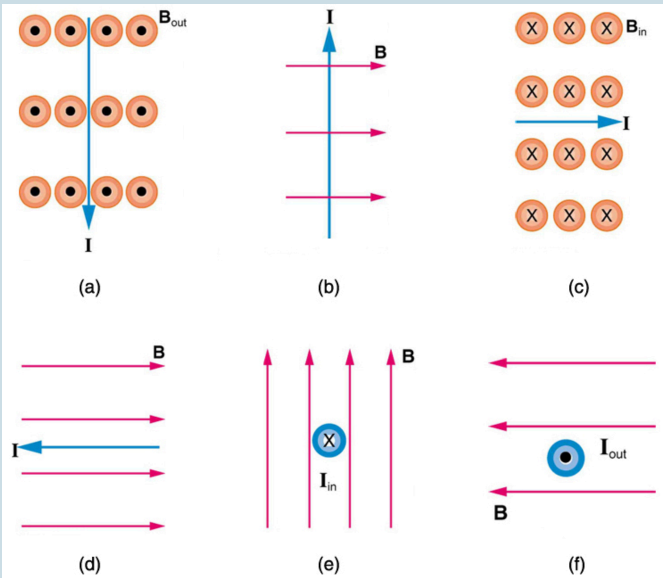


Figure 5.

2. What is the direction of a current that experiences the magnetic force shown in each of the three cases in Figure 6, assuming the current runs perpendicular to B?

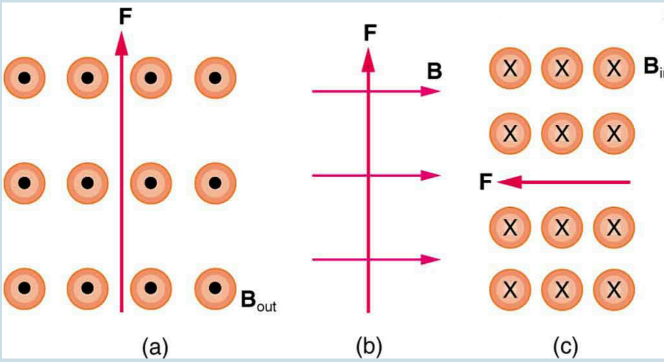


Figure 6.

3. What is the direction of the magnetic field that produces the magnetic force shown on the currents in each of the three cases in Figure 7, assuming **B** is perpendicular to **I**?

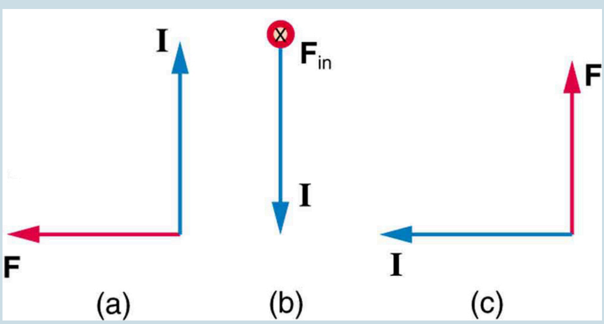


Figure 7.

4. (a) What is the force per meter on a lightning bolt at the equator that carries 20,000 A perpendicular to the

Earth's 3.00×10^{-5}-T field? (b) What is the direction of the force if the current is straight up and the Earth's field direction is due north, parallel to the ground?

5. (a) A DC power line for a light-rail system carries 1000 A at an angle of 30° to the Earth's 5.00×10^{-5}-T field. What is the force on a 100-m section of this line? (b) Discuss practical concerns this presents, if any.

6. What force is exerted on the water in an MHD drive utilizing a 25.0-cm-diameter tube, if 100-A current is passed across the tube that is perpendicular to a 2.00-T magnetic field? (The relatively small size of this force indicates the need for very large currents and magnetic fields to make practical MHD drives.)

7. A wire carrying a 30.0-A current passes between the poles of a strong magnet that is perpendicular to its field and experiences a 2.16-N force on the 4.00 cm of wire in the field. What is the average field strength?

8. (a) A 0.750-m-long section of cable carrying current to a car starter motor makes an angle of 60° with the Earth's 5.50×10^{-5} T field. What is the current when the wire experiences a force of 7.00×10^{-3} N? (b) If you run the wire between the poles of a strong horseshoe magnet, subjecting 5.00 cm of it to a 1.75-T field, what force is exerted on this segment of wire?

9. (a) What is the angle between a wire carrying an 8.00-A current and the 1.20-T field it is in if 50.0 cm of the wire experiences a magnetic force of 2.40 N? (b) What is the force on the wire if it is rotated to make an angle of 90° with the field?

10. The force on the rectangular loop of wire in the magnetic field in Figure 8 can be used to measure field strength. The field is uniform, and the plane of the loop is perpendicular to the field. (a) What is the direction of the magnetic force on the loop? Justify the claim that the forces on the sides of the loop are equal and opposite, independent of how much of the loop is in the field and do not affect the net force on the loop. (b) If a current of 5.00 A is used, what is the force per tesla on the 20.0-cm-wide loop?

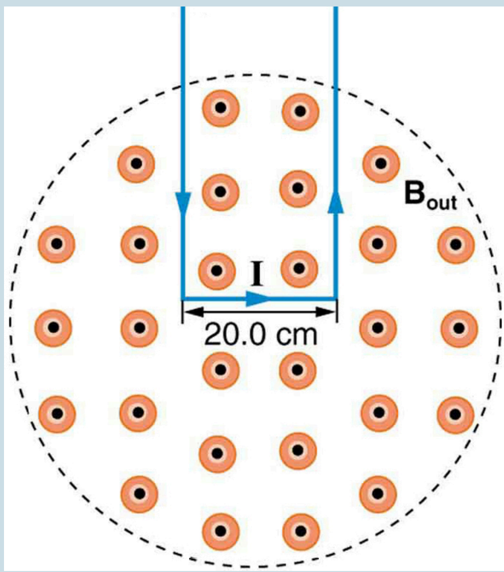


Figure 8. A rectangular loop of wire carrying a current is perpendicular to a magnetic field. The field is uniform in the region shown and is zero outside that region.

Selected Solutions to Problems & Exercises

1. (a) west (left)

(b) into page

(c) north (up)

(d) no force

(e) east (right)

(f) south (down)

3. (a) into page

(b) west (left)

(c) out of page

5. (a) 2.50 N (b) This is about half a pound of force per 100 m of wire, which is much less than the weight of the wire itself. Therefore, it does not cause any special concerns.

7. 1.80 T

9. (a) 30° (b) 4.80 N

50. Torque on a Current Loop: Motors and Meters

Learning Objectives

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

- Describe how motors and meters work in terms of torque on a current loop.
- Calculate the torque on a current-carrying loop in a magnetic field.

Motors are the most common application of magnetic force on current-carrying wires. Motors have loops of wire in a magnetic field. When current is passed through the loops, the magnetic field exerts torque on the loops, which rotates a shaft. Electrical energy is converted to mechanical work in the process. (See Figure 1.)

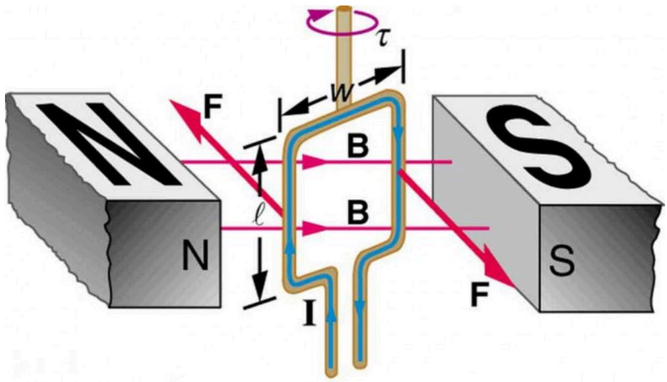


Figure 1. Torque on a current loop. A current-carrying loop of wire attached to a vertically rotating shaft feels magnetic forces that produce a clockwise torque as viewed from above.

Let us examine the force on each segment of the loop in Figure 1 to find the torques produced about the axis of the vertical shaft. (This will lead to a useful equation for the torque on the loop.) We take the magnetic field to be uniform over the rectangular loop, which has width w and height l. First, we note that the forces on the top and bottom segments are vertical and, therefore, parallel to the shaft, producing no torque. Those vertical forces are equal in magnitude and opposite in direction, so that they also produce no net force on the loop. Figure 2 shows views of the loop from above. Torque is defined as $\tau = rF \sin \theta$, where F is the force, r is the distance from the pivot that the force is applied, and θ is the angle between r and F. As seen in Figure 2(a), right hand rule 1 gives the forces on the sides to be equal in magnitude and opposite in direction, so that the net force is again zero. However, each force produces a clockwise torque. Since $r = w/2$, the torque on each vertical segment is $(w/2)F \sin \theta$, and the two add to give a total torque.

$$\tau = \frac{w}{2} F \sin \theta + \frac{w}{2} F \sin \theta = wF \sin \theta$$

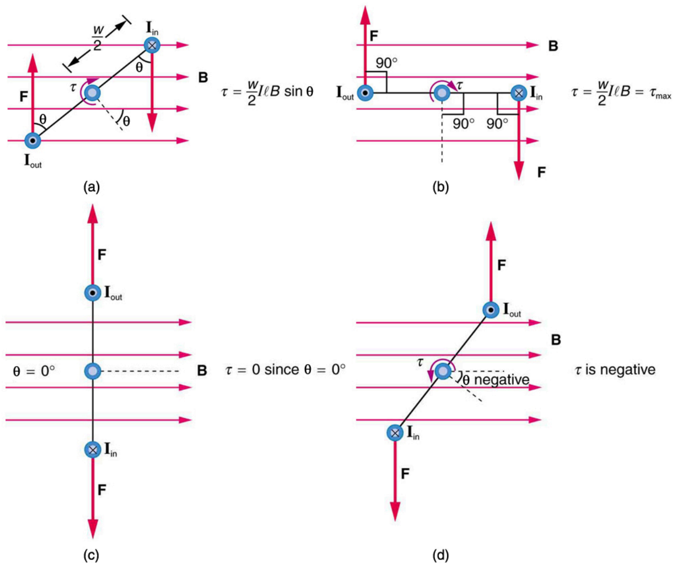


Figure 2. Top views of a current-carrying loop in a magnetic field. (a) The equation for torque is derived using this view. Note that the perpendicular to the loop makes an angle θ with the field that is the same as the angle between w/2 and F. (b) The maximum torque occurs when θ is a right angle and sin θ = 1. (c) Zero (minimum) torque occurs when θ is zero and sin θ = 0. (d) The torque reverses once the loop rotates past θ = 0.

Now, each vertical segment has a length *l* that is perpendicular to B, so that the force on each is $F = IlB$. Entering F into the expression for torque yields

$$\tau = wIlB \sin\theta.$$

If we have a multiple loop of N turns, we get N times the torque of one loop. Finally, note that the area of the loop is A = *wl*; the expression for the torque becomes

$$\tau = NIAB \sin\theta.$$

This is the torque on a current-carrying loop in a uniform magnetic field. This equation can be shown to be valid for a loop of any shape. The loop carries a current I, has N turns, each of area A, and the perpendicular to the loop makes an angle θ with the field B. The net force on the loop is zero.

Example 1. Calculating Torque on a Current-Carrying Loop in a Strong Magnetic Field

Find the maximum torque on a 100-turn square loop of a wire of 10.0 cm on a side that carries 15.0 A of current in a 2.00-T field.

Strategy

Torque on the loop can be found using $\tau = NIAB\sin\theta$. Maximum torque occurs when θ = 90° and sin θ = 1.

Solution

For sin θ = 1, the maximum torque is

$$\tau_{\text{max}} = NIAB$$

Entering known values yields

$$\begin{aligned}\tau_{\text{max}} &= (100)\,(15.0\text{ A})\left(0.100\text{ m}^2\right)(2.00\text{ T})\\ &= 30.0\text{ N}\cdot\text{m}\end{aligned}$$

.

Discussion

This torque is large enough to be useful in a motor.

The torque found in the preceding example is the maximum. As the coil rotates, the torque decreases to zero at $\theta = 0$. The torque then *reverses* its direction once the coil rotates past $\theta = 0$. (See Figure 2(d).) This means that, unless we do something, the coil will oscillate back and forth about equilibrium at $\theta = 0$. To get the coil to continue rotating in the same direction, we can reverse the current as it passes through $\theta = 0$ with automatic switches called *brushes*. (See Figure 3.)

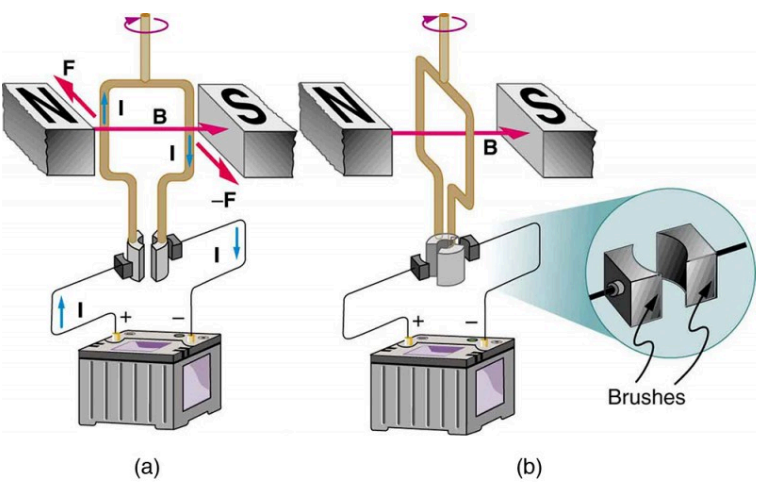


Figure 3. (a) As the angular momentum of the coil carries it through $\theta = 0$, the brushes reverse the current to keep the torque clockwise. (b) The coil will rotate continuously in the clockwise direction, with the current reversing each half revolution to maintain the clockwise torque.

Meters, such as those in analog fuel gauges on a car, are another common application of magnetic torque on a current-carrying loop. Figure 4 shows that a meter is very similar in construction to a motor. The meter in the figure has its magnets shaped to limit the effect of θ by making B perpendicular to the loop over a large angular range. Thus the torque is proportional to I and not θ. A linear spring exerts a counter-torque that balances the current-produced torque. This makes the needle deflection proportional to I. If an exact proportionality cannot be achieved, the gauge reading can be calibrated. To produce a galvanometer for use in analog voltmeters and ammeters that have a low resistance and respond to small currents, we use a large loop area A, high magnetic field B, and low-resistance coils.

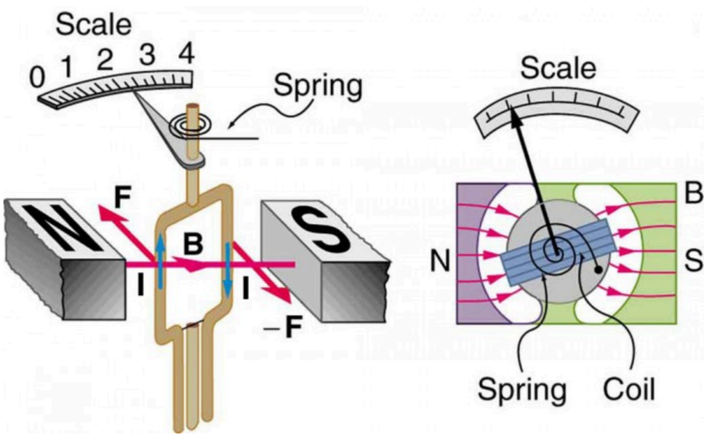


Figure 4. Meters are very similar to motors but only rotate through a part of a revolution. The magnetic poles of this meter are shaped to keep the component of B perpendicular to the loop constant, so that the torque does not depend on θ and the deflection against the return spring is proportional only to the current I.

Section Summary

- The torque τ on a current-carrying loop of any shape in a uniform magnetic field. is

$$\tau = NIAB\sin\theta,$$

where N is the number of turns, I is the current, A is the area of the loop, B is the magnetic field strength, and θ is the angle between the perpendicular to the loop and the magnetic field.

Conceptual Questions

1. Draw a diagram and use RHR-1 to show that the forces on the top and bottom segments of the motor's current loop in Figure 1 are vertical and produce no torque about the axis of rotation.

Problems & Exercises

1. (a) By how many percent is the torque of a motor decreased if its permanent magnets lose 5.0% of their strength? (b) How many percent would the current need to be increased to return the torque to original values?

2. (a) What is the maximum torque on a 150-turn square loop of wire 18.0 cm on a side that carries a 50.0-A current in a 1.60-T field? (b) What is the torque when θ is 10.9°?

3. Find the current through a loop needed to create a maximum torque of 9.00 N. The loop has 50 square turns that are 15.0 cm on a side and is in a uniform 0.800-T magnetic field.

4. Calculate the magnetic field strength needed on a 200-turn square loop 20.0 cm on a side to create a maximum torque of 300 N · m if the loop is carrying 25.0 A.

5. Since the equation for torque on a current-carrying loop is $\tau = NIAB\sin\theta$, the units of N · m must equal units of A · m^2 T. Verify this.

6. (a) At what angle θ is the torque on a current loop 90.0% of maximum? (b) 50.0% of maximum? (c) 10.0% of maximum?

7. A proton has a magnetic field due to its spin on its axis. The field is similar to that created by a circular current loop 0.650×10^{-15} m in radius with a current of 1.05×10^{4} A (no kidding). Find the maximum torque on a proton in a 2.50-T field. (This is a significant torque on a small particle.)

8. (a) A 200-turn circular loop of radius 50.0 cm is vertical, with its axis on an east-west line. A current of 100 A circulates clockwise in the loop when viewed from the east. The Earth's field here is due north, parallel to the ground, with a strength of 3.00×10^{-5} T. What are the direction and magnitude of the torque on the loop? (b) Does this device have any practical applications as a motor?

Glossary

motor:
: loop of wire in a magnetic field; when current is passed through the loops, the magnetic field exerts torque on the loops, which rotates a shaft; electrical energy is converted to mechanical work in the process

meter:

common application of magnetic torque on a current-carrying loop that is very similar in construction to a motor; by design, the torque is proportional to I and not θ, so the needle deflection is proportional to the current

Exercises

1. (a) τ decreases by 5.00% if B decreases by 5.00% (b) 5.26% increase

3. 10.0 A

5. $\mathbf{A} \cdot \mathbf{m}^2 \cdot \mathbf{T} = \mathbf{A} \cdot \mathbf{m}^2 \left(\frac{\mathbf{N}}{\mathbf{A} \cdot \mathbf{m}} \right) = \mathbf{N} \cdot \mathbf{m}$

7. 3.48×10^{-26} N · m

51. Magnetic Fields Produced by Currents: Ampere's Law

Learning Objectives

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

- Calculate current that produces a magnetic field.
- Use the right hand rule 2 to determine the direction of current or the direction of magnetic field loops.

How much current is needed to produce a significant magnetic field, perhaps as strong as the Earth's field? Surveyors will tell you that overhead electric power lines create magnetic fields that interfere with their compass readings. Indeed, when Oersted discovered in 1820 that a current in a wire affected a compass needle, he was not dealing with extremely large currents. How does the shape of wires carrying current affect the shape of the magnetic field created? We noted earlier that a current loop created a magnetic field similar to that of a bar magnet, but what about a straight wire or a toroid (doughnut)? How is the direction of a current-created field related to the direction of the current? Answers to these questions are explored in this section, together with a brief discussion of the law governing the fields created by currents.

Magnetic Field Created by a Long Straight Current-Carrying Wire: Right Hand Rule 2

Magnetic fields have both direction and magnitude. As noted before, one way to explore the direction of a magnetic field is with compasses, as shown for a long straight current-carrying wire in Figure 1. Hall probes can determine the magnitude of the field. The field around a long straight wire is found to be in circular loops. The *right hand rule 2* (RHR-2) emerges from this exploration and is valid for any current segment—*point the thumb in the direction of the current, and the fingers curl in the direction of the magnetic field loops* created by it.

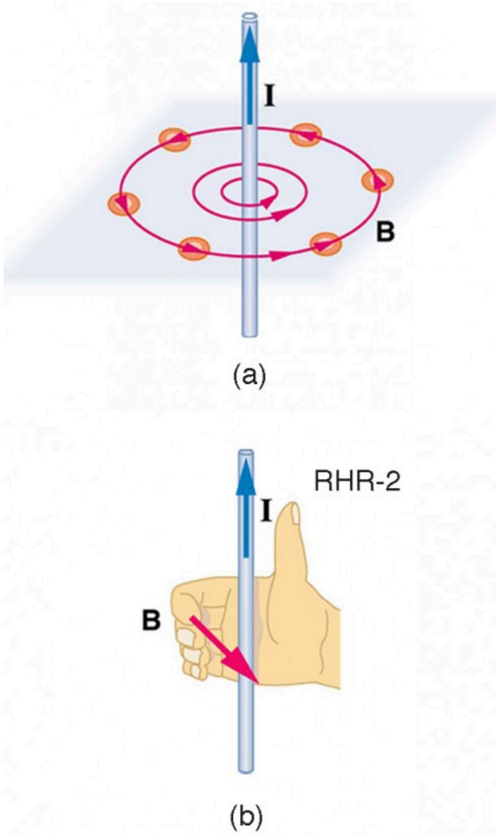


Figure 1. (a) Compasses placed near a long straight current-carrying wire indicate that field lines form circular loops centered on the wire. (b) Right hand rule 2 states that, if the right hand thumb points in the direction of the current, the fingers curl in the direction of the field. This rule is consistent with the field mapped for the long straight wire and is valid for any current segment.

The *magnetic field strength (magnitude) produced by a long straight current-carrying wire* is found by experiment to be

$$B = \frac{\mu_0 I}{2\pi r}\left(\text{long straight wire}\right),$$

where *I* is the current, *r* is the shortest distance to the wire, and the constant $\mu_0 = 4\pi \times 10^{-7} \mathrm{T} \cdot \mathrm{m/A}$ is the *permeability of free space*. (μ_0 is one of the basic constants in nature. We will see later that μ_0 is related to the speed of light.) Since the wire is very long, the magnitude of the field depends only on distance from the wire *r*, not on position along the wire.

Example 1. Calculating Current that Produces a Magnetic Field

Find the current in a long straight wire that would produce a magnetic field twice the strength of the Earth's at a distance of 5.0 cm from the wire.

Strategy

The Earth's field is about 5.0×10^{-5} T, and so here *B* due to the wire is taken to be 1.0×10^{-4} T. The equation $B = \frac{\mu_0 I}{2\pi r}$ can be used to find I, since all other quantities are known.

Solution

Solving for I and entering known values gives

$$\begin{aligned} I &= \frac{2\pi r B}{\mu_0} = \frac{2\pi\left(5.0\times10^{-2}\ \mathrm{m}\right)\left(1.0\times10^{-4}\ \mathrm{T}\right)}{4\pi\times10^{-7}\ \mathrm{T{\cdot}m/A}} \\ &= 25\ \mathrm{A} \end{aligned}$$

Discussion

So a moderately large current produces a significant magnetic field at a distance of 5.0 cm from a long straight wire. Note that the answer is stated to only two digits, since the Earth's field is specified to only two digits in this example.

Ampere's Law and Others

The magnetic field of a long straight wire has more implications than you might at first suspect. *Each segment of current produces a magnetic field like that of a long straight wire, and the total field of any shape current is the vector sum of the fields due to each segment.* The formal statement of the direction and magnitude of the field due to each segment is called the *Biot-Savart law*. Integral calculus is needed to sum the field for an arbitrary shape current. This results in a more complete law, called *Ampere's law*, which relates magnetic field and current in a general way. Ampere's law in turn is a part of *Maxwell's equations*, which give a complete theory of all electromagnetic phenomena. Considerations of how Maxwell's equations appear to different observers led to the modern theory of relativity, and the realization that electric and magnetic fields are different manifestations of the same thing. Most of this is beyond the scope of this text in both mathematical level, requiring calculus, and in the amount of space that can be devoted to it. But for the interested student, and particularly for those who continue in physics, engineering, or similar pursuits, delving into these matters further will reveal descriptions of nature that are elegant as well as profound. In this text, we shall keep the general features in mind, such as RHR-2 and the rules for magnetic field lines listed in Magnetic Fields and Magnetic Field Lines, while concentrating on the fields created in certain important situations.

Making Connections: Relativity

Hearing all we do about Einstein, we sometimes get the impression that he invented relativity out of nothing. On the contrary, one of Einstein's motivations was to solve difficulties in knowing how different observers see magnetic and electric fields.

Magnetic Field Produced by a Current-Carrying Circular Loop

The magnetic field near a current-carrying loop of wire is shown in Figure 2. Both the direction and the magnitude of the magnetic field produced by a current-carrying loop are complex. RHR-2 can be used to give the direction of the field near the loop, but mapping with compasses and the rules about field lines given in Magnetic Fields and Magnetic Field Lines are needed for more detail. There is a simple formula for the *magnetic field strength at the center of a circular loop*. It is

$$B=\frac{\mu_0 I}{2R}(\text{at center of loop}),$$

where R is the radius of the loop. This equation is very similar to that for a straight wire, but it is valid *only* at the center of a circular loop of wire. The similarity of the equations does indicate that similar field strength can be obtained at the center of a loop. One way to get a larger field is to have N loops; then, the field is $B = N\mu_0 I/(2R)$. Note that the larger the loop, the smaller the field at its center, because the current is farther away.

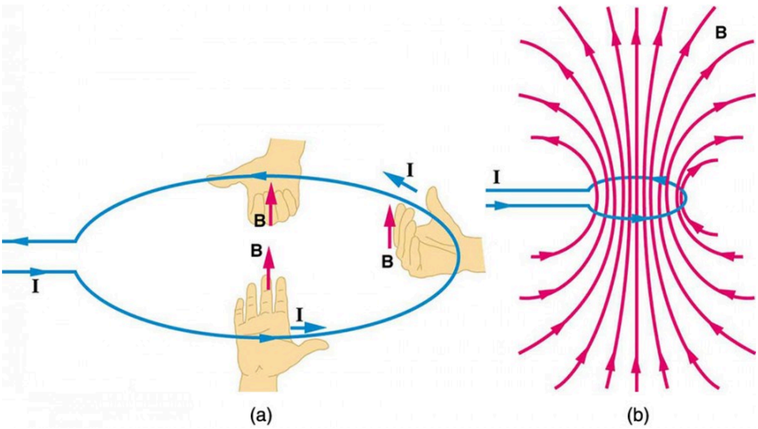


Figure 2. (a) RHR-2 gives the direction of the magnetic field inside and outside a current-carrying loop. (b) More detailed mapping with compasses or with a Hall probe completes the picture. The field is similar to that of a bar magnet.

Magnetic Field Produced by a Current-Carrying Solenoid

A *solenoid* is a long coil of wire (with many turns or loops, as opposed to a flat loop). Because of its shape, the field inside a solenoid can be very uniform, and also very strong. The field just outside the coils is nearly zero. Figure 3 shows how the field looks and how its direction is given by RHR-2.

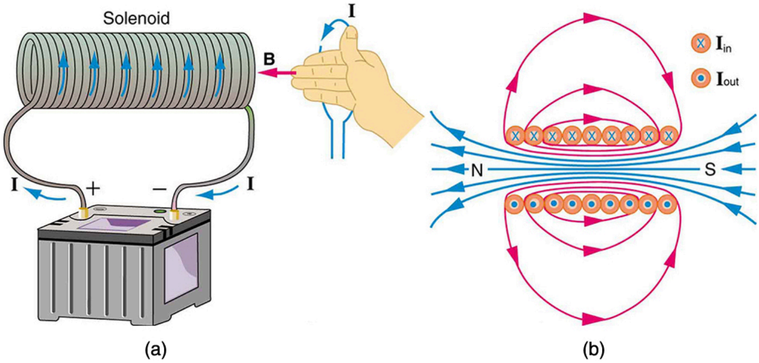


Figure 3. (a) Because of its shape, the field inside a solenoid of length l is remarkably uniform in magnitude and direction, as indicated by the straight and uniformly spaced field lines. The field outside the coils is nearly zero. (b) This cutaway shows the magnetic field generated by the current in the solenoid.

The magnetic field inside of a current-carrying solenoid is very uniform in direction and magnitude. Only near the ends does it begin to weaken and change direction. The field outside has similar complexities to flat loops and bar magnets, but the *magnetic field strength inside a solenoid* is simply

$$B = \mu_0 nI \text{ (inside a solenoid)},$$

where n is the number of loops per unit length of the solenoid (n = N/l, with N being the number of loops and l the length). Note that B is the field strength anywhere in the uniform region of the interior and not just at the center. Large uniform fields spread over a large volume are possible with solenoids, as Example 2 implies.

Example 2. Calculating Field Strength inside a Solenoid

What is the field inside a 2.00-m-long solenoid that has 2000 loops and carries a 1600-A current?

Strategy

To find the field strength inside a solenoid, we use $B=\mu_0 nI$. First, we note the number of loops per unit length is

$$n=\frac{N}{l}=\frac{2000}{2.00\text{ m}}=1000\text{ m}^{-1}=10\text{ cm}^{-1}.$$

Solution Substituting known values gives

$$\begin{aligned} B &= \mu_0 nI=\left(4\pi\times10^{-7}\text{ T}\cdot\text{m/A}\right)\left(1000\text{ m}^{-1}\right)\left(1600\text{ A}\right) \\ &= 2.01\text{ T} \end{aligned}$$

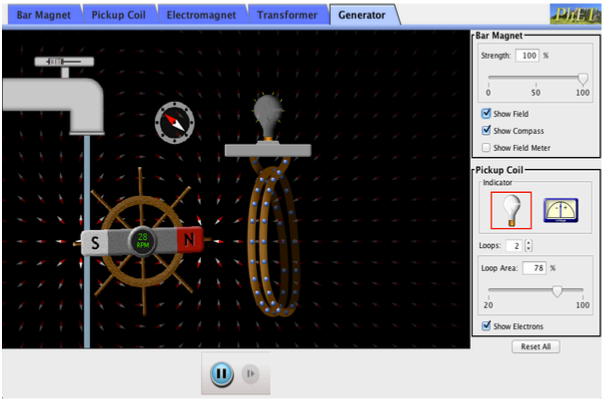
Discussion

This is a large field strength that could be established over a large-diameter solenoid, such as in medical uses of magnetic resonance imaging (MRI). The very large current is an indication that the fields of this strength are not easily achieved, however. Such a large current through 1000 loops squeezed into a meter's length would produce significant heating. Higher currents can be achieved by using superconducting wires, although this is expensive. There is an upper limit to the current, since the superconducting state is disrupted by very large magnetic fields.

There are interesting variations of the flat coil and solenoid. For example, the toroidal coil used to confine the reactive particles in tokamaks is much like a solenoid bent into a circle. The field inside a toroid is very strong but circular. Charged particles travel in circles, following the field lines, and collide with one another, perhaps inducing fusion. But the charged particles do not cross field lines and escape the toroid. A whole range of coil shapes are used to produce all sorts of magnetic field shapes. Adding ferromagnetic materials produces greater field strengths and can have a significant effect on the shape of the field. Ferromagnetic materials tend to trap magnetic fields (the field lines bend into the ferromagnetic material, leaving weaker fields outside it) and are used as shields for devices that are adversely affected by magnetic fields, including the Earth's magnetic field.

PhET Explorations: Generator

Generate electricity with a bar magnet! Discover the physics behind the phenomena by exploring magnets and how you can use them to make a bulb light.



Click to download the simulation. Run using Java.

Section Summary

- The strength of the magnetic field created by current in a long straight wire is given by
$$B = \frac{\mu_0 I}{2\pi r} (\text{long straight wire})$$
where I is the current, r is the shortest distance to the wire, and the constant $\mu_0 = 4\pi \times 10^{-7}\ \text{T} \cdot \text{m/A}$ is the permeability of free space.
- The direction of the magnetic field created by a long straight wire is given by right hand rule 2 (RHR-2): *Point the thumb of the right hand in the direction of current, and the fingers curl in the direction of the magnetic field loops* created by it.

- The magnetic field created by current following any path is the sum (or integral) of the fields due to segments along the path (magnitude and direction as for a straight wire), resulting in a general relationship between current and field known as Ampere's law.
- The magnetic field strength at the center of a circular loop is given by

  $$B = \frac{\mu_0 I}{2R} \text{(at center of loop)}$$

  where R is the radius of the loop. This equation becomes $B = \mu_0 nI/(2R)$ for a flat coil of N loops. RHR-2 gives the direction of the field about the loop. A long coil is called a solenoid.
- The magnetic field strength inside a solenoid is

  $$B = \mu_0 \text{nI (inside a solenoid)}$$

  where n is the number of loops per unit length of the solenoid. The field inside is very uniform in magnitude and direction.

Conceptual Questions

1. Make a drawing and use RHR-2 to find the direction of the magnetic field of a current loop in a motor (such as in Figure 1 from Torque on a Current Loop). Then show that the direction of the torque on the loop is the same as produced by like poles repelling and unlike poles attracting.

Glossary

right hand rule 2 (RHR-2):
: a rule to determine the direction of the magnetic field induced by a current-carrying wire: Point the thumb of the right hand in the direction of current, and the fingers curl in the direction of the magnetic field loops

magnetic field strength (magnitude) produced by a long straight current-carrying wire:
: defined as $B = \frac{\mu_0 I}{2\pi r}$, where *I* is the current, *r* is the shortest distance to the wire, and μ_0 is the permeability of free space

permeability of free space:
: the measure of the ability of a material, in this case free space, to support a magnetic field; the constant $\mu_0 = 4\pi \times 10^{-7} T \cdot \mathrm{m/A}$

magnetic field strength at the center of a circular loop:
: defined as $B = \frac{\mu_0 I}{2R}$ where R is the radius of the loop

solenoid:
: a thin wire wound into a coil that produces a magnetic field when an electric current is passed through it

magnetic field strength inside a solenoid:
: defined as $B = \mu_0 \mathrm{nI}$ where *n* is the number of loops per unit length of the solenoid *n* = N/*l*, with N being the number of loops and *l* the length)

Biot-Savart law:
: a physical law that describes the magnetic field generated by an electric current in terms of a specific equation

Ampere's law:
: the physical law that states that the magnetic field around an electric current is proportional to the current; each segment of current produces a magnetic field like that of a long straight

wire, and the total field of any shape current is the vector sum of the fields due to each segment

Maxwell's equations:
: a set of four equations that describe electromagnetic phenomena

52. Magnetic Force between Two Parallel Conductors

Learning Objectives

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

- Describe the effects of the magnetic force between two conductors.
- Calculate the force between two parallel conductors.

You might expect that there are significant forces between current-carrying wires, since ordinary currents produce significant magnetic fields and these fields exert significant forces on ordinary currents. But you might not expect that the force between wires is used to *define* the ampere. It might also surprise you to learn that this force has something to do with why large circuit breakers burn up when they attempt to interrupt large currents.

The force between two long straight and parallel conductors separated by a distance r can be found by applying what we have developed in preceding sections. Figure 1 shows the wires, their currents, the fields they create, and the subsequent forces they exert on one another. Let us consider the field produced by wire 1 and the force it exerts on wire 2 (call the force F_2). The field due to I_1 at a distance r is given to be

$$B_1 = \frac{\mu_0 I_1}{2\pi r}.$$

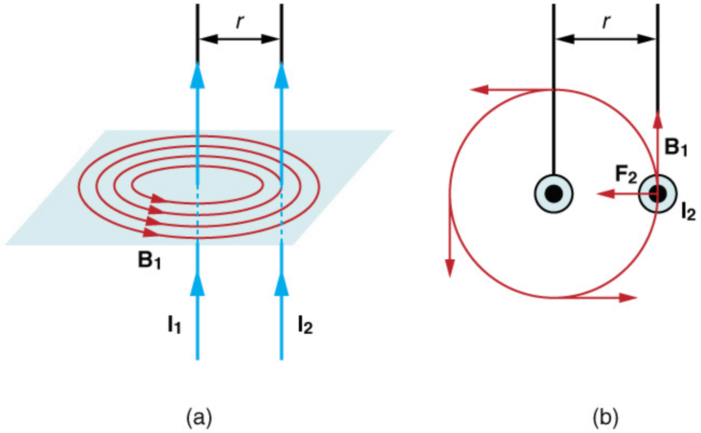


Figure 1. (a) The magnetic field produced by a long straight conductor is perpendicular to a parallel conductor, as indicated by RHR-2. (b) A view from above of the two wires shown in (a), with one magnetic field line shown for each wire. RHR-1 shows that the force between the parallel conductors is attractive when the currents are in the same direction. A similar analysis shows that the force is repulsive between currents in opposite directions.

This field is uniform along wire 2 and perpendicular to it, and so the force F_2 it exerts on wire 2 is given by $F = IlB \sin\theta$ with $\sin\theta = 1$:

$$F_2 = I_2 \mathrm{l} \mathrm{B}_1.$$

By Newton's third law, the forces on the wires are equal in magnitude, and so we just write F for the magnitude of F_2. (Note that $F_1 = -F_2$.) Since the wires are very long, it is convenient to think in terms of F/l, the force per unit length. Substituting the expression for B_1 into the last equation and rearranging terms gives

$$\frac{F}{l} = \frac{\mu_0 I_1 I_2}{2\pi \mathrm{r}}.$$

F/l is the force per unit length between two parallel currents I_1 and I_2 separated by a distance *r*. The force is attractive if the currents are in the same direction and repulsive if they are in opposite directions. This force is responsible for the *pinch effect* in electric arcs and plasmas. The force exists whether the currents are in wires or not. In an electric arc, where currents are moving parallel to one another, there is an attraction that squeezes currents into a smaller tube. In large circuit breakers, like those used in neighborhood power distribution systems, the pinch effect can concentrate an arc between plates of a switch trying to break a large current, burn holes, and even ignite the equipment. Another example of the pinch effect is found in the solar plasma, where jets of ionized material, such as solar flares, are shaped by magnetic forces.

The *operational definition of the ampere* is based on the force between current-carrying wires. Note that for parallel wires separated by 1 meter with each carrying 1 ampere, the force per meter is

$$\frac{F}{l} = \frac{\left(4\pi \times 10^{-7}\ \text{T} \cdot\ \text{m/A}\right)(1\ \text{A})^2}{(2\pi)\,(1\ \text{m})} = 2 \times 10^{-7}\ \text{N/m}$$

.

Since μ_0 is exactly $4\pi \times 10^{-7}$ T · m/A by definition, and because 1 T = 1 N/(A · m), the force per meter is exactly 2×10^{-7} N/m. This is the basis of the operational definition of the ampere.

The Ampere

The official definition of the ampere is: One ampere of current through each of two parallel conductors of infinite length, separated by one meter in empty space free of other magnetic fields, causes a force of exactly 2×10^{-7} N/m on each conductor.

Infinite-length straight wires are impractical and so, in practice, a current balance is constructed with coils of wire separated by a few centimeters. Force is measured to determine current. This also provides us with a method for measuring the coulomb. We measure the charge that flows for a current of one ampere in one second. That is, 1 C = 1 A · s. For both the ampere and the coulomb, the method of measuring force between conductors is the most accurate in practice.

Section Summary

- The force between two parallel currents I_1 and I_2, separated by a distance r, has a magnitude per unit length given by
$$\frac{F}{l} = \frac{\mu_0 I_1 I_2}{2\pi r}.$$
- The force is attractive if the currents are in the same direction, repulsive if they are in opposite directions.

Conceptual Questions

1. Is the force attractive or repulsive between the hot and neutral lines hung from power poles? Why?

2. If you have three parallel wires in the same plane, as in Figure 2, with currents in the outer two running in opposite directions, is it possible for the middle wire to be repelled by both? Attracted by both? Explain.

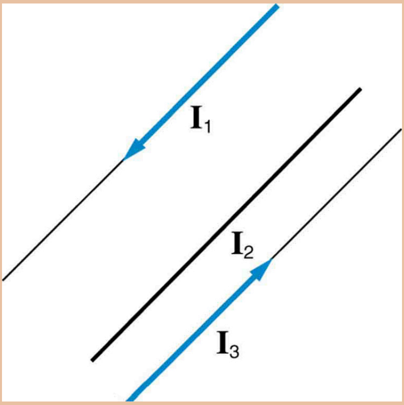


Figure 2. Three parallel coplanar wires with currents in the outer two in opposite directions.

3. Suppose two long straight wires run perpendicular to one another without touching. Does one exert a net force on the other? If so, what is its direction? Does one exert a net torque on the other? If so, what is its direction? Justify your responses by using the right hand rules.

4. Use the right hand rules to show that the force between the two loops in Figure 3 is attractive if the currents are in the same direction and repulsive if they are in opposite directions. Is this consistent with like poles of the loops repelling and unlike poles of the loops attracting? Draw sketches to justify your answers.

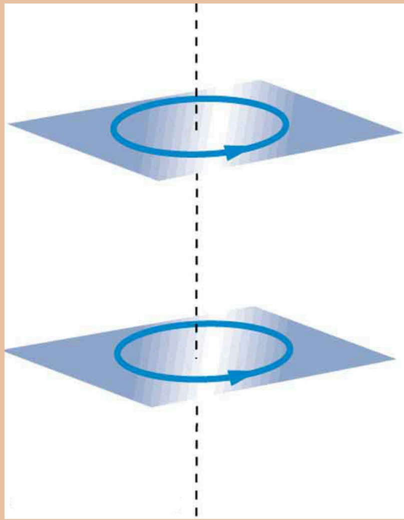


Figure 3. Two loops of wire carrying currents can exert forces and torques on one another.

5. If one of the loops in Figure 3 is tilted slightly relative to the other and their currents are in the same direction, what are the directions of the torques they exert on each other? Does this imply that the poles of the bar magnet-like fields they create will line up with each other if the loops are allowed to rotate?

6. Electric field lines can be shielded by the Faraday cage effect. Can we have magnetic shielding? Can we have gravitational shielding?

Exercises

1. (a) The hot and neutral wires supplying DC power to a light-rail commuter train carry 800 A and are separated by 75.0 cm. What is the magnitude and direction of the force between 50.0 m of these wires? (b) Discuss the practical consequences of this force, if any.

2. The force per meter between the two wires of a jumper cable being used to start a stalled car is 0.225 N/m. (a) What is the current in the wires, given they are separated by 2.00 cm? (b) Is the force attractive or repulsive?

3. A 2.50-m segment of wire supplying current to the motor of a submerged submarine carries 1000 A and feels a 4.00-N repulsive force from a parallel wire 5.00 cm away. What is the direction and magnitude of the current in the other wire?

4. The wire carrying 400 A to the motor of a commuter train feels an attractive force of 4.00×10^{-3} N/m due to a parallel wire carrying 5.00 A to a headlight. (a) How far apart are the wires? (b) Are the currents in the same direction?

5. An AC appliance cord has its hot and neutral wires separated by 3.00 mm and carries a 5.00-A current. (a) What is the average force per meter between the wires in the cord? (b) What is the maximum force per meter between the wires? (c) Are the forces attractive or repulsive? (d) Do appliance cords need any special design features to compensate for these forces?

6. Figure 4 shows a long straight wire near a rectangular

current loop. What is the direction and magnitude of the total force on the loop?

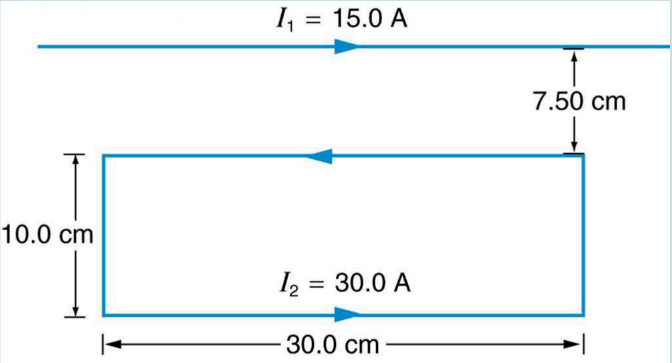


Figure 4.

7. Find the direction and magnitude of the force that each wire experiences in Figure 5(a) by, using vector addition.

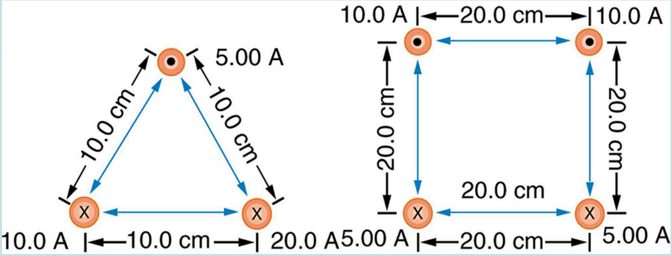


Figure 5.

8. Find the direction and magnitude of the force that each wire experiences in Figure 5(b), using vector addition.

Selected Solutions to Problems & Exercises

1. (a) 8.53 N, repulsive (b) This force is repulsive and therefore there is never a risk that the two wires will touch and short circuit.

3. 400 A in the opposite direction

5. (a) 1.67×10^{-3} N/m (b) 3.33×10^{-3} N/m (c) Repulsive (d) No, these are very small forces

7. (a) Top wire: 2.65×10^{-4} N/m s, 10.9° to left of up (b) Lower left wire: 3.61×10^{-4} N/m, 13.9° down from right (c) Lower right wire: 3.46×10^{-4} N/m, 30.0° down from left

53. More Applications of Magnetism

Learning Objective

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

- Describe some applications of magnetism.

Mass Spectrometry

The curved paths followed by charged particles in magnetic fields can be put to use. A charged particle moving perpendicular to a magnetic field travels in a circular path having a radius r.

$$r = \frac{mv}{qB}$$

It was noted that this relationship could be used to measure the mass of charged particles such as ions. A mass spectrometer is a device that measures such masses. Most mass spectrometers use magnetic fields for this purpose, although some of them have extremely sophisticated designs. Since there are five variables in the relationship, there are many possibilities. However, if v, q, and B can be fixed, then the radius of the path r is simply proportional to the mass m of the charged particle. Let us examine one such mass spectrometer that has a relatively simple design. (See Figure 1.) The process begins with an ion source, a device like an electron gun. The ion source gives ions their charge, accelerates them to some velocity v, and directs a beam of them into the next stage of the spectrometer. This next region is a *velocity selector* that only allows particles with a particular value of v to get through.

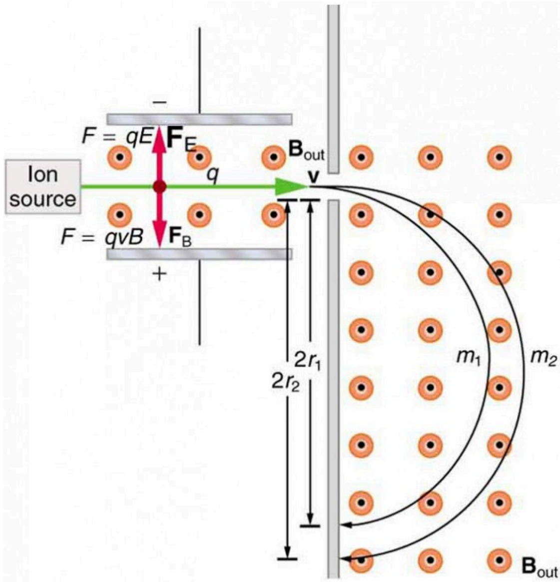


Figure 1. This mass spectrometer uses a velocity selector to fix v so that the radius of the path is proportional to mass.

The velocity selector has both an electric field and a magnetic field, perpendicular to one another, producing forces in opposite directions on the ions. Only those ions for which the forces balance travel in a straight line into the next region. If the forces balance, then the electric force $F = qE$ equals the magnetic force $F = qvB$, so that $qE = qvB$. Noting that q cancels, we see that

$$v = \frac{E}{B}$$

is the velocity particles must have to make it through the velocity selector, and further, that v can be selected by varying E and B. In the final region, there is only a uniform magnetic field, and so the charged particles move in circular arcs with radii proportional to particle mass. The paths also depend on charge q, but since q is in multiples of electron charges, it is easy to determine and to discriminate between ions in different charge states. Mass spectrometry today is used extensively in chemistry and biology laboratories to identify chemical and biological substances according to their mass-to-charge ratios. In medicine, mass spectrometers are used to measure the concentration of isotopes used as tracers. Usually, biological molecules such as proteins are very large, so they are broken down into smaller fragments before analyzing. Recently, large virus particles have been analyzed as a whole on mass spectrometers. Sometimes a gas chromatograph or high-performance liquid chromatograph provides an initial separation of the large molecules, which are then input into the mass spectrometer.

Cathode Ray Tubes—CRTs—and the Like

What do non-flat-screen TVs, old computer monitors, x-ray machines, and the 2-mile-long Stanford Linear Accelerator have in common? All of them accelerate electrons, making them different versions of the electron gun. Many of these devices use magnetic fields to steer the accelerated electrons. Figure 2 shows the construction of the type of cathode ray tube (CRT) found in some TVs, oscilloscopes, and old computer monitors. Two pairs of coils are used to steer the electrons, one vertically and the other horizontally, to their desired destination.

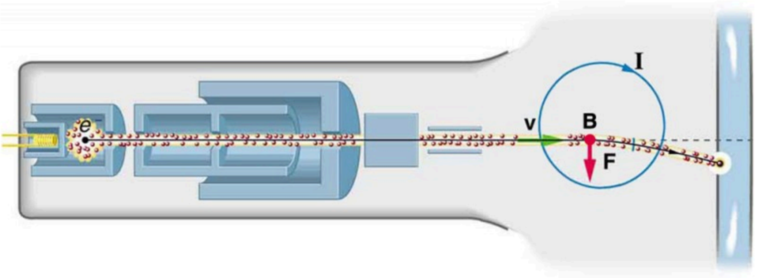


Figure 2. The cathode ray tube (CRT) is so named because rays of electrons originate at the cathode in the electron gun. Magnetic coils are used to steer the beam in many CRTs. In this case, the beam is moved down. Another pair of horizontal coils would steer the beam horizontally.

Magnetic Resonance Imaging

Magnetic resonance imaging (*MRI*) is one of the most useful and rapidly growing medical imaging tools. It non-invasively produces two-dimensional and three-dimensional images of the body that provide important medical information with none of the hazards of x-rays. MRI is based on an effect called *nuclear magnetic resonance* (*NMR*) in which an externally applied magnetic field interacts with the nuclei of certain atoms, particularly those of hydrogen (protons). These nuclei possess their own small magnetic fields, similar to those of electrons and the current loops discussed earlier in this module.

When placed in an external magnetic field, such nuclei experience a torque that pushes or aligns the nuclei into one of two new energy states—depending on the orientation of its spin (analogous to the N pole and S pole in a bar magnet). Transitions from the lower to higher energy state can be achieved by using an external radio frequency signal to "flip" the orientation of the small magnets. (This is actually a quantum mechanical process. The direction of the nuclear magnetic field is quantized as is energy in the radio waves. We will return to these topics in later chapters.) The specific frequency of the radio waves that are absorbed and reemitted depends sensitively on the type of nucleus, the chemical environment, and the external magnetic field strength. Therefore, this is a *resonance* phenomenon in which *nuclei* in a *magnetic* field act like resonators (analogous to those discussed in the treatment of sound in Oscillatory Motion and Waves) that absorb and reemit only certain frequencies. Hence, the phenomenon is named *nuclear magnetic resonance* (NMR).

NMR has been used for more than 50 years as an analytical tool. It was formulated in 1946 by F. Bloch and E. Purcell, with the 1952 Nobel Prize in Physics going to them for their work. Over the past two decades, NMR has been developed to produce detailed images in a process now called magnetic resonance imaging (MRI), a name coined to avoid the use of the word "nuclear" and the concomitant implication that nuclear radiation is involved. (It is not.) The 2003 Nobel Prize in Medicine went to P. Lauterbur and P. Mansfield for their work with MRI applications.

The largest part of the MRI unit is a superconducting magnet that creates a magnetic field, typically between 1 and 2 T in strength, over a relatively large volume. MRI images can be both highly detailed and informative about structures and organ functions. It is helpful that normal and non-normal tissues respond differently for slight changes in the magnetic field. In most medical images, the protons that are hydrogen nuclei are imaged. (About 2/3 of the atoms in the body are hydrogen.) Their location and density give a variety of medically useful information, such as organ function, the condition of tissue (as in the brain), and the shape of structures, such as vertebral disks and knee-joint surfaces. MRI can also be used to follow the movement of certain ions across membranes, yielding information on active transport, osmosis, dialysis, and other phenomena. With excellent spatial resolution, MRI can provide information about tumors, strokes, shoulder injuries, infections, etc.

An image requires position information as well as the density of a nuclear type (usually protons). By varying the magnetic field slightly over the volume to be imaged, the resonant frequency of the protons is made to vary with position. Broadcast radio frequencies are swept over an appropriate range and nuclei absorb and reemit them only if the nuclei are in a magnetic field with the correct strength. The imaging receiver gathers information through the body almost point by point, building up a tissue map. The reception of reemitted radio waves as a function of frequency thus gives position information.

These "slices" or cross sections through the body are only several mm thick. The intensity of the reemitted radio waves is proportional to the concentration of the nuclear type being flipped, as well as information on the chemical environment in that area of the body. Various techniques are available for enhancing contrast in images and for obtaining more information. Scans called T1, T2, or proton density scans rely on different relaxation mechanisms of nuclei. Relaxation refers to the time it takes for the protons to return to equilibrium after the external field is turned off. This time depends upon tissue type and status (such as inflammation).

While MRI images are superior to x rays for certain types of tissue and have none of the hazards of x rays, they do not completely supplant x-ray images. MRI is less effective than x rays for detecting breaks in bone, for example, and in imaging breast tissue, so the two diagnostic tools complement each other. MRI images are also expensive compared to simple x-ray images and tend to be used most often where they supply information not readily obtained from x rays. Another disadvantage of MRI is that the patient is totally enclosed with detectors close to the body for about 30 minutes or more, leading to claustrophobia. It is also difficult for the obese patient to be in the magnet tunnel. New "open-MRI" machines are now available in which the magnet does not completely surround the patient.

Over the last decade, the development of much faster scans, called "functional MRI" (fMRI), has allowed us to map the functioning of various regions in the brain responsible for thought and motor control. This technique measures the change in blood flow for activities (thought, experiences, action) in the brain. The nerve cells increase their consumption of oxygen when active. Blood hemoglobin releases oxygen to active nerve cells and has somewhat different magnetic properties when oxygenated than when deoxygenated. With MRI, we can measure this and detect a blood oxygen-dependent signal. Most of the brain scans today use fMRI.

Other Medical Uses of Magnetic Fields

Currents in nerve cells and the heart create magnetic fields like any other currents. These can be measured but with some difficulty since their strengths are about 10^{-6} to 10^{-8}*less* than the Earth's magnetic field. Recording of the heart's magnetic field as it beats is called a *magnetocardiogram* (MCG), while measurements of the brain's magnetic field is called a *magnetoencephalogram* (MEG). Both give information that differs from that obtained by measuring the electric fields of these organs (ECGs and EEGs), but they are not yet of sufficient importance to make these difficult measurements common.

In both of these techniques, the sensors do not touch the body. MCG can be used in fetal studies, and is probably more sensitive than echocardiography. MCG also looks at the heart's electrical activity whose voltage output is too small to be recorded by surface electrodes as in EKG. It has the potential of being a rapid scan for early diagnosis of cardiac ischemia (obstruction of blood flow to the heart) or problems with the fetus.

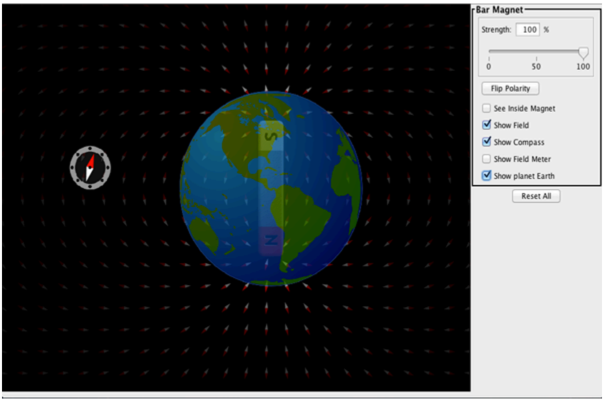
MEG can be used to identify abnormal electrical discharges in the brain that produce weak magnetic signals. Therefore, it looks at brain activity, not just brain structure. It has been used for studies of Alzheimer's disease and epilepsy. Advances in instrumentation to measure very small magnetic fields have allowed these two techniques to be used more in recent years. What is used is a sensor called a SQUID, for superconducting quantum interference device.

This operates at liquid helium temperatures and can measure magnetic fields thousands of times smaller than the Earth's.

Finally, there is a burgeoning market for magnetic cures in which magnets are applied in a variety of ways to the body, from magnetic bracelets to magnetic mattresses. The best that can be said for such practices is that they are apparently harmless, unless the magnets get close to the patient's computer or magnetic storage disks. Claims are made for a broad spectrum of benefits from cleansing the blood to giving the patient more energy, but clinical studies have not verified these claims, nor is there an identifiable mechanism by which such benefits might occur.

PhET Explorations: Magnet and Compass

Ever wonder how a compass worked to point you to the Arctic? Explore the interactions between a compass and bar magnet, and then add the Earth and find the surprising answer! Vary the magnet's strength, and see how things change both inside and outside. Use the field meter to measure how the magnetic field changes.



Click to download the simulation. Run using Java.

Section Summary

- Crossed (perpendicular) electric and magnetic fields act as a velocity filter, giving equal and opposite forces on any charge with velocity perpendicular to the fields and of magnitude
$$v = \frac{E}{B}.$$

Conceptual Questions

1. Measurements of the weak and fluctuating magnetic fields associated with brain activity are called magnetoencephalograms (MEGs). Do the brain's magnetic fields imply coordinated or uncoordinated nerve impulses? Explain.
2. Discuss the possibility that a Hall voltage would be generated on the moving heart of a patient during MRI imaging. Also discuss the same effect on the wires of a pacemaker. (The fact that patients with pacemakers are not given MRIs is significant.)
3. A patient in an MRI unit turns his head quickly to one side and experiences momentary dizziness and a strange taste in his mouth. Discuss the possible causes.
4. You are told that in a certain region there is either a uniform electric or magnetic field. What measurement or observation could you make to determine the type? (Ignore the Earth's magnetic field.)
5. An example of magnetohydrodynamics (MHD) comes from the flow of a river (salty water). This fluid interacts with the Earth's magnetic field to produce a potential difference between the two river banks. How would you go about calculating the potential difference?
6. Draw gravitational field lines between 2 masses, electric field lines between a positive and a negative charge, electric field lines between 2 positive charges and magnetic field lines around a magnet.

Qualitatively describe the differences between the fields and the entities responsible for the field lines.

Problems & Exercises

1. Indicate whether the magnetic field created in each of the three situations shown in Figure 4 is into or out of the page on the left and right of the current.

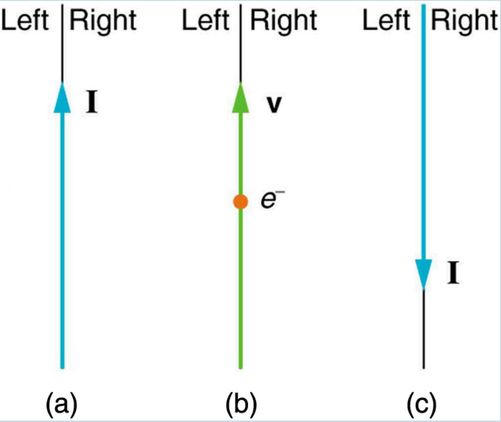


Figure 4.

2. What are the directions of the fields in the center of the loop and coils shown in Figure 5?

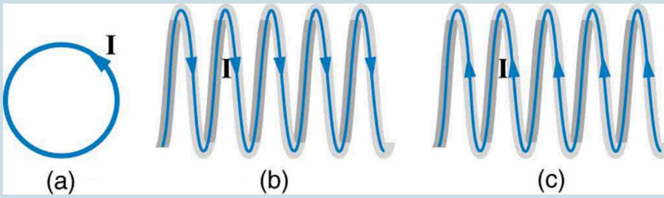


Figure 5.

3. What are the directions of the currents in the loop and coils shown in Figure 6?

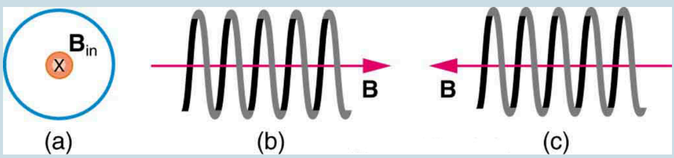


Figure 6.

4. To see why an MRI utilizes iron to increase the magnetic field created by a coil, calculate the current needed in a 400-loop-per-meter circular coil 0.660 m in radius to create a 1.20-T field (typical of an MRI instrument) at its center with no iron present. The magnetic field of a proton is approximately like that of a circular current loop 0.650×10^{-15} m in radius carrying 1.05×10^{4} A. What is the field at the center of such a loop?

5. Inside a motor, 30.0 A passes through a 250-turn circular loop that is 10.0 cm in radius. What is the magnetic field strength created at its center?

6. Nonnuclear submarines use batteries for power when submerged. (a) Find the magnetic field 50.0 cm from a straight wire carrying 1200 A from the batteries to the drive mechanism of a submarine. (b) What is the field if the wires to and from the drive mechanism are side by side? (c) Discuss the effects this could have for a compass on the submarine that is not shielded.

7. How strong is the magnetic field inside a solenoid with 10,000 turns per meter that carries 20.0 A?

8. What current is needed in the solenoid described in question 1 to produce a magnetic field 10^4 times the Earth's magnetic field of 5.00×10^{-5} T?

9. How far from the starter cable of a car, carrying 150 A, must you be to experience a field less than the Earth's (5.00×10^{-5} T)? Assume a long straight wire carries the current. (In practice, the body of your car shields the dashboard compass.)

10. Measurements affect the system being measured, such as the current loop in Figure 8 from Magnetic Force on a Current-Carrying Conductor. (a) Estimate the field the loop creates by calculating the field at the center of a circular loop 20.0 cm in diameter carrying 5.00 A. (b) What is the smallest field strength this loop can be used to measure, if its field must alter the measured field by less than 0.0100%?

11. Figure 7 shows a long straight wire just touching a loop carrying a current I_1. Both lie in the same plane. (a) What direction must the current I_2 in the straight wire have to

create a field at the center of the loop in the direction opposite to that created by the loop? (b) What is the ratio of I_1/I_2 that gives zero field strength at the center of the loop? (c) What is the direction of the field directly above the loop under this circumstance?

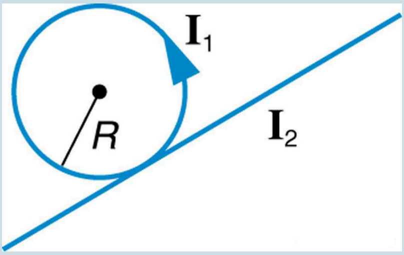


Figure 7.

12. Find the magnitude and direction of the magnetic field at the point equidistant from the wires in Figure 5(a) from [Magnetic Force between Two Parallel Conductors](#) (shown again below), using the rules of vector addition to sum the contributions from each wire.

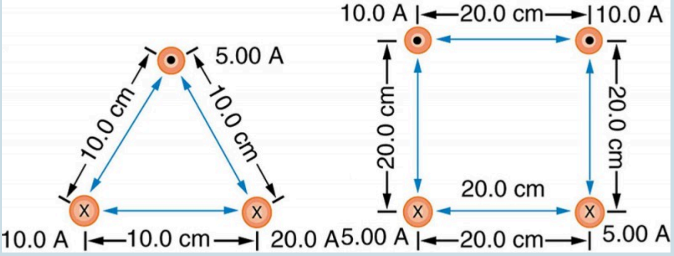


Figure 5 from Magnetic Force between Two Parallel Conductors.

13. Find the magnitude and direction of the magnetic field at the point equidistant from the wires in Figure 5(b) from Magnetic Force between Two Parallel Conductors (shown again above), using the rules of vector addition to sum the contributions from each wire.

14. What current is needed in the top wire in Figure 5(a) from Magnetic Force between Two Parallel Conductors (shown again above) to produce a field of zero at the point equidistant from the wires, if the currents in the bottom two wires are both 10.0 A into the page?

15. Calculate the size of the magnetic field 20 m below a high voltage power line. The line carries 450 MW at a voltage of 300,000 V.

16. **Integrated Concepts** (a) A pendulum is set up so that its bob (a thin copper disk) swings between the poles of a permanent magnet as shown in Figure 8. What is the magnitude and direction of the magnetic force on the bob at the lowest point in its path, if it has a positive 0.250 μC charge and is released from a height of 30.0 cm above its lowest point? The magnetic field strength is 1.50 T. (b) What is the acceleration of the bob at the bottom of its swing if its mass is 30.0 grams and it is hung from a flexible string? Be certain to include a free-body diagram as part of your analysis.

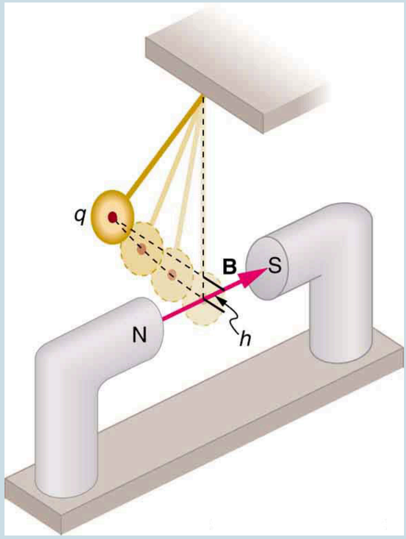


Figure 8.

17. **Integrated Concepts** (a) What voltage will accelerate electrons to a speed of 6.00×10^{-7} m/s? (b) Find the radius of curvature of the path of a *proton* accelerated through this potential in a 0.500-T field and compare this with the radius of curvature of an electron accelerated through the same potential.

18. **Integrated Concepts** Find the radius of curvature of the path of a 25.0-MeV proton moving perpendicularly to the 1.20-T field of a cyclotron.

19. **Integrated Concepts** To construct a nonmechanical water meter, a 0.500-T magnetic field is placed across the supply water pipe to a home and the Hall voltage is

recorded. (a) Find the flow rate in liters per second through a 3.00-cm-diameter pipe if the Hall voltage is 60.0 mV. (b) What would the Hall voltage be for the same flow rate through a 10.0-cm-diameter pipe with the same field applied?

20. **Integrated Concepts** (a) Calculate the maximum torque on a 50-turn, 1.50 cm radius circular current loop carrying 50 μA in a 0.500-T field. (b) If this coil is to be used in a galvanometer that reads 50 μA full scale, what force constant spring must be used, if it is attached 1.00 cm from the axis of rotation and is stretched by the 60° arc moved?

21. **Integrated Concepts** A current balance used to define the ampere is designed so that the current through it is constant, as is the distance between wires. Even so, if the wires change length with temperature, the force between them will change. What percent change in force per degree will occur if the wires are copper?

22. **Integrated Concepts** (a) Show that the period of the circular orbit of a charged particle moving perpendicularly to a uniform magnetic field is $T = 2\pi m/(qB)$. (b) What is the frequency f? (c) What is the angular velocity ω? Note that these results are independent of the velocity and radius of the orbit and, hence, of the energy of the particle. (Figure 9.)

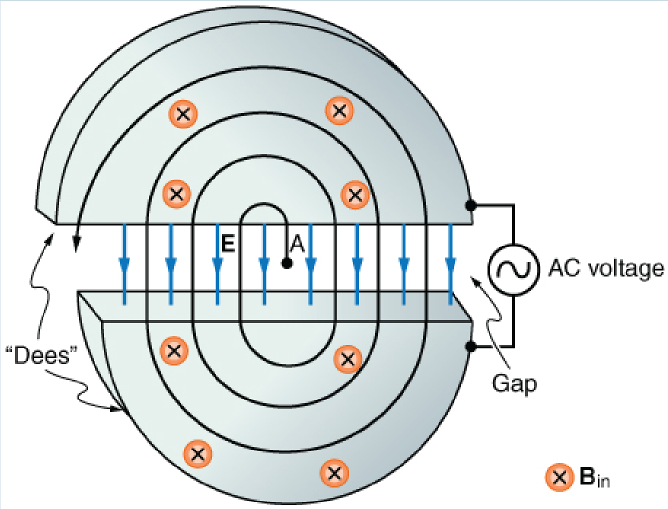


Figure 9. Cyclotrons accelerate charged particles orbiting in a magnetic field by placing an AC voltage on the metal Dees, between which the particles move, so that energy is added twice each orbit. The frequency is constant, since it is independent of the particle energy—the radius of the orbit simply increases with energy until the particles approach the edge and are extracted for various experiments and applications.

23. **Integrated Concepts** A cyclotron accelerates charged particles as shown in Figure 9. Using the results of the previous problem, calculate the frequency of the accelerating voltage needed for a proton in a 1.20-T field.

24. **Integrated Concepts** (a) A 0.140-kg baseball, pitched at 40.0 m/s horizontally and perpendicular to the Earth's horizontal 5.00×10^{-5} T field, has a 100-nC charge on it. What distance is it deflected from its path by the magnetic force, after traveling 30.0 m horizontally? (b) Would you

suggest this as a secret technique for a pitcher to throw curve balls?

25. **Integrated Concepts** (a) What is the direction of the force on a wire carrying a current due east in a location where the Earth's field is due north? Both are parallel to the ground. (b) Calculate the force per meter if the wire carries 20.0 A and the field strength is 3.00×10^{-5} T. (c) What diameter copper wire would have its weight supported by this force? (d) Calculate the resistance per meter and the voltage per meter needed.

26. **Integrated Concepts** One long straight wire is to be held directly above another by repulsion between their currents. The lower wire carries 100 A and the wire 7.50 cm above it is 10-gauge (2.588 mm diameter) copper wire. (a) What current must flow in the upper wire, neglecting the Earth's field? (b) What is the smallest current if the Earth's 3.00×10^{-5} T field is parallel to the ground and is not neglected? (c) Is the supported wire in a stable or unstable equilibrium if displaced vertically? If displaced horizontally?

27. **Unreasonable Results** (a) Find the charge on a baseball, thrown at 35.0 m/s perpendicular to the Earth's 5.00×10^{-5} T field, that experiences a 1.00-N magnetic force. (b) What is unreasonable about this result? (c) Which assumption or premise is responsible?

28. **Unreasonable Results** A charged particle having mass 6.64×10^{-27} kg (that of a helium atom) moving at 8.70×10^{5} m/s perpendicular to a 1.50-T magnetic field travels in a circular path of radius 16.0 mm. (a) What is the charge of

the particle? (b) What is unreasonable about this result? (c) Which assumptions are responsible?

29. **Unreasonable Results** An inventor wants to generate 120-V power by moving a 1.00-m-long wire perpendicular to the Earth's 5.00×10^{-5} T field. (a) Find the speed with which the wire must move. (b) What is unreasonable about this result? (c) Which assumption is responsible?

30. **Unreasonable Results** Frustrated by the small Hall voltage obtained in blood flow measurements, a medical physicist decides to increase the applied magnetic field strength to get a 0.500-V output for blood moving at 30.0 cm/s in a 1.50-cm-diameter vessel. (a) What magnetic field strength is needed? (b) What is unreasonable about this result? (c) Which premise is responsible?

31. **Unreasonable Results** A surveyor 100 m from a long straight 200-kV DC power line suspects that its magnetic field may equal that of the Earth and affect compass readings. (a) Calculate the current in the wire needed to create a 5.00×10^{-5} T field at this distance. (b) What is unreasonable about this result? (c) Which assumption or premise is responsible?

32. **Construct Your Own Problem** Consider a mass separator that applies a magnetic field perpendicular to the velocity of ions and separates the ions based on the radius of curvature of their paths in the field. Construct a problem in which you calculate the magnetic field strength needed to separate two ions that differ in mass, but not charge, and have the same initial velocity. Among the things to consider are the types of ions, the velocities they can be given before

entering the magnetic field, and a reasonable value for the radius of curvature of the paths they follow. In addition, calculate the separation distance between the ions at the point where they are detected.

33. **Construct Your Own Problem** Consider using the torque on a current-carrying coil in a magnetic field to detect relatively small magnetic fields (less than the field of the Earth, for example). Construct a problem in which you calculate the maximum torque on a current-carrying loop in a magnetic field. Among the things to be considered are the size of the coil, the number of loops it has, the current you pass through the coil, and the size of the field you wish to detect. Discuss whether the torque produced is large enough to be effectively measured. Your instructor may also wish for you to consider the effects, if any, of the field produced by the coil on the surroundings that could affect detection of the small field.

Glossary

magnetic resonance imaging (MRI):
: a medical imaging technique that uses magnetic fields create detailed images of internal tissues and organs

nuclear magnetic resonance (NMR):
: a phenomenon in which an externally applied magnetic field interacts with the nuclei of certain atoms

magnetocardiogram (MCG):
: a recording of the heart's magnetic field as it beats

magnetoencephalogram (MEG):
a measurement of the brain's magnetic field

Selected Solutions to Problems & Exercises

1. (a) right-into page, left-out of page (b) right-out of page, left-into page (c) right-out of page, left-into page

3. (a) clockwise (b) clockwise as seen from the left (c) clockwise as seen from the right

4. 1.01×10^{13} T

6. (a) 4.80×10^{-4} T (b) Zero (c) If the wires are not paired, the field is about 10 times stronger than Earth's magnetic field and so could severely disrupt the use of a compass.

8. 39.8 A

10. (a) 3.14×10^{-5} T (b) 0.314 T

12. 7.55×10^{-5} T, 23.4°

14. 10.0 A

16. (a) 9.09×10^{-7} N upward (b) $3.03 \times 10^{-5}\ \text{m/s}^2$

18. 60.2 cm

21. 17.0×10^{-4}%/°C

23. 18.3 MHz

25. (a) Straight up (b) 6.00×10^{-4} N/m (c) 94.1 μm (d) 2.47 Ω/m, 49.4 V/m

27. (a) 571 C (b) Impossible to have such a large separated charge on such a small object. (c) The 1.00-N force is much too great to be realistic in the Earth's field.

29. (a) 2.40×10^{6} m/s (b) The speed is too high to be

practical ≤ 1% speed of light (c) The assumption that you could reasonably generate such a voltage with a single wire in the Earth's field is unreasonable

31. (a) 25.0 kA (b) This current is unreasonably high. It implies a total power delivery in the line of 50.0×10^{9} W, which is much too high for standard transmission lines. (c)100 meters is a long distance to obtain the required field strength. Also coaxial cables are used for transmission lines so that there is virtually no field for DC power lines, because of cancellation from opposing currents. The surveyor's concerns are not a problem for his magnetic field measurements.

PART VII

ELECTROMAGNETIC INDUCTION, AC CIRCUITS, AND ELECTRICAL TECHNOLOGIES, PART II

54. Introduction to Electromagnetic Induction, AC Circuits, and Electrical Technologies

Figure 1. This wind turbine in the Thames Estuary in the UK is an example of induction at work. Wind pushes the blades of the turbine, spinning a shaft attached to magnets. The magnets spin around a conductive coil, inducing an electric current in the coil, and eventually feeding the electrical grid. (credit: phault, Flickr)

Nature's displays of symmetry are beautiful and alluring. A butterfly's wings exhibit an appealing symmetry in a complex system. (See Figure 2.) The laws of physics display symmetries at the most basic level—these symmetries are a source of wonder and imply deeper meaning. Since we place a high value on symmetry, we look for it when we explore nature. The remarkable thing is that we find it.

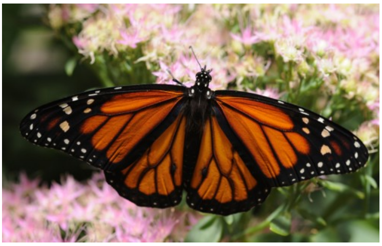


Figure 2. Physics, like this butterfly, has inherent symmetries. (credit: Thomas Bresson)

The hint of symmetry between electricity and magnetism found in the preceding chapter will be elaborated upon in this chapter. Specifically, we know that a current creates a magnetic field. If nature is symmetric here, then perhaps a magnetic field can create a current. The Hall effect is a voltage caused by a magnetic force. That voltage could drive a current. Historically, it was very shortly after Oersted discovered currents cause magnetic fields that other scientists asked the following question: Can magnetic fields cause currents? The answer was soon found by experiment to be yes. In 1831, some 12 years after Oersted's discovery, the English scientist Michael Faraday (1791–1862) and the American scientist Joseph Henry (1797–1878) independently demonstrated that magnetic fields can produce currents. The basic process of generating emfs (electromotive force) and, hence, currents with magnetic fields is known as *induction*; this process is also called magnetic induction to distinguish it from charging by induction, which utilizes the Coulomb force.

Today, currents induced by magnetic fields are essential to our technological society. The ubiquitous generator—found in automobiles, on bicycles, in nuclear power plants, and so on—uses magnetism to generate current. Other devices that use magnetism to induce currents include pickup coils in electric guitars, transformers of every size, certain microphones, airport security gates, and damping mechanisms on sensitive chemical balances.

Not so familiar perhaps, but important nevertheless, is that the behavior of AC circuits depends strongly on the effect of magnetic fields on currents.

55. Induced Emf and Magnetic Flux

Learning Objectives

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

- Calculate the flux of a uniform magnetic field through a loop of arbitrary orientation.
- Describe methods to produce an electromotive force (emf) with a magnetic field or magnet and a loop of wire.

The apparatus used by Faraday to demonstrate that magnetic fields can create currents is illustrated in Figure 1. When the switch is closed, a magnetic field is produced in the coil on the top part of the iron ring and transmitted to the coil on the bottom part of the ring. The galvanometer is used to detect any current induced in the coil on the bottom. It was found that each time the switch is closed, the galvanometer detects a current in one direction in the coil on the bottom. (You can also observe this in a physics lab.) Each time the switch is opened, the galvanometer detects a current in the opposite direction. Interestingly, if the switch remains closed or open for any length of time, there is no current through the galvanometer. *Closing and opening the switch* induces the current. It is the *change* in magnetic field that creates the current. More basic than the current that flows is the emfthat causes it. The current is a result of an *emf induced by a changing magnetic field*, whether or not there is a path for current to flow.

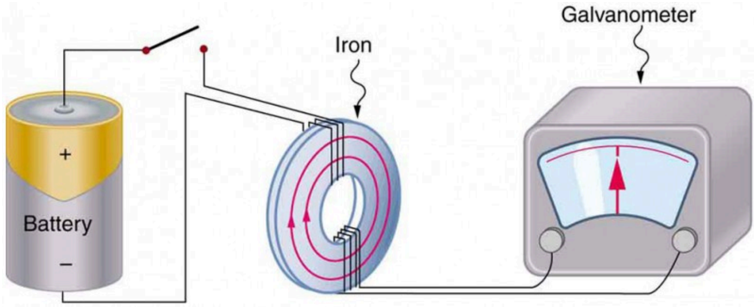


Figure 1. Faraday's apparatus for demonstrating that a magnetic field can produce a current. A change in the field produced by the top coil induces an emf and, hence, a current in the bottom coil. When the switch is opened and closed, the galvanometer registers currents in opposite directions. No current flows through the galvanometer when the switch remains closed or open.

An experiment easily performed and often done in physics labs is illustrated in Figure 2. An emf is induced in the coil when a bar magnet is pushed in and out of it. Emfs of opposite signs are produced by motion in opposite directions, and the emfs are also reversed by reversing poles. The same results are produced if the coil is moved rather than the magnet—it is the relative motion that is important. The faster the motion, the greater the emf, and there is no emf when the magnet is stationary relative to the coil.

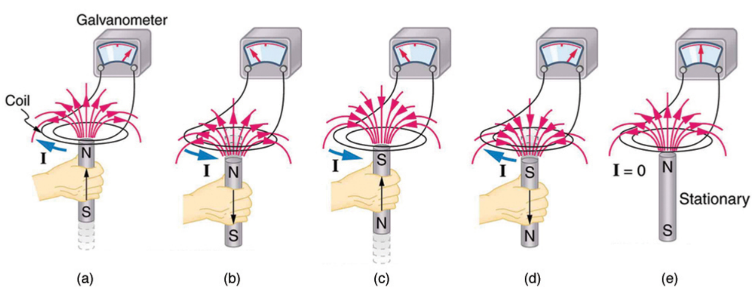


Figure 2. Movement of a magnet relative to a coil produces emfs as shown. The same emfs are produced if the coil is moved relative to the magnet. The greater the speed, the greater the magnitude of the emf, and the emf is zero when there is no motion.

The method of inducing an emf used in most electric generators is shown in Figure 3. A coil is rotated in a magnetic field, producing an alternating current emf, which depends on rotation rate and other factors that will be explored in later sections. Note that the generator is remarkably similar in construction to a motor (another symmetry).

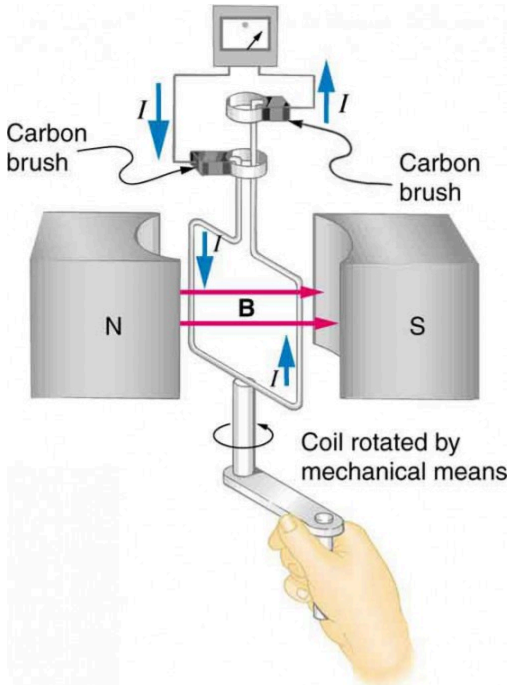


Figure 3. Rotation of a coil in a magnetic field produces an emf. This is the basic construction of a generator, where work done to turn the coil is converted to electric energy. Note the generator is very similar in construction to a motor.

So we see that changing the magnitude or direction of a magnetic field produces an emf. Experiments revealed that there is a crucial quantity called the *magnetic flux*, Φ, given by

$$\Phi = BA \cos \theta,$$

where B is the magnetic field strength over an area A, at an angle θ with the perpendicular to the area as shown in Figure 5. **Any change in magnetic flux Φ induces an emf.** This process is defined to be *electromagnetic induction.* Units of magnetic flux Φ are $T \cdot m^2$. As seen in Figure 4, $B \cos \theta = B_\perp$, which is the component of B perpendicular to the area A. Thus magnetic flux is $\Phi = B_\perp A$, the product of the area and the component of the magnetic field perpendicular to it.

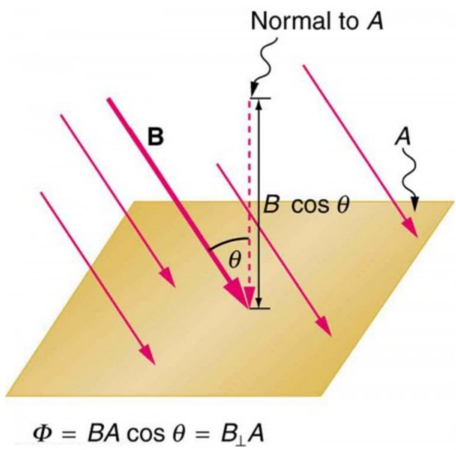


Figure 4. Magnetic flux Φ is related to the magnetic field and the area over which it exists. The flux $\Phi = BA \cos \theta$ is related to induction; any change in Φ induces an emf.

All induction, including the examples given so far, arises from some change in magnetic flux Φ. For example, Faraday changed B and hence Φ when opening and closing the switch in his apparatus (shown in Figure 1). This is also true for the bar magnet and coil shown in Figure 2. When rotating the coil of a generator, the angle θ and, hence, Φ is changed. Just how great an emf and what direction it takes depend on the change in Φ and how rapidly the change is made, as examined in the next section.

Section Summary

- The crucial quantity in induction is magnetic flux Φ, defined to be $\Phi = BA \cos\theta$, where B is the magnetic field strength over an area A at an angle θ with the perpendicular to the area.
- Units of magnetic flux Φ are $T \cdot m^2$.
- Any change in magnetic flux Φ induces an emf—the process is defined to be electromagnetic induction.

Conceptual Questions

1. How do the multiple-loop coils and iron ring in the version of Faraday's apparatus shown in Figure 1 enhance the observation of induced emf?

2. When a magnet is thrust into a coil as in Figure 2(a), what is the direction of the force exerted by the coil on the magnet? Draw a diagram showing the direction of the current induced in the coil and the magnetic field it produces, to justify your response. How does the magnitude of the force depend on the resistance of the galvanometer?

3. Explain how magnetic flux can be zero when the magnetic field is not zero.

4. Is an emf induced in the coil in Figure 5 when it is

stretched? If so, state why and give the direction of the induced current.

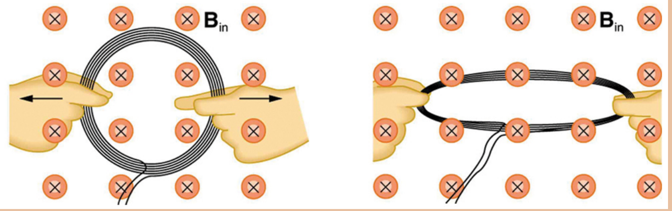


Figure 5. A circular coil of wire is stretched in a magnetic field.

Problems & Exercises

1. What is the value of the magnetic flux at coil 2 in Figure 6 due to coil 1?

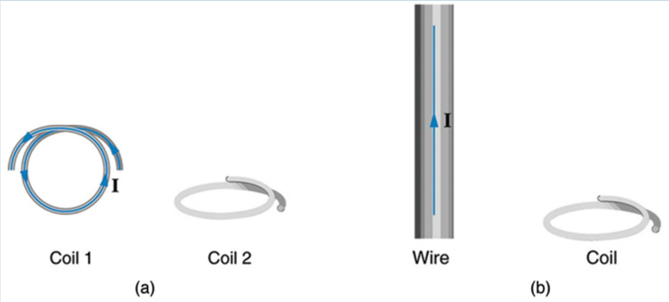


Figure 6. (a) The planes of the two coils are perpendicular. (b) The wire is perpendicular to the plane of the coil.

2. What is the value of the magnetic flux through the coil in Figure 6(b) due to the wire?

Glossary

magnetic flux:
: the amount of magnetic field going through a particular area, calculated with $\Phi = BA \cos \theta$ where B is the magnetic field strength over an area A at an angle θ with the perpendicular to the area

electromagnetic induction:
: the process of inducing an emf (voltage) with a change in magnetic flux

Exercises

1. Zero

56. Faraday's Law of Induction: Lenz's Law

Learning Objectives

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

- Calculate emf, current, and magnetic fields using Faraday's Law.
- Explain the physical results of Lenz's Law

Faraday's and Lenz's Law

Faraday's experiments showed that the emf induced by a change in magnetic flux depends on only a few factors. First, emf is directly proportional to the change in flux $\Delta\Phi$. Second, emf is greatest when the change in time Δt is smallest—that is, emf is inversely proportional to Δt. Finally, if a coil has N turns, an emf will be produced that is N times greater than for a single coil, so that emf is directly proportional to N. The equation for the emf induced by a change in magnetic flux is

$$\text{emf} = -N\frac{\Delta\Phi}{\Delta t}.$$

This relationship is known as *Faraday's law of induction*. The units for emf are volts, as is usual. The minus sign in Faraday's law of induction is very important. The minus means that *the emf creates a* *current I and magnetic field* B *that oppose the change in flux* $\Delta\Phi$*—this is known as Lenz's law*. The direction (given by the minus sign) of the emfis so important that it is called *Lenz's law* after the Russian Heinrich Lenz (1804–1865), who, like Faraday and Henry,independently investigated aspects of induction. Faraday was aware of the direction, but Lenz stated it so clearly that he is credited for its discovery. (See Figure 1.)

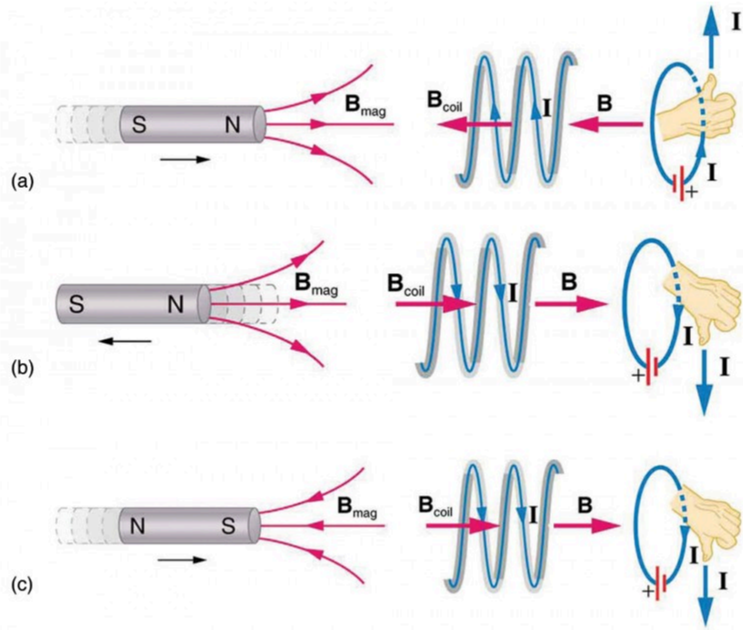


Figure 1. (a) When this bar magnet is thrust into the coil, the strength of the magnetic field increases in the coil. The current induced in the coil creates another field, in the opposite direction of the bar magnet's to oppose the increase. This is one aspect of Lenz's law—induction opposes any change in flux. (b) and (c) are two other situations. Verify for yourself that the direction of the induced Bcoil shown indeed opposes the change in flux and that the current direction shown is consistent with RHR-2.

Problem-Solving Strategy for Lenz's Law

To use Lenz's law to determine the directions of the induced magnetic fields, currents, and emfs:

1. Make a sketch of the situation for use in visualizing and recording directions.
2. Determine the direction of the magnetic field B.
3. Determine whether the flux is increasing or decreasing.
4. Now determine the direction of the induced magnetic field B. It opposes the *change* in flux by adding or subtracting from the original field.
5. Use RHR-2 to determine the direction of the induced current I that is responsible for the induced magnetic field B.
6. The direction (or polarity) of the induced emf will now drive a current in this direction and can be represented as current emerging from the positive terminal of the emf and returning to its negative terminal.

For practice, apply these steps to the situations shown in Figure 1 and to others that are part of the following text material.

Applications of Electromagnetic Induction

There are many applications of Faraday's Law of induction, as we will explore in this chapter and others. At this juncture, let us mention several that have to do with data storage and magnetic fields. A very important application has to do with audio and video *recording tapes*. A plastic tape, coated with iron oxide, moves past a recording head. This recording head is basically a round iron ring about which is wrapped a coil of wire—an electromagnet (Figure 2). A signal in the form of a varying input current from a microphone or camera goes to the recording head. These signals (which are a function of the signal amplitude and frequency) produce varying magnetic fields at the recording head. As the tape moves past the recording head, the magnetic field orientations of the iron oxide molecules on the tape are changed thus recording the signal. In the playback mode, the magnetized tape is run past another head, similar in structure to the recording head. The different magnetic field orientations of the iron oxide molecules on the tape induces an emf in the coil of wire in the playback head. This signal then is sent to a loudspeaker or video player.

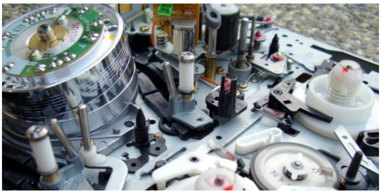


Figure 2. Recording and playback heads used with audio and video magnetic tapes. (credit: Steve Jurvetson)

Similar principles apply to computer hard drives, except at a much faster rate. Here recordings are on a coated, spinning disk. Read heads historically were made to work on the principle of induction. However, the input information is carried in digital rather than analog form – a series of 0's or 1's are written upon the spinning hard drive. Today, most hard drive readout devices do not work on the principle of induction, but use a technique known as *giant magnetoresistance*. (The discovery that weak changes in a magnetic field in a thin film of iron and chromium could bring about much larger changes in electrical resistance was one of the first large successes of nanotechnology.) Another application of induction is found on the magnetic stripe on the back of your personal credit card as used at the grocery store or the ATM machine. This works on the same principle as the audio or video tape mentioned in the last paragraph in which a head reads personal information from your card.

Another application of electromagnetic induction is when electrical signals need to be transmitted across a barrier. Consider the *cochlear implant* shown below. Sound is picked up by a microphone on the outside of the skull and is used to set up a varying magnetic field. A current is induced in a receiver secured in the bone beneath the skin and transmitted to electrodes in the inner ear. Electromagnetic induction can be used in other instances where electric signals need to be conveyed across various media.

Figure 3. Electromagnetic induction used in transmitting electric currents across mediums. The device on the baby's head induces an electrical current in a receiver secured in the bone beneath the skin. (credit: Bjorn Knetsch)

Another contemporary area of research in which electromagnetic induction is being successfully implemented (and with substantial potential) is transcranial magnetic simulation. A host of disorders, including depression and hallucinations can be traced to irregular localized electrical activity in the brain. In *transcranial magnetic stimulation*, a rapidly varying and very localized magnetic field is placed close to certain sites identified in the brain. Weak electric currents are induced in the identified sites and can result in recovery of electrical functioning in the brain tissue.

Sleep apnea ("the cessation of breath") affects both adults and infants (especially premature babies and it may be a cause of sudden infant deaths [SID]). In such individuals, breath can stop repeatedly during their sleep. A cessation of more than 20 seconds can be very dangerous. Stroke, heart failure, and tiredness are just some of the possible consequences for a person having sleep apnea. The concern in infants is the stopping of breath for these longer times. One type of monitor to alert parents when a child is not breathing uses electromagnetic induction. A wire wrapped around the infant's chest has an alternating current running through it. The expansion and contraction of the infant's chest as the infant breathes changes the area through the coil. A pickup coil located nearby has an alternating current induced in it due to the changing magnetic field of the initial wire. If the child stops breathing, there will be a change in the induced current, and so a parent can be alerted.

Making Connections: Conservation of Energy

Lenz's law is a manifestation of the conservation of energy. The induced emf produces a current that opposes the change in flux, because a change in flux means a change in energy. Energy can enter or leave, but not instantaneously. Lenz's law is a consequence. As the change begins, the law says induction opposes and, thus, slows the change. In fact, if the induced emf were in the same direction as the change in flux, there would be a positive feedback that would give us free energy from no apparent source—conservation of energy would be violated.

Example 1. Calculating Emf: How Great Is the Induced Emf?

Calculate the magnitude of the induced emf when the magnet in Figure 1(a) is thrust into the coil, given the following information: the single loop coil has a radius of 6.00 cm and the average value of $B \cos \theta$ (this is given, since the bar magnet's field is complex) increases from 0.0500 T to 0.250 T in 0.100 s.

Strategy

To find the *magnitude* of emf, we use Faraday's law of induction as stated by $\mathrm{emf} = -N\frac{\Delta\Phi}{\Delta t}$, but without the minus sign that indicates direction:

$$\mathrm{emf} = N\frac{\Delta\Phi}{\Delta t}.$$

Solution

We are given that $N = 1$ and $\Delta t = 0.100$ s, but we must determine the change in flux $\Delta\Phi$ before we can find emf. Since the area of the loop is fixed, we see that

$$\Delta\Phi = \Delta(BA \cos \theta) = A\Delta(B \cos \theta).$$

Now $\Delta(B \cos \theta) = 0.200$ T, since it was given that $B \cos$

θ changes from 0.0500 to 0.250 T. The area of the loop is A = πr2 = (3.14...)(0.060 m)^2 = 1.13 × 10^{-2} m^2. Thus,

$$\Delta\Phi = \left(1.13 \times 10^{-2}\ \mathrm{m}^2\right)(0.200\ \mathrm{T}).$$

Entering the determined values into the expression for emf gives

$$\mathrm{Emf} = N\frac{\Delta\Phi}{\Delta t} = \frac{\left(1.13 \times 10^{-2}\ \mathrm{m}^2\right)(0.200\ \mathrm{T})}{0.100\ \mathrm{s}} = 22.6\ \mathrm{mV}$$

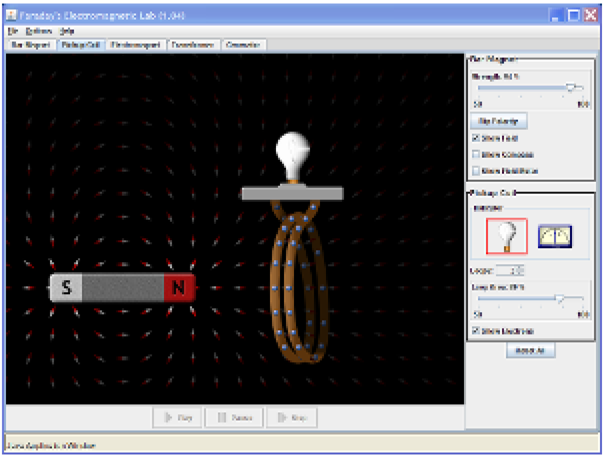
.

Discussion

While this is an easily measured voltage, it is certainly not large enough for most practical applications. More loops in the coil, a stronger magnet, and faster movement make induction the practical source of voltages that it is.

PhET Explorations: Faraday's Electromagnetic Lab

Play with a bar magnet and coils to learn about Faraday's law. Move a bar magnet near one or two coils to make a light bulb glow. View the magnetic field lines. A meter shows the direction and magnitude of the current. View the magnetic field lines or use a meter to show the direction and magnitude of the current. You can also play with electromagnets, generators and transformers!



Click to download the simulation. Run using Java.

Section Summary

- Faraday's law of induction states that the emfinduced by a change in magnetic flux is

$$\text{emf} = -N\frac{\Delta\Phi}{\Delta t}$$

  when flux changes by $\Delta\Phi$ in a time Δt.
- If emf is induced in a coil, N is its number of turns.
- The minus sign means that the emf creates a current I and

magnetic field B that *oppose the change in flux* $\Delta\Phi$—this opposition is known as Lenz's law.

Conceptual Questions

1. A person who works with large magnets sometimes places her head inside a strong field. She reports feeling dizzy as she quickly turns her head. How might this be associated with induction?
2. A particle accelerator sends high-velocity charged particles down an evacuated pipe. Explain how a coil of wire wrapped around the pipe could detect the passage of individual particles. Sketch a graph of the voltage output of the coil as a single particle passes through it.

Problems & Exercises

1. Referring to Figure 5(a), what is the direction of the current induced in coil 2: (a) If the current in coil 1 increases? (b) If the current in coil 1 decreases? (c) If the current in coil 1 is constant? Explicitly show how you follow the steps in the *Problem-Solving Strategy for Lenz's Law* above.

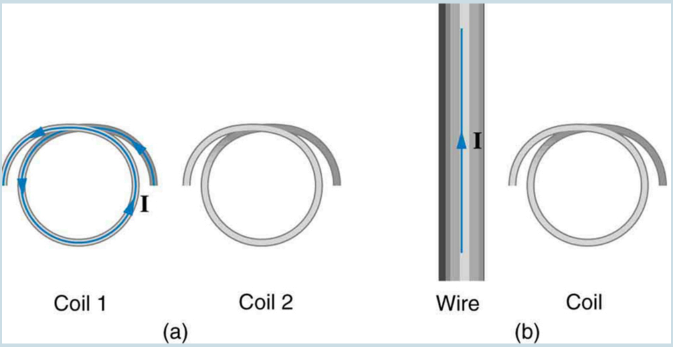


Figure 5. (a) The coils lie in the same plane. (b) The wire is in the plane of the coil.

2. Referring to Figure 5(b), what is the direction of the current induced in the coil: (a) If the current in the wire increases? (b) If the current in the wire decreases? (c) If the current in the wire suddenly changes direction? Explicitly show how you follow the steps in the *Problem-Solving Strategy for Lenz's Law* above.

3. Referring to Figure 6, what are the directions of the currents in coils 1, 2, and 3 (assume that the coils are lying in the plane of the circuit): (a) When the switch is first closed? (b) When the switch has been closed for a long time? (c) Just after the switch is opened?

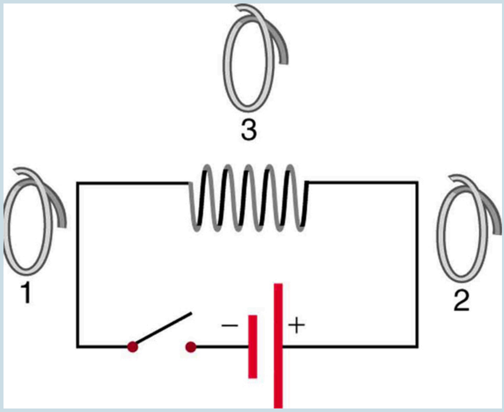


Figure 6.

4. Repeat the previous problem with the battery reversed.

5. Verify that the units of $\Delta\Phi/\Delta t$ are volts. That is, show that $1\ \text{T}\cdot\text{m}^2/\text{s} = 1\ \text{V}$.

6. Suppose a 50-turn coil lies in the plane of the page in a uniform magnetic field that is directed into the page. The coil originally has an area of $0.250\ \text{m}^2$. It is stretched to have no area in 0.100 s. What is the direction and magnitude of the induced emf if the uniform magnetic field has a strength of 1.50 T?

7. (a) An MRI technician moves his hand from a region of very low magnetic field strength into an MRI scanner's 2.00 T field with his fingers pointing in the direction of the field. Find the average emf induced in his wedding ring, given its diameter is 2.20 cm and assuming it takes 0.250 s to move

it into the field. (b) Discuss whether this current would significantly change the temperature of the ring.

8. **Integrated Concepts** Referring to the situation in the previous problem: (a) What current is induced in the ring if its resistance is0.0100 Ω? (b) What average power is dissipated? (c) What magnetic field is induced at the center of the ring? (d) What is the direction of the induced magnetic field relative to the MRI's field?

9. An emf is induced by rotating a 1000-turn, 20.0 cm diameter coil in the Earth's 5.00×10^{-5} T magnetic field. What average emf is induced, given the plane of the coil is originally perpendicular to the Earth's field and is rotated to be parallel to the field in 10.0 ms?

10. A 0.250 m radius, 500-turn coil is rotated one-fourth of a revolution in 4.17 ms, originally having its plane perpendicular to a uniform magnetic field. (This is 60 rev/s.) Find the magnetic field strength needed to induce an average emf of 10,000 V.

11. **Integrated Concepts** Approximately how does the emf induced in the loop in Figure 5(b) depend on the distance of the center of the loop from the wire?

12. **Integrated Concepts** (a) A lightning bolt produces a rapidly varying magnetic field. If the bolt strikes the earth vertically and acts like a current in a long straight wire, it will induce a voltage in a loop aligned like that in Figure 5(b). What voltage is induced in a 1.00 m diameter loop 50.0 m from a 2.00 × 106 lightning strike, if the current falls to

zero in25.0 μs? (b) Discuss circumstances under which such a voltage would produce noticeable consequences.

Glossary

Faraday's law of induction:
the means of calculating the emf in a coil due to changing magnetic flux, given by $\text{emf} = -N\dfrac{\Delta\Phi}{\Delta t}$

Lenz's law:
the minus sign in Faraday's law, signifying that the emf induced in a coil opposes the change in magnetic flux

Selected Solutions to Problems & Exercises

1. (a) CCW (b) CW (c) No current induced

3. (a) 1 CCW, 2 CCW, 3 CW (b) 1, 2, and 3 no current induced (c) 1 CW, 2 CW, 3 CCW

7. (a) 3.04 mV (b) As a lower limit on the ring, estimate R = 1.00 mΩ. The heat transferred will be 2.31 mJ. This is not a significant amount of heat.

9. 0.157 V

11. proportional to $\dfrac{1}{r}$

57. Video: Electromagnetic Induction

Watch the following Physics Concept Trailer to see how cochlear implants work because of electromagnetic induction.

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://library.achievingthedream.org/austinccphysics2/?p=81#oembed-1

58. Motional Emf

Learning Objective

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

- Calculate emf, force, magnetic field, and work due to the motion of an object in a magnetic field.

As we have seen, any change in magnetic flux induces an emf opposing that change—a process known as induction. Motion is one of the major causes of induction. For example, a magnet moved toward a coil induces an emf, and a coil moved toward a magnet produces a similar emf. In this section, we concentrate on motion in a magnetic field that is stationary relative to the Earth, producing what is loosely called *motional emf*. One situation where motional emf occurs is known as the Hall effect and has already been examined. Charges moving in a magnetic field experience the magnetic force $F = qvB \sin \theta$, which moves opposite charges in opposite directions and produces an em f $= Blv$. We saw that the Hall effect has applications, including measurements of B and v. We will now see that the Hall effect is one aspect of the broader phenomenon of induction, and we will find that motional emf can be used as a power source. Consider the situation shown in Figure 1. A rod is moved at a speed v along a pair of conducting rails separated by a distance l in a uniform magnetic field B. The rails are stationary relative to B and are connected to a stationary resistor R. The resistor could be anything from a light bulb to a voltmeter. Consider the area enclosed by the moving rod, rails, and resistor. B is perpendicular to this area, and the area is increasing as the rod moves. Thus the magnetic flux enclosed by the rails, rod, and resistor is increasing. When flux changes, an emf is induced according to Faraday’s law of induction.

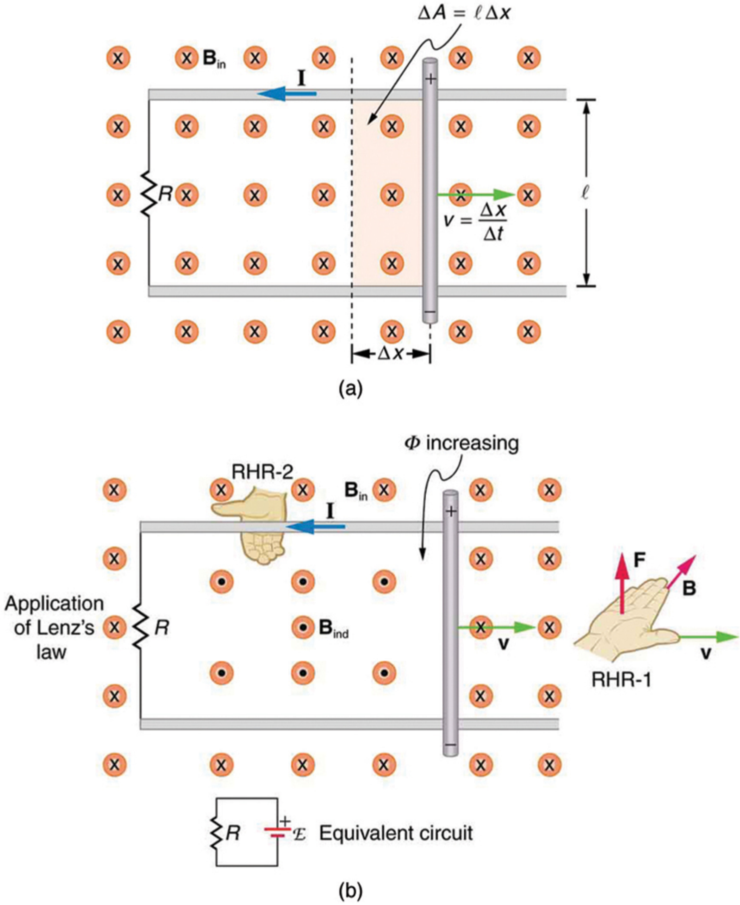


Figure 1. (a) A motional emf = Blv is induced between the rails when this rod moves to the right in the uniform magnetic field. The magnetic field B is into the page, perpendicular to the moving rod and rails and, hence, to the area enclosed by them. (b) Lenz's law gives the directions of the induced field and current, and the polarity of the induced emf. Since the flux is increasing, the induced field is in the opposite direction, or out of the page. RHR-2 gives the current direction shown, and the polarity of the rod will drive such a current. RHR-1 also indicates the same polarity for the rod. (Note that the script E symbol used in the equivalent circuit at the bottom of part (b) represents emf.)

To find the magnitude of emf induced along the moving rod, we use Faraday's law of induction without the sign:

$$\text{emf} = \text{N}\frac{\Delta\Phi}{\Delta t}.$$

Here and below, "emf" implies the magnitude of the emf. In this equation, $N = 1$ and the flux $\Phi = BA\cos\theta$. We have $\theta = 0°$ and $\cos\theta = 1$, since B is perpendicular to A. Now $\Delta\Phi = \Delta(BA) = B\Delta A$, since B is uniform. Note that the area swept out by the rod is $\Delta A = l\Delta x$. Entering these quantities into the expression for emf yields

$$\text{emf} = \frac{B\Delta A}{\Delta t} = B\frac{\ell\Delta x}{\Delta t}.$$

Finally, note that $\Delta x/\Delta t = v$, the velocity of the rod. Entering this into the last expression shows that

$$\text{emf} = B\ell v \quad (B, \ell, \text{ and } v \text{ perpendicular})$$

is the motional emf. This is the same expression given for the Hall effect previously.

Making Connections: Unification of Forces

There are many connections between the electric force and the magnetic force. The fact that a moving electric field produces a magnetic field and, conversely, a moving magnetic field produces an electric field is part of why electric and magnetic forces are now considered to be different manifestations of the same force. This classic unification of electric and magnetic forces into what is called the electromagnetic force is the inspiration for contemporary efforts to unify other basic forces.

To find the direction of the induced field, the direction of the current, and the polarity of the induced emf, we apply Lenz's law as explained in Faraday's Law of Induction: Lenz's Law. (See Figure 1(b).) Flux is increasing, since the area enclosed is increasing. Thus the induced field must oppose the existing one and be out of the page. And so the RHR-2 requires that I be counterclockwise, which in turn means the top of the rod is positive as shown.

Motional emf also occurs if the magnetic field moves and the rod (or other object) is stationary relative to the Earth (or some observer). We have seen an example of this in the situation where a moving magnet induces an emf in a stationary coil. It is the relative motion that is important. What is emerging in these observations is a connection between magnetic and electric fields. A moving magnetic field produces an electric field through its induced emf. We already have seen that a moving electric field produces a magnetic field—moving charge implies moving electric field and moving charge produces a magnetic field.

Motional emfs in the Earth's weak magnetic field are not ordinarily very large, or we would notice voltage along metal rods, such as a screwdriver, during ordinary motions. For example, a simple calculation of the motional emf of a 1 m rod moving at 3.0 m/s perpendicular to the Earth's field gives emf = Blv = $(5.0 \times 10^{-5}\ \text{T})(1.0\ \text{m})(3.0\ \text{m/s}) = 150\ \mu\text{V}$. This small value is consistent with experience. There is a spectacular exception, however. In 1992 and 1996, attempts were made with the space shuttle to create large motional emfs. The Tethered Satellite was to be let out on a 20 km length of wire as shown in Figure 2, to create a 5 kV emf by moving at orbital speed through the Earth's field. This emf could be used to convert some of the shuttle's kinetic and potential energy into electrical energy if a complete circuit could be made. To complete the circuit, the stationary ionosphere was to supply a return path for the current to flow. (The ionosphere is the rarefied and partially ionized atmosphere at orbital altitudes. It conducts because of the ionization. The ionosphere serves the same function as the stationary rails and connecting resistor in Figure 1, without which there would not be a complete circuit.) Drag on the current in the cable due to the magnetic force $F = IlB \sin \theta$ does the work that reduces the shuttle's kinetic and potential energy and allows it to be converted to electrical energy. The tests were both unsuccessful. In the first, the cable hung up and could only be extended a couple of hundred meters; in the second, the cable broke when almost fully extended. The following example indicates feasibility in principle.

Example 1. Calculating the Large Motional Emf of an Object in Orbit

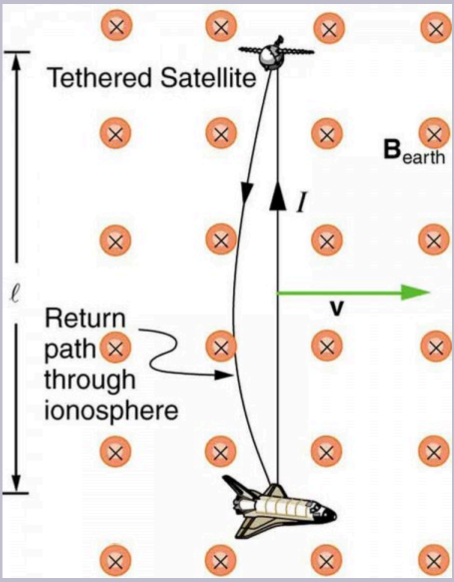


Figure 2. Motional emf as electrical power conversion for the space shuttle is the motivation for the Tethered Satellite experiment. A 5 kV emf was predicted to be induced in the 20 km long tether while moving at orbital speed in the Earth's magnetic field. The circuit is completed by a return path through the stationary ionosphere.

Calculate the motional emf induced along a 20.0 km long conductor moving at an orbital speed of 7.80 km/s perpendicular to the Earth's 5.00×10^{-5} T magnetic field.

Strategy

This is a straightforward application of the expression for motional emf — emf = Blv.

Solution

Entering the given values into emf = Blv gives

$$\begin{aligned} \text{emf} &= B\ell v \\ &= \left(5.00 \times 10^{-5}\ \text{T}\right)\left(2.0 \times 10^{4}\ \text{m}\right)\left(7.80 \times 10^{3}\ \text{m/s}\right) \\ &= 7.80 \times 10^{3}\ \text{V} \end{aligned}$$

.

Discussion

The value obtained is greater than the 5 kV measured voltage for the shuttle experiment, since the actual orbital motion of the tether is not perpendicular to the Earth's field. The 7.80 kV value is the maximum emf obtained when $\theta = 90°$ and $\sin\theta = 1$.

Section Summary

- An emf induced by motion relative to a magnetic field B is called a *motional emf* and is given by

$$\text{emf} = B\ell v \quad (B, \ell, \text{ and } v \text{ perpendicular})$$

where l is the length of the object moving at speed v relative to the field.

Conceptual Questions

1. Why must part of the circuit be moving relative to other parts, to have usable motional emf? Consider, for example, that the rails in Figure 1 are stationary relative to the magnetic field, while the rod moves.
2. A powerful induction cannon can be made by placing a metal cylinder inside a solenoid coil. The cylinder is forcefully expelled when solenoid current is turned on rapidly. Use Faraday's and Lenz's laws to explain how this works. Why might the cylinder get live/hot when the cannon is fired?
3. An induction stove heats a pot with a coil carrying an alternating current located beneath the pot (and without a hot surface). Can the stove surface be a conductor? Why won't a coil carrying a direct current work?
4. Explain how you could thaw out a frozen water pipe by wrapping a coil carrying an alternating current around it. Does it matter whether or not the pipe is a conductor? Explain.

Problems & Exercises

1. Use Faraday's law, Lenz's law, and RHR-1 to show that the magnetic force on the current in the moving rod in Figure 1 is in the opposite direction of its velocity.

2. If a current flows in the Satellite Tether shown in Figure 2, use Faraday's law, Lenz's law, and RHR-1 to show that there is a magnetic force on the tether in the direction opposite to its velocity.

3. (a) A jet airplane with a 75.0 m wingspan is flying at 280 m/s. What emf is induced between wing tips if the vertical component of the Earth's field is 3.00×10^{-5} T? (b) Is an emf of this magnitude likely to have any consequences? Explain.

4. (a) A nonferrous screwdriver is being used in a 2.00 T magnetic field. What maximum emf can be induced along its 12.0 cm length when it moves at 6.00 m/s? (b) Is it likely that this emf will have any consequences or even be noticed?

5. At what speed must the sliding rod in Figure 1 move to produce an emf of 1.00 V in a 1.50 T field, given the rod's length is 30.0 cm?

6. The 12.0 cm long rod in Figure 1 moves at 4.00 m/s. What is the strength of the magnetic field if a 95.0 V emf is induced?

7. Prove that when B, l, and v are not mutually perpendicular, motional emf is given by emf = $Blv \sin \theta$. If v is perpendicular to B, then θ is the angle between l and B. If

l is perpendicular to B, then θ is the angle between v and B.

8. In the August 1992 space shuttle flight, only 250 m of the conducting tether considered in Example 1 (above) could be let out. A 40.0 V motional emf was generated in the Earth's 5.00×10^{-5} T field, while moving at 7.80×10^{3} m/s. What was the angle between the shuttle's velocity and the Earth's field, assuming the conductor was perpendicular to the field?

9. **Integrated Concepts** Derive an expression for the current in a system like that in Figure 1, under the following conditions. The resistance between the rails is R, the rails and the moving rod are identical in cross section A and have the same resistivity ρ. The distance between the rails is l, and the rod moves at constant speed v perpendicular to the uniform field B. At time zero, the moving rod is next to the resistance R.

10. **Integrated Concepts** The Tethered Satellite in Figure 2 has a mass of 525 kg and is at the end of a 20.0 km long, 2.50 mm diameter cable with the tensile strength of steel. (a) How much does the cable stretch if a 100 N force is exerted to pull the satellite in? (Assume the satellite and shuttle are at the same altitude above the Earth.) (b) What is the effective force constant of the cable? (c) How much energy is stored in it when stretched by the 100 N force?

11. **Integrated Concepts** The Tethered Satellite discussed in this module is producing 5.00 kV, and a current of 10.0 A flows. (a) What magnetic drag force does this produce if the system is moving at 7.80 km/s? (b) How much kinetic energy is removed from the system in 1.00 h, neglecting any

change in altitude or velocity during that time? (c) What is the change in velocity if the mass of the system is 100,000 kg? (d) Discuss the long term consequences (say, a week-long mission) on the space shuttle's orbit, noting what effect a decrease in velocity has and assessing the magnitude of the effect.

Selected Solutions to Problems & Exercises

1. (a) 0.630 V (b) No, this is a very small emf.

5. 2.22 m/s

11.(a) 10.0 N (b) 2.81×10^{8} J (c) 0.36 m/s (d) For a week-long mission (168 hours), the change in velocity will be 60 m/s, or approximately 1%. In general, a decrease in velocity would cause the orbit to start spiraling inward because the velocity would no longer be sufficient to keep the circular orbit. The long-term consequences are that the shuttle would require a little more fuel to maintain the desired speed, otherwise the orbit would spiral slightly inward.

59. Eddy Currents and Magnetic Damping

Learning Objectives

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

- Explain the magnitude and direction of an induced eddy current, and the effect this will have on the object it is induced in.
- Describe several applications of magnetic damping.

Eddy Currents and Magnetic Damping

As discussed in Motional Emf, motional emf is induced when a conductor moves in a magnetic field or when a magnetic field moves relative to a conductor. If motional emf can cause a current loop in the conductor, we refer to that current as an *eddy current*. Eddy currents can produce significant drag, called *magnetic damping*, on the motion involved. Consider the apparatus shown in Figure 1, which swings a pendulum bob between the poles of a strong magnet. (This is another favorite physics lab activity.) If the bob is metal, there is significant drag on the bob as it enters and leaves the field, quickly damping the motion. If, however, the bob is a slotted metal plate, as shown in Figure 1(b), there is a much smaller effect due to the magnet. There is no discernible effect on a bob made of an insulator. Why is there drag in both directions, and are there any uses for magnetic drag?

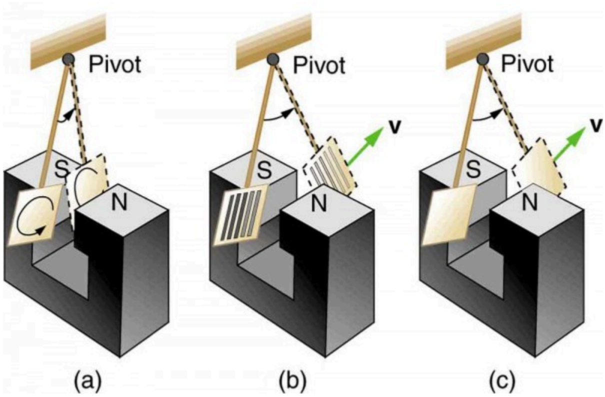


Figure 1. A common physics demonstration device for exploring eddy currents and magnetic damping. (a) The motion of a metal pendulum bob swinging between the poles of a magnet is quickly damped by the action of eddy currents. (b) There is little effect on the motion of a slotted metal bob, implying that eddy currents are made less effective. (c) There is also no magnetic damping on a nonconducting bob, since the eddy currents are extremely small.

Figure 2 shows what happens to the metal plate as it enters and leaves the magnetic field. In both cases, it experiences a force opposing its motion. As it enters from the left, flux increases, and so an eddy current is set up (Faraday's law) in the counterclockwise direction (Lenz's law), as shown. Only the right-hand side of the current loop is in the field, so that there is an unopposed force on it to the left (RHR-1). When the metal plate is completely inside the field, there is no eddy current if the field is uniform, since the flux remains constant in this region. But when the plate leaves the field on the right, flux decreases, causing an eddy current in the clockwise direction that, again, experiences a force to the left, further slowing the motion. A similar analysis of what happens when the plate swings from the right toward the left shows that its motion is also damped when entering and leaving the field.

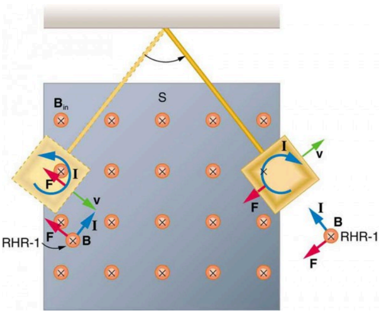


Figure 2. A more detailed look at the conducting plate passing between the poles of a magnet. As it enters and leaves the field, the change in flux produces an eddy current. Magnetic force on the current loop opposes the motion. There is no current and no magnetic drag when the plate is completely inside the uniform field.

When a slotted metal plate enters the field, as shown in Figure 3, an emf is induced by the change in flux, but it is less effective because the slots limit the size of the current loops. Moreover, adjacent loops have currents in opposite directions, and their effects cancel. When an insulating material is used, the eddy current is extremely small, and so magnetic damping on insulators is negligible. If eddy currents are to be avoided in conductors, then they can be slotted or constructed of thin layers of conducting material separated by insulating sheets.

Applications of Magnetic Damping

One use of magnetic damping is found in sensitive laboratory balances. To have maximum sensitivity and accuracy, the balance must be as friction-free as possible. But if it is friction-free, then it will oscillate for a very long time. Magnetic damping is a simple and ideal

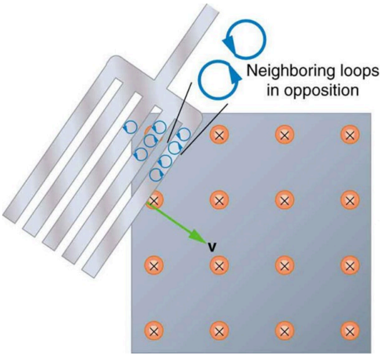


Figure 3. Eddy currents induced in a slotted metal plate entering a magnetic field form small loops, and the forces on them tend to cancel, thereby making magnetic drag almost zero.

solution. With magnetic damping, drag is proportional to speed and becomes zero at zero velocity. Thus the oscillations are quickly damped, after which the damping force disappears, allowing the balance to be very sensitive. (See Figure 4.) In most balances, magnetic damping is accomplished with a conducting disc that rotates in a fixed field.

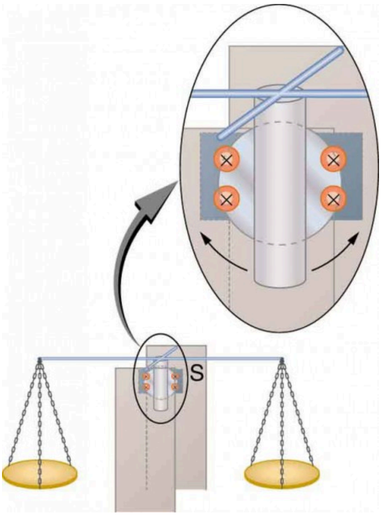


Figure 4. Magnetic damping of this sensitive balance slows its oscillations. Since Faraday's law of induction gives the greatest effect for the most rapid change, damping is greatest for large oscillations and goes to zero as the motion stops.

Since eddy currents and magnetic damping occur only in conductors, recycling centers can use magnets to separate metals from other materials. Trash is dumped in batches down a ramp, beneath which lies a powerful magnet. Conductors in the trash are slowed by magnetic damping while nonmetals in the trash move on, separating from the metals. (See Figure 5.) This works for all metals, not just ferromagnetic ones. A magnet can separate out the ferromagnetic materials alone by acting on stationary trash.

Other major applications of eddy currents are in metal detectors and braking systems in trains and roller coasters. Portable metal detectors (Figure 6) consist of a primary coil carrying an alternating current and a secondary coil in which a current is induced. An eddy current will be induced in a piece of metal close to the detector which will cause a change in the induced current within the secondary coil, leading to some sort of signal like a shrill noise. Braking using eddy currents is safer because factors such as rain do not affect the braking and the braking is smoother. However, eddy currents cannot bring the motion to a complete stop, since the force produced decreases with speed. Thus, speed can be reduced from say 20 m/s to 5 m/s, but another form of braking is needed to completely stop the vehicle. Generally, powerful rare earth magnets such as neodymium magnets are used in roller coasters. Figure 7 shows rows of magnets in such an application. The vehicle has metal fins (normally containing copper) which pass through the magnetic field slowing the vehicle down in much the same way as with the pendulum bob shown in Figure 1.

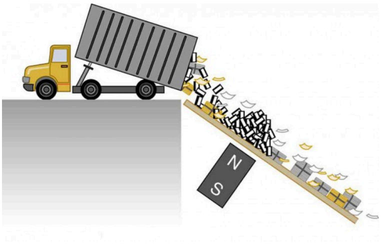


Figure 5. Metals can be separated from other trash by magnetic drag. Eddy currents and magnetic drag are created in the metals sent down this ramp by the powerful magnet beneath it. Nonmetals move on.

Figure 6. A soldier in Iraq uses a metal detector to search for explosives and weapons. (credit: U.S. Army)

Figure 7. The rows of rare earth magnets (protruding horizontally) are used for magnetic braking in roller coasters. (credit: Stefan Scheer, Wikimedia Commons)

Induction cooktops have electromagnets under their surface. The magnetic field is varied rapidly producing eddy currents in the base of the pot, causing the pot and its contents to increase in temperature. Induction cooktops have high efficiencies and good response times but the base of the pot needs to be ferromagnetic, iron or steel for induction to work.

Section Summary

- Current loops induced in moving conductors are called eddy currents.
- They can create significant drag, called magnetic damping.

Conceptual Questions

1. Explain why magnetic damping might not be effective on an object made of several thin conducting layers separated by insulation.
2. Explain how electromagnetic induction can be used to detect metals? This technique is particularly important in detecting buried landmines for disposal, geophysical prospecting and at airports.

Problems & Exercises

1. Make a drawing similar to Figure 2, but with the pendulum moving in the opposite direction. Then use Faraday's law, Lenz's law, and RHR-1 to show that magnetic force opposes motion.

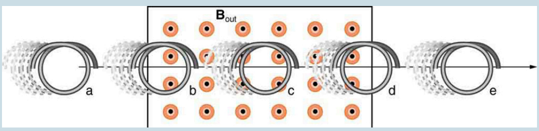


Figure 8. A coil is moved into and out of a region of uniform magnetic field.

2. A coil is moved through a magnetic field as shown in Figure 8. The field is uniform inside the rectangle and zero outside. What is the direction of the induced current and what is the direction of the magnetic force on the coil at each position shown?

Glossary

eddy current:
a current loop in a conductor caused by motional emf

magnetic damping:
the drag produced by eddy currents

60. Electric Generators

Learning Objectives

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

- Calculate the emf induced in a generator.
- Calculate the peak emf which can be induced in a particular generator system.

Electric generators induce an emf by rotating a coil in a magnetic field, as briefly discussed in Induced Emf and Magnetic Flux. We will now explore generators in more detail. Consider the following example.

Example 1. Calculating the Emf Induced in a Generator Coil

The generator coil shown in Figure 1 is rotated through one-fourth of a revolution (from $\theta = 0°$ to $\theta = 90°$) in 15.0 ms. The 200-turn circular coil has a 5.00 cm radius and is in a uniform 1.25 T magnetic field. What is the average emf induced?

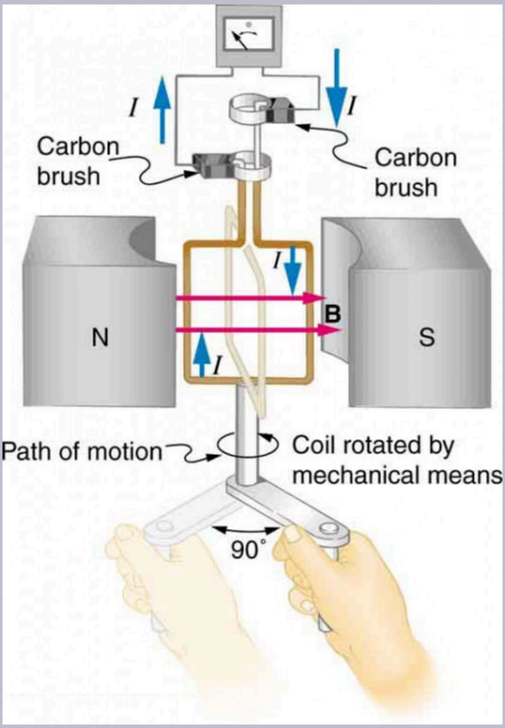


Figure 1. When this generator coil is rotated through one-fourth of a revolution, the magnetic flux Φ changes from its maximum to zero, inducing an emf.

Strategy

We use Faraday's law of induction to find the average emf induced over a time Δt:

$$\text{emf} = -N\frac{\Delta\Phi}{\Delta t}.$$

We know that N = 200 and Δt = 15.0 ms, and so we must determine the change in flux $\Delta\Phi$ to find emf.

Solution

Since the area of the loop and the magnetic field strength are constant, we see that

$$\Delta\Phi = \Delta\left(BA\cos\theta\right) = AB\Delta\left(\cos\theta\right).$$

Now, Δ (cos θ) = −1.0, since it was given that θ goes from 0° to 90° . Thus $\Delta\Phi$ = −AB, and

$$\text{emf} = N\frac{AB}{\Delta t}.$$

The area of the loop is $A = \pi r^2 = (3.14...)(0.0500\text{m})^2 = 7.85 \times 10^{-3}\ \text{m}^2$. Entering this value gives

$$\text{emf} = 200\frac{\left(7.85\times10^{-3}\ \text{m}^2\right)(1.25\ \text{T})}{15.0\times10^{-3}\ \text{s}} = 131\ \text{V}.$$

Discussion

This is a practical average value, similar to the 120 V used in household power.

The emf calculated in Example 1 above is the average over one-fourth of a revolution. What is the emf at any given instant? It varies with the angle between the magnetic field and a perpendicular to the coil. We can get an expression for emf as a function of time by considering the motional emf on a rotating rectangular coil of width w and height l in a uniform magnetic field, as illustrated in Figure 2.

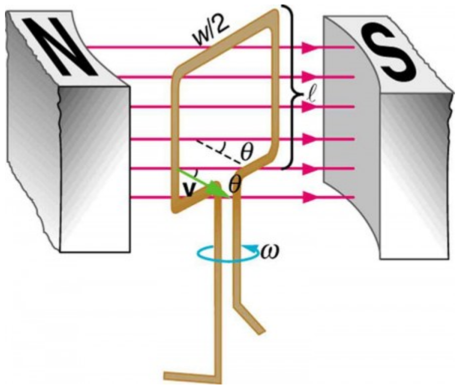


Figure 2. A generator with a single rectangular coil rotated at constant angular velocity in a uniform magnetic field produces an emf that varies sinusoidally in time. Note the generator is similar to a motor, except the shaft is rotated to produce a current rather than the other way around.

Charges in the wires of the loop experience the magnetic force, because they are moving in a magnetic field. Charges in the vertical wires experience forces parallel to the wire, causing currents. But those in the top and bottom segments feel a force perpendicular to the wire, which does not cause a current. We can thus find the induced emf by considering only the side wires. Motional emf is given to be emf = Blv, where the velocity v is perpendicular to the magnetic field B. Here the velocity is at an angle θ with B, so that its component perpendicular to B is v sin θ (see Figure 2). Thus in this case the emf induced on each side is emf = Blv sin θ, and they are in the same direction. The total emf around the loop is then

$$\text{emf} = 2B\ell v \sin\theta.$$

This expression is valid, but it does not give emf as a function of time. To find the time dependence of emf, we assume the coil rotates at a constant angular velocity ω. The angle θ is related to angular velocity by $\theta = \omega t$, so that

$$\text{emf} = 2B\ell v \sin \omega t.$$

Now, linear velocity v is related to angular velocity ω by $v = r\omega$. Here $r = w/2$, so that $v = (w/2)\omega$, and

$$\text{emf} = 2B\ell \frac{w}{2} \omega \sin \omega t = (\ell w) B\omega \sin \omega t.$$

Noting that the area of the loop is $A = lw$, and allowing for N loops, we find that

$$\text{emf} = NAB\omega \sin \omega t$$

is the *emf induced in a generator coil* of N turns and area A rotating at a constant angular velocity ω in a uniform magnetic field B. This can also be expressed as

$$\text{emf} = \text{emf}_0 \sin \omega t,$$

where

$$\text{emf}_0 = NAB\omega$$

is the maximum *(peak) emf*. Note that the frequency of the oscillation is $f = \omega/2\pi$, and the period is $T = 1/f = 2\pi/\omega$. Figure 3 shows a graph of emf as a function of time, and it now seems reasonable that AC voltage is sinusoidal.

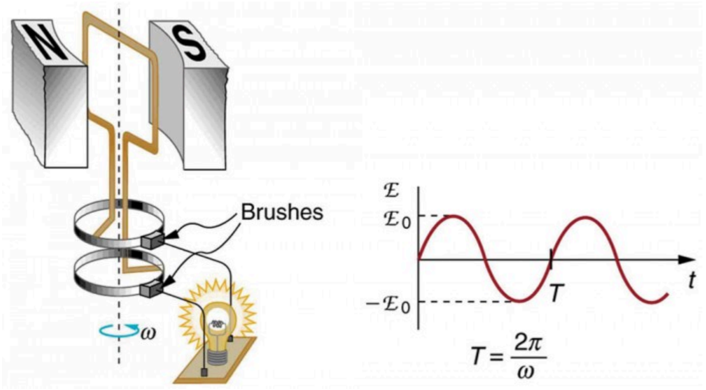


Figure 3. The emf of a generator is sent to a light bulb with the system of rings and brushes shown. The graph gives the emf of the generator as a function of time. emf0 is the peak emf. The period is T = 1/f = 2π/ω, where f is the frequency. Note that the script E stands for emf.

The fact that the peak emf, $emf_0 = NAB\omega$, makes good sense. The greater the number of coils, the larger their area, and the stronger the field, the greater the output voltage. It is interesting that the faster the generator is spun (greater ω), the greater the emf. This is noticeable on bicycle generators—at least the cheaper varieties. One of the authors as a juvenile found it amusing to ride his bicycle fast enough to burn out his lights, until he had to ride home lightless one dark night. Figure 4 shows a scheme by which a generator can be made to produce pulsed DC. More elaborate arrangements of multiple coils and split rings can produce smoother DC, although electronic rather than mechanical means are usually used to make ripple-free DC.

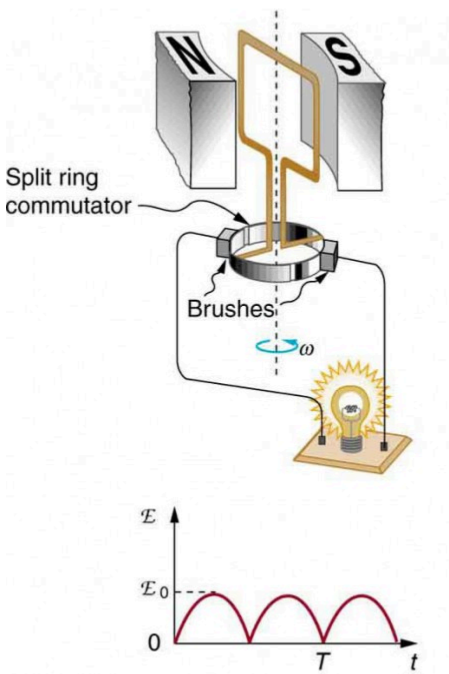


Figure 4. Split rings, called commutators, produce a pulsed DC emf output in this configuration.

Example 2. Calculating the Maximum Emf of a Generator

Calculate the maximum emf, emf_0, of the generator that was the subject of Example 1.

Strategy

Once ω, the angular velocity, is determined, $emf_0 = NAB\omega$ can be used to find emf_0. All other quantities are known.

Solution

Angular velocity is defined to be the change in angle per unit time:

$$\omega = \frac{\Delta\theta}{\Delta t}.$$

One-fourth of a revolution is π/2 radians, and the time is 0.0150 s; thus,

$$\begin{aligned} \omega &= \frac{\pi/2 \text{ rad}}{0.0150 \text{ s}} \\ &= 104.7 \text{ rad/s} \end{aligned}.$$

104.7 rad/s is exactly 1000 rpm. We substitute this value for ω and the information from the previous example into $emf_0 = NAB\omega$, yielding

$$\begin{aligned}\text{emf}_0 &= NAB\omega\\ &= 200\left(7.85\times10^{-3}\ \text{m}^2\right)(1.25\ \text{T})(104.7\ \text{rad/s})\\ &= 206\ \text{V}\end{aligned}$$

.

Discussion

The maximum emf is greater than the average emf of 131 V found in the previous example, as it should be.

In real life, electric generators look a lot different than the figures in this section, but the principles are the same. The source of mechanical energy that turns the coil can be falling water (hydropower), steam produced by the burning of fossil fuels, or the kinetic energy of wind. Figure 5 shows a cutaway view of a steam turbine; steam moves over the blades connected to the shaft, which rotates the coil within the generator.

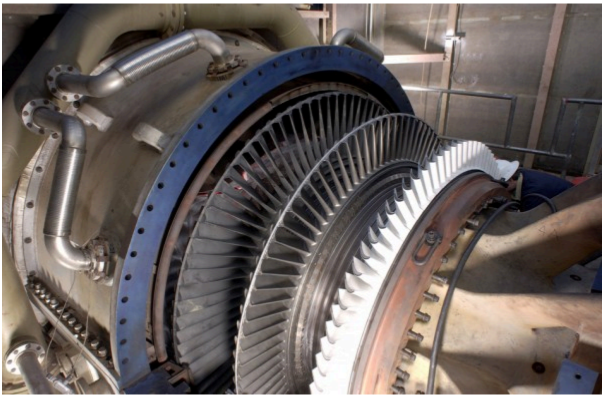


Figure 5. Steam turbine/generator. The steam produced by burning coal impacts the turbine blades, turning the shaft which is connected to the generator. (credit: Nabonaco, Wikimedia Commons)

Generators illustrated in this section look very much like the motors illustrated previously. This is not coincidental. In fact, a motor becomes a generator when its shaft rotates. Certain early automobiles used their starter motor as a generator. In Back Emf, we shall further explore the action of a motor as a generator.

Section Summary

- An electric generator rotates a coil in a magnetic field, inducing an emf given as a function of time by
$$\text{emf} = 2B\ell v \sin \omega t,$$
where A is the area of an N-turn coil rotated at a constant angular velocity ω in a uniform magnetic field B.
- The peak emf emf_0 of a generator is
$$\text{emf}_0 = NAB\omega$$

Conceptual Questions

1. Using RHR-1, show that the emfs in the sides of the generator loop in Figure 4 are in the same sense and thus add.
2. The source of a generator's electrical energy output is the work done to turn its coils. How is the work needed to turn the generator related to Lenz's law?

Problems & Exercises

1. Calculate the peak voltage of a generator that rotates its 200-turn, 0.100 m diameter coil at 3600 rpm in a 0.800 T field.

2. At what angular velocity in rpm will the peak voltage of a generator be 480 V, if its 500-turn, 8.00 cm diameter coil rotates in a 0.250 T field?

3. What is the peak emf generated by rotating a 1000-turn, 20.0 cm diameter coil in the Earth's 5.00×10^{-5} T magnetic field, given the plane of the coil is originally perpendicular to the Earth's field and is rotated to be parallel to the field in 10.0 ms?

4. What is the peak emf generated by a 0.250 m radius, 500-turn coil is rotated one-fourth of a revolution in 4.17 ms, originally having its plane perpendicular to a uniform magnetic field. (This is 60 rev/s.)

5. (a) A bicycle generator rotates at 1875 rad/s, producing an 18.0 V peak emf. It has a 1.00 by 3.00 cm rectangular coil in a 0.640 T field. How many turns are in the coil? (b) Is this number of turns of wire practical for a 1.00 by 3.00 cm coil?

6. **Integrated Concepts** This problem refers to the bicycle generator considered in the previous problem. It is driven by a 1.60 cm diameter wheel that rolls on the outside rim of the bicycle tire. (a) What is the velocity of the bicycle if the generator's angular velocity is 1875 rad/s? (b) What is the maximum emf of the generator when the bicycle moves at

10.0 m/s, noting that it was 18.0 V under the original conditions? (c) If the sophisticated generator can vary its own magnetic field, what field strength will it need at 5.00 m/s to produce a 9.00 V maximum emf?

7. (a) A car generator turns at 400 rpm when the engine is idling. Its 300-turn, 5.00 by 8.00 cm rectangular coil rotates in an adjustable magnetic field so that it can produce sufficient voltage even at low rpms. What is the field strength needed to produce a 24.0 V peak emf? (b) Discuss how this required field strength compares to those available in permanent and electromagnets.

8. Show that if a coil rotates at an angular velocity ω, the period of its AC output is $2\pi/\omega$.

9. A 75-turn, 10.0 cm diameter coil rotates at an angular velocity of 8.00 rad/s in a 1.25 T field, starting with the plane of the coil parallel to the field. (a) What is the peak emf? (b) At what time is the peak emf first reached? (c) At what time is the emf first at its most negative? (d) What is the period of the AC voltage output?

10. (a) If the emf of a coil rotating in a magnetic field is zero at $t = 0$, and increases to its first peak at $t = 0.100$ ms, what is the angular velocity of the coil? (b) At what time will its next maximum occur? (c) What is the period of the output? (d) When is the output first one-fourth of its maximum? (e) When is it next one-fourth of its maximum?

11. **Unreasonable Results** A 500-turn coil with a $0.250\ \text{m}^2$ area is spun in the Earth's 5.00×10^{-5} T field, producing a 12.0 kV maximum emf. (a) At what angular velocity must the

coil be spun? (b) What is unreasonable about this result? (c) Which assumption or premise is responsible?

Glossary

electric generator:
: a device for converting mechanical work into electric energy; it induces an emf by rotating a coil in a magnetic field

emf induced in a generator coil:
: $\text{emf} = NAB\omega \sin \omega t$, where A is the area of an N-turn coil rotated at a constant angular velocity ω in a uniform magnetic field B, over a period of time t

peak emf:
: $\text{emf}_0 = NAB\omega$

Selected Solutions to Problems & Exercises

1. 474 V

3. 0.247 V

5. (a) 50 (b) yes

7. (a) 0.477 T (b) This field strength is small enough that it can be obtained using either a permanent magnet or an electromagnet.

9. (a) 5.89 V (b) At $t = 0$ (c) 0.393 s (d) 0.785 s

11. (a) 1.92×10^6 rad/s (b) This angular velocity is

unreasonably high, higher than can be obtained for any mechanical system. (c) The assumption that a voltage as great as 12.0 kV could be obtained is unreasonable.

61. Back Emf

Learning Objective

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

- Explain what back emf is and how it is induced.

It has been noted that motors and generators are very similar. Generators convert mechanical energy into electrical energy, whereas motors convert electrical energy into mechanical energy. Furthermore, motors and generators have the same construction. When the coil of a motor is turned, magnetic flux changes, and an emf (consistent with Faraday's law of induction) is induced. The motor thus acts as a generator whenever its coil rotates. This will happen whether the shaft is turned by an external input, like a belt drive, or by the action of the motor itself. That is, when a motor is doing work and its shaft is turning, an emf is generated. Lenz's law tells us the emf opposes any change, so that the input emf that powers the motor will be opposed by the motor's self-generated emf, called the *back emf* of the motor. (See Figure 1.)

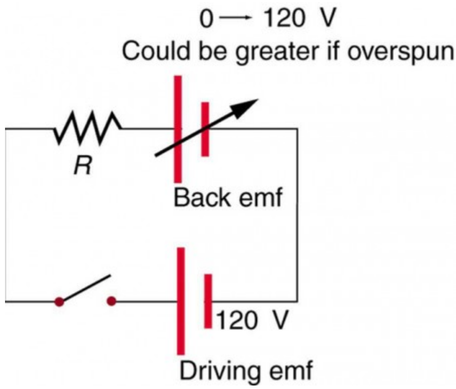


Figure 1. The coil of a DC motor is represented as a resistor in this schematic. The back emf is represented as a variable emf that opposes the one driving the motor. Back emf is zero when the motor is not turning, and it increases proportionally to the motor's angular velocity.

Back emf is the generator output of a motor, and so it is proportional to the motor's angular velocity ω. It is zero when the motor is first turned on, meaning that the coil receives the full driving voltage and the motor draws maximum current when it is on but not turning. As the motor turns faster and faster, the back emf grows, always opposing the driving emf, and reduces the voltage across the coil and the amount of current it draws. This effect is noticeable in a number of situations. When a vacuum cleaner, refrigerator, or washing machine is first turned on, lights in the same circuit dim briefly due to the IR drop produced in feeder lines by the large current drawn by the motor. When a motor first comes on, it draws more current than when it runs at its normal operating speed. When a mechanical load is placed on the motor, like an electric wheelchair going up a hill, the motor slows, the back emf drops, more current flows, and more work can be done. If the motor runs at too low a speed, the larger current can overheat it (via resistive power in the coil, $P=I^2R$), perhaps even burning it out.

On the other hand, if there is no mechanical load on the motor, it will increase its angular velocity ω until the back emf is nearly equal to the driving emf. Then the motor uses only enough energy to overcome friction.

Consider, for example, the motor coils represented in Figure 1. The coils have a 0.400 Ω equivalent resistance and are driven by a 48.0 V emf. Shortly after being turned on, they draw a current $I = V/R = (48.0\text{V})/(0.400\ \Omega) = 120\ \text{A}$ and, thus, dissipate $P = I^2R = 5.76$ kW of energy as heat transfer. Under normal operating conditions for this motor, suppose the back emf is 40.0 V. Then at operating speed, the total voltage across the coils is 8.0 V (48.0 V minus the 40.0 V back emf), and the current drawn is $I = V/R = (8.0\ \text{V})/(0.400\ \Omega) = 20\ \text{A}$. Under normal load, then, the power dissipated is $P = IV = (20\ \text{A})/(8.0\ \text{V}) = 160\ \text{W}$. The latter will not cause a problem for this motor, whereas the former 5.76 kW would burn out the coils if sustained.

Section Summary

- Any rotating coil will have an induced emf—in motors, this is called back emf, since it opposes the emf input to the motor.

Conceptual Questions

1. Suppose you find that the belt drive connecting a powerful motor to an air conditioning unit is broken and the motor is running freely. Should you be worried that the motor is consuming a great deal of energy for no useful purpose? Explain why or why not.

Problems & Exercises

1. Suppose a motor connected to a 120 V source draws 10.0 A when it first starts. (a) What is its resistance? (b) What current does it draw at its normal operating speed when it develops a 100 V back emf?

2. A motor operating on 240 V electricity has a 180 V back emf at operating speed and draws a 12.0 A current. (a) What is its resistance? (b) What current does it draw when it is first started?

3. What is the back emf of a 120 V motor that draws 8.00 A at its normal speed and 20.0 A when first starting?

4. The motor in a toy car operates on 6.00 V, developing a 4.50 V back emf at normal speed. If it draws 3.00 A at normal speed, what current does it draw when starting?

5. **Integrated Concepts** The motor in a toy car is powered by four batteries in series, which produce a total emf of 6.00 V. The motor draws 3.00 A and develops a 4.50 V back emf at normal speed. Each battery has a 0.100 Ω internal resistance. What is the resistance of the motor?

Glossary

back emf:
: the emf generated by a running motor, because it consists of a coil turning in a magnetic field; it opposes the voltage

powering the motor

Exercises

1. (a) 12.00 Ω (b) 1.67 A

3. 72.0 V

5. 0.100 Ω

62. Transformers

Learning Objectives

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

- Explain how a transformer works.
- Calculate voltage, current, and/or number of turns given the other quantities.

Transformers do what their name implies—they transform voltages from one value to another (The term voltage is used rather than emf, because transformers have internal resistance). For example, many cell phones, laptops, video games, and power tools and small appliances have a transformer built into their plug-in unit (like that in Figure 1) that changes 120 V or 240 V AC into whatever voltage the device uses. Transformers are also used at several points in the power distribution systems, such as illustrated in Figure 2. Power is sent long distances at high voltages, because less current is required for a given amount of power, and this means less line loss, as was discussed previously. But high voltages pose greater hazards, so that transformers are employed to produce lower voltage at the user's location.

Figure 1. The plug-in transformer has become increasingly familiar with the proliferation of electronic devices that operate on voltages other than common 120 V AC. Most are in the 3 to 12 V range. (credit: Shop Xtreme)

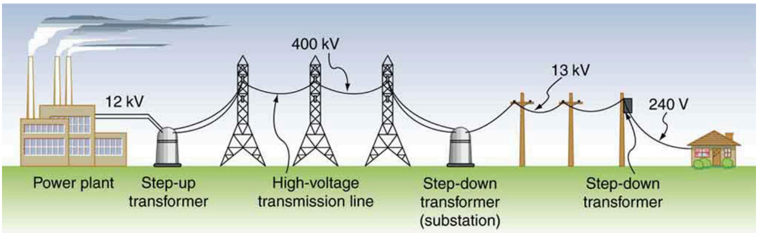


Figure 2. Transformers change voltages at several points in a power distribution system. Electric power is usually generated at greater than 10 kV, and transmitted long distances at voltages over 200 kV—sometimes as great as 700 kV—to limit energy losses. Local power distribution to neighborhoods or industries goes through a substation and is sent short distances at voltages ranging from 5 to 13 kV. This is reduced to 120, 240, or 480 V for safety at the individual user site.

The type of transformer considered in this text—see Figure 3—is based on Faraday's law of induction and is very similar in construction to the apparatus Faraday used to demonstrate magnetic fields could cause currents. The two coils are called the *primary* and *secondary coils*. In normal use, the input voltage is placed on the primary, and the secondary produces the transformed output voltage. Not only does the iron core trap the magnetic field created by the primary coil, its magnetization increases the field strength. Since the input voltage is AC, a time-varying magnetic flux is sent to the secondary, inducing its AC output voltage.

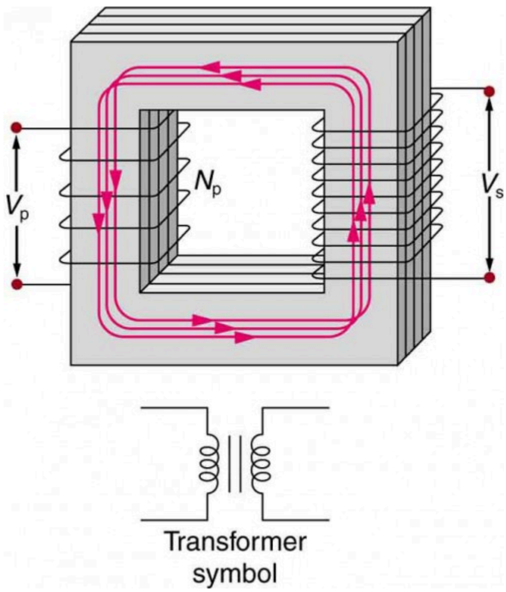


Figure 3. A typical construction of a simple transformer has two coils wound on a ferromagnetic core that is laminated to minimize eddy currents. The magnetic field created by the primary is mostly confined to and increased by the core, which transmits it to the secondary coil. Any change in current in the primary induces a current in the secondary.

For the simple transformer shown in Figure 3, the output voltage V_s depends almost entirely on the input voltage V_p and the ratio of the number of loops in the primary and secondary coils. Faraday's law of induction for the secondary coil gives its induced output voltage V_s to be

$$V_\mathrm{s} = -N_\mathrm{s}\frac{\Delta\Phi}{\Delta t},$$

where N_s is the number of loops in the secondary coil and $\Delta\Phi/\Delta t$ is the rate of change of magnetic flux. Note that the output voltage equals the induced emf (V_s=emf_s), provided coil resistance is small (a reasonable assumption for transformers). The cross-sectional area of the coils is the same on either side, as is the magnetic field strength, and so $\Delta\Phi/\Delta t$ is the same on either side. The input primary voltage V_p is also related to changing flux by

$$V_p = -N_\mathrm{p}\frac{\Delta\Phi}{\Delta t}.$$

The reason for this is a little more subtle. Lenz's law tells us that the primary coil opposes the change in flux caused by the input voltage V_p, hence the minus sign (This is an example of *self-inductance*, a topic to be explored in some detail in later sections). Assuming negligible coil resistance, Kirchhoff's loop rule tells us that the induced emf exactly equals the input voltage. Taking the ratio of these last two equations yields a useful relationship:

$$\frac{V_\mathrm{s}}{V_\mathrm{p}} = \frac{N_\mathrm{s}}{N_\mathrm{p}}$$

This is known as the *transformer equation*, and it simply states that the ratio of the secondary to primary voltages in a transformer equals the ratio of the number of loops in their coils. The output voltage of a transformer can be less than, greater than, or equal to the input voltage, depending on the ratio of the number of loops in their coils. Some transformers even provide a variable output by allowing connection to be made at different points on the secondary coil. A *step-up transformer* is one that increases voltage, whereas a *step-down transformer* decreases voltage. Assuming, as we have, that resistance is negligible, the electrical power output of a transformer equals its input. This is nearly true in practice—transformer efficiency often exceeds 99%. Equating the power input and output,

$$P_p = I_pV_p = I_sV_s = P_s.$$

Rearranging terms gives

$$\frac{V_s}{V_p} = \frac{I_p}{I_s}.$$

Combining this with $\frac{V_s}{V_p} = \frac{N_s}{N_p}$, we find that

$$\frac{I_s}{I_p} = \frac{N_p}{N_s}$$

is the relationship between the output and input currents of a transformer. So if voltage increases, current decreases. Conversely, if voltage decreases, current increases.

Example 1. Calculating Characteristics of a Step-Up Transformer

A portable x-ray unit has a step-up transformer, the 120 V input of which is transformed to the 100 kV output needed by the x-ray tube. The primary has 50 loops and draws a current of 10.00 A when in use. (a) What is the number of loops in the secondary? (b) Find the current output of the secondary.

Strategy and Solution for (a)

We solve $\frac{V_s}{V_p} = \frac{N_s}{N_p}$ for N_s for N_s, the number of loops in the secondary, and enter the known values. This gives

$$\begin{aligned} N_s &= N_p \frac{V_s}{V_p} \\ &= (50)\frac{100{,}000 \text{ V}}{120 \text{ V}} = 4.17 \times 10^4 \end{aligned}.$$

Discussion for (a)

A large number of loops in the secondary (compared with the primary) is required to produce such a large voltage. This would be true for neon sign transformers and those supplying high voltage inside TVs and CRTs.

Strategy and Solution for (b)

We can similarly find the output current of the secondary by solving $\frac{I_s}{I_p} = \frac{N_p}{N_s}$ for I_s for I_s and entering known values. This gives

$$\begin{aligned} I_s &= I_p \frac{N_p}{N_s} \\ &= (10.00 \text{ A})\frac{50}{4.17 \times 10^4} = 12.0 \text{ mA} \end{aligned}.$$

Discussion for (b)

As expected, the current output is significantly less than the input. In certain spectacular demonstrations, very large voltages are used to produce long arcs, but they are relatively safe because the transformer output does not supply a large current. Note that the power input here is $P_p = I_p V_p = (10.00\ \text{A})(120\ \text{V}) = 1.20\ \text{kW}$. This equals the power output $P_p = I_s V_s = (12.0\ \text{mA})(100\ \text{kV}) = 1.20\ \text{kW}$, as we assumed in the derivation of the equations used.

The fact that transformers are based on Faraday's law of induction makes it clear why we cannot use transformers to change DC voltages. If there is no change in primary voltage, there is no voltage induced in the secondary. One possibility is to connect DC to the primary coil through a switch. As the switch is opened and closed, the secondary produces a voltage like that in Figure 4. This is not really a practical alternative, and AC is in common use wherever it is necessary to increase or decrease voltages.

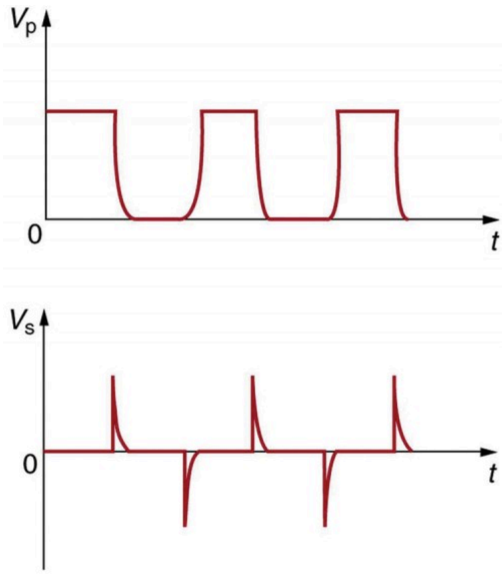


Figure 4. Transformers do not work for pure DC voltage input, but if it is switched on and off as on the top graph, the output will look something like that on the bottom graph. This is not the sinusoidal AC most AC appliances need.

Example 2. Calculating Characteristics of a Step-Down Transformer

A battery charger meant for a series connection of ten nickel-cadmium batteries (total emf of 12.5 V DC) needs to have a 15.0 V output to charge the batteries. It uses a step-down transformer with a 200-loop primary and a 120 V input. (a) How many loops should there be in the secondary coil? (b) If the charging current is 16.0 A, what is the input current?

Strategy and Solution for (a)

You would expect the secondary to have a small number of loops. Solving $\frac{V_s}{V_p} = \frac{N_s}{N_p}$ for N_s for N_s and entering known values gives

$$\begin{aligned} N_s &= N_p \frac{V_s}{V_p} \\ &= (200) \frac{15.0 \text{ V}}{120 \text{ V}} = 25 \end{aligned}$$

Strategy and Solution for (b)

The current input can be obtained by solving $\frac{I_s}{I_p} = \frac{N_p}{N_s}$ for I_p and entering known values. This gives

$$\begin{aligned} I_p &= I_s \frac{N_s}{N_p} \\ &= (16.0 \text{ A}) \frac{25}{200} = 2.00 \text{ A} \end{aligned}$$

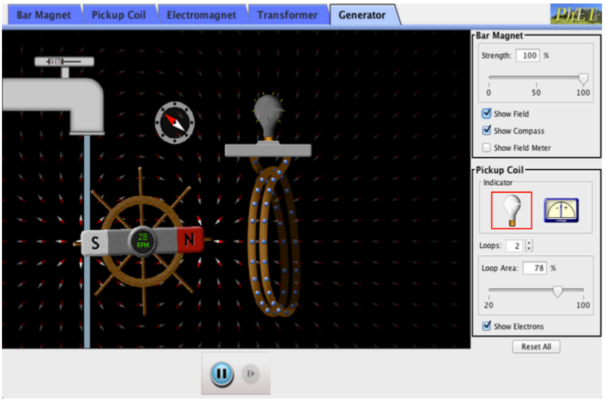
Discussion

The number of loops in the secondary is small, as expected for a step-down transformer. We also see that a small input current produces a larger output current in a step-down transformer. When transformers are used to operate large magnets, they sometimes have a small number of very heavy loops in the secondary. This allows the secondary to have low internal resistance and produce large currents. Note again that this solution is based on the assumption of 100% efficiency—or power out equals power in ($P_p = P_s$)—reasonable for good transformers. In this case the primary and secondary power is 240 W. (Verify this for yourself as a consistency check.) Note that the Ni-Cd batteries need to be charged from a DC power source (as would a 12 V battery). So the AC output of the secondary coil needs to be converted into DC. This is done using something called a rectifier, which uses devices called diodes that allow only a one-way flow of current.

Transformers have many applications in electrical safety systems, which are discussed in Electrical Safety: Systems and Devices.

PhET Explorations: Generator

Generate electricity with a bar magnet! Discover the physics behind the phenomena by exploring magnets and how you can use them to make a bulb light.



Click to download the simulation. Run using Java.

Section Summary

- Transformers use induction to transform voltages from one value to another.
- For a transformer, the voltages across the primary and secondary coils are related by
$$\frac{V_s}{V_p} = \frac{N_s}{N_p},$$
where V_p and V_s are the voltages across primary and secondary coils having N_p and N_s turns.
- The currents I_p and I_s in the primary and secondary coils are related by $\frac{I_s}{I_p} = \frac{N_p}{N_s}$.

- A step-up transformer increases voltage and decreases current, whereas a step-down transformer decreases voltage and increases current.

Conceptual Questions

1. Explain what causes physical vibrations in transformers at twice the frequency of the AC power involved.

Problems & Exercises

1. A plug-in transformer, like that in Figure 4, supplies 9.00 V to a video game system. (a) How many turns are in its secondary coil, if its input voltage is 120 V and the primary coil has 400 turns? (b) What is its input current when its output is 1.30 A?

2. An American traveler in New Zealand carries a transformer to convert New Zealand's standard 240 V to 120 V so that she can use some small appliances on her trip. (a) What is the ratio of turns in the primary and secondary coils of her transformer? (b) What is the ratio of input to output current? (c) How could a New Zealander traveling in the United States use this same transformer to power her 240 V appliances from 120 V?

3. A cassette recorder uses a plug-in transformer to convert 120 V to 12.0 V, with a maximum current output of

200 mA. (a) What is the current input? (b) What is the power input? (c) Is this amount of power reasonable for a small appliance?

4. (a) What is the voltage output of a transformer used for rechargeable flashlight batteries, if its primary has 500 turns, its secondary 4 turns, and the input voltage is 120 V? (b) What input current is required to produce a 4.00 A output? (c) What is the power input?

5. (a) The plug-in transformer for a laptop computer puts out 7.50 V and can supply a maximum current of 2.00 A. What is the maximum input current if the input voltage is 240 V? Assume 100% efficiency. (b) If the actual efficiency is less than 100%, would the input current need to be greater or smaller? Explain.

6. A multipurpose transformer has a secondary coil with several points at which a voltage can be extracted, giving outputs of 5.60, 12.0, and 480 V. (a) The input voltage is 240 V to a primary coil of 280 turns. What are the numbers of turns in the parts of the secondary used to produce the output voltages? (b) If the maximum input current is 5.00 A, what are the maximum output currents (each used alone)?

7. A large power plant generates electricity at 12.0 kV. Its old transformer once converted the voltage to 335 kV. The secondary of this transformer is being replaced so that its output can be 750 kV for more efficient cross-country transmission on upgraded transmission lines. (a) What is the ratio of turns in the new secondary compared with the old secondary? (b) What is the ratio of new current output to old output (at 335 kV) for the same power? (c) If the

upgraded transmission lines have the same resistance, what is the ratio of new line power loss to old?

8. If the power output in the previous problem is 1000 MW and line resistance is 2.00 Ω, what were the old and new line losses?

9. **Unreasonable Results** The 335 kV AC electricity from a power transmission line is fed into the primary coil of a transformer. The ratio of the number of turns in the secondary to the number in the primary is $N_s/N_p = 1000$. (a) What voltage is induced in the secondary? (b) What is unreasonable about this result? (c) Which assumption or premise is responsible?

10. **Construct Your Own Problem** Consider a double transformer to be used to create very large voltages. The device consists of two stages. The first is a transformer that produces a much larger output voltage than its input. The output of the first transformer is used as input to a second transformer that further increases the voltage. Construct a problem in which you calculate the output voltage of the final stage based on the input voltage of the first stage and the number of turns or loops in both parts of both transformers (four coils in all). Also calculate the maximum output current of the final stage based on the input current. Discuss the possibility of power losses in the devices and the effect on the output current and power.

Glossary

transformer:
: a device that transforms voltages from one value to another using induction

transformer equation:
: the equation showing that the ratio of the secondary to primary voltages in a transformer equals the ratio of the number of loops in their coils; $\frac{V_s}{V_p} = \frac{N_s}{N_p}$

step-up transformer:
: a transformer that increases voltage

step-down transformer:
: a transformer that decreases voltsage

Selected Solutions to Problems & Exercises

1. (a) 30.0 (b) 9.75×10^{-2} A

3. (a) 20.0 mA (b) 2.40 W (c) Yes, this amount of power is quite reasonable for a small appliance.

5. (a) 0.063 A (b) Greater input current needed.

7. (a) 2.2 (b) 0.45 (c) 0.20, or 20.0%

9. (a) 335 MV (b) way too high, well beyond the breakdown voltage of air over reasonable distances (c) input voltage is too high

63. Electrical Safety: Systems and Devices

Learning Objectives

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

- Explain how various modern safety features in electric circuits work, with an emphasis on how induction is employed.

Electricity has two hazards. A *thermal hazard* occurs when there is electrical overheating. A *shock hazard* occurs when electric current passes through a person. Both hazards have already been discussed. Here we will concentrate on systems and devices that prevent electrical hazards. Figure 1 shows the schematic for a simple AC circuit with no safety features. This is not how power is distributed in practice. Modern household and industrial wiring requires the *three-wire system*, shown schematically in Figure 2, which has several safety features. First is the familiar *circuit breaker* (or *fuse*) to prevent thermal overload. Second, there is a protective *case* around the appliance, such as a toaster or refrigerator. The case's safety feature is that it prevents a person from touching exposed wires and coming into electrical contact with the circuit, helping prevent shocks.

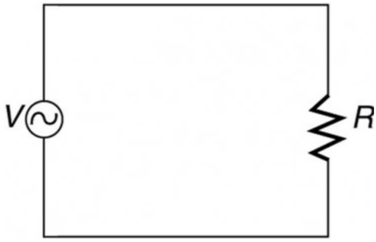


Figure 1. Schematic of a simple AC circuit with a voltage source and a single appliance represented by the resistance R. There are no safety features in this circuit.

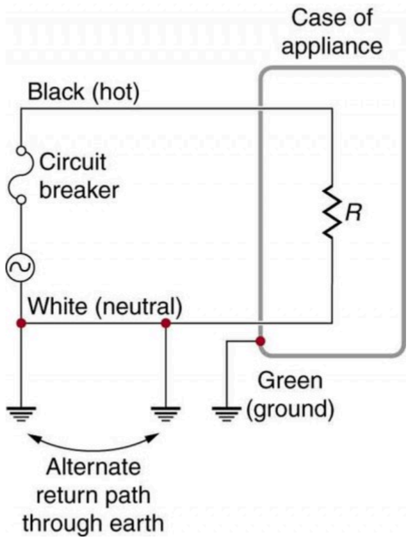


Figure 2. The three-wire system connects the neutral wire to the earth at the voltage source and user location, forcing it to be at zero volts and supplying an alternative return path for the current through the earth. Also grounded to zero volts is the case of the appliance. A circuit breaker or fuse protects against thermal overload and is in series on the active (live/hot) wire. Note that wire insulation colors vary with region and it is essential to check locally to determine which color codes are in use (and even if they were followed in the particular installation).

There are *three connections to earth or ground* (hereafter referred to as "earth/ground") shown in Figure 2. Recall that an earth/ground connection is a low-resistance path directly to the earth. The two earth/ground connections on the *neutral wire* force it to be at zero volts relative to the earth, giving the wire its name. This wire is therefore safe to touch even if its insulation, usually white, is missing. The neutral wire is the return path for the current to follow to complete the circuit. Furthermore, the two earth/ground connections supply an alternative path through the earth, a good conductor, to complete the circuit. The earth/ground connection closest to the power source could be at the generating plant, while the other is at the user's location. The third earth/ground is to the case of the appliance, through the green *earth/ground wire*, forcing the case, too, to be at zero volts. The *live* or *hot wire* (hereafter referred to as "live/hot") supplies voltage and current to operate the appliance. Figure 3 shows a more pictorial version of how the three-wire system is connected through a three-prong plug to an appliance.

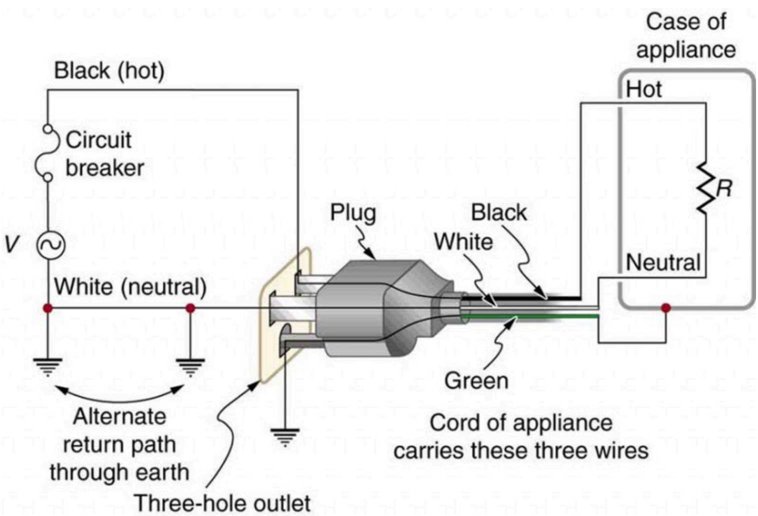


Figure 3. The standard three-prong plug can only be inserted in one way, to assure proper function of the three-wire system.

A note on insulation color-coding: Insulating plastic is color-coded to identify live/hot, neutral and ground wires but these codes vary around the world. Live/hot wires may be brown, red, black, blue or grey. Neutral wire may be blue, black or white. Since the same color may be used for live/hot or neutral in different parts of the world, it is essential to determine the color code in your region. The only exception is the earth/ground wire which is often green but may be yellow or just bare wire. Striped coatings are sometimes used for the benefit of those who are colorblind. The three-wire system replaced the older two-wire system, which lacks an earth/ground wire. Under ordinary circumstances, insulation on the live/hot and neutral wires prevents the case from being directly in the circuit, so that the earth/ground wire may seem like double protection. Grounding the case solves more than one problem, however. The simplest problem is worn insulation on the live/hot wire that allows it to contact the case, as shown in Figure 4. Lacking an earth/ground connection (some people cut the third prong off the plug because they only have outdated two hole receptacles), a severe shock is possible. This is particularly dangerous in the kitchen, where a good connection to earth/ground is available through water on the floor or a water faucet. With the earth/ground connection intact, the circuit breaker will trip, forcing repair of the appliance. Why are some appliances still sold with two-prong plugs? These have nonconducting cases, such as power tools with impact resistant plastic cases, and are called *doubly insulated*. Modern two-prong plugs can be inserted into the asymmetric standard outlet in only one way, to ensure proper connection of live/hot and neutral wires.

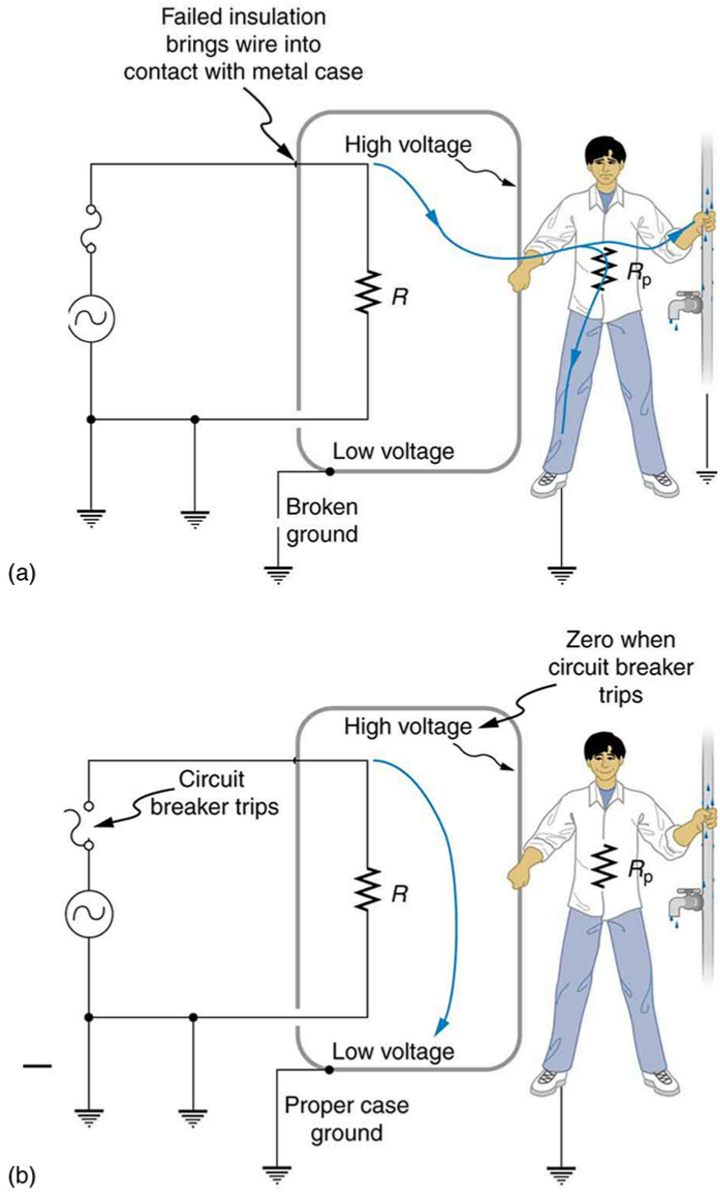


Figure 4. Worn insulation allows the live/hot wire to come into direct contact with the metal case of this appliance. (a) The earth/ground connection being broken, the person is severely shocked. The appliance may operate normally in this situation. (b) With a proper earth/ground, the circuit breaker trips,

forcing repair of the appliance.

Electromagnetic induction causes a more subtle problem that is solved by grounding the case. The AC current in appliances can induce an emf on the case. If grounded, the case voltage is kept near zero, but if the case is not grounded, a shock can occur as pictured in Figure 5. Current driven by the induced case emf is called a *leakage current*, although current does not necessarily pass from the resistor to the case.

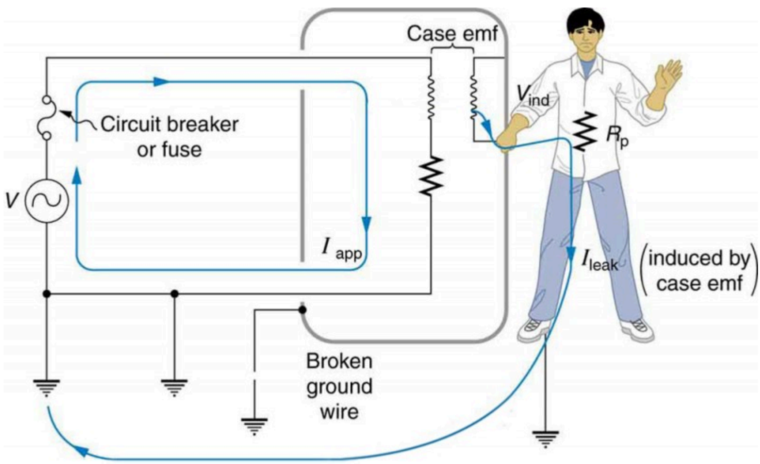


Figure 5. AC currents can induce an emf on the case of an appliance. The voltage can be large enough to cause a shock. If the case is grounded, the induced emf is kept near zero.

A *ground fault interrupter* (GFI) is a safety device found in updated kitchen and bathroom wiring that works based on electromagnetic induction. GFIs compare the currents in the live/hot and neutral wires. When live/hot and neutral currents are not equal, it is almost always because current in the neutral is less than in the live/hot wire. Then some of the current, again called a leakage current, is returning to the voltage source by a path other than through the neutral wire. It is assumed that this path presents a hazard, such as shown in Figure 6. GFIs are usually set to interrupt the circuit if the leakage current is greater than 5 mA, the accepted maximum harmless shock. Even if the leakage current goes safely to earth/ground through an intact earth/ground wire, the GFI will trip, forcing repair of the leakage.

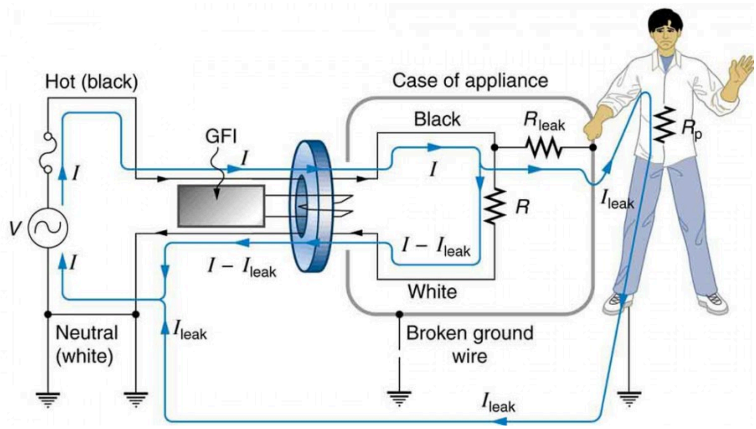


Figure 6. A ground fault interrupter (GFI) compares the currents in the live/hot and neutral wires and will trip if their difference exceeds a safe value. The leakage current here follows a hazardous path that could have been prevented by an intact earth/ground wire.

Figure 7 shows how a GFI works. If the currents in the live/hot and neutral wires are equal, then they induce equal and opposite emfs in the coil. If not, then the circuit breaker will trip.

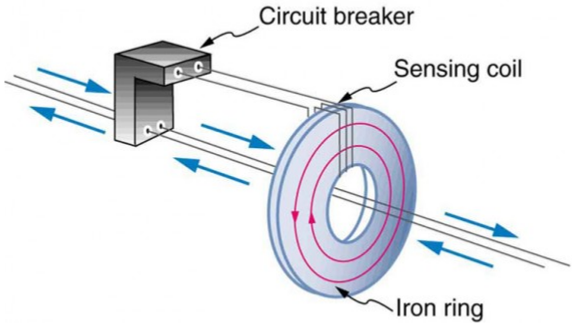


Figure 7. A GFI compares currents by using both to induce an emf in the same coil. If the currents are equal, they will induce equal but opposite emfs.

Another induction-based safety device is the *isolation transformer*, shown in Figure 8. Most isolation transformers have equal input and output voltages. Their function is to put a large resistance between the original voltage source and the device being operated. This prevents a complete circuit between them, even in the circumstance shown. There is a complete circuit through the appliance. But there is not a complete circuit for current to flow through the person in the figure, who is touching only one of the transformer's output wires, and neither output wire is grounded. The appliance is isolated from the original voltage source by the high resistance of the material between the transformer coils, hence the name isolation transformer. For current to flow through the person, it must pass through the high-resistance material between the coils, through the wire, the person, and back through the earth—a path with such a large resistance that the current is negligible.

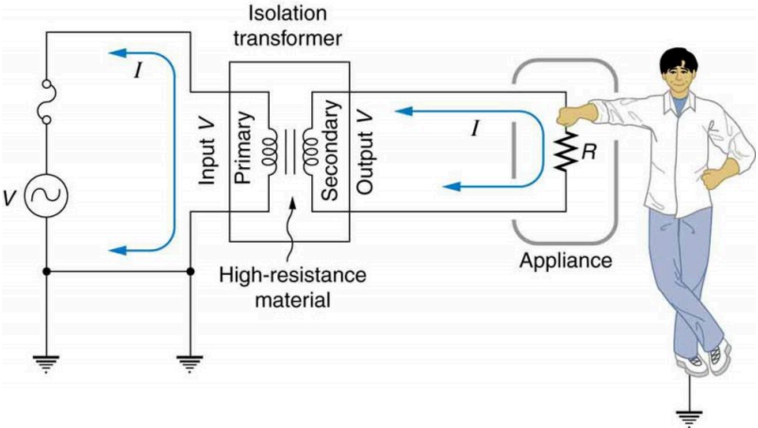


Figure 8. An isolation transformer puts a large resistance between the original voltage source and the device, preventing a complete circuit between them.

The basics of electrical safety presented here help prevent many electrical hazards. Electrical safety can be pursued to greater depths. There are, for example, problems related to different earth/ground connections for appliances in close proximity. Many other examples are found in hospitals. Microshock-sensitive patients, for instance, require special protection. For these people, currents as low as 0.1 mA may cause ventricular fibrillation. The interested reader can use the material presented here as a basis for further study.

Section Summary

- Electrical safety systems and devices are employed to prevent thermal and shock hazards.
- Circuit breakers and fuses interrupt excessive currents to prevent thermal hazards.

- The three-wire system guards against thermal and shock hazards, utilizing live/hot, neutral, and earth/ground wires, and grounding the neutral wire and case of the appliance.
- A ground fault interrupter (GFI) prevents shock by detecting the loss of current to unintentional paths.
- An isolation transformer insulates the device being powered from the original source, also to prevent shock.
- Many of these devices use induction to perform their basic function.

Conceptual Questions

1. Does plastic insulation on live/hot wires prevent shock hazards, thermal hazards, or both?
2. Why are ordinary circuit breakers and fuses ineffective in preventing shocks?
3. A GFI may trip just because the live/hot and neutral wires connected to it are significantly different in length. Explain why.

Problems & Exercises

1. **Integrated Concepts** A short circuit to the grounded metal case of an appliance occurs as shown in Figure 9. The person touching the case is wet and only has a 3.00 kΩ resistance to earth/ground. (a) What is the voltage on the case if 5.00 mA flows through the person? (b) What is the

current in the short circuit if the resistance of the earth/ground wire is 0.200 Ω? (c) Will this trigger the 20.0 A circuit breaker supplying the appliance?

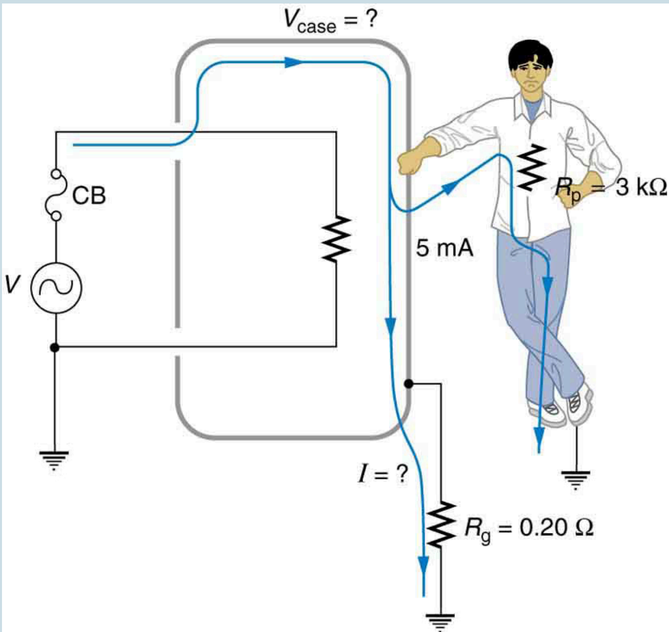


Figure 9. A person can be shocked even when the case of an appliance is grounded. The large short circuit current produces a voltage on the case of the appliance, since the resistance of the earth/ground wire is not zero.

Glossary

thermal hazard:
: the term for electrical hazards due to overheating

shock hazard:
: the term for electrical hazards due to current passing through a human

three-wire system:
: the wiring system used at present for safety reasons, with live, neutral, and ground wires

Selected Solutions to Problems & Exercises

1. (a) 15.0 V (b) 75.0 A (c) yes

PART VIII
ELECTROMAGNETIC WAVES

64. Introduction to Electromagnetic Waves

Figure 1. Human eyes detect these orange "sea goldie" fish swimming over a coral reef in the blue waters of the Gulf of Eilat (Red Sea) using visible light. (credit: Daviddarom, Wikimedia Commons)

The beauty of a coral reef, the warm radiance of sunshine, the sting of sunburn, the X-ray revealing a broken bone, even microwave popcorn—all are brought to us by *electromagnetic waves*. The list of the various types of electromagnetic waves, ranging from radio transmission waves to nuclear gamma-ray (γ-ray) emissions, is interesting in itself.

Even more intriguing is that all of these widely varied phenomena are different manifestations of the same thing—electromagnetic waves. (See Figure 2.) What are electromagnetic waves? How are they created, and how do they travel? How can we understand and organize their widely varying properties? What is their relationship to electric and magnetic effects? These and other questions will be explored.

Misconception Alert: Sound Waves vs. Radio Waves

Many people confuse sound waves with *radio waves*, one type of electromagnetic (EM) wave. However, sound and radio waves are completely different phenomena. Sound creates pressure variations (waves) in matter, such as air or water, or your eardrum. Conversely, radio waves are *electromagnetic waves*, like visible light, infrared, ultraviolet, X-rays, and gamma rays. EM waves don't need a medium in which to propagate; they can travel through a vacuum, such as outer space.

A radio works because sound waves played by the D.J. at the radio station are converted into electromagnetic waves, then encoded and transmitted in the radio-frequency range. The radio in your car receives the radio waves, decodes the information, and uses a speaker to change it back into a sound wave, bringing sweet music to your ears.

Discovering a New Phenomenon

Figure 2. The electromagnetic waves sent and received by this 50-foot radar dish antenna at Kennedy Space Center in Florida are not visible, but help track expendable launch vehicles with high-definition imagery. The first use of this C-band radar dish was for the launch of the Atlas V rocket sending the New Horizons probe toward Pluto. (credit: NASA)

It is worth noting at the outset that the general phenomenon of electromagnetic waves was predicted by theory before it was realized that light is a form of electromagnetic wave. The prediction was made by James Clerk Maxwell in the mid-19th century when he formulated a single theory combining all the electric and magnetic effects known by scientists at that time. "Electromagnetic waves" was the name he gave to the phenomena his theory predicted.

Such a theoretical prediction followed by experimental verification is an indication of the power of science in general, and physics in particular. The underlying connections and unity of physics allow certain great minds to solve puzzles without having all the pieces. The prediction of electromagnetic waves is one of the most spectacular examples of this power. Certain others, such as the prediction of antimatter, will be discussed in later modules.

65. Maxwell's Equations: Electromagnetic Waves Predicted and Observed

Learning Objective

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

- Restate Maxwell's equations.

The Scotsman James Clerk Maxwell (1831–1879) is regarded as the greatest theoretical physicist of the 19th century. (See Figure 1.) Although he died young, Maxwell not only formulated a complete electromagnetic theory, represented by *Maxwell's equations*, he also developed the kinetic theory of gases and made significant contributions to the understanding of color vision and the nature of Saturn's rings.

This black and white engraving shows physicist James Clerk Maxwell as a Victorian era gentleman dressed in bowtie, vest, and jacket, and sporting a full, graying beard and moustache.

Figure 1. James Clerk Maxwell, a 19th-century physicist, developed a theory that explained the relationship between electricity and magnetism and correctly predicted that visible light is caused by electromagnetic waves. (credit: G. J. Stodart)

Maxwell brought together all the work that had been done by brilliant physicists such as Oersted, Coulomb, Gauss, and Faraday, and added his own insights to develop the overarching theory of electromagnetism. Maxwell's equations are paraphrased here in words because their mathematical statement is beyond the level of this text. However, the equations illustrate how apparently simple mathematical statements can elegantly unite and express a multitude of concepts—why mathematics is the language of science.

Maxwell's Equations

1. *Electric field lines* originate on positive charges and terminate on negative charges. The electric field is defined as the force per unit charge on a test charge, and the strength of the force is related to the electric constant ε_0, also known as the permittivity of free space. From Maxwell's first equation we obtain a special form of Coulomb's law known as Gauss's law for electricity.
2. *Magnetic field lines* are continuous, having no beginning or end. No magnetic monopoles are known to exist. The strength of the magnetic force is related to the magnetic constant μ_0, also known as the permeability of free space. This second of Maxwell's equations is known as Gauss's law for magnetism.
3. A changing magnetic field induces an electromotive force (emf) and, hence, an electric field. The direction of the emf opposes the change. This third of Maxwell's equations is Faraday's law of induction, and includes Lenz's law.
4. Magnetic fields are generated by moving charges or by changing electric fields. This fourth

of Maxwell's equations encompasses Ampere's law and adds another source of magnetism—changing electric fields.

Maxwell's equations encompass the major laws of electricity and magnetism. What is not so apparent is the symmetry that Maxwell introduced in his mathematical framework. Especially important is his addition of the hypothesis that changing electric fields create magnetic fields. This is exactly analogous (and symmetric) to Faraday's law of induction and had been suspected for some time, but fits beautifully into Maxwell's equations.

Symmetry is apparent in nature in a wide range of situations. In contemporary research, symmetry plays a major part in the search for sub-atomic particles using massive multinational particle accelerators such as the new Large Hadron Collider at CERN.

Making Connections: Unification of Forces

Maxwell's complete and symmetric theory showed that electric and magnetic forces are not separate, but different manifestations of the same thing—the electromagnetic force. This classical unification of forces is one motivation for current attempts to unify the four basic forces in nature—the gravitational, electrical, strong, and weak nuclear forces.

Since changing electric fields create relatively weak magnetic fields, they could not be easily detected at the time of Maxwell's hypothesis. Maxwell realized, however, that oscillating charges, like those in AC circuits, produce changing electric fields. He predicted that these changing fields would propagate from the source like waves generated on a lake by a jumping fish.

The waves predicted by Maxwell would consist of oscillating electric and magnetic fields—defined to be an electromagnetic wave (EM wave). Electromagnetic waves would be capable of exerting forces on charges great distances from their source, and they might thus be detectable. Maxwell calculated that electromagnetic waves would propagate at a speed given by the equation

$$c = \frac{1}{\sqrt{\mu_0 \epsilon_0}}$$

When the values for μ_0 and ε_0 are entered into the equation for c , we find that

$$c = \frac{1}{\sqrt{\left(8.85 \times 10^{-12} \frac{\mathrm{C}^2}{\mathrm{N \cdot m^2}}\right)\left(4\pi \times 10^{-7} \frac{\mathrm{T \cdot m}}{\mathrm{A}}\right)}} = 300 \times 10^8 \ \mathrm{m/s}$$

which is the speed of light. In fact, Maxwell concluded that light is an electromagnetic wave having such wavelengths that it can be detected by the eye.

Other wavelengths should exist—it remained to be seen if they did. If so, Maxwell's theory and remarkable predictions would be verified, the greatest triumph of physics since Newton. Experimental verification came within a few years, but not before Maxwell's death.

Hertz's Observations

The German physicist Heinrich Hertz (1857–1894) was the first to generate and detect certain types of electromagnetic waves in the laboratory. Starting in 1887, he performed a series of experiments that not only confirmed the existence of electromagnetic waves, but also verified that they travel at the speed of light.

Hertz used an AC *RLC* (resistor-inductor-capacitor) circuit that resonates at a known frequency $f_0 = \dfrac{1}{2\pi\sqrt{LC}}$ and connected it to a loop of wire as shown in Figure 2. High voltages induced across the gap in the loop produced sparks that were visible evidence of the current in the circuit and that helped generate electromagnetic waves.

Across the laboratory, Hertz had another loop attached to another *RLC* circuit, which could be tuned (as the dial on a radio) to the same resonant frequency as the first and could, thus, be made to receive electromagnetic waves. This loop also had a gap across which sparks were generated, giving solid evidence that electromagnetic waves had been received.

The circuit diagram shows a simple circuit containing an alternating voltage source, a resistor R, capacitor C and a transformer, which provides the impedance. The transformer is shown to consist of two coils separated by a core. In parallel with the transformer is connected a wire loop labeled as Loop one Transmitter with a small gap that creates sparks across the gap. The sparks create electromagnetic waves, which are transmitted through the air to a similar loop next to it labeled as Loop two Receiver. These waves induce sparks in Loop two, and are detected by the tuner shown as a rectangular box connected to it.

Figure 2. The apparatus used by Hertz in 1887 to generate and detect electromagnetic waves. An RLC circuit connected to the first loop caused sparks across a gap in the wire loop and generated electromagnetic waves. Sparks across a gap in the second loop located across the laboratory gave evidence that the waves had been received.

Hertz also studied the reflection, refraction, and interference patterns of the electromagnetic waves he generated, verifying their wave character. He was able to determine wavelength from the interference patterns, and knowing their frequency, he could calculate the propagation speed using the equation $v = f\lambda$ (velocity—or speed—equals frequency times wavelength). Hertz was thus able to prove that electromagnetic waves travel at the speed of light. The SI unit for frequency, the hertz (1 Hz = 1 cycle/sec), is named in his honor.

Section Summary

- Electromagnetic waves consist of oscillating electric and magnetic fields and propagate at the speed of light *c*. They were predicted by Maxwell, who also showed that
$$c = \frac{1}{\sqrt{\mu_0 \epsilon_0}},$$
where μ_0 is the permeability of free space and ε_0 is the permittivity of free space.
- Maxwell's prediction of electromagnetic waves resulted from his formulation of a complete and symmetric theory of electricity and magnetism, known as Maxwell's equations.
- These four equations are paraphrased in this text, rather than presented numerically, and encompass the major laws of electricity and magnetism. First is Gauss's law for electricity, second is Gauss's law for magnetism, third is Faraday's law of induction, including Lenz's law, and fourth is Ampere's law in a symmetric formulation that adds another source of magnetism—changing electric fields.

Problems & Exercises

1. Verify that the correct value for the speed of light c is obtained when numerical values for the permeability and permittivity of free space (μ_0 and ε_0) are entered into the equation $c = \frac{1}{\sqrt{\mu_0 \epsilon_0}}$.
2. Show that, when SI units for μ_0 and ε_0 are entered, the units given by the right-hand side of the equation in the problem above are m/s.

Glossary

electromagnetic waves: radiation in the form of waves of electric and magnetic energy

Maxwell's equations: a set of four equations that comprise a complete, overarching theory of electromagnetism

RLC circuit: an electric circuit that includes a resistor, capacitor and inductor

hertz: an SI unit denoting the frequency of an electromagnetic wave, in cycles per second

speed of light: in a vacuum, such as space, the speed of light is a constant 3×10^8 m/s

electromotive force (emf): energy produced per unit charge, drawn from a source that produces an electrical current

electric field lines: a pattern of imaginary lines that extend between an electric source and charged objects in the surrounding area, with arrows pointed away from positively charged objects and toward negatively charged objects. The more lines in the pattern, the stronger the electric field in that region

magnetic field lines: a pattern of continuous, imaginary lines that emerge from and enter into opposite magnetic poles. The density of the lines indicates the magnitude of the magnetic field

66. Production of Electromagnetic Waves

Learning Objectives

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

- Describe the electric and magnetic waves as they move out from a source, such as an AC generator.
- Explain the mathematical relationship between the magnetic field strength and the electrical field strength.
- Calculate the maximum strength of the magnetic field in an electromagnetic wave, given the maximum electric field strength.

We can get a good understanding of *electromagnetic waves* (EM) by considering how they are produced. Whenever a current varies, associated electric and magnetic fields vary, moving out from the source like waves. Perhaps the easiest situation to visualize is a varying current in a long straight wire, produced by an AC generator at its center, as illustrated in Figure 1.

A long straight gray wire with an A C generator at its center, functioning as a broadcast antenna for electromagnetic waves, is shown. The wave distributions at four different times are shown in four different parts. Part a of the diagram shows a long straight gray wire with an A C generator at its center. The time is marked t equals zero. The bottom part of the antenna is positive and the upper end of the antenna is negative. An electric field E acting upward is shown by an upward arrow. Part b of the diagram shows a long straight gray wire with an A C generator at its center. The time is marked t equals capital T divided by four. The antenna has no polarity marked and a wave is shown to emerge from the A C source. An electric field E acting upward as shown by an upward arrow. The electric field E propagates away from the antenna at the speed of light, forming part of the electromagnetic wave from the A C source. A quarter portion of the wave is shown above the horizontal axis. Part c of the diagram shows a long straight gray wire with an A C generator at its center. The time is marked t equals capital T divided by two. The bottom part of the antenna is negative and the upper end of the antenna is positive and a wave is shown to emerge from the A C source. The electric field E propagates away from the antenna at the speed of light, forming part of the electromagnetic wave from the A C source. A quarter portion of the wave is shown below the horizontal axis and a quarter portion of the wave is above the horizontal axis. Part d of the diagram shows a long straight gray wire with an AC generator at its center. The time is marked t equals capital T. The bottom part of the antenna is positive and the upper end of the antenna is negative. A wave is shown to emerge from the A C source. The electric field E propagates away from the antenna at the speed of light, forming part of the electromagnetic wave from the A C source. A quarter portion of the wave is shown above the horizontal axis followed by a half wave below the horizontal axis and then again a quarter of a wave above the horizontal axis.

Figure 1. This long straight gray wire with an AC generator at its center becomes a broadcast antenna for electromagnetic waves. Shown here are the charge distributions at four different times. The electric field (E) propagates

away from the antenna at the speed of light, forming part of an electromagnetic wave.

The *electric field* (**E**) shown surrounding the wire is produced by the charge distribution on the wire. Both the **E** and the charge distribution vary as the current changes. The changing field propagates outward at the speed of light.

There is an associated *magnetic field* (**B**) which propagates outward as well (see Figure 2). The electric and magnetic fields are closely related and propagate as an electromagnetic wave. This is what happens in broadcast antennae such as those in radio and TV stations.

Closer examination of the one complete cycle shown in Figure 1 reveals the periodic nature of the generator-driven charges oscillating up and down in the antenna and the electric field produced. At time t=0, there is the maximum separation of charge, with negative charges at the top and positive charges at the bottom, producing the maximum magnitude of the electric field (or E-field) in the upward direction. One-fourth of a cycle later, there is no charge separation and the field next to the antenna is zero, while the maximum E-field has moved away at speed c.

As the process continues, the charge separation reverses and the field reaches its maximum downward value, returns to zero, and rises to its maximum upward value at the end of one complete cycle. The outgoing wave has an *amplitude* proportional to the maximum separation of charge. Its *wavelength* (λ) is proportional to the period of the oscillation and, hence, is smaller for short periods or high frequencies. (As usual, wavelength and *frequency*(f) are inversely proportional.)

Electric and Magnetic Waves: Moving Together

Following Ampere's law, current in the antenna produces a magnetic field, as shown in Figure 2. The relationship between **E** and **B** is shown at one instant in Figure 2a. As the current varies, the magnetic field varies in magnitude and direction.

Part a of the diagram shows a long straight gray wire with an A C generator at its center, functioning as a broadcast antenna. The antenna has a current I flowing vertically upward. The bottom end of the antenna is negative and the upper end of the antenna is positive. An electric field is shown to act vertically downward. The magnetic field lines B produced in the antenna are circular in direction around the wire. Part b of the diagram shows a long straight gray wire with an A C generator at its center, functioning as a broadcast antenna. The electric field E and magnetic field B near the wire are shown perpendicular to each other. Part c of the diagram shows a long straight gray wire with an A C generator at its center, functioning as a broadcast antenna. The current is shown to flow in the antenna. The magnetic field varies with the current and propagates away from the antenna as a sine wave in the horizontal plane. The vibrations in the wave are marked as small arrows along the wave.

*Figure 2. (a) The current in the antenna produces the circular magnetic field lines. The current (I) produces the separation of charge along the wire, which in turn creates the electric field as shown. (b) The electric and magnetic fields (**E** and **B**) near the wire are perpendicular; they are shown here for one point in space. (c) The magnetic field varies with current and propagates away from the antenna at the speed of light.*

The magnetic field lines also propagate away from the antenna at the speed of light, forming the other part of the electromagnetic wave, as seen in Figure 2b. The magnetic part of the wave has the same period and wavelength as the electric part, since they are both produced by the same movement and separation of charges in the antenna.

The electric and magnetic waves are shown together at one instant in time in Figure 3. The electric and magnetic fields produced by a long straight wire antenna are exactly in phase. Note that they are perpendicular to one another and to the direction of propagation, making this a *transverse wave*

A part of the electromagnetic wave sent out from the antenna at one instant in time is shown. The wave is shown with the variation of two components, E and B, moving with velocity c. E is a sine wave in one plane with small arrows showing the vibrations of particles in the plane. B is a sine wave in a plane perpendicular to the E wave. The B wave has arrows to show the vibrations of particles in the plane. The waves are shown intersecting each other at the junction of the planes because E and B are perpendicular to each other. E and B are in phase, and they are perpendicular to one another and to the direction of propagation.

*Figure 3. A part of the electromagnetic wave sent out from the antenna at one instant in time. The electric and magnetic fields (**E** and **B**) are in phase, and they are perpendicular to one another and the direction of propagation. For clarity, the waves are shown only along one direction, but they propagate out in other directions too.*

Electromagnetic waves generally propagate out from a source in all directions, sometimes forming a complex radiation pattern. A linear antenna like this one will not radiate parallel to its length, for example. The wave is shown in one direction from the antenna in Figure 3 to illustrate its basic characteristics.

Instead of the AC generator, the antenna can also be driven by an AC circuit. In fact, charges radiate whenever they are accelerated. But while a current in a circuit needs a complete path, an antenna has a varying charge distribution forming a *standing wave*, driven by the AC. The dimensions of the antenna are critical for determining the frequency of the radiated electromagnetic waves. This is a *resonant* phenomenon and when we tune radios or TV, we vary electrical properties to achieve appropriate resonant conditions in the antenna.

Receiving Electromagnetic Waves

Electromagnetic waves carry energy away from their source, similar to a sound wave carrying energy away from a standing wave on a guitar string. An antenna for receiving EM signals works in reverse. And like antennas that produce EM waves, receiver antennas are specially designed to resonate at particular frequencies.

An incoming electromagnetic wave accelerates electrons in the antenna, setting up a standing wave. If the radio or TV is switched on, electrical components pick up and amplify the signal formed by the accelerating electrons. The signal is then converted to audio and/or video format. Sometimes big receiver dishes are used to focus the signal onto an antenna.

In fact, charges radiate whenever they are accelerated. When designing circuits, we often assume that energy does not quickly escape AC circuits, and mostly this is true. A broadcast antenna is specially designed to enhance the rate of electromagnetic radiation, and shielding is necessary to keep the radiation close to zero. Some familiar phenomena are based on the production of electromagnetic waves by varying currents. Your microwave oven, for example, sends electromagnetic waves, called microwaves, from a concealed antenna that has an oscillating current imposed on it.

Relating *E*-Field and *B*-Field Strengths

There is a relationship between the E− and B-field strengths in an electromagnetic wave. This can be understood by again considering the antenna just described. The stronger the E-field created by a separation of charge, the greater the current and, hence, the greater the B-field created.

Since current is directly proportional to voltage (Ohm's law) and voltage is directly proportional to E-field strength, the two should be directly proportional. It can be shown that the magnitudes of the fields do have a constant ratio, equal to the speed of light. That is, $\frac{E}{B} = c$ is the ratio of E-field strength to B-field strength in any electromagnetic wave. This is true at all times and at all locations in space. A simple and elegant result.

Example 1. Calculating B-Field Strength in an Electromagnetic Wave

What is the maximum strength of the B-field in an electromagnetic wave that has a maximum E-field strength of 1000 V/m?

Strategy

To find the B-field strength, we rearrange the above equation to solve for B, yielding

$$B = \frac{E}{c}.$$

Solution

We are given E, and c is the speed of light. Entering these into the expression for B yields

$$B = \frac{1000\ \text{V/m}}{3.00 \times 10^{8}\ \text{m/s}} = 3.33 \times 10^{-6}\ \text{T},$$

Where T stands for Tesla, a measure of magnetic field strength.

Discussion

The B-field strength is less than a tenth of the Earth's admittedly weak magnetic field. This means that a relatively strong electric field of 1000 V/m is accompanied by a relatively weak magnetic field. Note that as this wave spreads out, say with distance from an antenna, its field strengths become progressively weaker.

The result of this example is consistent with the statement made in the module Maxwell's Equations: Electromagnetic Waves Predicted and Observed that changing electric fields create relatively weak magnetic fields. They can be detected in electromagnetic waves, however, by taking advantage of the phenomenon of resonance, as Hertz did. A system with the same natural frequency as the electromagnetic wave can be made to oscillate. All radio and TV receivers use this principle to pick up and then amplify weak electromagnetic waves, while rejecting all others not at their resonant frequency.

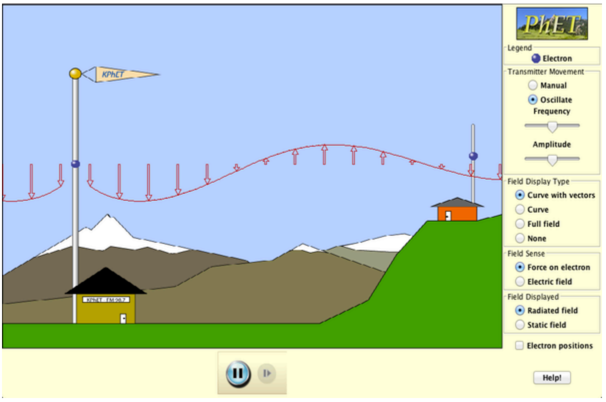
Take-Home Experiment: Antennas

For your TV or radio at home, identify the antenna, and sketch its shape. If you don't have cable, you might have an outdoor or indoor TV antenna. Estimate its size. If the TV signal is between 60 and 216 MHz for basic channels, then what is the wavelength of those EM waves?

Try tuning the radio and note the small range of frequencies at which a reasonable signal for that station is received. (This is easier with digital readout.) If you have a car with a radio and extendable antenna, note the quality of reception as the length of the antenna is changed.

PhET Explorations: Radio Waves and Electromagnetic Fields

Broadcast radio waves from KPhET. Wiggle the transmitter electron manually or have it oscillate automatically. Display the field as a curve or vectors. The strip chart shows the electron positions at the transmitter and at the receiver.



Click to download the simulation. Run using Java.

Section Summary

- Electromagnetic waves are created by oscillating charges (which radiate whenever accelerated) and have the same frequency as the oscillation.
- Since the electric and magnetic fields in most electromagnetic waves are perpendicular to the direction in which the wave moves, it is ordinarily a transverse wave.
- The strengths of the electric and magnetic parts of the wave are related by $\frac{E}{B} = \mathrm{c}$,
  which implies that the magnetic field B is very weak relative to the electric field E.

Conceptual Questions

1. The direction of the electric field shown in each part of Figure 1 is that produced by the charge distribution in the wire. Justify the direction shown in each part, using the Coulomb force law and the definition of $\mathbf{E} = \frac{\mathbf{F}}{q}$, where q is a positive test charge.
2. Is the direction of the magnetic field shown in Figure 2a consistent with the right-hand rule for current (RHR-2) in the direction shown in the figure?
3. Why is the direction of the current shown in each part of Figure 2 opposite to the electric field produced by the wire's charge separation?
4. In which situation shown in Figure 4 will the electromagnetic wave be more successful in inducing a current in the wire? Explain.

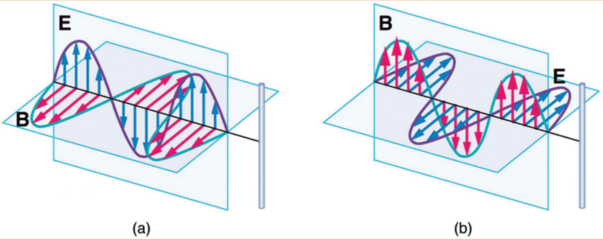


Figure 4. Electromagnetic waves approaching long straight wires.

5. In which situation shown in Figure 5 will the

electromagnetic wave be more successful in inducing a current in the loop? Explain.

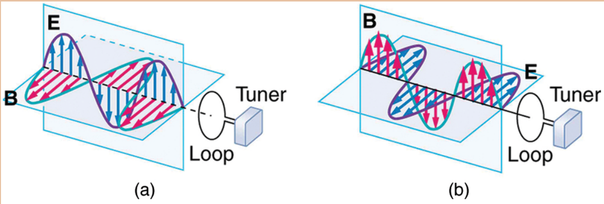


Figure 5. Electromagnetic waves approaching a wire loop.

6. Should the straight wire antenna of a radio be vertical or horizontal to best receive radio waves broadcast by a vertical transmitter antenna? How should a loop antenna be aligned to best receive the signals? (Note that the direction of the loop that produces the best reception can be used to determine the location of the source. It is used for that purpose in tracking tagged animals in nature studies, for example.)
7. Under what conditions might wires in a DC circuit emit electromagnetic waves?
8. Give an example of interference of electromagnetic waves.
9. Figure 6 shows the interference pattern of two radio antennas broadcasting the same signal. Explain how this is analogous to the interference pattern for sound produced by two speakers. Could this be used to make a directional antenna system that broadcasts preferentially in certain directions? Explain.

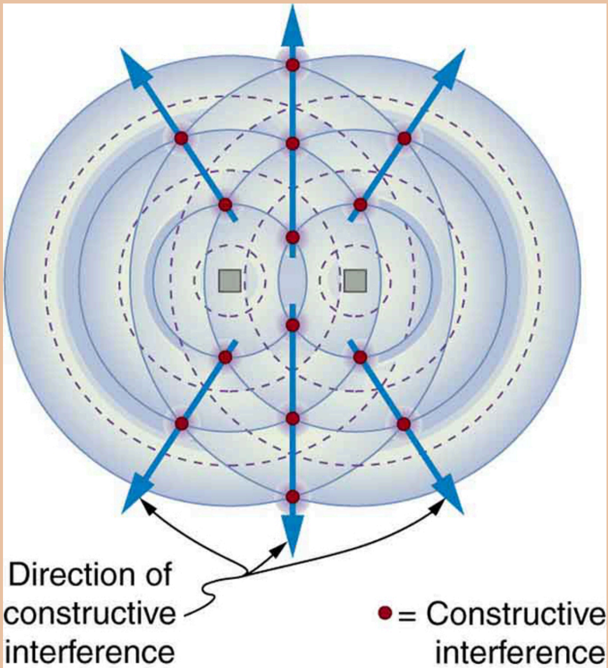


Figure 6. An overhead view of two radio broadcast antennas sending the same signal, and the interference pattern they produce.

10. Can an antenna be any length? Explain your answer.

Problems & Exercises

1. What is the maximum electric field strength in an electromagnetic wave that has a maximum magnetic field strength of 5.00×10^{-4} T (about 10 times the Earth's)?
2. The maximum magnetic field strength of an electromagnetic field is 5×10^{-6} T. Calculate the maximum electric field strength if the wave is traveling in a medium in which the speed of the wave is
0.75 *c*.
3. Verify the units obtained for magnetic field strength B in Example 1 (using the equation $B = \frac{E}{c}$) are in fact teslas (T).

Glossary

electric field: a vector quantity (**E**); the lines of electric force per unit charge, moving radially outward from a positive charge and in toward a negative charge

electric field strength: the magnitude of the electric field, denoted E-field

magnetic field: a vector quantity (**B**); can be used to determine the magnetic force on a moving charged particle

magnetic field strength: the magnitude of the magnetic field, denoted B-field

transverse wave: a wave, such as an electromagnetic wave, which oscillates perpendicular to the axis along the line of travel

standing wave: a wave that oscillates in place, with nodes where no motion happens

wavelength: the distance from one peak to the next in a wave

amplitude: the height, or magnitude, of an electromagnetic wave

frequency: the number of complete wave cycles (up-down-up) passing a given point within one second (cycles/second)

resonant: a system that displays enhanced oscillation when subjected to a periodic disturbance of the same frequency as its natural frequency

oscillate: to fluctuate back and forth in a steady beat

Selected Solutions to Problems & Exercises

1. 150 kV/m

67. The Electromagnetic Spectrum

Learning Objectives

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

- List three “rules of thumb” that apply to the different frequencies along the electromagnetic spectrum.
- Explain why the higher the frequency, the shorter the wavelength of an electromagnetic wave.
- Draw a simplified electromagnetic spectrum, indicating the relative positions, frequencies, and spacing of the different types of radiation bands.
- List and explain the different methods by which electromagnetic waves are produced across the spectrum.

In this module we examine how electromagnetic waves are classified into categories such as radio, infrared, ultraviolet, and so on, so that we can understand some of their similarities as well as some of their differences. We will also find that there are many connections with previously discussed topics, such as wavelength and resonance. A brief overview of the production and utilization of electromagnetic waves is found in Table 1.

Table 1. Electromagnetic Waves				
Type of EM wave	**Production**	**Applications**	**Life sciences aspect**	**Issues**
Radio & TV	Accelerating charges	Communications remote controls	MRI	Requires controls for band use
Microwaves	Accelerating charges & thermal agitation	Communications, ovens, radar	Deep heating	Cell phone use
Infrared	Thermal agitations & electronic transitions	Thermal imaging, heating	Absorbed by atmosphere	Greenhouse effect
Visible light	Thermal agitations & electronic transitions	All pervasive	Photosynthesis, Human vision	
Ultraviolet	Thermal agitations & electronic transitions	Sterilization, Cancer control	Vitamin D production	Ozone depletion, Cancer causing
X-rays	Inner electronic transitions and fast collisions	Medical Security	Medical diagnosis, Cancer therapy	Cancer causing
Gamma rays	Nuclear decay	Nuclear medicine, Security	Medical diagnosis, Cancer therapy	Cancer causing, Radiation damage

Connections: Waves

There are many types of waves, such as water waves and even earthquakes. Among the many shared attributes of waves are propagation speed, frequency, and wavelength. These are always related by the expression $v_W = f\lambda$. This module concentrates on EM waves, but other modules contain examples of all of these characteristics for sound waves and submicroscopic particles.

As noted before, an electromagnetic wave has a frequency and a wavelength associated with it and travels at the speed of light, or c. The relationship among these wave characteristics can be described by $v_W = f\lambda$, where v_W is the propagation speed of the wave, f is the frequency, and λ is the wavelength. Here $v_W = c$, so that for all electromagnetic waves, $c = f\lambda$.

Thus, for all electromagnetic waves, the greater the frequency, the smaller the wavelength.

Figure 1 shows how the various types of electromagnetic waves are categorized according to their wavelengths and frequencies—that is, it shows the electromagnetic spectrum. Many of the characteristics of the various types of electromagnetic waves are related to their frequencies and wavelengths, as we shall see.

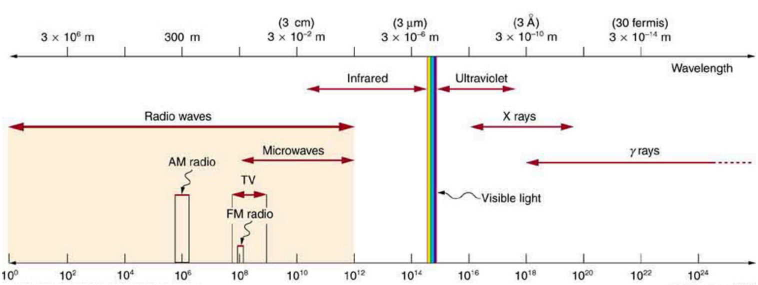


Figure 1. The electromagnetic spectrum, showing the major categories of electromagnetic waves. The range of frequencies and wavelengths is remarkable. The dividing line between some categories is distinct, whereas other categories overlap.

Electromagnetic Spectrum: Rules of Thumb

Three rules that apply to electromagnetic waves in general are as follows:

- High-frequency electromagnetic waves are more energetic and are more able to penetrate than low-frequency waves.
- High-frequency electromagnetic waves can carry more information per unit time than low-frequency waves.
- The shorter the wavelength of any electromagnetic wave probing a material, the smaller the detail it is possible to resolve.

Note that there are exceptions to these rules of thumb.

Transmission, Reflection, and Absorption

What happens when an electromagnetic wave impinges on a material? If the material is transparent to the particular frequency, then the wave can largely be transmitted. If the material is opaque to the frequency, then the wave can be totally reflected. The wave can also be absorbed by the material, indicating that there is some interaction between the wave and the material, such as the thermal agitation of molecules.

Of course it is possible to have partial transmission, reflection, and absorption. We normally associate these properties with visible light, but they do apply to all electromagnetic waves. What is not obvious is that something that is transparent to light may be opaque at other frequencies. For example, ordinary glass is transparent to visible light but largely opaque to ultraviolet radiation. Human skin is opaque to visible light—we cannot see through people—but transparent to X-rays.

Radio and TV Waves

The broad category of *radio waves* is defined to contain any electromagnetic wave produced by currents in wires and circuits. Its name derives from their most common use as a carrier of audio information (i.e., radio). The name is applied to electromagnetic waves of similar frequencies regardless of source. Radio waves from outer space, for example, do not come from alien radio stations. They are created by many astronomical phenomena, and their study has revealed much about nature on the largest scales.

There are many uses for radio waves, and so the category is divided into many subcategories, including microwaves and those electromagnetic waves used for AM and FM radio, cellular telephones, and TV.

The lowest commonly encountered radio frequencies are produced by high-voltage AC power transmission lines at frequencies of 50 or 60 Hz. (See Figure 2.) These extremely long wavelength electromagnetic waves (about 6000 km!) are one means of energy loss in long-distance power transmission.

A high-voltage traction power line is shown to the side of a roadway. The power line in the photo has two transmission poles supporting the cables.

Figure 2. This high-voltage traction power line running to Eutingen Railway Substation in Germany radiates electromagnetic waves with very long wavelengths. (credit: Zonk43, Wikimedia Commons)

There is an ongoing controversy regarding potential health hazards associated with exposure to these electromagnetic fields (E-fields). Some people suspect that living near such transmission lines may cause a variety of illnesses, including cancer. But demographic data are either inconclusive or simply do not support the hazard theory. Recent reports that have looked at many European and American epidemiological studies have found no increase in risk for cancer due to exposure to E-fields.

Extremely low frequency (*ELF*) radio waves of about 1 kHz are used to communicate with submerged submarines. The ability of radio waves to penetrate salt water is related to their wavelength (much like ultrasound penetrating tissue)—the longer the wavelength, the farther they penetrate. Since salt water is a good conductor, radio waves are strongly absorbed by it, and very long wavelengths are needed to reach a submarine under the surface. (See Figure 3.)

AM radio waves are used to carry commercial radio signals in the frequency range from 540 to 1600 kHz. The abbreviation AM stands for *amplitude modulation*, which is the method for placing information on these waves. (See Figure 4.) A *carrier wave* having the basic frequency of the radio station, say 1530 kHz, is varied or modulated in amplitude by an audio signal. The resulting wave has a constant frequency, but a varying amplitude.

The picture of a submarine under water is shown. The submarine is shown to receive extremely low frequency signals shown as a curvy line from the ocean surface to the submarine in the ocean depth.

Figure 3. Very long wavelength radio waves are needed to reach this submarine, requiring extremely low frequency signals (ELF). Shorter wavelengths do not penetrate to any significant depth.

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.

Part a of the diagram shows a carrier wave along the horizontal axis. The wave is shown to have a high frequency as the vibrations are closely spaced. The wave has constant amplitude represented by uniform height of crest and trough. Part b of the diagram shows an audio wave with a lower frequency. The wave is on the upper side of horizontal axis. The amplitude of the wave is not uniform. It has an initial small rise and fall followed by a steep rise and a gradual fall in the wave. Part c of the diagram shows the amplitude modulated wave. It is the resultant wave obtained by mixing of the waves in part a and part b. The amplitude of the resultant wave is non uniform, similar to the audio wave. The frequency of the amplitude modulated wave is equal to the frequency of the carrier wave. The wave spreads on both sides of the horizontal axis.

Figure 4. Amplitude modulation for AM radio. (a) A carrier wave at the station's basic frequency. (b) An audio signal at much lower audible frequencies. (c) The amplitude of the carrier is modulated by the audio signal without changing its basic frequency.

FM Radio Waves

FM radio waves are also used for commercial radio transmission, but in the frequency range of 88 to 108 MHz. FM stands for *frequency modulation*, another method of carrying information. (See Figure 5.) Here a carrier wave having the basic frequency of the radio station, perhaps 105.1 MHz, is modulated in frequency by the audio signal, producing a wave of constant amplitude but varying frequency.

Part a of the diagram shows a carrier wave along the horizontal axis. The wave is shown to have a high frequency as the vibrations are closely spaced. The wave has constant amplitude represented by uniform height of crest and trough. Part b of the diagram shows an audio wave with a lower frequency as shown by widely spaced vibrations. The wave has constant amplitude, represented by uniform length of crest and trough. Part c shows the frequency modulated wave obtained from waves in part a and part b. The amplitude of the resultant wave is similar to the source waves but the frequency varies. Frequency maxima are shown as closely spaced vibrations and frequency minima are shown as widely spaced vibrations. These maxima and minima are shown to alternate.

Figure 5. Frequency modulation for FM radio. (a) A carrier wave at the station's basic frequency. (b) An audio signal at much lower audible frequencies. (c) The frequency of the carrier is modulated by the audio signal without changing its amplitude.

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 1. Calculating Wavelengths of Radio Waves

Calculate the wavelengths of a 1530-kHz AM radio signal, a 105.1-MHz FM radio signal, and a 1.90-GHz cell phone signal.

Strategy

The relationship between wavelength and frequency is $c = f\lambda$, where $c = 3.00 \times 10^8$ m/s is the speed of light (the speed of light is only very slightly smaller in air than it is in a vacuum). We can rearrange this equation to find the wavelength for all three frequencies.

Solution

Rearranging gives $\lambda = \frac{c}{f}$.

For the f = 1530 kHz AM radio signal:

$$\begin{aligned}\lambda &= \frac{3.00 \times 10^8 \text{ m/s}}{1530 \times 10^3 \text{ cycles/s}} \\ &= 196 \text{ m}\end{aligned}$$

For the f = 105.1 MHz FM radio signal:

$$\begin{aligned}\lambda &= \frac{3.00 \times 10^8 \text{ m/s}}{105.1 \times 10^6 \text{ cycles/s}} \\ &= 2.85 \text{ m}\end{aligned}$$

And for the f = 1.90 GHz cell phone:

$$\begin{aligned}\lambda &= \frac{3.00 \times 10^8 \text{ m/s}}{1.90 \times 10^9 \text{ cycles/s}} \\ &= 0.158 \text{ m}\end{aligned}$$

Discussion

These wavelengths are consistent with the spectrum in Figure 1. The wavelengths are also related to other properties of these electromagnetic waves, as we shall see.

The wavelengths found in the preceding example are representative of AM, FM, and cell phones, and account for some of the differences in how they are broadcast and how well they travel. The most efficient length for a linear antenna, such as discussed in Production of Electromagnetic Waves, is $\frac{\lambda}{2}$, half the wavelength of the electromagnetic wave. Thus a very large antenna is needed to efficiently broadcast typical AM radio with its carrier wavelengths on the order of hundreds of meters.

One benefit to these long AM wavelengths is that they can go over and around rather large obstacles (like buildings and hills), just as ocean waves can go around large rocks. FM and TV are best received when there is a line of sight between the broadcast antenna and receiver, and they are often sent from very tall structures. FM, TV, and mobile phone antennas themselves are much smaller than those used for AM, but they are elevated to achieve an unobstructed line of sight. (See Figure 6.)

The first photograph shows a large tower used to broadcast TV signals. The tower is alternately painted red and white along the length. The antennas are shown as small structures on top of the tower. The second photograph shows a photo of a mobile phone tower. The tower has two ring shaped structures at its top most point.

Figure 6. (a) A large tower is used to broadcast TV signals. The actual antennas are small structures on top of the tower—they are placed at great heights to have a clear line of sight over a large broadcast area. (credit: Ozizo, Wikimedia Commons) (b) The NTT Dokomo mobile phone tower at Tokorozawa City, Japan. (credit: tokoroten, Wikimedia Commons)

Radio Wave Interference

Astronomers and astrophysicists collect signals from outer space using electromagnetic waves. A common problem for astrophysicists is the "pollution" from electromagnetic radiation pervading our surroundings from communication systems in general. Even everyday gadgets like our car keys having the facility to lock car doors remotely and being able to turn TVs on and off using remotes involve radio-wave frequencies. In order to prevent interference between all these electromagnetic signals, strict regulations are drawn up for different organizations to utilize different radio frequency bands.

One reason why we are sometimes asked to switch off our mobile phones (operating in the range of 1.9 GHz) on airplanes and in hospitals is that important communications or medical equipment often uses similar radio frequencies and their operation can be affected by frequencies used in the communication devices.

For example, radio waves used in magnetic resonance imaging (MRI) have frequencies on the order of 100 MHz, although this varies significantly depending on the strength of the magnetic field used and the nuclear type being scanned. MRI is an important medical imaging and research tool, producing highly detailed two- and three-dimensional images. Radio waves are broadcast, absorbed, and reemitted in a resonance process that is sensitive to the density of nuclei (usually protons or hydrogen nuclei).

The wavelength of 100-MHz radio waves is 3 m, yet using the sensitivity of the resonant frequency to the magnetic field strength, details smaller than a millimeter can be imaged. This is a good example of an exception to a rule of thumb (in this case, the rubric that details much smaller than the probe's wavelength cannot be detected). The intensity of the radio waves used in MRI presents little or no hazard to human health.

Microwaves

Microwaves are the highest-frequency electromagnetic waves that can be produced by currents in macroscopic circuits and devices. Microwave frequencies range from about 10^9 Hz to the highest practical *LC* resonance at nearly 10^{12} Hz. Since they have high frequencies, their wavelengths are short compared with those of other radio waves—hence the name "microwave."

Microwaves can also be produced by atoms and molecules. They are, for example, a component of electromagnetic radiation generated by *thermal agitation*. The thermal motion of atoms and molecules in any object at a temperature above absolute zero causes them to emit and absorb radiation.

Since it is possible to carry more information per unit time on high frequencies, microwaves are quite suitable for communications. Most satellite-transmitted information is carried on microwaves, as are land-based long-distance transmissions. A clear line of sight between transmitter and receiver is needed because of the short wavelengths involved.

A photograph of the surface of planet Venus is shown. The lava flows on Venus are shown as orange red color of the surface.

Figure 7. An image of Sif Mons with lava flows on Venus, based on Magellan synthetic aperture radar data combined with radar altimetry to produce a three-dimensional map of the surface. The Venusian atmosphere is opaque to visible light, but not to the microwaves that were used to create this image. (credit: NSSDC, NASA/JPL)

Radar is a common application of microwaves that was first developed in World War II. By detecting and timing microwave echoes, radar systems can determine the distance to objects as diverse as clouds and aircraft. A Doppler shift in the radar echo can be used to determine the speed of a car or the intensity of a rainstorm. Sophisticated radar systems are used to map the Earth and other planets, with a resolution limited by wavelength. (See Figure 7.) The shorter the wavelength of any probe, the smaller the detail it is possible to observe.

Heating with Microwaves

How does the ubiquitous microwave oven produce microwaves electronically, and why does food absorb them preferentially? Microwaves at a frequency of 2.45 GHz are produced by accelerating electrons. The microwaves are then used to induce an alternating electric field in the oven.

Water and some other constituents of food have a slightly negative charge at one end and a slightly positive charge at one end (called polar molecules). The range of microwave frequencies is specially selected so that the polar molecules, in trying to keep orienting themselves with the electric field, absorb these energies and increase their temperatures—called dielectric heating.

The energy thereby absorbed results in thermal agitation heating food and not the plate, which does not contain water. Hot spots in the food are related to constructive and destructive interference patterns. Rotating antennas and food turntables help spread out the hot spots.

Another use of microwaves for heating is within the human body. Microwaves will penetrate more than shorter wavelengths into tissue and so can accomplish “deep heating” (called microwave diathermy). This is used for treating muscular pains, spasms, tendonitis, and rheumatoid arthritis.

Take-Home Experiment—Microwave Ovens

1. Look at the door of a microwave oven. Describe the structure of the door. Why is there a metal grid on the door? How does the size of the holes in the grid compare with the wavelengths of microwaves used in microwave ovens? What is this wavelength?
2. Place a glass of water (about 250 ml) in the microwave and heat it for 30 seconds. Measure the temperature gain (the ΔT). Assuming that the power output of the oven is 1000 W, calculate the efficiency of the heat-transfer process.
3. Remove the rotating turntable or moving plate and place a cup of water in several places along a line parallel with the opening. Heat for 30 seconds and measure the ΔT for each position. Do you see cases of destructive interference?

Microwaves generated by atoms and molecules far away in time and space can be received and detected by electronic circuits. Deep space acts like a blackbody with a 2.7 K temperature, radiating most of its energy in the microwave frequency range. In 1964, Penzias and Wilson detected this radiation and eventually recognized that it was the radiation of the Big Bang's cooled remnants.

Infrared Radiation

The microwave and infrared regions of the electromagnetic spectrum overlap (see Figure 1). *Infrared radiation* is generally produced by thermal motion and the vibration and rotation of atoms and molecules. Electronic transitions in atoms and molecules can also produce infrared radiation.

The range of infrared frequencies extends up to the lower limit of visible light, just below red. In fact, infrared means "below red." Frequencies at its upper limit are too high to be produced by accelerating electrons in circuits, but small systems, such as atoms and molecules, can vibrate fast enough to produce these waves.

Water molecules rotate and vibrate particularly well at infrared frequencies, emitting and absorbing them so efficiently that the emissivity for skin is $e = 0.97$ in the infrared. Night-vision scopes can detect the infrared emitted by various warm objects, including humans, and convert it to visible light.

We can examine radiant heat transfer from a house by using a camera capable of detecting infrared radiation. Reconnaissance satellites can detect buildings, vehicles, and even individual humans by their infrared emissions, whose power radiation is proportional to the fourth power of the absolute temperature. More mundanely, we use infrared lamps, some of which are called quartz heaters, to preferentially warm us because we absorb infrared better than our surroundings.

The Sun radiates like a nearly perfect blackbody (that is, it has e =

1), with a 6000 K surface temperature. About half of the solar energy arriving at the Earth is in the infrared region, with most of the rest in the visible part of the spectrum, and a relatively small amount in the ultraviolet. On average, 50 percent of the incident solar energy is absorbed by the Earth.

The relatively constant temperature of the Earth is a result of the energy balance between the incoming solar radiation and the energy radiated from the Earth. Most of the infrared radiation emitted from the Earth is absorbed by CO_2 and H_2O in the atmosphere and then radiated back to Earth or into outer space. This radiation back to Earth is known as the greenhouse effect, and it maintains the surface temperature of the Earth about 40°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

V*isible light* is the narrow segment of the electromagnetic spectrum to which the normal human eye responds. Visible light is produced by vibrations and rotations of atoms and molecules, as well as by electronic transitions within atoms and molecules. The receivers or detectors of light largely utilize electronic transitions. We say the atoms and molecules are excited when they absorb and relax when they emit through electronic transitions.

Figure 8 shows this part of the spectrum, together with the colors associated with particular pure wavelengths. We usually refer to visible light as having wavelengths of between 400 nm and 750 nm. (The retina of the eye actually responds to the lowest ultraviolet frequencies, but these do not normally reach the retina because they are absorbed by the cornea and lens of the eye.)

Red light has the lowest frequencies and longest wavelengths, while violet has the highest frequencies and shortest wavelengths. Blackbody radiation from the Sun peaks in the visible part of the spectrum but is more intense in the red than in the violet, making the Sun yellowish in appearance.

The visible strip of the electromagnetic spectrum is highlighted and shown in the picture. The wave length range is from eight hundred nanometers on the left to three hundred nanometers on the right. The divisions between infrared, visible, and ultraviolet are not perfectly distinct. The colors in the visible strip are also not perfectly distinct; they are marked as bands labeled from red on the left to violet on the right.

Figure 8. A small part of the electromagnetic spectrum that includes its visible components. The divisions between infrared, visible, and ultraviolet are not perfectly distinct, nor are those between the seven rainbow colors.

Living things—plants and animals—have evolved to utilize and respond to parts of the electromagnetic spectrum they are embedded in. Visible light is the most predominant and we enjoy the beauty of nature through visible light. Plants are more selective. Photosynthesis makes use of parts of the visible spectrum to make sugars.

Example 2. Integrated Concept Problem: Correcting Vision with Lasers

During laser vision correction, a brief burst of 193-nm ultraviolet light is projected onto the cornea of a patient. It makes a spot 0.80 mm in diameter and evaporates a layer of cornea 0.30 μm thick. Calculate the energy absorbed, assuming the corneal tissue has the same properties as water; it is initially at 34°C. Assume the evaporated tissue leaves at a temperature of 100°C.

Strategy

The energy from the laser light goes toward raising the temperature of the tissue and also toward evaporating it. Thus we have two amounts of heat to add together. Also, we need to find the mass of corneal tissue involved.

Solution

To Figure out the heat required to raise the temperature of the tissue to 100°C, we can apply concepts of thermal energy. We know that

$$Q = mc\Delta T,$$

where Q is the heat required to raise the temperature, ΔT is the desired change in temperature, *m* is the mass of tissue to be heated, and *c* is the specific heat of water equal to 4186 J/kg/K.

Without knowing the mass *m* at this point, we have

$$Q = m(4186\ \text{J/kg/K})(100°\text{C}-34°\text{C}) = m(276{,}276\ \text{J/kg}) = m(276\ \text{kJ/kg}).$$

The latent heat of vaporization of water is 2256 kJ/kg, so that the energy needed to evaporate mass *m* is

$$Q_v = mL_v = m(2256\ \text{kJ/kg}).$$

To find the mass *m*, we use the equation $\rho = \dfrac{m}{V}$, where ρ is the density of the tissue and V is its volume. For this case,

$$\begin{aligned} m &= \rho V \\ &= \left(1000\ \mathrm{kg/m^3}\right)\left(\mathrm{area} \times \mathrm{thickness}\left(m^3\right)\right) \\ &= \left(1000\ \mathrm{kg/m^3}\right)\left(\pi\left(0.80 \times 10^{-3}\ \mathrm{m}\right)^2/4\right)\left(0.30 \times 10^{-6}\ \mathrm{m}\right) \\ &= 0.151 \times 10^{-9}\ \mathrm{kg} \end{aligned}$$

Therefore, the total energy absorbed by the tissue in the eye is the sum of Q and Q_V:

$$Q_{tot} = m(c\Delta T + L_V) = (0.151 \times 10^{-9}\ \mathrm{kg})(276 \cdot \mathrm{kJ/kg} + 2256\ \mathrm{kJ/kg}) = 382 \times 10^{-9}\ \mathrm{kJ}.$$

Discussion

The lasers used for this eye surgery are excimer lasers, whose light is well absorbed by biological tissue. They evaporate rather than burn the tissue, and can be used for precision work. Most lasers used for this type of eye surgery have an average power rating of about one watt. For our example, if we assume that each laser burst from this pulsed laser lasts for 10 ns, and there are 400 bursts per second, then the average power is $Q_{tot} \times 400 = 150$ mW.

Optics is the study of the behavior of visible light and other forms of electromagnetic waves. Optics falls into two distinct categories. When electromagnetic radiation, such as visible light, interacts with objects that are large compared with its wavelength, its motion can be represented by straight lines like rays. Ray optics is the study of such situations and includes lenses and mirrors.

When electromagnetic radiation interacts with objects about the same size as the wavelength or smaller, its wave nature becomes apparent. For example, observable detail is limited by the wavelength, and so visible light can never detect individual atoms, because they are so much smaller than its wavelength. Physical or wave optics is the study of such situations and includes all wave characteristics.

Take-Home Experiment: Colors That Match

When you light a match you see largely orange light; when you light a gas stove you see blue light. Why are the colors different? What other colors are present in these?

Ultraviolet Radiation

Ultraviolet means "above violet." The electromagnetic frequencies of *ultraviolet radiation* (UV) extend upward from violet, the highest-frequency visible light. Ultraviolet is also produced by atomic and molecular motions and electronic transitions. The wavelengths of ultraviolet extend from 400 nm down to about 10 nm at its highest frequencies, which overlap with the lowest X-ray frequencies. It was recognized as early as 1801 by Johann Ritter that the solar spectrum had an invisible component beyond the violet range.

Solar UV radiation is broadly subdivided into three regions: UV-A (320–400 nm), UV-B (290–320 nm), and UV-C (220–290 nm), ranked from long to shorter wavelengths (from smaller to larger energies). Most UV-B and all UV-C is absorbed by ozone (O_3) molecules in the upper atmosphere. Consequently, 99% of the solar UV radiation reaching the Earth's surface is UV-A.

Human Exposure to UV Radiation

It is largely exposure to UV-B that causes skin cancer. It is estimated that as many as 20% of adults will develop skin cancer over the course of their lifetime. Again, treatment is often successful if caught early. Despite very little UV-B reaching the Earth's surface, there are substantial increases in skin-cancer rates in countries such as Australia, indicating how important it is that UV-B and UV-C continue to be absorbed by the upper atmosphere.

All UV radiation can damage collagen fibers, resulting in an acceleration of the aging process of skin and the formation of wrinkles. Because there is so little UV-B and UV-C reaching the Earth's surface, sunburn is caused by large exposures, and skin cancer from repeated exposure. Some studies indicate a link between overexposure to the Sun when young and melanoma later in life.

The tanning response is a defense mechanism in which the body produces pigments to absorb future exposures in inert skin layers above living cells. Basically UV-B radiation excites DNA molecules, distorting the DNA helix, leading to mutations and the possible formation of cancerous cells.

Repeated exposure to UV-B may also lead to the formation of cataracts in the eyes—a cause of blindness among people living in the equatorial belt where medical treatment is limited. Cataracts, clouding in the eye's lens and a loss of vision, are age related; 60% of those between the ages of 65 and 74 will develop cataracts. However, treatment is easy and successful, as one replaces the lens of the eye with a plastic lens. Prevention is important. Eye protection from UV is more effective with plastic sunglasses than those made of glass.

A major acute effect of extreme UV exposure is the suppression of the immune system, both locally and throughout the body.

Low-intensity ultraviolet is used to sterilize haircutting implements, implying that the energy associated with ultraviolet is deposited in a manner different from lower-frequency electromagnetic waves. (Actually this is true for all electromagnetic waves with frequencies greater than visible light.)

Flash photography is generally not allowed of precious artworks and colored prints because the UV radiation from the flash can cause photo-degradation in the artworks. Often artworks will have an extra-thick layer of glass in front of them, which is especially designed to absorb UV radiation.

UV Light and the Ozone Layer

If all of the Sun's ultraviolet radiation reached the Earth's surface, there would be extremely grave effects on the biosphere from the severe cell damage it causes. However, the layer of ozone (O_3) in our upper atmosphere (10 to 50 km above the Earth) protects life by absorbing most of the dangerous UV radiation.

Unfortunately, today we are observing a depletion in ozone concentrations in the upper atmosphere. This depletion has led to the formation of an "ozone hole" in the upper atmosphere. The hole is more centered over the southern hemisphere, and changes with the seasons, being largest in the spring. This depletion is attributed to the breakdown of ozone molecules by refrigerant gases called chlorofluorocarbons (CFCs).

The UV radiation helps dissociate the CFC's, releasing highly reactive chlorine (Cl) atoms, which catalyze the destruction of the ozone layer. For example, the reaction of $CFCl_3$ with a photon of light (*hv*) can be written as

$$CFCl_3 + hv \rightarrow CFCl_2 + Cl.$$

The Cl atom then catalyzes the breakdown of ozone as follows:

$Cl + O_3 \rightarrow ClO + O_2$ and $ClO + O_3 \rightarrow Cl + 2O_2$.

A single chlorine atom could destroy ozone molecules for up to two years before being transported down to the surface. The CFCs are relatively stable and will contribute to ozone depletion for years to come. CFCs are found in refrigerants, air conditioning systems, foams, and aerosols.

International concern over this problem led to the establishment of the "Montreal Protocol" agreement (1987) to phase out CFC production in most countries. However, developing-country participation is needed if worldwide production and elimination of CFCs is to be achieved. Probably the largest contributor to CFC emissions today is India. But the protocol seems to be working, as there are signs of an ozone recovery. (See Figure 9.)

The map shows the variation in concentration of ozone over Antarctica. The scale for the total ozone level is depicted below the graph in Dobson units. The values are marked in colors of spectrum with the lowest value is marked in violet and the maximum value in red. The Antarctica region is marked in violet showing lesser ozone concentration and more ultraviolet rays. The region around Antarctica is in green, showing slightly higher concentration of ozone.

Figure 9. This map of ozone concentration over Antarctica in October 2011 shows severe depletion suspected to be caused by CFCs. Less dramatic but more general depletion has been observed over northern latitudes, suggesting the effect is global. With less ozone, more ultraviolet radiation from the Sun reaches the surface, causing more damage. (credit: NASA Ozone Watch)

Benefits of UV Light

Besides the adverse effects of ultraviolet radiation, there are also benefits of exposure in nature and uses in technology. Vitamin D production in the skin (epidermis) results from exposure to UVB radiation, generally from sunlight. A number of studies indicate lack of vitamin D can result in the development of a range of cancers (prostate, breast, colon), so a certain amount of UV exposure is helpful. Lack of vitamin D is also linked to osteoporosis. Exposures (with no sunscreen) of 10 minutes a day to arms, face, and legs might be sufficient to provide the accepted dietary level. However, in the winter time north of about 37º latitude, most UVB gets blocked by the atmosphere.

UV radiation is used in the treatment of infantile jaundice and in some skin conditions. It is also used in sterilizing workspaces and tools, and killing germs in a wide range of applications. It is also used as an analytical tool to identify substances.

When exposed to ultraviolet, some substances, such as minerals, glow in characteristic visible wavelengths, a process called fluorescence. So-called black lights emit ultraviolet to cause posters and clothing to fluoresce in the visible. Ultraviolet is also used in special microscopes to detect details smaller than those observable with longer-wavelength visible-light microscopes.

Things Great and Small: A Submicroscopic View of X-Ray Production

An atom is shown. The nucleus is in the center as a cluster of small spheres packed together. Four electron orbits are shown around the nucleus. The one close to the nucleus is circular. All the other orbits are elliptical in nature and inclined at various angles. An electron, represented as a tiny sphere, is shown to strike the atom. An electron is shown knocked out from the closest orbit. A second image of the same atom illustrates another electron striking innermost orbit; a wavy red arrow representing an x ray is shooting away from the innermost orbit.

Figure 10. Artist's conception of an electron ionizing an atom followed by the recapture of an electron and emission of an X-ray. An energetic electron strikes an atom and knocks an electron out of one of the orbits closest to the nucleus. Later, the atom captures another electron, and the energy released by its fall into a low orbit generates a high-energy EM wave called an X-ray.

X-rays can be created in a high-voltage discharge. They are emitted in the material struck by electrons in the discharge current. There are two mechanisms by which the electrons create X-rays.

The first method is illustrated in Figure 10. An electron is accelerated in an evacuated tube by a high positive voltage. The electron strikes a metal plate (e.g., copper) and produces X-rays. Since this is a high-voltage discharge, the electron gains sufficient energy to ionize the atom.

In the case shown, an inner-shell electron (one in an orbit relatively close to and tightly bound to the nucleus) is ejected. A short time later, another electron is captured and falls into the orbit in a single great plunge. The energy released by this fall is given to an EM wave known as an X-ray. Since the orbits of the atom are unique to the type of atom, the energy of the X-ray is characteristic of the atom, hence the name characteristic X-ray.

A picture showing an electron represented as a tiny sphere shown to strike the atoms in the material represented as spheres slightly larger in size than the electron. A ray of X ray is shown to come out from the material shown by a wavy arrow.

Figure 11. Artist's conception of an electron being slowed by collisions in a material and emitting X-ray radiation. This energetic electron makes numerous collisions with electrons and atoms in a material it penetrates. An accelerated charge radiates EM waves, a second method by which X-rays are created.

The second method by which an energetic electron creates an X-ray when it strikes a material is illustrated in Figure 11. The electron interacts with charges in the material as it penetrates. These collisions transfer kinetic energy from the electron to the electrons and atoms in the material.

A loss of kinetic energy implies an acceleration, in this case decreasing the electron's velocity. Whenever a charge is accelerated, it radiates EM waves. Given the high energy of the electron, these EM waves can have high energy. We call them X-rays. Since the process is random, a broad spectrum of X-ray energy is emitted that is more characteristic of the electron energy than the type of material the electron encounters. Such EM radiation is called “bremsstrahlung” (German for “braking radiation”).

X-Rays

In the 1850s, scientists (such as Faraday) began experimenting with high-voltage electrical discharges in tubes filled with rarefied gases. It was later found that these discharges created an invisible, penetrating form of very high frequency electromagnetic radiation. This radiation was called an X-*ray*, because its identity and nature were unknown.

As described in “Things Great and Small” feature, there are two methods by which X-rays are created—both are submicroscopic processes and can be caused by high-voltage discharges. While the low-frequency end of the X-ray range overlaps with the ultraviolet, X-rays extend to much higher frequencies (and energies).

X-rays have adverse effects on living cells similar to those of ultraviolet radiation, and they have the additional liability of being more penetrating, affecting more than the surface layers of cells. Cancer and genetic defects can be induced by exposure to X-rays. Because of their effect on rapidly dividing cells, X-rays can also be used to treat and even cure cancer.

The widest use of X-rays is for imaging objects that are opaque to visible light, such as the human body or aircraft parts. In humans, the risk of cell damage is weighed carefully against the benefit of the diagnostic information obtained. However, questions have risen in recent years as to accidental overexposure of some people during CT scans—a mistake at least in part due to poor monitoring of radiation dose.

The ability of X-rays to penetrate matter depends on density, and so an X-ray image can reveal very detailed density information. Figure 12 shows an example of the simplest type of X-ray image, an X-ray shadow on film. The amount of information in a simple X-ray image is impressive, but more sophisticated techniques, such as CT scans, can reveal three-dimensional information with details smaller than a millimeter.

The use of X-ray technology in medicine is called radiology—an established and relatively cheap tool in comparison to more sophisticated technologies. Consequently, X-rays are widely available and used extensively in medical diagnostics. During World War I, mobile X-ray units, advocated by Madame Marie Curie, were used to diagnose soldiers.

An X ray image of the chest is shown. It shows the section of the heart with artificial heart valves, a pacemaker, and the wires used to close the sternum.

Figure 12. This shadow X-ray image shows many interesting features, such as artificial heart valves, a pacemaker, and the wires used to close the sternum. (credit: P. P. Urone)

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.

A skeletal image of a human body is shown. The image represents gamma rays emitted by nuclei in a compound that is concentrated in the bones and eliminated through the kidneys. Some parts of the image are darker than the others. The ribs are shown darker than the leg and hand bones.

Figure 13. This is an image of the γ-rays emitted by nuclei in a compound that is concentrated in the bones and eliminated through the kidneys. Bone cancer is evidenced by nonuniform concentration in similar structures. For example, some ribs are darker than others. (credit: P. P. Urone)

In fact, γ rays are any electromagnetic radiation emitted by a nucleus. This can be from natural nuclear decay or induced nuclear processes in nuclear reactors and weapons. The lower end of the γ-ray frequency range overlaps the upper end of the X-ray range, but γ rays can have the highest frequency of any electromagnetic radiation.

Gamma rays have characteristics identical to X-rays of the same frequency—they differ only in source. At higher frequencies, γ rays are more penetrating and more damaging to living tissue. They have many of the same uses as X-rays, including cancer therapy. Gamma radiation from radioactive materials is used in nuclear medicine.

Figure 13 shows a medical image based on γ rays. Food spoilage can be greatly inhibited by exposing it to large doses of γ radiation, thereby obliterating responsible microorganisms. Damage to food cells through irradiation occurs as well, and the long-term hazards of consuming radiation-preserved food are unknown and controversial for some groups. Both X-ray and γ-ray technologies are also used in scanning luggage at airports.

Detecting Electromagnetic Waves from Space

A final note on star gazing. The entire electromagnetic spectrum is used by researchers for investigating stars, space, and time. As noted earlier, Penzias and Wilson detected microwaves to identify the background radiation originating from the Big Bang. Radio telescopes such as the Arecibo Radio Telescope in Puerto Rico and Parkes Observatory in Australia were designed to detect radio waves.

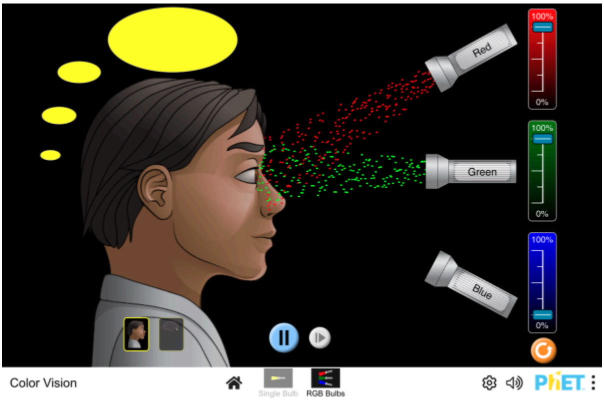
Infrared telescopes need to have their detectors cooled by liquid nitrogen to be able to gather useful signals. Since infrared radiation is predominantly from thermal agitation, if the detectors were not cooled, the vibrations of the molecules in the antenna would be stronger than the signal being collected.

The most famous of these infrared sensitive telescopes is the James Clerk Maxwell Telescope in Hawaii. The earliest telescopes, developed in the seventeenth century, were optical telescopes, collecting visible light. Telescopes in the ultraviolet, X-ray, and γ-ray regions are placed outside the atmosphere on satellites orbiting the Earth.

The Hubble Space Telescope (launched in 1990) gathers ultraviolet radiation as well as visible light. In the X-ray region, there is the Chandra X-ray Observatory (launched in 1999), and in the γ-ray region, there is the new Fermi Gamma-ray Space Telescope (launched in 2008—taking the place of the Compton Gamma Ray Observatory, 1991–2000.).

PhET Explorations: Color Vision

Make a whole rainbow by mixing red, green, and blue light. Change the wavelength of a monochromatic beam or filter white light. View the light as a solid beam, or see the individual photons.



Click to run the simulation.

Section Summary

- The relationship among the speed of propagation, wavelength, and frequency for any wave is given by $v_W = f\lambda$, so that for electromagnetic waves, $c = f\lambda$, where f is the frequency, λ is the wavelength, and *c* is the speed of light.

- The electromagnetic spectrum is separated into many categories and subcategories, based on the frequency and wavelength, source, and uses of the electromagnetic waves.
- Any electromagnetic wave produced by currents in wires is classified as a radio wave, the lowest frequency electromagnetic waves. Radio waves are divided into many types, depending on their applications, ranging up to microwaves at their highest frequencies.
- Infrared radiation lies below visible light in frequency and is produced by thermal motion and the vibration and rotation of atoms and molecules. Infrared's lower frequencies overlap with the highest-frequency microwaves.
- Visible light is largely produced by electronic transitions in atoms and molecules, and is defined as being detectable by the human eye. Its colors vary with frequency, from red at the lowest to violet at the highest.
- Ultraviolet radiation starts with frequencies just above violet in the visible range and is produced primarily by electronic transitions in atoms and molecules.
- X-rays are created in high-voltage discharges and by electron bombardment of metal targets. Their lowest frequencies overlap the ultraviolet range but extend to much higher values, overlapping at the high end with gamma rays.
- Gamma rays are nuclear in origin and are defined to include the highest-frequency electromagnetic radiation of any type.

Conceptual Questions

1. If you live in a region that has a particular TV station, you can sometimes pick up some of its audio

portion on your FM radio receiver. Explain how this is possible. Does it imply that TV audio is broadcast as FM?

2. Explain why people who have the lens of their eye removed because of cataracts are able to see low-frequency ultraviolet.
3. How do fluorescent soap residues make clothing look "brighter and whiter" in outdoor light? Would this be effective in candlelight?
4. Give an example of resonance in the reception of electromagnetic waves.
5. Illustrate that the size of details of an object that can be detected with electromagnetic waves is related to their wavelength, by comparing details observable with two different types (for example, radar and visible light or infrared and X-rays).
6. Why don't buildings block radio waves as completely as they do visible light?
7. Make a list of some everyday objects and decide whether they are transparent or opaque to each of the types of electromagnetic waves.
8. Your friend says that more patterns and colors can be seen on the wings of birds if viewed in ultraviolet light. Would you agree with your friend? Explain your answer.
9. The rate at which information can be transmitted on an electromagnetic wave is proportional to the frequency of the wave. Is this consistent with the fact that laser telephone transmission at visible frequencies carries far more conversations per optical fiber than conventional electronic transmission in a wire? What is the implication for

ELF radio communication with submarines?

10. Give an example of energy carried by an electromagnetic wave.
11. In an MRI scan, a higher magnetic field requires higher frequency radio waves to resonate with the nuclear type whose density and location is being imaged. What effect does going to a larger magnetic field have on the most efficient antenna to broadcast those radio waves? Does it favor a smaller or larger antenna?
12. Laser vision correction often uses an excimer laser that produces 193-nm electromagnetic radiation. This wavelength is extremely strongly absorbed by the cornea and ablates it in a manner that reshapes the cornea to correct vision defects. Explain how the strong absorption helps concentrate the energy in a thin layer and thus give greater accuracy in shaping the cornea. Also explain how this strong absorption limits damage to the lens and retina of the eye.

Problems & Exercises

1. (a) Two microwave frequencies are authorized for use in microwave ovens: 900 and 2560 MHz. Calculate the wavelength of each. (b) Which frequency would produce smaller hot spots in foods due to interference effects?
2. (a) Calculate the range of wavelengths for AM radio

given its frequency range is 540 to 1600 kHz. (b) Do the same for the FM frequency range of 88.0 to 108 MHz.

3. A radio station utilizes frequencies between commercial AM and FM. What is the frequency of a 11.12-m-wavelength channel?
4. Find the frequency range of visible light, given that it encompasses wavelengths from 380 to 760 nm.
5. Combing your hair leads to excess electrons on the comb. How fast would you have to move the comb up and down to produce red light?
6. Electromagnetic radiation having a 15.0−μm wavelength is classified as infrared radiation. What is its frequency?
7. Approximately what is the smallest detail observable with a microscope that uses ultraviolet light of frequency 1.20×10^{15} Hz?
8. A radar used to detect the presence of aircraft receives a pulse that has reflected off an object 6×10^{-5} s after it was transmitted. What is the distance from the radar station to the reflecting object?
9. Some radar systems detect the size and shape of objects such as aircraft and geological terrain. Approximately what is the smallest observable detail utilizing 500-MHz radar?
10. Determine the amount of time it takes for X-rays of frequency 3×10^{18} Hz to travel (a) 1 mm and (b) 1 cm.
11. If you wish to detect details of the size of atoms (about 1×10^{-10} m) with electromagnetic radiation, it must have a wavelength of about this size. (a) What is its frequency? (b) What type of electromagnetic radiation might this be?

12. If the Sun suddenly turned off, we would not know it until its light stopped coming. How long would that be, given that the Sun is 1.50×10^{11} m away?
13. Distances in space are often quoted in units of light years, the distance light travels in one year. (a) How many meters is a light year? (b) How many meters is it to Andromeda, the nearest large galaxy, given that it is 2.00×10^{6} light years away? (c) The most distant galaxy yet discovered is 12.0×10^{9} light years away. How far is this in meters?
14. A certain 50.0-Hz AC power line radiates an electromagnetic wave having a maximum electric field strength of 13.0 kV/m. (a) What is the wavelength of this very low frequency electromagnetic wave? (b) What is its maximum magnetic field strength?
15. During normal beating, the heart creates a maximum 4.00-mV potential across 0.300 m of a person's chest, creating a 1.00-Hz electromagnetic wave. (a) What is the maximum electric field strength created? (b) What is the corresponding maximum magnetic field strength in the electromagnetic wave? (c) What is the wavelength of the electromagnetic wave?
16. (a) The ideal size (most efficient) for a broadcast antenna with one end on the ground is one-fourth the wavelength $\left(\frac{\lambda}{4}\right)$ of the electromagnetic radiation being sent out. If a new radio station has such an antenna that is 50.0 m high, what frequency does it broadcast most efficiently? Is this in the AM or

FM band? (b) Discuss the analogy of the fundamental resonant mode of an air column closed at one end to the resonance of currents on an antenna that is one-fourth their wavelength.

17. (a) What is the wavelength of 100-MHz radio waves used in an MRI unit? (b) If the frequencies are swept over a$\pm$1.00 range centered on 100 MHz, what is the range of wavelengths broadcast?
18. (a) What is the frequency of the 193-nm ultraviolet radiation used in laser eye surgery? (b) Assuming the accuracy with which this EM radiation can ablate the cornea is directly proportional to wavelength, how much more accurate can this UV be than the shortest visible wavelength of light?
19. TV-reception antennas for VHF are constructed with cross wires supported at their centers, as shown in [link]. The ideal length for the cross wires is one-half the wavelength to be received, with the more expensive antennas having one for each channel. Suppose you measure the lengths of the wires for particular channels and find them to be 1.94 and 0.753 m long, respectively. What are the frequencies for these channels?

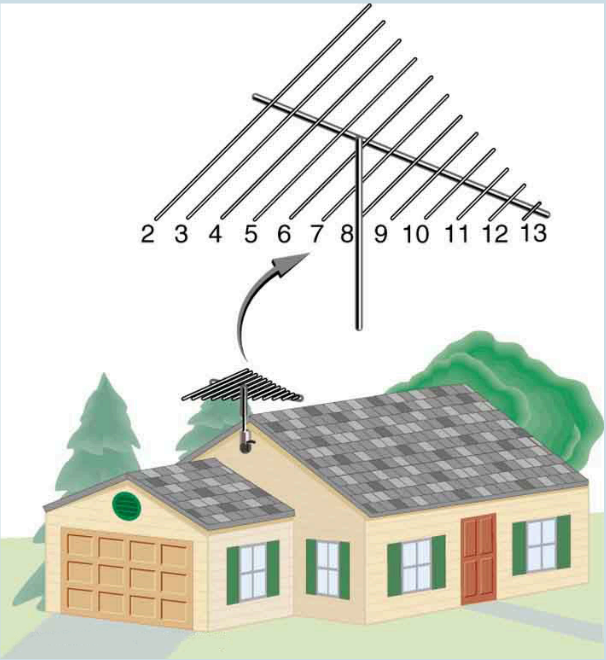


Figure 14. A television reception antenna has cross wires of various lengths to most efficiently receive different wavelengths.

20. Conversations with astronauts on lunar walks had an echo that was used to estimate the distance to the Moon. The sound spoken by the person on Earth was transformed into a radio signal sent to the Moon, and transformed back into sound on a speaker inside the astronaut's space suit. This sound was picked up by the microphone in the space suit (intended for the astronaut's voice) and sent back to Earth as a radio echo of sorts. If the round-trip time was 2.60 s, what

was the approximate distance to the Moon, neglecting any delays in the electronics?

21. Lunar astronauts placed a reflector on the Moon's surface, off which a laser beam is periodically reflected. The distance to the Moon is calculated from the round-trip time. (a) To what accuracy in meters can the distance to the Moon be determined, if this time can be measured to 0.100 ns? (b) What percent accuracy is this, given the average distance to the Moon is 3.84×10^8 m?
22. Radar is used to determine distances to various objects by measuring the round-trip time for an echo from the object. (a) How far away is the planet Venus if the echo time is 1000 s? (b) What is the echo time for a car 75.0 m from a Highway Police radar unit? (c) How accurately (in nanoseconds) must you be able to measure the echo time to an airplane 12.0 km away to determine its distance within 10.0 m?
23. **Integrated Concepts.** (a) Calculate the ratio of the highest to lowest frequencies of electromagnetic waves the eye can see, given the wavelength range of visible light is from 380 to 760 nm. (b) Compare this with the ratio of highest to lowest frequencies the ear can hear.
24. **Integrated Concepts.** (a) Calculate the rate in watts at which heat transfer through radiation occurs (almost entirely in the infrared) from 1.0 m^2 of the Earth's surface at night. Assume the emissivity is 0.90, the temperature of the Earth is 15ºC, and that of outer space is 2.7 K. (b) Compare the intensity of this radiation with that coming to the Earth from the Sun during the day, which averages about 800 W/m^2, only

half of which is absorbed. (c) What is the maximum magnetic field strength in the outgoing radiation, assuming it is a continuous wave?

Glossary

electromagnetic spectrum: the full range of wavelengths or frequencies of electromagnetic radiation

radio waves: electromagnetic waves with wavelengths in the range from 1 mm to 100 km; they are produced by currents in wires and circuits and by astronomical phenomena

microwaves: electromagnetic waves with wavelengths in the range from 1 mm to 1 m; they can be produced by currents in macroscopic circuits and devices

thermal agitation: the thermal motion of atoms and molecules in any object at a temperature above absolute zero, which causes them to emit and absorb radiation

radar: a common application of microwaves. Radar can determine the distance to objects as diverse as clouds and aircraft, as well as determine the speed of a car or the intensity of a rainstorm

infrared radiation (IR): a region of the electromagnetic spectrum with a frequency range that extends from just below the red region of the visible light spectrum up to the microwave region, or from 0.74μm to 300μm

ultraviolet radiation (UV): electromagnetic radiation in the range extending upward in frequency from violet light and overlapping with the lowest X-ray frequencies, with wavelengths from 400 nm down to about 10 nm

visible light: the narrow segment of the electromagnetic spectrum to which the normal human eye responds

amplitude modulation (AM): a method for placing information on electromagnetic waves by modulating the amplitude of a carrier wave with an audio signal, resulting in a wave with constant frequency but varying amplitude

extremely low frequency (ELF): electromagnetic radiation with wavelengths usually in the range of 0 to 300 Hz, but also about 1kHz

carrier wave: an electromagnetic wave that carries a signal by modulation of its amplitude or frequency

frequency modulation (FM): a method of placing information on electromagnetic waves by modulating the frequency of a carrier wave with an audio signal, producing a wave of constant amplitude but varying frequency

TV: video and audio signals broadcast on electromagnetic waves

very high frequency (VHF): TV channels utilizing frequencies in the two ranges of 54 to 88 MHz and 174 to 222 MHz

ultra-high frequency (UHF): TV channels in an even higher frequency range than VHF, of 470 to 1000 MHz

X-ray: invisible, penetrating form of very high frequency electromagnetic radiation, overlapping both the ultraviolet range and the γ-ray range

gamma ray: (γ ray); extremely high frequency electromagnetic radiation emitted by the nucleus of an atom, either from natural nuclear decay or induced nuclear processes in nuclear reactors and weapons. The lower end of the γ-ray frequency range overlaps the upper end of the X-ray range, but γ rays can have the highest frequency of any electromagnetic radiation

Selected Solutions to Problems & Exercises

1. (a) 33.3 cm (900 MHz) 11.7 cm (2560 MHz); (b) The microwave oven with the smaller wavelength would

produce smaller hot spots in foods, corresponding to the one with the frequency 2560 MHz.

3. 26.96 MHz

5. 5.0 × 10^{14} Hz

7.

$$\lambda = \frac{c}{f} = \frac{3.00 \times 10^8 \text{ m/s}}{1.20 \times 10^{15} \text{ Hz}} = 2.50 \times 10^{-7} \text{ m}$$

9. 0.600 m

11. (a)

$$f = \frac{c}{\lambda} = \frac{3.00 \times 10^8 \text{ m/s}}{1 \times 10^{-10} \text{ m}} = 3 \times 10^{18} \text{ Hz}$$

; (b) X-rays

14. (a) 6.00 × 10^6 m; (b) 4.33 × 10^{-5} T

16. (a) 1.50 × 10 6 Hz, AM band; (b) The resonance of currents on an antenna that is 1/4 their wavelength is analogous to the fundamental resonant mode of an air column closed at one end, since the tube also has a length equal to 1/4 the wavelength of the fundamental oscillation.

18. (a) 1.55 × 10^{15} Hz; (b) The shortest wavelength of visible light is 380 nm, so that

$$\begin{aligned} & \frac{\lambda_{\text{visible}}}{\lambda_{\text{UV}}} \\ = \; & \frac{380 \text{ nm}}{193 \text{ nm}} \\ = \; & 1.97 \end{aligned}$$

In other words, the UV radiation is 97% more accurate than the shortest wavelength of visible light, or almost twice as accurate!

20. 3.90 × 10^8 m

22. (a) 1.50×10^{11} m; (b) 0.500 μs; (c) 66.7 ns

24. (a) -3.5×10^{2} W/m^2; (b) 88%; (c) 1.7 μT

68. Energy in Electromagnetic Waves

Learning Objectives

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

- Explain how the energy and amplitude of an electromagnetic wave are related.
- Given its power output and the heating area, calculate the intensity of a microwave oven's electromagnetic field, as well as its peak electric and magnetic field strengths

Anyone who has used a microwave oven knows there is energy in *electromagnetic waves*. Sometimes this energy is obvious, such as in the warmth of the summer sun. Other times it is subtle, such as the unfelt energy of gamma rays, which can destroy living cells.

Electromagnetic waves can bring energy into a system by virtue of their *electric and magnetic fields*. These fields can exert forces and move charges in the system and, thus, do work on them. If the frequency of the electromagnetic wave is the same as the natural frequencies of the system (such as microwaves at the resonant frequency of water molecules), the transfer of energy is much more efficient.

Making Connections: Waves and Particles

The behavior of electromagnetic radiation clearly exhibits wave characteristics. But we shall find in later modules that at high frequencies, electromagnetic radiation also exhibits particle characteristics. These particle characteristics will be used to explain more of the properties of the electromagnetic spectrum and to introduce the formal study of modern physics.

Another startling discovery of modern physics is that particles, such as electrons and protons, exhibit wave characteristics. This simultaneous sharing of wave and particle properties for all submicroscopic entities is one of the great symmetries in nature.

The propagation of two electromagnetic waves is shown in three dimensional planes. The first wave shows with the variation of two components E and B. E is a sine wave in one plane with small arrows showing the vibrations of particles in the plane. B is a sine wave in a plane perpendicular to the E wave. The B wave has arrows to show the vibrations of particles in the plane. The waves are shown intersecting each other at the junction of the planes because E and B are perpendicular to each other. The direction of propagation of wave is shown perpendicular to E and B waves. The energy carried is given as E sub u. The second wave shows with the variation of the components two E and two B, that is, E and B waves with double the amplitude of the first case. Two E is a sine wave in one plane with small arrows showing the vibrations of particles in the plane. Two B is a sine wave in a plane perpendicular to the two E wave. The two B wave has arrows to show the vibrations of particles in the plane. The waves are shown intersecting each other at the junction of the planes because two E and two B waves are perpendicular to each other. The direction of propagation of wave is shown perpendicular to two E and two B waves. The energy carried is given as four E sub u.

Figure 1. Energy carried by a wave is proportional to its amplitude squared. With electromagnetic waves, larger E-fields and B-fields exert larger forces and can do more work.

But there is energy in an electromagnetic wave, whether it is absorbed or not. Once created, the fields carry energy away from a source. If absorbed, the field strengths are diminished and anything left travels on. Clearly, the larger the strength of the electric and magnetic fields, the more work they can do and the greater the energy the electromagnetic wave carries.

A wave's energy is proportional to its *amplitude* squared (E^2 or B^2). This is true for waves on guitar strings, for water waves, and for sound waves, where amplitude is proportional to pressure. In electromagnetic waves, the amplitude is the *maximum field strength* of the electric and magnetic fields. (See Figure 1.)

Thus the energy carried and the *intensity* I of an electromagnetic wave is proportional to E^2 and B^2. In fact, for a continuous sinusoidal electromagnetic wave, the average intensity I_{ave} is given by

$$I_{\text{ave}} = \frac{c\epsilon_0 E_0^2}{2}$$

where c is the speed of light, ε_0 is the permittivity of free space, and E_0 is the maximum electric field strength; intensity, as always, is power per unit area (here in W/m^2).

The average intensity of an electromagnetic wave I_{ave} can also be expressed in terms of the magnetic field strength by using the relationship $B = \frac{E}{c}$, and the fact that $\epsilon_0 = \frac{1}{\mu_0 c^2}$, where μ_0 is the permeability of free space. Algebraic manipulation produces the relationship

$$I_{\text{ave}} = \frac{cB_0^2}{2\mu_0},$$

where B_0 is the maximum magnetic field strength.

One more expression for I_{ave} in terms of both electric and magnetic field strengths is useful. Substituting the fact that $c \cdot B_0 = E_0$, the previous expression becomes

$$I_{\text{ave}} = \frac{E_0 B_0^2}{2\mu_0}.$$

Whichever of the three preceding equations is most convenient can be used, since they are really just different versions of the same principle: Energy in a wave is related to amplitude squared. Furthermore, since these equations are based on the assumption that the electromagnetic waves are sinusoidal, peak intensity is twice the average; that is, $I_0 = 2I_{ave}$.

Example 1. Calculate Microwave Intensities and Fields

On its highest power setting, a certain microwave oven projects 1.00 kW of microwaves onto a 30.0 by 40.0 cm area.

1. What is the intensity in W/m^2?
2. Calculate the peak electric field strength E_0 in these waves.
3. What is the peak magnetic field strength B_0?

Strategy

In Part 1, we can find intensity from its definition as power per unit area. Once the intensity is known, we can use the equations below to find the field strengths asked for in Parts 2 and 3.

Solution for Part 1

Entering the given power into the definition of intensity, and noting the area is 0.300 by 0.400 m, yields

$$I = \frac{P}{A} = \frac{1.00\ \text{kW}}{0.300\ \text{m} \times 0.400\ \text{m}}$$

Here $I = I_{ave}$, so that

$$I_{\text{ave}} = \frac{1000\ \text{W}}{0.120\ \text{m}^2} = 8.33 \times 10^3\ \text{W/m}^2$$

Note that the peak intensity is twice the average: $I_0 = 2I_{\text{ave}} = 1.67 \times 10^4\ \text{W/m}^2$.

Solution for Part 2

To find E_0, we can rearrange the first equation given above for I_{ave} to give

$$E_0 = \left(\frac{2I_{\text{ave}}}{c\epsilon_0}\right)^{1/2}$$

Entering known values gives

$$\begin{aligned} E_0 &= \sqrt{\frac{2\left(8.33\times10^3\ \text{W/m}^2\right)}{\left(3.00\times10^8\ \text{m/s}\right)\left(8.85\times10^{-12}\ \text{C}^2/\text{N}\cdot\text{m}^2\right)}} \\ &= 2.51 \times 10^3\ \text{V/m} \end{aligned}$$

Solution for Part 3

Perhaps the easiest way to find magnetic field strength, now that the electric field strength is known, is to use the relationship given by $B_0 = \dfrac{E_0}{c}$.

Entering known values gives

$$\begin{aligned} B_0 &= \frac{2.51\times10^3\ \text{V/m}}{3.0\times10^8\ \text{m/s}} \\ &= 8.35 \times 10^{-6}\ \text{T} \end{aligned}$$

Discussion

As before, a relatively strong electric field is accompanied by a relatively weak magnetic field in an electromagnetic wave, since $B = \frac{E}{c}$, and *c* is a large number.

Section Summary

- The energy carried by any wave is proportional to its amplitude squared. For electromagnetic waves, this means intensity can be expressed as $I_{\text{ave}} = \frac{\mathrm{c}\epsilon_0 E_0^2}{2}$, where I_{ave} is the average intensity in W/m^2, and E_0 is the maximum electric field strength of a continuous sinusoidal wave.
- This can also be expressed in terms of the maximum magnetic field strength B_0 as $I_{\text{ave}} = \frac{\mathrm{cB}_0^2}{2\mu_0}$ and in terms of both electric and magnetic fields as $I_{\text{ave}} = \frac{E_0 B_0}{2\mu_0}$.
- The three expressions for I_{ave} are all equivalent.

Problems & Exercises

1. What is the intensity of an electromagnetic wave with a peak electric field strength of 125 V/m?
2. Find the intensity of an electromagnetic wave having a peak magnetic field strength of 4.00×10^{-9} T.
3. Assume the helium-neon lasers commonly used in student physics laboratories have power outputs of 0.250 mW. (a) If such a laser beam is projected onto a circular spot 1.00 mm in diameter, what is its intensity? (b) Find the peak magnetic field strength. (c) Find the peak electric field strength.
4. An AM radio transmitter broadcasts 50.0 kW of power uniformly in all directions. (a) Assuming all of the radio waves that strike the ground are completely absorbed, and that there is no absorption by the atmosphere or other objects, what is the intensity 30.0 km away? (Hint: Half the power will be spread over the area of a hemisphere.) (b) What is the maximum electric field strength at this distance?
5. Suppose the maximum safe intensity of microwaves for human exposure is taken to be 1.00 W/m2. (a) If a radar unit leaks 10.0 W of microwaves (other than those sent by its antenna) uniformly in all directions, how far away must you be to be exposed to an intensity considered to be safe? Assume that the power spreads uniformly over the area of a sphere with no complications from absorption or reflection. (b) What is the maximum electric field strength at the safe intensity? (Note that early radar units leaked more than modern ones do. This caused identifiable

health problems, such as cataracts, for people who worked near them.)

6. A 2.50-m-diameter university communications satellite dish receives TV signals that have a maximum electric field strength (for one channel) of $7.50\ \mu V/m$. (See Figure 2.) (a) What is the intensity of this wave? (b) What is the power received by the antenna? (c) If the orbiting satellite broadcasts uniformly over an area of $1.50 \times 10^{13}\ m^2$ (a large fraction of North America), how much power does it radiate?

Figure 2. Satellite dishes receive TV signals sent from orbit. Although the signals are quite weak, the receiver can detect them by being tuned to resonate at their frequency.

7. Lasers can be constructed that produce an extremely high intensity electromagnetic wave for a brief time—called pulsed lasers. They are used to ignite nuclear fusion, for example. Such a laser may

produce an electromagnetic wave with a maximum electric field strength of 1.00×10^{11} V/m for a time of 1.00 ns. (a) What is the maximum magnetic field strength in the wave? (b) What is the intensity of the beam? (c) What energy does it deliver on a 1.00-mm^2 area?

8. Show that for a continuous sinusoidal electromagnetic wave, the peak intensity is twice the average intensity ($I_0 = 2I_{ave}$), using either the fact that $E_0 = \sqrt{2}E_{rms}$, or $B_0 = \sqrt{2}B_{rms}$, where rms means average (actually root mean square, a type of average).
9. Suppose a source of electromagnetic waves radiates uniformly in all directions in empty space where there are no absorption or interference effects. (a) Show that the intensity is inversely proportional to r^2, the distance from the source squared. (b) Show that the magnitudes of the electric and magnetic fields are inversely proportional to r.
10. **Integrated Concepts.** An *LC* circuit with a 5.00-pF capacitor oscillates in such a manner as to radiate at a wavelength of 3.30 m. (a) What is the resonant frequency? (b) What inductance is in series with the capacitor?
11. **Integrated Concepts.** What capacitance is needed in series with an 800-μH inductor to form a circuit that radiates a wavelength of 196 m?
12. **Integrated Concepts.** Police radar determines the speed of motor vehicles using the same Doppler-shift technique employed for ultrasound in medical diagnostics. Beats are produced by mixing the double Doppler-shifted echo with the original frequency. If

1.50×10^9-Hz microwaves are used and a beat frequency of 150 Hz is produced, what is the speed of the vehicle? (Assume the same Doppler-shift formulas are valid with the speed of sound replaced by the speed of light.)

13. **Integrated Concepts.**Assume the mostly infrared radiation from a heat lamp acts like a continuous wave with wavelength 1.50μm. (a) If the lamp's 200-W output is focused on a person's shoulder, over a circular area 25.0 cm in diameter, what is the intensity in W/m^2? (b) What is the peak electric field strength? (c) Find the peak magnetic field strength. (d) How long will it take to increase the temperature of the 4.00-kg shoulder by 2.00° C, assuming no other heat transfer and given that its specific heat is 3.47×10^3 J/kg·°C?
14. **Integrated Concepts.** On its highest power setting, a microwave oven increases the temperature of 0.400 kg of spaghetti by 45.0°C in 120 s. (a) What was the rate of power absorption by the spaghetti, given that its specific heat is 3.76×10^3 J/kg · °C? (b) Find the average intensity of the microwaves, given that they are absorbed over a circular area 20.0 cm in diameter. (c) What is the peak electric field strength of the microwave? (d) What is its peak magnetic field strength?
15. **Integrated Concepts.** Electromagnetic radiation from a 5.00-mW laser is concentrated on a 1.00-mm^2 area. (a) What is the intensity in W/m^2? (b) Suppose a 2.00-nC static charge is in the beam. What is the maximum electric force it experiences? (c) If the static charge moves at 400 m/s, what maximum

magnetic force can it feel?

16. **Integrated Concepts.** A 200-turn flat coil of wire 30.0 cm in diameter acts as an antenna for FM radio at a frequency of 100 MHz. The magnetic field of the incoming electromagnetic wave is perpendicular to the coil and has a maximum strength of 1.00×10^{-12} T. (a) What power is incident on the coil? (b) What average emf is induced in the coil over one-fourth of a cycle? (c) If the radio receiver has an inductance of 2.50 μH, what capacitance must it have to resonate at 100 MHz?
17. **Integrated Concepts.** If electric and magnetic field strengths vary sinusoidally in time, being zero at $t = 0$, then $E = E_0 \sin 2\pi ft$ and $B = B_0 \sin 2\pi ft$. Let $f = 1.00$ GHz here. (a) When are the field strengths first zero? (b) When do they reach their most negative value? (c) How much time is needed for them to complete one cycle?
18. **Unreasonable Results.** A researcher measures the wavelength of a 1.20-GHz electromagnetic wave to be 0.500 m. (a) Calculate the speed at which this wave propagates. (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?
19. **Unreasonable Results.** The peak magnetic field strength in a residential microwave oven is 9.20×10^{-5} T. (a) What is the intensity of the microwave? (b) What is unreasonable about this result? (c) What is wrong about the premise?
20. **Unreasonable Results.** An LC circuit containing a 2.00-H inductor oscillates at such a frequency that it radiates at a 1.00-m wavelength. (a) What is the

capacitance of the circuit? (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

21. **Unreasonable Results.** An LC circuit containing a 1.00-pF capacitor oscillates at such a frequency that it radiates at a 300-nm wavelength. (a) What is the inductance of the circuit? (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?
22. **Create Your Own Problem.** Consider electromagnetic fields produced by high voltage power lines. Construct a problem in which you calculate the intensity of this electromagnetic radiation in W/m2 based on the measured magnetic field strength of the radiation in a home near the power lines. Assume these magnetic field strengths are known to average less than a μT. The intensity is small enough that it is difficult to imagine mechanisms for biological damage due to it. Discuss how much energy may be radiating from a section of power line several hundred meters long and compare this to the power likely to be carried by the lines. An idea of how much power this is can be obtained by calculating the approximate current responsible for μT fields at distances of tens of meters.
23. **Create Your Own Problem.** Consider the most recent generation of residential satellite dishes that are a little less than half a meter in diameter. Construct a problem in which you calculate the power received by the dish and the maximum electric field strength of the microwave signals for a single channel received by the dish. Among the things to be

considered are the power broadcast by the satellite and the area over which the power is spread, as well as the area of the receiving dish.

Glossary

maximum field strength: the maximum amplitude an electromagnetic wave can reach, representing the maximum amount of electric force and/or magnetic flux that the wave can exert

intensity: the power of an electric or magnetic field per unit area, for example, Watts per square meter

Selected Solutions to Problems & Exercises

1.

$$\begin{aligned} I &= \frac{c\epsilon_0 E_0^2}{2} \\ &= \frac{\left(3.00\times10^8\text{m/s}\right)\left(8.85\times10^{-12}\text{C}^2/\text{N}\cdot\text{m}^2\right)\left(125\text{ V/m}\right)^2}{2} \\ &= 20.7\text{ W/m}^2 \end{aligned}$$

3. (a)

$$I = \frac{P}{A} = \frac{P}{\pi r^2} = \frac{0.250\times10^{-3}\text{W}}{\pi\left(0.500\times10^{-3}\text{m}\right)^2} = 318\text{ W/m}^2$$

(b)

$$I_{\text{ave}} = \frac{\mathrm{cB}_0^2}{2\mu_0} \Rightarrow B_0 = \left(\frac{2\mu_0 I}{c}\right)^{1/2}$$

$$= \left(\frac{2\left(4\pi\times10^{-7}\mathrm{T{\cdot}m/A}\right)\left(318.3\ \mathrm{W/m^2}\right)}{3.00\times10^{8}\mathrm{m/s}}\right)^{1/2}$$

$$= 1.63\times10^{-6}\mathrm{T}$$

(c)

$$E_0 = \mathrm{cB}_0 = \left(3.00\times10^{8}\mathrm{m/s}\right)\left(1.633\times10^{-6}\mathrm{T}\right)$$

$$= 4.90\times10^{2}\mathrm{V/m}$$

5. (a) 89.2 cm; (b) 27.4 V/m

7. (a) 333 T; (b) 1.33×10^{19} W/m^2; (c) 13.3 kJ

9. (a) $I = \frac{P}{A} = \frac{P}{4\pi r^2} \propto \frac{1}{r^2}$; (b)

$$I \propto {E_0}^2, {B_0}^2 \Rightarrow {E_0}^2, {B_0}^2 \propto \frac{1}{r^2} \Rightarrow E_0, B_0 \propto \frac{1}{r}$$

11. 13.5 pF

13. (a) 4.07 kW/m^2; (b) 1.75 kV/m; (c) 5.84 μT; (d) 2 min 19 s

15. (a) 5.00×10^{3} W/m^2; (b) 3.88×10^{-6} N; (c) 5.18×10^{-12} N

17. (a) $t = 0$; (b) 7.50×10^{-10} s; (c) 1.00×10^{-9} s

19. (a) 1.01×10^{6} W/m^2; (b) Much too great for an oven; (c) The assumed magnetic field is unreasonably large.

21. (a) 2.53×10^{-20} H; (b) L is much too small; (c) The wavelength is unreasonably small.

PART IX
GEOMETRIC OPTICS

69. Introduction to Geometric Optics

People in white clothing covered from head to toe and wearing blue colored gloves are working in a research laboratory setting, one person holding a flash light and analyzing and another reading a manuscript and so on. Their images are seen on a smooth colored glass top of a work table.

Figure 1. Image seen as a result of reflection of light on a plane smooth surface. (credit: NASA Goddard Photo and Video, via Flickr)

Light from this page or screen is formed into an image by the lens of your eye, much as the lens of the camera that made this photograph. Mirrors, like lenses, can also form images that in turn are captured by your eye.

Our lives are filled with light. Through vision, the most valued of our senses, light can evoke spiritual emotions, such as when we view a magnificent sunset or glimpse a rainbow breaking through the clouds. Light can also simply amuse us in a theater, or warn us to stop at an intersection. It has innumerable uses beyond vision. Light can carry telephone signals through glass fibers or cook a meal in a solar oven. Life itself could not exist without light's energy. From photosynthesis in plants to the sun warming a cold-blooded animal, its supply of energy is vital.

Figure 2. Double Rainbow over the bay of Pocitos in Montevideo, Uruguay. (credit: Madrax, Wikimedia Commons)

We already know that visible light is the type of electromagnetic waves to which our eyes respond. That knowledge still leaves many questions regarding the nature of light and vision. What is color, and how do our eyes detect it? Why do diamonds sparkle? How does light travel? How do lenses and mirrors form images? These are but a few of the questions that are answered by the study of optics. Optics is the branch of physics that deals with the behavior of visible light and other electromagnetic waves. In particular, optics is concerned with the generation and propagation of light and its interaction with matter. What we have already learned about the generation of light in our study of heat transfer by radiation will be expanded upon in later topics, especially those on atomic physics. Now, we will concentrate on the propagation of light and its interaction with matter.

It is convenient to divide optics into two major parts based on the size of objects that light encounters. When light interacts with an object that is several times as large as the light's wavelength, its observable behavior is like that of a ray; it does not prominently display its wave characteristics. We call this part of optics "geometric optics." This chapter will concentrate on such situations. When light interacts with smaller objects, it has very prominent wave characteristics, such as constructive and destructive interference. Wave Optics will concentrate on such situations.

70. The Ray Aspect of Light

Learning Objectives

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

- List the ways by which light travels from a source to another location.

There are three ways in which light can travel from a source to another location. (See Figure 1.) It can come directly from the source through empty space, such as from the Sun to Earth. Or light can travel through various media, such as air and glass, to the person. Light can also arrive after being reflected, such as by a mirror. In all of these cases, light is modeled as traveling in straight lines called rays. Light may change direction when it encounters objects (such as a mirror) or in passing from one material to another (such as in passing from air to glass), but it then continues in a straight line or as a ray. The word *ray* comes from mathematics and here means a straight line that originates at some point. It is acceptable to visualize light rays as laser rays (or even science fiction depictions of ray guns).

Ray

The word “ray” comes from mathematics and here means a straight line that originates at some point.

Experiments, as well as our own experiences, show that when light interacts with objects several times as large as its wavelength, it travels in straight lines and acts like a ray. Its wave characteristics are not pronounced in such situations. Since the wavelength of light is less than a micron (a thousandth of a millimeter), it acts like a ray in the many common situations in which it encounters objects larger than a micron. For example, when light encounters anything we can observe with unaided eyes, such as a mirror, it acts like a ray, with only subtle wave characteristics. We will concentrate on the ray characteristics in this chapter.

Part a shows the suns rays shining through the atmosphere to a house on the earth. The second image shows a woman looking out the window at a car and arrows depicting the light rays hitting the car and the window, then her eyes are shown.

Figure 1. Three methods for light to travel from a source to another location. (a) Light reaches the upper atmosphere of Earth traveling through empty space directly from the source. (b) Light can reach a person in one of two ways. It can travel through media like air and glass. It can also reflect from an object like a mirror. In the situations shown here, light interacts with objects large enough that it travels in straight lines, like a ray.

Since light moves in straight lines, changing directions when it interacts with materials, it is described by geometry and simple trigonometry. This part of optics, where the ray aspect of light dominates, is therefore called *geometric optics*. There are two laws that govern how light changes direction when it interacts with matter. These are the law of reflection, for situations in which light bounces off matter, and the law of refraction, for situations in which light passes through matter.

Geometric Optics

The part of optics dealing with the ray aspect of light is called geometric optics.

Section Summary

A straight line that originates at some point is called a ray.
The part of optics dealing with the ray aspect of light is called geometric optics.
Light can travel in three ways from a source to another location: (1) directly from the source through empty space; (2) through various media; (3) after being reflected from a mirror.

Problems & Exercises

Suppose a man stands in front of a mirror as shown in Figure 2. His eyes are 1.65 m above the floor, and the top of his head is 0.13 m higher. Find the height above the floor of the top and bottom of the smallest mirror in which he can see both the top of his head and his feet. How is this distance related to the man's height?

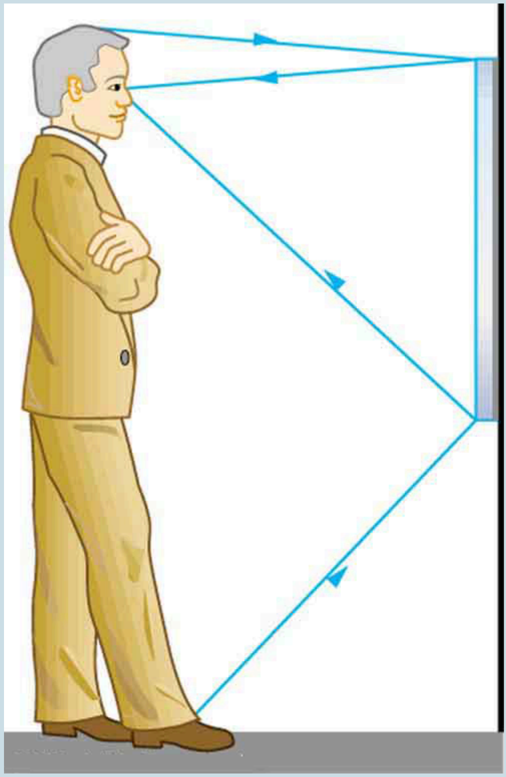


Figure 2. A man standing in front of a mirror on a wall at a distance of several feet. The mirror's top is at eye level, but its bottom is only waist high. Arrows illustrate how the man can see his reflection from head to toe in the mirror.

Glossary

ray: straight line that originates at some point

geometric optics: part of optics dealing with the ray aspect of light

Solution to Problems & Exercises

Top 1.715 m from floor, bottom 0.825 m from floor. Height of mirror is 0.890 m, or precisely one-half the height of the person.

71. The Law of Reflection

Learning Objectives

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

- Explain reflection of light from polished and rough surfaces.

Whenever we look into a mirror, or squint at sunlight glinting from a lake, we are seeing a reflection. When you look at this page, too, you are seeing light reflected from it. Large telescopes use reflection to form an image of stars and other astronomical objects.

A light ray is incident on a smooth surface and is falling obliquely, making an angle theta i relative to a perpendicular line drawn to the surface at the point where the incident ray strikes. The light ray gets reflected making an angle theta r with the same perpendicular drawn to the surface.

Figure 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.

The law of reflection is illustrated in Figure 1, which also shows how the angles are measured relative to the perpendicular to the surface at the point where the light ray strikes. We expect to see reflections from smooth surfaces, but Figure 2 illustrates how a rough surface reflects light. Since the light strikes different parts of the surface at different angles, it is reflected in many different directions, or diffused. Diffused light is what allows us to see a sheet of paper from any angle, as illustrated in Figure 3. Many objects, such as people, clothing, leaves, and walls, have rough surfaces and can be seen from all sides. A mirror, on the other hand, has a smooth surface (compared with the wavelength of light) and reflects light at specific angles, as illustrated in Figure 4. When the moon reflects from a lake, as shown in Figure 5, a combination of these effects takes place.

Parallel light rays falling on a rough surface get scattered at different angles.

Figure 2. Light is diffused when it reflects from a rough surface. Here many parallel rays are incident, but they are reflected at many different angles since the surface is rough.

Light from a flashlight falls on a sheet of paper and the light gets reflected at different angles as the surface is rough.

Figure 3. When a sheet of paper is illuminated with many parallel incident rays, it can be seen at many different angles, because its surface is rough and diffuses the light.

A flashlight casting light on a mirror, which is smooth; the mirror reflects light only in one direction at a particular angle.

Figure 4. A mirror illuminated by many parallel rays reflects them in only one direction, since its surface is very smooth. Only the observer at a particular angle will see the reflected light.

A dark night is lit by moonlight. The moonlight is falling on the lake and as it hits, the lake's shiny surface reflects it. A bright strip of moonlight is seen reflecting from the lake on a dark background reflecting the night sky.

Figure 5. Moonlight is spread out when it is reflected by the lake, since the surface is shiny but uneven. (credit: Diego Torres Silvestre, Flickr)

The law of reflection is very simple: The angle of reflection equals the angle of incidence.

The Law of Reflection

The angle of reflection equals the angle of incidence.

When we see ourselves in a mirror, it appears that our image is actually behind the mirror. This is illustrated in Figure 6. We see the light coming from a direction determined by the law of reflection. The angles are such that our image is exactly the same distance behind the mirror as we stand away from the mirror. If the mirror is on the wall of a room, the images in it are all behind the mirror, which can make the room seem bigger. Although these mirror images make objects appear to be where they cannot be (like behind a solid wall), the images are not figments of our imagination. Mirror images can be photographed and videotaped by instruments and look just as they do with our eyes (optical instruments themselves). The precise manner in which images are formed by mirrors and lenses will be treated in later sections of this chapter.

A girl stands in front of a mirror and looks into the mirror for her image. The light rays from her feet and head fall on the mirror and get reflected following the law of reflection: the angle of incidence theta is equal to the angle of reflection theta.

Figure 6. Our image in a mirror is behind the mirror. The two rays shown are those that strike the mirror at just the correct angles to be reflected into the eyes of the person. The image appears to be in the direction the rays are coming from when they enter the eyes.

Take-Home Experiment: Law of Reflection

Take a piece of paper and shine a flashlight at an angle at the paper, as shown in Figure 3. Now shine the flashlight at a mirror at an angle. Do your observations confirm the predictions in Figure 3 and Figure 4? Shine the flashlight on various surfaces and determine whether the reflected light is diffuse or not. You can choose a shiny metallic lid of a pot or your skin. Using the mirror and flashlight, can you confirm the law of reflection? You will need to draw lines on a piece of paper showing the incident and reflected rays. (This part works even better if you use a laser pencil.)

Section Summary

- The angle of reflection equals the angle of incidence.
- A mirror has a smooth surface and reflects light at specific angles.
- Light is diffused when it reflects from a rough surface.
- Mirror images can be photographed and videotaped by instruments.

Conceptual Question

1. Using the law of reflection, explain how powder

takes the shine off of a person's nose. What is the name of the optical effect?

Problems & Exercises

1. Show that when light reflects from two mirrors that meet each other at a right angle, the outgoing ray is parallel to the incoming ray, as illustrated in the following figure.

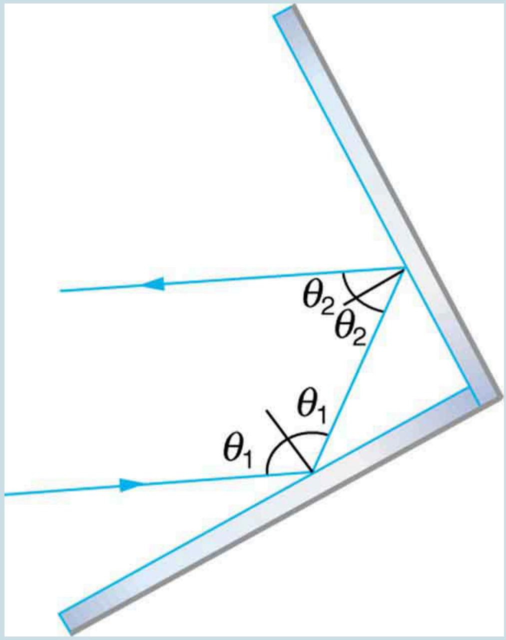


Figure 7. A corner reflector sends the reflected ray back in a direction parallel to the incident ray, independent of incoming direction.

2. Light shows staged with lasers use moving mirrors to swing beams and create colorful effects. Show that a light ray reflected from a mirror changes direction by 2θ when the mirror is rotated by an angle θ.
3. A flat mirror is neither converging nor diverging. To prove this, consider two rays originating from the same point and diverging at an angle θ. Show that after striking a plane mirror, the angle between their directions remains θ.

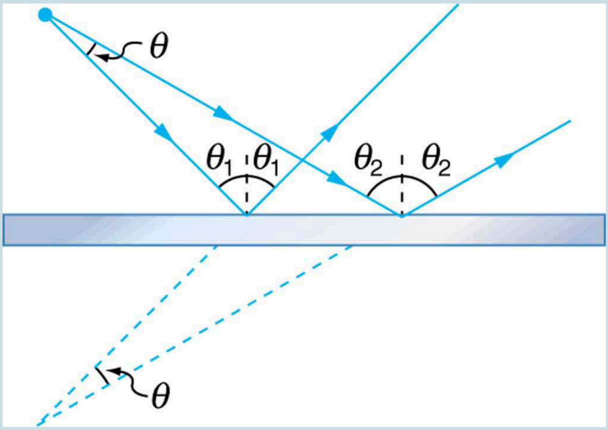


Figure 8. A flat mirror neither converges nor diverges light rays. Two rays continue to diverge at the same angle after reflection.

Glossary

mirror: smooth surface that reflects light at specific angles, forming an image of the person or object in front of it

law of reflection: angle of reflection equals the angle of incidence

72. The Law of Refraction

Learning Objective

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

- Determine the index of refraction, given the speed of light in a medium.

It is easy to notice some odd things when looking into a fish tank. For example, you may see the same fish appearing to be in two different places. (See Figure 1.) This is because light coming from the fish to us changes direction when it leaves the tank, and in this case, it can travel two different paths to get to our eyes. The changing of a light ray's direction (loosely called bending) when it passes through variations in matter is called *refraction*. Refraction is responsible for a tremendous range of optical phenomena, from the action of lenses to voice transmission through optical fibers.

Refraction

The changing of a light ray's direction (loosely called bending) when it passes through variations in matter is called refraction.

Speed of Light

The speed of light *c* not only affects refraction, it is one of the central concepts of Einstein's theory of relativity. As the accuracy of the measurements of the speed of light were improved, *c* was found not to depend on the velocity of the source or the observer. However, the speed of light does vary in a precise manner with the material it traverses. These facts have far-reaching implications, as we will see in the chapter Special Relativity. It makes connections between space and time and alters our expectations that all observers measure the same time for the same event, for example. The speed of light is so important that its value in a vacuum is one of the most fundamental constants in nature as well as being one of the four fundamental SI units.

A person looks at a fish tank and he sees the same fish in two different directions at the edge of the water tank facing him.

Figure 1. Looking at the fish tank as shown, we can see the same fish in two different locations, because light changes directions when it passes from water to air. In this case, the light can reach the observer by two different paths, and so the fish seems to be in two different places. This bending of light is called refraction and is responsible for many optical phenomena.

Why does light change direction when passing from one material (medium) to another? It is because light changes speed when going from one material to another. So before we study the law of refraction, it is useful to discuss the speed of light and how it varies in different media.

The Speed of Light

Early attempts to measure the speed of light, such as those made by Galileo, determined that light moved extremely fast, perhaps instantaneously. The first real evidence that light traveled at a finite speed came from the Danish astronomer Ole Roemer in the late 17th century. Roemer had noted that the average orbital period of one of Jupiter's moons, as measured from Earth, varied depending on whether Earth was moving toward or away from Jupiter. He correctly concluded that the apparent change in period was due to the change in distance between Earth and Jupiter and the time it took light to travel this distance. From his 1676 data, a value of the speed of light was calculated to be 2.26×10^8 m/s (only 25% different from today's accepted value). In more recent times, physicists have measured the speed of light in numerous ways and with increasing accuracy. One particularly direct method, used in 1887 by the American physicist Albert Michelson (1852–1931), is illustrated in Figure 2. Light reflected from a rotating set of mirrors was reflected from a stationary mirror 35 km away and returned to the rotating mirrors. The time for the light to travel can be determined by how fast the mirrors must rotate for the light to be returned to the observer's eye.

In stage one of the figure, the light falling from a source on an eight-sided mirror is viewed by an observer; in stage two, the mirror is made to rotate and the reflected light falling onto a stationary mirror kept at a certain distance of 35 kilometers is viewed by an observer. In stage three, the observer can see the reflected ray only when the mirror has rotated into the correct position just as the ray returns.

Figure 2. A schematic of early apparatus used by Michelson and others to determine the speed of light. As the mirrors rotate, the reflected ray is only briefly directed at the stationary mirror. The returning ray will be reflected into the observer's eye only if the next mirror has rotated into the correct position just as the ray returns. By measuring the correct rotation rate, the time for the round trip can be measured and the speed of light calculated. Michelson's calculated value of the speed of light was only 0.04% different from the value used today.

The speed of light is now known to great precision. In fact, the speed of light in a vacuum *c* is so important that it is accepted as one of the basic physical quantities and has the fixed value $c = 2.9972458 \times 10^8$ m/s $\approx 3.00 \times 10^8$ m/s, where the approximate value of 3.00×10^8 m/s is used whenever three-digit accuracy is sufficient. The speed of light through matter is less than it is in a vacuum, because light interacts with atoms in a material. The speed of light depends strongly on the type of material, since its interaction with different atoms, crystal lattices, and other substructures varies. We define the *index of refraction* *n* of a material to be $n = \frac{c}{v}$, where *v* is the observed speed of light in the material. Since the speed of light is always less than *c* in matter and equals *c* only in a vacuum, the index of refraction is always greater than or equal to one.

Value of the Speed of Light

$$c = 2.9972458 \times 10^8 \text{ m/s} \approx 3.00 \times 10^8 \text{ m/s}$$

Index of Refraction

$$n = \frac{c}{v}$$

That is, $n \geq 1$. Table 1 gives the indices of refraction for some representative substances. The values are listed for a particular wavelength of light, because they vary slightly with wavelength. (This can have important effects, such as colors produced by a prism.) Note that for gases, n is close to 1.0. This seems reasonable, since atoms in gases are widely separated and light travels at c in the vacuum between atoms. It is common to take $n = 1$ for gases unless great precision is needed. Although the speed of light v in a medium varies considerably from its value c in a vacuum, it is still a large speed.

Table 1. Index of Refraction in Various Media	
Medium	***n***
Gases at 0°C, 1 atm	
Air	1.000293
Carbon dioxide	1.00045
Hydrogen	1.000139
Oxygen	1.000271
Liquids at 20°C	
Benzene	1.501
Carbon disulfide	1.628
Carbon tetrachloride	1.461
Ethanol	1.361
Glycerine	1.473
Water, fresh	1.333
Solids at 20°C	
Diamond	2.419
Fluorite	1.434
Glass, crown	1.52
Glass, flint	1.66
Ice at 20°C	1.309
Polystyrene	1.49
Plexiglas	1.51
Quartz, crystalline	1.544
Quartz, fused	1.458
Sodium chloride	1.544
Zircon	1.923

Example 1. Speed of Light in Matter

Calculate the speed of light in zircon, a material used in jewelry to imitate diamond.

Strategy

The speed of light in a material, v, can be calculated from the index of refraction n of the material using the equation $n = \frac{c}{v}$.

Solution

The equation for index of refraction states that $n = \frac{c}{v}$. Rearranging this to determine v gives $v = \frac{c}{n}$.

The index of refraction for zircon is given as 1.923 in Table 1, and c is given in the equation for speed of light. Entering these values in the last expression gives

$$\begin{aligned} v &= \frac{3.00 \times 10^8 \text{ m/s}}{1.923} \\ &= 1.56 \times 10^8 \text{ m/s} \end{aligned}$$

Discussion

This speed is slightly larger than half the speed of light in a vacuum and is still high compared with speeds we normally experience. The only substance listed in Table 1 that has a greater index of refraction than zircon is diamond. We shall see later that the large index of refraction for zircon makes it sparkle more than glass, but less than diamond.

Law of Refraction

Figure 3 shows how a ray of light changes direction when it passes from one medium to another. As before, the angles are measured relative to a perpendicular to the surface at the point where the light ray crosses it. (Some of the incident light will be reflected from the surface, but for now we will concentrate on the light that is transmitted.) The change in direction of the light ray depends on how the speed of light changes. The change in the speed of light is related to the indices of refraction of the media involved. In the situations shown in Figure 3, medium 2 has a greater index of refraction than medium 1. This means that the speed of light is less in medium 2 than in medium 1. Note that as shown in Figure 3a, the direction of the ray moves closer to the perpendicular when it slows down. Conversely, as shown in Figure 3b, the direction of the ray moves away from the perpendicular when it speeds up. The path is exactly reversible. In both cases, you can imagine what happens by thinking about pushing a lawn mower from a footpath onto grass, and vice versa. Going from the footpath to grass, the front wheels are slowed and pulled to the side as shown. This is the same change in direction as for light when it goes from a fast medium to a slow one. When going from the grass to the footpath, the front wheels can move faster and the mower changes direction as shown. This, too, is the same change in direction as for light going from slow to fast.

The figures compare the working of a lawn mower to that of the refraction phenomenon. In figure (a) the lawn mower goes from a sidewalk to grass, it slows down and bends towards a perpendicular drawn at the point of contact of the mower with the surface of separation. An imaginary line along the mower when it is on sidewalk is taken to be the incident ray and the angle which the mower makes with the perpendicular is taken to be theta one. As it goes into the grass, the mower turns and the imaginary line moves towards the perpendicular line drawn and makes an angle theta two with it. The imaginary line drawn along the mower when the mower is in the grass is taken to be the refracted ray. Sidewalk is taken to be a medium of refractive index n one and that of grass to be taken as n two. In figure (b), the situation is the reverse of what has happened in figure (a). The mower moves from grass to sidewalk and the ray of light moves away from the perpendicular when it speeds up.

Figure 3. The change in direction of a light ray depends on how the speed of light changes when it crosses from one medium to another. The speed of light is greater in medium 1 than in medium 2 in the situations shown here. (a) A ray of light moves closer to the perpendicular when it slows down. This is analogous to what happens when a lawn mower goes from a footpath to grass. (b) A ray of light moves away from the perpendicular when it speeds up. This is analogous to what happens when a lawn mower goes from grass to footpath. The paths are exactly reversible.

The amount that a light ray changes its direction depends both on the incident angle and the amount that the speed changes. For a ray at a given incident angle, a large change in speed causes a large change in direction, and thus a large change in angle. The exact mathematical relationship is the *law of refraction*, or “Snell’s Law,” which is stated in equation form as $n_1 \sin\theta_1 = n_2 \sin\theta_2$.

Here n_1 and n_2 are the indices of refraction for medium 1 and 2, and θ_1 and θ_2 are the angles between the rays and the perpendicular in medium 1 and 2, as shown in Figure 3. The incoming ray is called the incident ray and the outgoing ray the refracted ray, and the associated angles the incident angle and the refracted angle. The law of refraction is also called Snell’s law after the Dutch mathematician Willebrord Snell (1591–1626), who discovered it in 1621. Snell’s experiments showed that the law of refraction was obeyed and that a characteristic index of refraction n could be assigned to a given medium. Snell was not aware that the speed of light varied in different media, but through experiments he was able to determine indices of refraction from the way light rays changed direction.

The Law of Refraction

$$n_1 \sin\theta_1 = n_2 \sin\theta_2$$

Take-Home Experiment: A Broken Pencil

A classic observation of refraction occurs when a pencil is placed in a glass half filled with water. Do this and observe the shape of the pencil when you look at the pencil sideways, that is, through air, glass, water. Explain your observations. Draw ray diagrams for the situation.

Example 2. Determine the Index of Refraction from Refraction Data

Find the index of refraction for medium 2 in Figure 3a, assuming medium 1 is air and given the incident angle is 30.0° and the angle of refraction is 22.0°.

Strategy

The index of refraction for air is taken to be 1 in most cases (and up to four significant figures, it is 1.000). Thus n_1=1.00 here. From the given information, θ_1 = 30.0° and θ_2 = 22.0°. With this information, the only unknown in Snell's law is n_2, so that it can be used to find this unknown.

Solution

Snell's law is $n_1 \sin\theta_1 = n_2 \sin\theta_2$.

Rearranging to isolate n_2 gives

$$n_2 = n_1 \frac{\sin\theta_1}{\sin\theta_2}$$

Entering known values,

$$\begin{aligned} n_2 &= 1.00\frac{\sin 30.0^\circ}{\sin 22.0^\circ} = \frac{0.500}{0.375} \\ &= 1.33 \end{aligned}$$

Discussion

This is the index of refraction for water, and Snell could have determined it by measuring the angles and performing this calculation. He would then have found 1.33 to be the appropriate index of refraction for water in all other situations, such as when a ray passes from water to glass. Today we can verify that the index of refraction is related to the speed of light in a medium by measuring that speed directly.

Example 3. A Larger Change in Direction

Suppose that in a situation like that in Example 2, light goes from air to diamond and that the incident angle is 30.0°. Calculate the angle of refraction θ_2 in the diamond.

Strategy

Again the index of refraction for air is taken to be n_1 = 1.00, and we are given θ_1 = 30.0°. We can look up the index of refraction for diamond in Table 1, finding n_2 = 2.419. The only unknown in Snell's law is θ_2, which we wish to determine.

Solution

Solving Snell's law for sin θ_2 yields

$$\sin\theta_2 = \frac{n_1}{n_2}\sin\theta_1.$$

Entering known values,

$$\sin\theta_2 = \frac{1.00}{2.419}\sin 30.0^\circ = (0.413)\,(0.500) = 0.207$$

.

The angle is thus $\theta_2 = \sin^{-1} 0.207 = 11.9°$.

Discussion

For the same 30° angle of incidence, the angle of refraction in diamond is significantly smaller than in water (11.9° rather than 22°—see the preceding example). This means there is a larger change in direction in diamond. The cause of a large change in direction is a large change in the index of refraction (or speed). In general, the larger the change in speed, the greater the effect on the direction of the ray.

Section Summary

- The changing of a light ray's direction when it passes through variations in matter is called refraction.
  The speed of light in vacuum $c = 2.9972458 \times 10^8$ m/s $\approx 3.00 \times 10^8$ m/s.
  Index of refraction $n = \frac{c}{v}$, where v is the speed of light in the material, c is the speed of light in vacuum, and n is the index of refraction.
  Snell's law, the law of refraction, is stated in equation form as $n_1 \sin \theta_1 = n_2 \sin\theta_2$.

Conceptual Questions

1. Diffusion by reflection from a rough surface is described in this chapter. Light can also be diffused by refraction. Describe how this occurs in a specific situation, such as light interacting with crushed ice.
2. Why is the index of refraction always greater than or equal to 1?
3. Does the fact that the light flash from lightning reaches you before its sound prove that the speed of light is extremely large or simply that it is greater than the speed of sound? Discuss how you could use this effect to get an estimate of the speed of light.
4. Will light change direction toward or away from the perpendicular when it goes from air to water? Water to glass? Glass to air?
5. Explain why an object in water always appears to be

at a depth shallower than it actually is? Why do people sometimes sustain neck and spinal injuries when diving into unfamiliar ponds or waters?

6. Explain why a person's legs appear very short when wading in a pool. Justify your explanation with a ray diagram showing the path of rays from the feet to the eye of an observer who is out of the water.
7. Why is the front surface of a thermometer curved as shown?

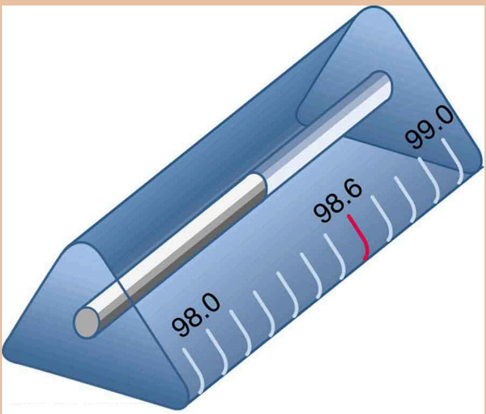


Figure 4. The curved surface of the thermometer serves a purpose.

8. Suppose light were incident from air onto a material that had a negative index of refraction, say −1.3; where does the refracted light ray go?

Problems & Exercises

1. What is the speed of light in water? In glycerine?
2. What is the speed of light in air? In crown glass?
3. Calculate the index of refraction for a medium in which the speed of light is 2.012×10^8 m/s, and identify the most likely substance based on Table 1.
4. In what substance in Table 1 is the speed of light 2.290×10^8 m/s?
5. There was a major collision of an asteroid with the Moon in medieval times. It was described by monks at Canterbury Cathedral in England as a red glow on and around the Moon. How long after the asteroid hit the Moon, which is 3.84×10^5 km away, would the light first arrive on Earth?
6. A scuba diver training in a pool looks at his instructor as shown in Figure 5. What angle does the ray from the instructor's face make with the perpendicular to the water at the point where the ray enters? The angle between the ray in the water and the perpendicular to the water is 25.0º.

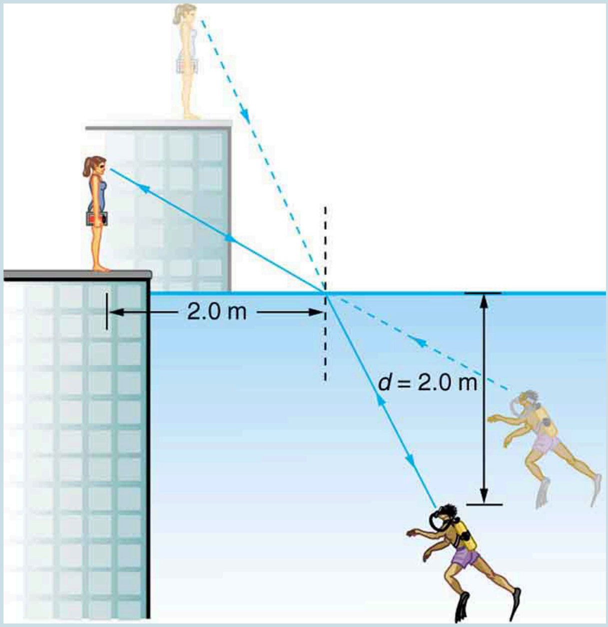


Figure 5. A scuba diver in a pool and his trainer look at each other.

7. Components of some computers communicate with each other through optical fibers having an index of refraction $n = 1.55$. What time in nanoseconds is required for a signal to travel 0.200 m through such a fiber?
8. (a) Using information in Figure 5, find the height of the instructor's head above the water, noting that you will first have to calculate the angle of incidence. (b) Find the apparent depth of the diver's head below water as seen by the instructor.

9. Suppose you have an unknown clear substance immersed in water, and you wish to identify it by finding its index of refraction. You arrange to have a beam of light enter it at an angle of 45.0°, and you observe the angle of refraction to be 40.3°. What is the index of refraction of the substance and its likely identity?
10. On the Moon's surface, lunar astronauts placed a corner reflector, off which a laser beam is periodically reflected. The distance to the Moon is calculated from the round-trip time. What percent correction is needed to account for the delay in time due to the slowing of light in Earth's atmosphere? Assume the distance to the Moon is precisely 3.84×10^8 m, and Earth's atmosphere (which varies in density with altitude) is equivalent to a layer 30.0 km thick with a constant index of refraction $n = 1.000293$.
11. Suppose Figure 6 represents a ray of light going from air through crown glass into water, such as going into a fish tank. Calculate the amount the ray is displaced by the glass (Δx), given that the incident angle is 40.0° and the glass is 1.00 cm thick.

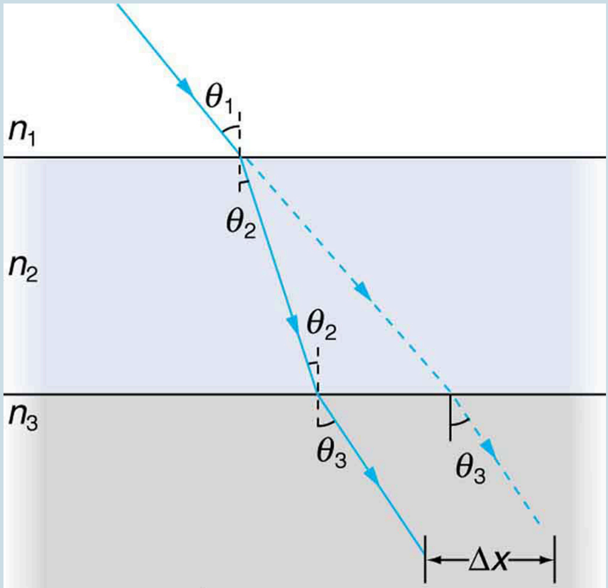


Figure 6. A ray of light passes from one medium to a third by traveling through a second. The final direction is the same as if the second medium were not present, but the ray is displaced by Δx (shown exaggerated).

12. Figure 6 shows a ray of light passing from one medium into a second and then a third. Show that θ_3 is the same as it would be if the second medium were not present (provided total internal reflection does not occur).
13. **Unreasonable Results.** Suppose light travels from water to another substance, with an angle of incidence of 10.0° and an angle of refraction of 14.9°. (a) What is the index of refraction of the other substance? (b) What is unreasonable about this

result? (c) Which assumptions are unreasonable or inconsistent?

14. **Construct Your Own Problem.** Consider sunlight entering the Earth's atmosphere at sunrise and sunset—that is, at a 90° incident angle. Taking the boundary between nearly empty space and the atmosphere to be sudden, calculate the angle of refraction for sunlight. This lengthens the time the Sun appears to be above the horizon, both at sunrise and sunset. Now construct a problem in which you determine the angle of refraction for different models of the atmosphere, such as various layers of varying density. Your instructor may wish to guide you on the level of complexity to consider and on how the index of refraction varies with air density.
15. **Unreasonable Results.** Light traveling from water to a gemstone strikes the surface at an angle of 80.0° and has an angle of refraction of 15.2°. (a) What is the speed of light in the gemstone? (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

Glossary

refraction: changing of a light ray's direction when it passes through variations in matter

index of refraction: for a material, the ratio of the speed of light in vacuum to that in the material

Selected Solutions to Problems & Exercises

1. 2.25×10^8 m/s in water; 2.04×10^8 m/s in glycerine

3. 1.490, polystyrene

5. 1.28 s

7. 1.03 ns

9. $n = 1.46$, fused quartz

13. (a) 0.898; (b) Can't have $n < 1.00$ since this would imply a speed greater than c; (c) Refracted angle is too big relative to the angle of incidence.

15. (a) $\frac{c}{5.00}$; (b) Speed of light too slow, since index is much greater than that of diamond; (c) Angle of refraction is unreasonable relative to the angle of incidence.

73. Total Internal Reflection

Learning Objectives

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

- Explain the phenomenon of total internal reflection.
- Describe the workings and uses of fiber optics.
- Analyze the reason for the sparkle of diamonds.

A good-quality mirror may reflect more than 90% of the light that falls on it, absorbing the rest. But it would be useful to have a mirror that reflects all of the light that falls on it. Interestingly, we can produce *total reflection* using an aspect of *refraction*.

Consider what happens when a ray of light strikes the surface between two materials, such as is shown in Figure 1a. Part of the light crosses the boundary and is refracted; the rest is reflected. If, as shown in the figure, the index of refraction for the second medium is less than for the first, the ray bends away from the perpendicular. (Since $n_1 > n_2$, the angle of refraction is greater than the angle of incidence—that is, $\theta_1 > \theta_2$.) Now imagine what happens as the incident angle is increased. This causes θ_2 to increase also. The largest the angle of refraction θ_2 can be is 90°, as shown in Figure 1b.The *critical angle* θ_c for a combination of materials is defined to be the incident angle θ_1 that produces an angle of refraction of 90°. That is, θ_c is the incident angle for which $\theta_2 = 90°$. If the incident angle θ_1 is greater than the critical angle, as shown in Figure 1c, then all of the light is reflected back into medium 1, a condition called *total internal reflection.*

Critical Angle

The incident angle θ_1 that produces an angle of refraction of 90º is called the critical angle, θ_c.

In the first figure, an incident ray at an angle theta 1 with a perpendicular line drawn at the point of incidence travels from n1 to n2. The incident ray suffers both refraction and reflection. The angle of refraction is theta 2. In the second figure, as theta 1 is increased, the angle of refraction theta 2 becomes 90 degrees and the angle of reflection corresponding to 90 degrees is theta c. In the third figure, theta c greater than theta i, total internal reflection takes place and instead of refraction, reflection takes place and the light ray travels back into medium n1.

Figure 1. (a) A ray of light crosses a boundary where the speed of light increases and the index of refraction decreases. That is, $n_2 < n_1$. The ray bends away from the perpendicular. (b) The critical angle θc is the one for which the angle of refraction is. (c) Total internal reflection occurs when the incident angle is greater than the critical angle.

Snell's law states the relationship between angles and indices of refraction. It is given by

$$n_1 \sin \theta_1 = n_2 \sin \theta_2.$$

When the incident angle equals the critical angle ($\theta_1 = \theta_c$), the angle of refraction is 90º ($\theta_2 = 90º$). Noting that sin 90º = 1, Snell's law in this case becomes

$n_1 \sin \theta_1 = n_2$.

The critical angle θ_c for a given combination of materials is thus

$$\theta_c = \sin^{-1}\left(\frac{n_2}{n_1}\right) \text{ for } n_1 > n_2.$$

Total internal reflection occurs for any incident angle greater than the critical angle θ_c, and it can only occur when the second medium has an index of refraction less than the first. Note the above equation is written for a light ray that travels in medium 1 and reflects from medium 2, as shown in the figure.

Example 1. How Big is the Critical Angle Here?

What is the critical angle for light traveling in a polystyrene (a type of plastic) pipe surrounded by air?

Strategy

The index of refraction for polystyrene is found to be 1.49 in Figure 2, and the index of refraction of air can be taken to be 1.00, as before. Thus, the condition that the second medium (air) has an index of refraction less than the first (plastic) is satisfied, and the equation $\theta_c = \sin^{-1}\left(\frac{n_2}{n_1}\right)$ can be used to find the critical angle θ_c. Here, then, $n_2 = 1.00$ and $n_1 = 1.49$.

Solution

The critical angle is given by

$$\theta_c = \sin^{-1}\left(\frac{n_2}{n_1}\right)$$

Substituting the identified values gives

$$\begin{aligned}\theta_c &= \sin^{-1}\left(\tfrac{1.00}{1.49}\right)\\ &= \sin^{-1}(0.671)\\ &= 42.2^\circ\end{aligned}$$

Discussion

This means that any ray of light inside the plastic that strikes the surface at an angle greater than 42.2° will be totally reflected. This will make the inside surface of the clear plastic a perfect mirror for such rays without any need for the silvering used on common mirrors. Different combinations of materials have different critical angles, but any combination with $n_1 > n_2$ can produce total internal reflection. The same calculation as made here shows that the critical angle for a ray going from water to air is 48.6°, while that from diamond to air is 24.4°, and that from flint glass to crown glass is 66.3°. There is no total reflection for rays going in the other direction—for example, from air to water—since the condition that the second medium must have a smaller index of refraction is not satisfied. A number of interesting applications of total internal reflection follow.

Fiber Optics: Endoscopes to Telephones

Fiber optics is one application of total internal reflection that is in wide use. In communications, it is used to transmit telephone, internet, and cable TV signals. *Fiber optics* employs the transmission of light down fibers of plastic or glass. Because the fibers are thin, light entering one is likely to strike the inside surface at an angle greater than the critical angle and, thus, be totally reflected (See Figure 2.) The index of refraction outside the fiber must be smaller than inside, a condition that is easily satisfied by coating the outside of the fiber with a material having an appropriate refractive index. In fact, most fibers have a varying refractive index to allow more light to be guided along the fiber through total internal refraction. Rays are reflected around corners as shown, making the fibers into tiny light pipes.

Light ray enters an S-shaped tube and undergoes multiple reflections, finally emerging through the other end.

Figure 2. Light entering a thin fiber may strike the inside surface at large or grazing angles and is completely reflected if these angles exceed the critical angle. Such rays continue down the fiber, even following it around corners, since the angles of reflection and incidence remain large.

Bundles of fibers can be used to transmit an image without a lens, as illustrated in Figure 3. The output of a device called an *endoscope* is shown in Figure 3b. Endoscopes are used to explore the body through various orifices or minor incisions. Light is transmitted down one fiber bundle to illuminate internal parts, and the reflected light is transmitted back out through another to be observed. Surgery can be performed, such as arthroscopic surgery on the knee joint, employing cutting tools attached to and observed with the endoscope. Samples can also be obtained, such as by lassoing an intestinal polyp for external examination.

Picture (a) shows how an image A is transmitted through a bundle of parallel fibers. Picture (b) shows an endoscope image.

Figure 3. (a) An image is transmitted by a bundle of fibers that have fixed neighbors. (b) An endoscope is used to probe the body, both transmitting light to the interior and returning an image such as the one shown. (credit: Med_Chaos, Wikimedia Commons)

Fiber optics has revolutionized surgical techniques and observations within the body. There are a host of medical diagnostic and therapeutic uses. The flexibility of the fiber optic bundle allows it to navigate around difficult and small regions in the body, such as the intestines, the heart, blood vessels, and joints. Transmission of an intense laser beam to burn away obstructing plaques in major arteries as well as delivering light to activate chemotherapy drugs are becoming commonplace. Optical fibers have in fact enabled microsurgery and remote surgery where the incisions are small and the surgeon's fingers do not need to touch the diseased tissue.

The image shows a bundle fiber with a medium of refractive index n sub 1 inside surrounded by a medium n sub 2. Medium n sub 2 is made up of cladding material and n sub 1 is the core.

Figure 4. Fibers in bundles are clad by a material that has a lower index of refraction than the core to ensure total internal reflection, even when fibers are in contact with one another. This shows a single fiber with its cladding.

Fibers in bundles are surrounded by a cladding material that has a lower index of refraction than the core. (See Figure 4.) The cladding prevents light from being transmitted between fibers in a bundle. Without cladding, light could pass between fibers in contact, since their indices of refraction are identical. Since no light gets into the cladding (there is total internal reflection back into the core), none can be transmitted between clad fibers that are in contact with one another. The cladding prevents light from escaping out of the fiber; instead most of the light is propagated along the length of the fiber, minimizing the loss of signal and ensuring that a quality image is formed at the other end. The cladding and an additional protective layer make optical fibers flexible and durable.

Cladding

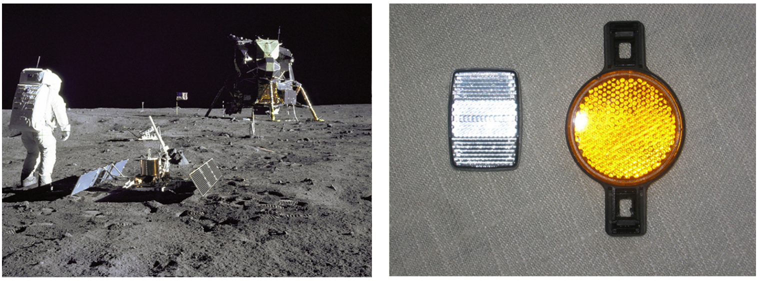
The cladding prevents light from being transmitted between fibers in a bundle.

Special tiny lenses that can be attached to the ends of bundles of fibers are being designed and fabricated. Light emerging from a fiber bundle can be focused and a tiny spot can be imaged. In some cases the spot can be scanned, allowing quality imaging of a region inside the body. Special minute optical filters inserted at the end of the fiber bundle have the capacity to image tens of microns below the surface without cutting the surface—non-intrusive diagnostics. This is particularly useful for determining the extent of cancers in the stomach and bowel.

Most telephone conversations and Internet communications are now carried by laser signals along optical fibers. Extensive optical fiber cables have been placed on the ocean floor and underground to enable optical communications. Optical fiber communication systems offer several advantages over electrical (copper) based systems, particularly for long distances. The fibers can be made so transparent that light can travel many kilometers before it becomes dim enough to require amplification—much superior to copper conductors. This property of optical fibers is called *low loss*. Lasers emit light with characteristics that allow far more conversations in one fiber than are possible with electric signals on a single conductor. This property of optical fibers is called *high bandwidth*. Optical signals in one fiber do not produce undesirable effects in other adjacent fibers. This property of optical fibers is called *reduced crosstalk*. We shall explore the unique characteristics of laser radiation in a later chapter.

Corner Reflectors and Diamonds

A light ray that strikes an object consisting of two mutually perpendicular reflecting surfaces is reflected back exactly parallel to the direction from which it came. This is true whenever the reflecting surfaces are perpendicular, and it is independent of the angle of incidence. Such an object, shown in Figure 5, is called a *corner reflector*, since the light bounces from its inside corner. Many inexpensive reflector buttons on bicycles, cars, and warning signs have corner reflectors designed to return light in the direction from which it originated. It was more expensive for astronauts to place one on the moon. Laser signals can be bounced from that corner reflector to measure the gradually increasing distance to the moon with great precision.



(a) (b)

Figure 5. (a) Astronauts placed a corner reflector on the moon to measure its gradually increasing orbital distance. (credit: NASA) (b) The bright spots on these bicycle safety reflectors are reflections of the flash of the camera that took this picture on a dark night. (credit: Julo, Wikimedia Commons)

Corner reflectors are perfectly efficient when the conditions for total internal reflection are satisfied. With common materials, it is easy to obtain a critical angle that is less than 45°. One use of these perfect mirrors is in binoculars, as shown in Figure 6. Another use is in periscopes found in submarines.

The picture shows binoculars with prisms inside. The light through one of the object lenses enters through the first prism and suffers total internal reflection and then falls on the second prism and gets total internally reflected and emerges out through one of the eyepiece lenses.

Figure 6. These binoculars employ corner reflectors with total internal reflection to get light to the observer's eyes.

The Sparkle of Diamonds

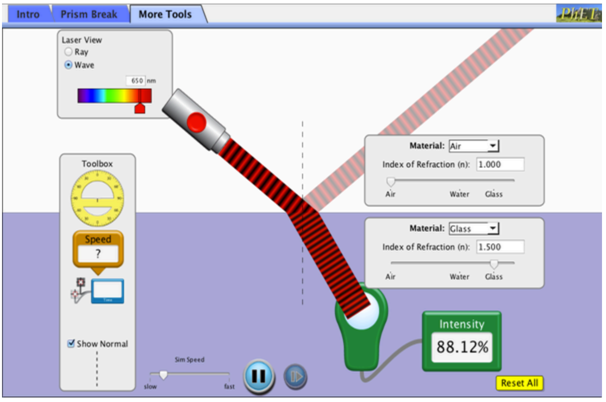
Total internal reflection, coupled with a large index of refraction, explains why diamonds sparkle more than other materials. The critical angle for a diamond-to-air surface is only 24.4°, and so when light enters a diamond, it has trouble getting back out. (See Figure 7.) Although light freely enters the diamond, it can exit only if it makes an angle less than 24.4°. Facets on diamonds are specifically intended to make this unlikely, so that the light can exit only in certain places. Good diamonds are very clear, so that the light makes many internal reflections and is concentrated at the few places it can exit—hence the sparkle. (Zircon is a natural gemstone that has an exceptionally large index of refraction, but not as large as diamond, so it is not as highly prized. Cubic zirconia is manufactured and has an even higher index of refraction (≈2.17), but still less than that of diamond.) The colors you see emerging from a sparkling diamond are not due to the diamond's color, which is usually nearly colorless. Those colors result from dispersion, the topic of Dispersion: The Rainbow and Prisms. Colored diamonds get their color from structural defects of the crystal lattice and the inclusion of minute quantities of graphite and other materials. The Argyle Mine in Western Australia produces around 90% of the world's pink, red, champagne, and cognac diamonds, while around 50% of the world's clear diamonds come from central and southern Africa.

A light ray falls onto one of the faces of a diamond, gets refracted, falls on another face and gets totally internally reflected, and this reflected ray further undergoes multiple reflections when it falls on other faces.

Figure 7. Light cannot easily escape a diamond, because its critical angle with air is so small. Most reflections are total, and the facets are placed so that light can exit only in particular ways—thus concentrating the light and making the diamond sparkle.

PhET Explorations: Bending Light

Explore bending of light between two media with different indices of refraction. See how changing from air to water to glass changes the bending angle. Play with prisms of different shapes and make rainbows.



Click to download the simulation. Run using Java.

Section Summary

- The incident angle that produces an angle of refraction of 90° is called critical angle.
- Total internal reflection is a phenomenon that occurs at the

boundary between two mediums, such that if the incident angle in the first medium is greater than the critical angle, then all the light is reflected back into that medium.

- Fiber optics involves the transmission of light down fibers of plastic or glass, applying the principle of total internal reflection.
- Endoscopes are used to explore the body through various orifices or minor incisions, based on the transmission of light through optical fibers.
- Cladding prevents light from being transmitted between fibers in a bundle.
- Diamonds sparkle due to total internal reflection coupled with a large index of refraction.

Conceptual Questions

1. A ring with a colorless gemstone is dropped into water. The gemstone becomes invisible when submerged. Can it be a diamond? Explain.
2. A high-quality diamond may be quite clear and colorless, transmitting all visible wavelengths with little absorption. Explain how it can sparkle with flashes of brilliant color when illuminated by white light.
3. Is it possible that total internal reflection plays a role in rainbows? Explain in terms of indices of refraction and angles, perhaps referring to Figure 8. Some of us have seen the formation of a double rainbow. Is it physically possible to observe a triple rainbow?

Figure 8. Double rainbows are not a very common observance. (credit: InvictusOU812, Flickr)

4. The most common type of mirage is an illusion that light from faraway objects is reflected by a pool of water that is not really there. Mirages are generally observed in deserts, when there is a hot layer of air near the ground. Given that the refractive index of air is lower for air at higher temperatures, explain how mirages can be formed.

Problems & Exercises

1. Verify that the critical angle for light going from water to air is 48.6°, as discussed at the end of

Example 1, regarding the critical angle for light traveling in a polystyrene (a type of plastic) pipe surrounded by air.

2. (a) At the end of Example 1, it was stated that the critical angle for light going from diamond to air is 24.4°. Verify this. (b) What is the critical angle for light going from zircon to air?
3. An optical fiber uses flint glass clad with crown glass. What is the critical angle?
4. At what minimum angle will you get total internal reflection of light traveling in water and reflected from ice?
5. Suppose you are using total internal reflection to make an efficient corner reflector. If there is air outside and the incident angle is 45.0°, what must be the minimum index of refraction of the material from which the reflector is made?
6. You can determine the index of refraction of a substance by determining its critical angle. (a) What is the index of refraction of a substance that has a critical angle of 68.4° when submerged in water? What is the substance, based on Figure 9? (b) What would the critical angle be for this substance in air?

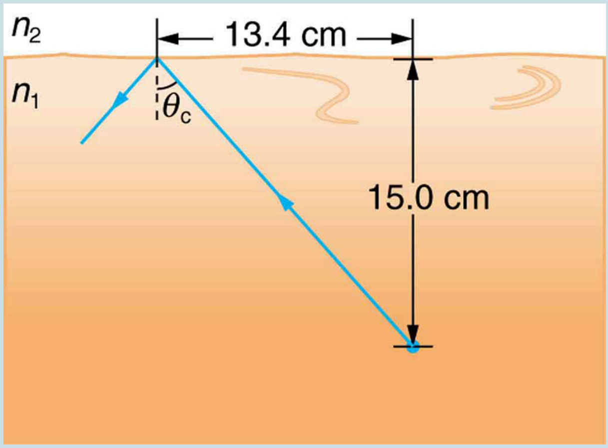


Figure 9. A light ray inside a liquid strikes the surface at the critical angle and undergoes total internal reflection.

7. A ray of light, emitted beneath the surface of an unknown liquid with air above it, undergoes total internal reflection as shown in Figure 9. What is the index of refraction for the liquid and its likely identification?
8. A light ray entering an optical fiber surrounded by air is first refracted and then reflected as shown in Figure 10. Show that if the fiber is made from crown glass, any incident ray will be totally internally reflected.

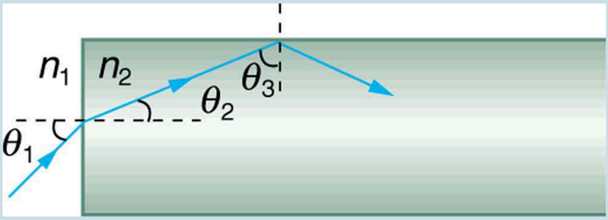


Figure 10. A light ray enters the end of a fiber, the surface of which is perpendicular to its sides. Examine the conditions under which it may be totally internally reflected.

Glossary

critical angle: incident angle that produces an angle of refraction of 90º

fiber optics: transmission of light down fibers of plastic or glass, applying the principle of total internal reflection

corner reflector: an object consisting of two mutually perpendicular reflecting surfaces, so that the light that enters is reflected back exactly parallel to the direction from which it came

zircon: natural gemstone with a large index of refraction

Selected Solutions to Problems & Exercises

3. 66.3º

5. > 1.414

7. 1.50, benzene

74. Dispersion: The Rainbow and Prisms

Learning Objective

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

- Explain the phenomenon of dispersion and discuss its advantages and disadvantages.

Part a of this figure shows the colors produced by a rainbow. Part b shows the colors produced by a prism.

Figure 1. The colors of the rainbow (a) and those produced by a prism (b) are identical. (credit: Alfredo55, Wikimedia Commons; NASA)

Everyone enjoys the spectacle of a rainbow glimmering against a dark stormy sky. How does sunlight falling on clear drops of rain get broken into the rainbow of colors we see? The same process causes white light to be broken into colors by a clear glass prism or a diamond. (See Figure 1.)

We see about six colors in a rainbow—red, orange, yellow, green, blue, and violet; sometimes indigo is listed, too. Those colors are associated with different wavelengths of light, as shown in Figure 2. When our eye receives pure-wavelength light, we tend to see only one of the six colors, depending on wavelength. The thousands of other hues we can sense in other situations are our eye's response to various mixtures of wavelengths. White light, in particular, is a fairly uniform mixture of all visible wavelengths. Sunlight, considered to be white, actually appears to be a bit yellow because of its mixture of wavelengths, but it does contain all visible wavelengths. The sequence of colors in rainbows is the same sequence as the colors plotted versus wavelength in Figure 2. What this implies is that white light is spread out according to wavelength in a rainbow. *Dispersion* is defined as the spreading of white light into its full spectrum of wavelengths. More technically, dispersion occurs whenever there is a process that changes the direction of light in a manner that depends on wavelength. Dispersion, as a general phenomenon, can occur for any type of wave and always involves wavelength-dependent processes.

A continuous distribution of colors with their range of wavelength lambda in nanometers, starting with infrared at 800 nanometers. Following infrared is the visible region with red at 700 nanometers, orange, yellow at 600 nanometers, green, blue at 500 nanometers, and violet at 400 nanometers. The distribution ends with ultraviolet at 300 nanometers.

Figure 2. Even though rainbows are associated with seven colors, the rainbow is a continuous distribution of colors according to wavelengths.

Dispersion

Dispersion is defined to be the spreading of white light into its full spectrum of wavelengths.

Refraction is responsible for dispersion in rainbows and many other situations. The angle of refraction depends on the index of refraction, as we saw in The Law of Refraction. We know that the index of refraction n depends on the medium. But for a given medium, n also depends on wavelength. (See Table 1. Note that, for a given medium, n increases as wavelength decreases and is greatest for violet light. Thus violet light is bent more than red light, as shown for a prism in Figure 3b, and the light is dispersed into the same sequence of wavelengths as seen in Figure 1 and Figure 2.

Table 1. Index of Refraction *n* in Selected Media at Various Wavelengths

Medium	Red (660 nm)	Orange (610 nm)	Yellow (580 nm)	Green (550 nm)	Blue (470 nm)	Violet (410 nm)
Water	1.331	1.332	1.333	1.335	1.338	1.342
Diamond	2.410	2.415	2.417	2.426	2.444	2.458
Glass, crown	1.512	1.514	1.518	1.519	1.524	1.530
Glass, flint	1.662	1.665	1.667	1.674	1.684	1.698
Polystyrene	1.488	1.490	1.492	1.493	1.499	1.506
Quartz, fused	1.455	1.456	1.458	1.459	1.462	1.468

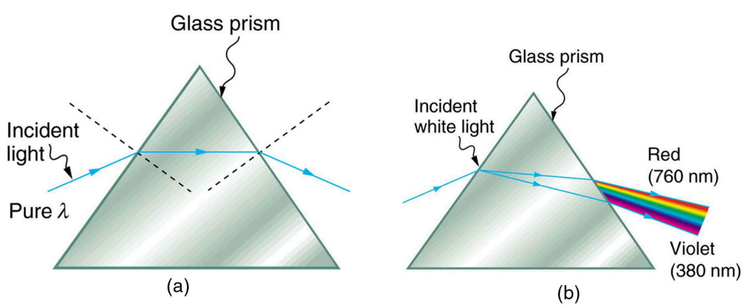


Figure 3. (a) A pure wavelength of light falls onto a prism and is refracted at both surfaces. (b) White light is dispersed by the prism (shown exaggerated). Since the index of refraction varies with wavelength, the angles of refraction vary with wavelength. A sequence of red to violet is produced, because the index of refraction increases steadily with decreasing wavelength.

Making Connections: Dispersion

Any type of wave can exhibit dispersion. Sound waves, all types of electromagnetic waves, and water waves can be dispersed according to wavelength. Dispersion occurs whenever the speed of propagation depends on wavelength, thus separating and spreading out various wavelengths. Dispersion may require special circumstances and can result in spectacular displays such as in the production of a rainbow. This is also true for sound, since all frequencies ordinarily travel at the same speed. If you listen to sound through a long tube, such as a vacuum cleaner hose, you can easily hear it is dispersed by interaction with the tube. Dispersion, in fact, can reveal a great deal about what the wave has encountered that disperses its wavelengths. The dispersion of electromagnetic radiation from outer space, for example, has revealed much about what exists between the stars—the so-called empty space.

Rainbows are produced by a combination of refraction and reflection. You may have noticed that you see a rainbow only when you look away from the sun. Light enters a drop of water and is reflected from the back of the drop, as shown in Figure 4. The light is refracted both as it enters and as it leaves the drop. Since the index of refraction of water varies with wavelength, the light is dispersed, and a rainbow is observed, as shown in Figure 5a. (There is no dispersion caused by reflection at the back surface, since the law of reflection does not depend on wavelength.) The actual rainbow of colors seen by an observer depends on the myriad of rays being refracted and reflected toward the observer's eyes from numerous drops of water. The effect is most spectacular when the background is dark, as in stormy weather, but can also be observed in waterfalls and lawn sprinklers. The arc of a rainbow comes from the need to be looking at a specific angle relative to the direction of the sun, as illustrated in Figure 5b. (If there are two reflections of light within the water drop, another "secondary" rainbow is produced. This rare event produces an arc that lies above the primary rainbow arc—see Figure 5c.)

Sun light incident on a spherical water droplet gets refracted at various angles. The refracted rays further undergo total internal reflection and when they leave the water droplet, a sequence of colors ranging from violet to red is formed.

Figure 4. Part of the light falling on this water drop enters and is reflected from the back of the drop. This light is refracted and dispersed both as it enters and as it leaves the drop.

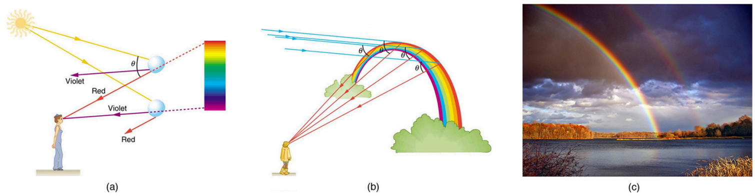


Figure 5. (a) Different colors emerge in different directions, and so you must look at different locations to see the various colors of a rainbow. (b) The arc of a rainbow results from the fact that a line between the observer and any point on the arc must make the correct angle with the parallel rays of sunlight to receive the refracted rays. (c) Double rainbow. (credit: Nicholas, Wikimedia Commons)

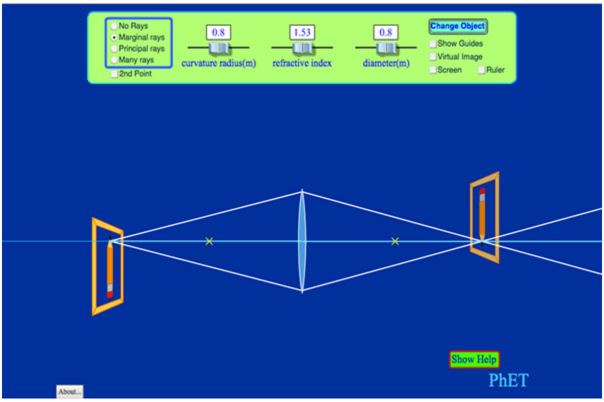
Rainbows

Rainbows are produced by a combination of refraction and reflection.

Dispersion may produce beautiful rainbows, but it can cause problems in optical systems. White light used to transmit messages in a fiber is dispersed, spreading out in time and eventually overlapping with other messages. Since a laser produces a nearly pure wavelength, its light experiences little dispersion, an advantage over white light for transmission of information. In contrast, dispersion of electromagnetic waves coming to us from outer space can be used to determine the amount of matter they pass through. As with many phenomena, dispersion can be useful or a nuisance, depending on the situation and our human goals.

PhET Explorations: Geometric Optics

How does a lens form an image? See how light rays are refracted by a lens. Watch how the image changes when you adjust the focal length of the lens, move the object, move the lens, or move the screen.



Click to run the simulation.

Section Summary

- The spreading of white light into its full spectrum of wavelengths is called dispersion.
- Rainbows are produced by a combination of refraction and

reflection and involve the dispersion of sunlight into a continuous distribution of colors.

- Dispersion produces beautiful rainbows but also causes problems in certain optical systems.

Problems & Exercises

1. (a) What is the ratio of the speed of red light to violet light in diamond, based on [link]? (b) What is this ratio in polystyrene? (c) Which is more dispersive?
2. A beam of white light goes from air into water at an incident angle of 75.0º. At what angles are the red (660 nm) and violet (410 nm) parts of the light refracted?
3. By how much do the critical angles for red (660 nm) and violet (410 nm) light differ in a diamond surrounded by air?
4. (a) A narrow beam of light containing yellow (580 nm) and green (550 nm) wavelengths goes from polystyrene to air, striking the surface at a 30.0º incident angle. What is the angle between the colors when they emerge? (b) How far would they have to travel to be separated by 1.00 mm?
5. A parallel beam of light containing orange (610 nm) and violet (410 nm) wavelengths goes from fused quartz to water, striking the surface between them at a 60.0º incident angle. What is the angle between the two colors in water?
6. A ray of 610 nm light goes from air into fused quartz at an incident angle of 55.0º. At what incident angle

must 470 nm light enter flint glass to have the same angle of refraction?

7. A narrow beam of light containing red (660 nm) and blue (470 nm) wavelengths travels from air through a 1.00 cm thick flat piece of crown glass and back to air again. The beam strikes at a 30.0º incident angle. (a) At what angles do the two colors emerge? (b) By what distance are the red and blue separated when they emerge?
8. A narrow beam of white light enters a prism made of crown glass at a 45.0º incident angle, as shown in Figure 6. At what angles, θ_R and θ_V, do the red (660 nm) and violet (410 nm) components of the light emerge from the prism?

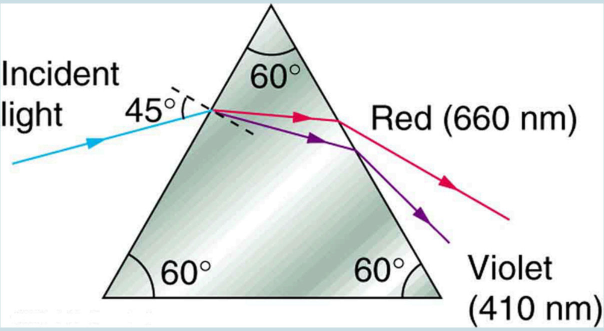


Figure 6. This prism will disperse the white light into a rainbow of colors. The incident angle is 45.0º, and the angles at which the red and violet light emerge are θ_R and θ_V.

Glossary

dispersion: spreading of white light into its full spectrum of wavelengths

rainbow: dispersion of sunlight into a continuous distribution of colors according to wavelength, produced by the refraction and reflection of sunlight by water droplets in the sky

Selected Solutions to Problems & Exercises

2. 46.5°, red; 46.0°, violet

4. (a) 0.043°; (b) 1.33 m

6. 71.3°

8. 53.5°, red; 55.2°, violet

75. Image Formation by Lenses

Learning Objectives

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

- List the rules for ray tracking for thin lenses.
- Illustrate the formation of images using the technique of ray tracking.
- Determine power of a lens given the focal length.

Lenses are found in a huge array of optical instruments, ranging from a simple magnifying glass to the eye to a camera's zoom lens. In this section, we will use the law of refraction to explore the properties of lenses and how they form images.

The word *lens* derives from the Latin word for a lentil bean, the shape of which is similar to the convex lens in Figure 1. The convex lens shown has been shaped so that all light rays that enter it parallel to its axis cross one another at a single point on the opposite side of the lens. (The axis is defined to be a line normal to the lens at its center, as shown in Figure 1.) Such a lens is called a *converging (or convex) lens* for the converging effect it has on light rays. An expanded view of the path of one ray through the lens is shown, to illustrate how the ray changes direction both as it enters and as it leaves the lens. Since the index of refraction of the lens is greater than that of air, the ray moves towards the perpendicular as it enters and away from the perpendicular as it leaves. (This is in accordance with the law of refraction.) Due to the lens's shape, light is thus bent toward the axis at both surfaces. The point at which the rays cross is defined to be the *focal point* F of the lens. The distance from the center of the lens to its focal point is defined to be the *focal length f* of the lens. Figure 2 shows how a converging lens, such as that in a magnifying glass, can converge the nearly parallel light rays from the sun to a small spot.

The Figure on the right shows a convex lens. Three rays heading from left to right, 1, 2, and 3, are considered. Ray 2 falls on the axis and rays 1 and 3 are parallel to the axis. The distance from the center of the lens to the focal point F is small f on the right side of the lens. Rays 1 and 3 after refraction converge at F on the axis. Ray 2 on the axis goes undeviated. The Figure on the left shows an expanded view of refraction for ray 1. The angle of incidence is theta 1 and angle of refraction theta 2 and a dotted line is the perpendicular drawn to the surface of the lens at the point of incidence. The ray after the refraction at the second surface emerges with an angle equal to theta 1 prime with the perpendicular drawn at that point. The perpendiculars are shown as dotted lines.

Figure 1. Rays of light entering a converging lens parallel to its axis converge at its focal point F. (Ray 2 lies on the axis of the lens.) The distance from the center of the lens to the focal point is the lens's focal length f. An expanded view of the path taken by ray 1 shows the perpendiculars and the angles of incidence and refraction at both surfaces.

Converging or Convex Lens

The lens in which light rays that enter it parallel to its axis cross one another at a single point on the opposite side with a converging effect is called converging lens.

Focal Point F

The point at which the light rays cross is called the focal point F of the lens.

Focal Length f

The distance from the center of the lens to its focal point is called focal length f.

A person's hand is holding a magnifying glass to focus the sunlight to a point. The magnifying glass focuses the sunlight to burn paper.

Figure 2. Sunlight focused by a converging magnifying glass can burn paper. Light rays from the sun are nearly parallel and cross at the focal point of the lens. The more powerful the lens, the closer to the lens the rays will cross.

The greater effect a lens has on light rays, the more powerful it is said to be. For example, a powerful converging lens will focus parallel light rays closer to itself and will have a smaller focal length than a weak lens. The light will also focus into a smaller and more intense spot for a more powerful lens. The *power* P of a lens is defined to be the inverse of its focal length. In equation form, this is

$$P = \frac{1}{f}.$$

Power P

The **power** P of a lens is defined to be the inverse of its focal length. In equation form, this is $P = \frac{1}{f}$, where f is the focal length of the lens, which must be given in meters (and not cm or mm). The power of a lens P has the unit diopters (D), provided that the focal length is given in meters. That is, $1\text{D} = \frac{1}{\text{m}}$, or 1m^{-1}. (Note that this power (optical power, actually) is not the same as power in watts defined in the chapter Work, Energy, and Energy Resources. It is a concept related to the effect of optical devices on light.) Optometrists prescribe common spectacles and contact lenses in units of diopters.

Example 1. What is the Power of a Common Magnifying Glass?

Suppose you take a magnifying glass out on a sunny day and you find that it concentrates sunlight to a small spot 8.00 cm away from the lens. What are the focal length and power of the lens?

Strategy

The situation here is the same as those shown in Figure 1 and Figure 2. The Sun is so far away that the Sun's rays are nearly parallel when they reach Earth. The magnifying glass is a convex (or converging) lens, focusing the nearly parallel rays of sunlight. Thus the focal length of the lens is the distance from the lens to the spot, and its power is the inverse of this distance (in m).

Solution

The focal length of the lens is the distance from the center of the lens to the spot, given to be 8.00 cm. Thus,

$$f = 8.00 \text{ cm}.$$

To find the power of the lens, we must first convert the focal length to meters; then, we substitute this value into the equation for power. This gives

$$P = \frac{1}{f} = \frac{1}{0.0800\ \text{m}} = 12.5\ \text{D}.$$

Discussion

This is a relatively powerful lens. The power of a lens in diopters should not be confused with the familiar concept of power in watts. It is an unfortunate fact that the word "power" is used for two completely different concepts. If you examine a prescription for eyeglasses, you will note lens powers given in diopters. If you examine the label on a motor, you will note energy consumption rate given as a power in watts.

Figure 3 shows a concave lens and the effect it has on rays of light that enter it parallel to its axis (the path taken by ray 2 in the Figure is the axis of the lens). The concave lens is a *diverging lens*, because it causes the light rays to bend away (diverge) from its axis. In this case, the lens has been shaped so that all light rays entering it parallel to its axis appear to originate from the same point, F, defined to be the focal point of a diverging lens. The distance from the center of the lens to the focal point is again called the focal length f of the lens. Note that the focal length and power of a diverging lens are defined to be negative.

For example, if the distance to F in Figure 3 is 5.00 cm, then the focal length is $f = -5.00$ cm and the power of the lens is P = −20 D. An expanded view of the path of one ray through the lens is shown in the Figure to illustrate how the shape of the lens, together with the law of refraction, causes the ray to follow its particular path and be diverged.

The Figure on the top shows an expanded view of refraction for ray 1 falling on a concave lens. The angle of incidence is theta 1 and angle of refraction theta 2. The ray after the refraction at the second surface emerges with an angle equal to theta 1 prime with the perpendicular drawn at that point. Perpendiculars are shown as dotted lines. The Figure at the bottom shows a concave lens. Three rays, 1, 2, and 3, are considered. Ray 2 falls on the axis and rays 1 and 3 are parallel to the axis. Rays 1 and 3 after refraction appear to come from a point F on the axis. The distance from the center of the lens to F is small f and is measured from the same side as the incident rays. Ray 2 on the axis goes undeviated.

Figure 3. Rays of light entering a diverging lens parallel to its axis are diverged, and all appear to originate at its focal point F. The dashed lines are not rays—they indicate the directions from which the rays appear to come. The focal length f of a diverging lens is negative. An expanded view of the path taken by ray 1 shows the perpendiculars and the angles of incidence and refraction at both surfaces.

Diverging Lens

A lens that causes the light rays to bend away from its axis is called a diverging lens.

As noted in the initial discussion of the law of refraction in The Law of Refraction, the paths of light rays are exactly reversible. This means that the direction of the arrows could be reversed for all of the rays in Figure 1 and Figure 3. For example, if a point light source is placed at the focal point of a convex lens, as shown in Figure 4, parallel light rays emerge from the other side.

Three light rays coming from a light bulb filament are incident on a convex lens and the rays after refraction are rendered parallel.

Figure 4. A small light source, like a light bulb filament, placed at the focal point of a convex lens, results in parallel rays of light emerging from the other side. The paths are exactly the reverse of those shown in Figure 1. This technique is used in lighthouses and sometimes in traffic lights to produce a directional beam of light from a source that emits light in all directions.

Ray Tracing and Thin Lenses

Ray tracing is the technique of determining or following (tracing) the paths that light rays take. For rays passing through matter, the law of refraction is used to trace the paths. Here we use ray tracing to help us understand the action of lenses in situations ranging from forming images on film to magnifying small print to correcting nearsightedness. While ray tracing for complicated lenses, such as those found in sophisticated cameras, may require computer techniques, there is a set of simple rules for tracing rays through thin lenses.

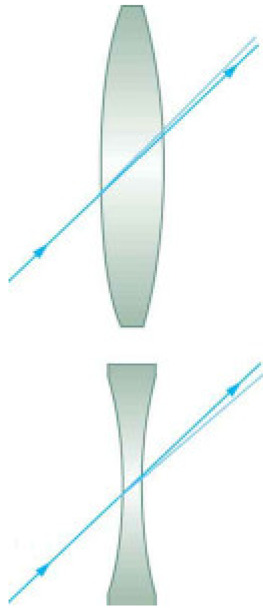


Figure 6. The light ray through the center of a thin lens is deflected by a negligible amount and is assumed to emerge parallel to its original path (shown as a shaded line).

A *thin lens* is defined to be one whose thickness allows rays to refract, as illustrated in Figure 1, but does not allow properties such as dispersion and aberrations. An ideal thin lens has two refracting surfaces but the lens is thin enough to assume that light rays bend only once. A thin symmetrical lens has two focal points, one on either side and both at the same distance from the lens. (See Figure 6.)

Another important characteristic of a thin lens is that light rays through its center are deflected by a negligible amount, as seen in Figure 5.

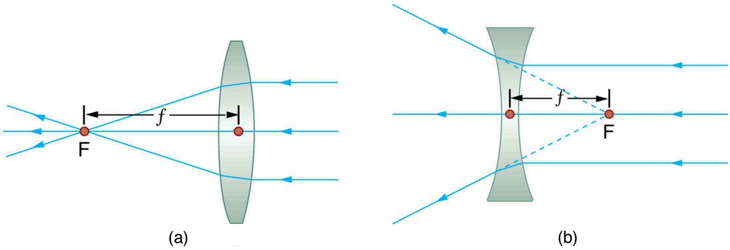


Figure 6. Thin lenses have the same focal length on either side. (a) Parallel light rays entering a converging lens from the right cross at its focal point on the left. (b) Parallel light rays entering a diverging lens from the right seem to come from the focal point on the right.

Thin Lens

A thin lens is defined to be one whose thickness allows rays to refract but does not allow properties such as dispersion and aberrations.

Take-Home Experiment: A Visit to the Optician

Look through your eyeglasses (or those of a friend) backward and forward and comment on whether they act like thin lenses.

Using paper, pencil, and a straight edge, ray tracing can accurately describe the operation of a lens. The rules for ray tracing for thin lenses are based on the illustrations already discussed:

1. A ray entering a converging lens parallel to its axis passes through the focal point F of the lens on the other side. (See rays 1 and 3 in Figure 1.)
2. A ray entering a diverging lens parallel to its axis seems to come from the focal point F. (See rays 1 and 3 in Figure 2.)
3. A ray passing through the center of either a converging or a diverging lens does not change direction. (See Figure 5, and see ray 2 in Figure 1 and Figure 2.)
4. A ray entering a converging lens through its focal point exits parallel to its axis. (The reverse of rays 1 and 3 in Figure 1.)
5. A ray that enters a diverging lens by heading toward the focal point on the opposite side exits parallel to the axis. (The reverse of rays 1 and 3 in Figure 2.)

Rules for Ray Tracing

1. A ray entering a converging lens parallel to its axis passes through the focal point F of the lens on the other side.
2. A ray entering a diverging lens parallel to its axis seems to come from the focal point F.
3. A ray passing through the center of either a converging or a diverging lens does not change direction.
4. A ray entering a converging lens through its

focal point exits parallel to its axis.

5. A ray that enters a diverging lens by heading toward the focal point on the opposite side exits parallel to the axis.

Image Formation by Thin Lenses

In some circumstances, a lens forms an obvious image, such as when a movie projector casts an image onto a screen. In other cases, the image is less obvious. Where, for example, is the image formed by eyeglasses? We will use ray tracing for thin lenses to illustrate how they form images, and we will develop equations to describe the image formation quantitatively.

Consider an object some distance away from a converging lens, as shown in Figure 7. To find the location and size of the image formed, we trace the paths of selected light rays originating from one point on the object, in this case the top of the person's head. The Figure shows three rays from the top of the object that can be traced using the ray tracing rules given above. (Rays leave this point going in many directions, but we concentrate on only a few with paths that are easy to trace.) The first ray is one that enters the lens parallel to its axis and passes through the focal point on the other side (rule 1). The second ray passes through the center of the lens without changing direction (rule 3). The third ray passes through the nearer focal point on its way into the lens and leaves the lens parallel to its axis (rule 4). The three rays cross at the same point on the other side of the lens. The image of the top of the person's head is located at this point. All rays that come from the same point on the top of the person's head are refracted in such a

First of four images shows an incident ray 1 coming from an object (a girl) placed on the axis. After refraction, the ray passes through F on other side of the lens. Second of four images shows an incident ray 2 passing through the center without any deviation. Third of four images shows an incident ray passing through F, which after refraction goes parallel to the axis. Fourth image shows a combination of all three rays, 1, 2, and 3, incident on a convex lens; after refraction, they converge or cross at a point below the axis at some distance from F. Here the height of the object h sub o is the height of the girl above the axis and h sub i is the height of the image below the axis. The distance from the center to point F is small f. The distance from the center to the girl is d sub o and that to the image is d sub i.

Figure 7. Ray tracing is used to locate the image formed by a lens. Rays originating from the same point on the object are traced—the three chosen rays each follow one of the rules for ray tracing, so that their paths are easy to determine. The image is located at the point where the rays cross. In this case, a real image—one that can be projected on a screen—is formed.

way as to cross at the point shown. Rays from another point on the object, such as her belt buckle, will also cross at another common point, forming a complete image, as shown. Although three rays are traced in Figure 7, only two are necessary to locate the image. It is best to trace rays for which there are simple ray tracing rules. Before applying ray tracing to other situations, let us consider the example shown in Figure 7 in more detail.

The image formed in Figure 7 is a *real image*, meaning that it can be projected. That is, light rays from one point on the object actually cross at the location of the image and can be projected onto a screen, a piece of film, or the retina of an eye, for example. Figure 8 shows how such an image would be projected onto film by a camera lens. This Figure also shows how a real image is projected onto the retina by the lens of an eye. Note that the image is there whether it is projected onto a screen or not.

Real Image

The image in which light rays from one point on the object actually cross at the location of the image and can be projected onto a screen, a piece of film, or the retina of an eye is called a real image.

Several important distances appear in Figure 7. We define d_o to be the object distance, the distance of an object from the center of a lens. *Image distance* d_i is defined to be the distance of the image from the center of a lens. The height of the object and height of the image are given the symbols h_o and h_i, respectively. Images that appear upright relative to the object have heights that are positive and those that are inverted have negative heights. Using the rules of ray tracing and making a scale drawing with paper and pencil, like that in Figure 7, we can accurately describe the location and size of an image. But the real benefit of ray tracing is in visualizing how images are formed in a variety of situations. To obtain numerical information, we use a pair of equations that can be derived from a geometric analysis of ray tracing for thin lenses. The *thin lens equations* are

Figure (a) shows incident rays coming from an object (a girl) and falling on a convex lens in a camera. The rays after refraction produce an inverted, real, and diminished image on the film of the camera. Figure (b) shows the same object in front of a human eye. The rays from the object fall on the convex lens and on refraction produce a real, inverted, and diminished image on the retina of the eyeball.

Figure 8. Real images can be projected. (a) A real image of the person is projected onto film. (b) The converging nature of the multiple surfaces that make up the eye result in the projection of a real image on the retina.

$$\frac{1}{d_o}+\frac{1}{d_i}=\frac{1}{f}\text{ and }\frac{h_i}{h_o}=\frac{d_i}{d_o}=m.$$

We define the ratio of image height to object height $\left(\frac{h_i}{h_o}\right)$ to be the *magnification m*. (The minus sign in the equation above will be discussed shortly.) The thin lens equations are broadly applicable to all situations involving thin lenses (and "thin" mirrors, as we will see later). We will explore many features of image formation in the following worked examples.

Image Distance

The distance of the image from the center of the lens is called image distance.

Thin Lens Equations and Magnification

$$\frac{1}{d_{\mathrm{o}}}+\frac{1}{d_{\mathrm{i}}}=\frac{1}{f}$$

$$\frac{h_{\mathrm{i}}}{h_{\mathrm{o}}}=\frac{d_{\mathrm{i}}}{d_{\mathrm{o}}}=m$$

Example 2. Finding the Image of a Light Bulb Filament by Ray Tracing and by the Thin Lens Equations

A clear glass light bulb is placed 0.750 m from a convex lens having a 0.500 m focal length, as shown in Figure 9.

Use ray tracing to get an approximate location for the image. Then use the thin lens equations to calculate both the location of the image and its magnification. Verify that ray tracing and the thin lens equations produce consistent results.

A light bulb at d sub o equals 0.75 m is placed in front of a convex lens of f equals 0.50 meter. The convex lens produces a real, inverted, and enlarged image on a screen at d sub I equals 1.50 meters.

Figure 9. A light bulb placed 0.750 m from a lens having a 0.500 m focal length produces a real image on a poster board as discussed in the example above. Ray tracing predicts the image location and size.

Strategy and Concept

Since the object is placed farther away from a converging lens than the focal length of the lens, this situation is analogous to those illustrated in Figure 7 and Figure 8. Ray tracing to scale should produce similar results for d_i. Numerical solutions for d_i and m can be obtained using the thin lens equations, noting that d_o = 0.750 m and f = 0.500 m.

Solutions (Ray Tracing)

The ray tracing to scale in Figure 9 shows two rays from a point on the bulb's filament crossing about 1.50 m on the far side of the lens. Thus the image distance d_i is about 1.50 m. Similarly, the image height based on ray tracing is greater than the object height by about a factor of 2, and the image is inverted. Thus m is about −2. The minus sign indicates that the image is inverted.

The thin lens equations can be used to find d_i from the given information:

$$\frac{1}{d_o} + \frac{1}{d_i} = \frac{1}{f}.$$

Rearranging to isolate d_i gives

$$\frac{1}{d_i} = \frac{1}{f} - \frac{1}{d_o}.$$

Entering known quantities gives a value for $\frac{1}{d_i}$:

$$\frac{1}{d_i} = \frac{1}{0.500\text{ m}} - \frac{1}{0.750\text{ m}} = \frac{0.667}{\text{m}}.$$

This must be inverted to find d_i:

$$d_i = \frac{\text{m}}{0.667} = 1.50\text{ m}.$$

Note that another way to find d_i is to rearrange the equation:

$$\frac{1}{d_i} = \frac{1}{f} - \frac{1}{d_o}.$$

This yields the equation for the image distance as:

$$d_i = \frac{f d_o}{d_o - f}$$

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

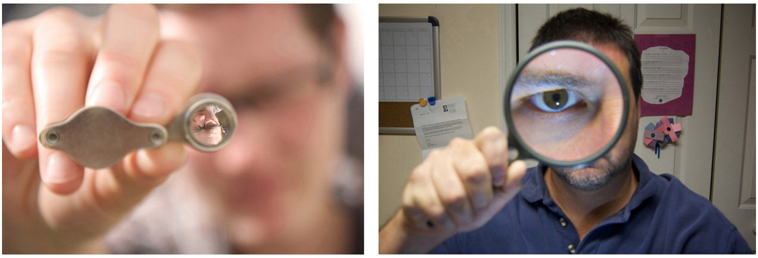
$$m = -\frac{d_\text{i}}{d_\text{o}} = -\frac{1.50\text{ m}}{0.750\text{ m}} = -2.00$$

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 $d_\text{o} > f$ and f is positive, as in Figure 10a. (A summary of the three cases or types of image formation appears at the end of this section.)

A different type of image is formed when an object, such as a person's face, is held close to a convex lens. The image is upright and larger than the object, as seen in Figure 10b, and so the lens is called a magnifier. If you slowly pull the magnifier away from the face, you will see that the magnification steadily increases until the image begins to blur. Pulling the magnifier even farther away produces an inverted image as seen in Figure 10a. The distance at which the image blurs, and beyond which it inverts, is the focal length of the lens. To use a convex lens as a magnifier, the object must be closer to the converging lens than its focal length. This is called a *case 2* image. A case 2 image is formed when $d_o < f$ and f is positive.



(a) (b)

Figure 10. (a) When a converging lens is held farther away from the face than the lens's focal length, an inverted image is formed. This is a case 1 image. Note that the image is in focus but the face is not, because the image is much closer to the camera taking this photograph than the face. (credit: DaMongMan, Flickr) (b) A magnified image of a face is produced by placing it closer to the converging lens than its focal length. This is a case 2 image. (credit: Casey Fleser, Flickr)

Figure 11. Ray tracing predicts the image location and size for an object held closer to a converging lens than its focal length. Ray 1 enters parallel to the axis and exits through the focal point on the opposite side, while ray 2 passes through the center of the lens without changing path. The two rays continue to diverge on the other side of the lens, but both appear to come from a common point, locating the upright, magnified, virtual image. This is a case 2 image.

Figure 11 uses ray tracing to show how an image is formed when an object is held closer to a converging lens than its focal length. Rays coming from a common point on the object continue to diverge after passing through the lens, but all appear to originate from a point at the location of the image. The image is on the same side of the lens as the object and is farther away from the lens than the object. This image, like all case 2 images, cannot be projected and, hence, is called a *virtual image*.

Light rays only appear to originate at a virtual image; they do not actually pass through that location in space. A screen placed at the location of a virtual image will receive only diffuse light from the object, not focused rays from the lens. Additionally, a screen placed on the opposite side of the lens will receive rays that are still diverging, and so no image will be projected on it. We can see the magnified image with our eyes, because the lens of the eye converges the rays into a real image projected on our retina. Finally, we note that a virtual image is upright and larger than the object, meaning that the magnification is positive and greater than 1.

Virtual Image

An image that is on the same side of the lens as the object and cannot be projected on a screen is called a virtual image.

Example 3. Image Produced by a Magnifying Glass

Suppose the book page in Figure 11a is held 7.50 cm from a convex lens of focal length 10.0 cm, such as a typical magnifying glass might have. What magnification is produced?

Strategy and Concept

We are given that d_o = 7.50 cm and f = 10.0 cm, so we have a situation where the object is placed closer to the lens than its focal length. We therefore expect to get a case 2 virtual image with a positive magnification that is greater than 1. Ray tracing produces an image like that shown in Figure 11, but we will use the thin lens equations to get numerical solutions in this example.

Solution

To find the magnification m, we try to use magnification equation, $m = -\frac{d_i}{d_o}$. We do not have a value for d_i, so that we must first find the location of the image using lens equation. (The procedure is the same as followed in the preceding example, where d_o and f were known.) Rearranging the magnification equation to isolate d_i gives

$$\frac{1}{d_i} = \frac{1}{f} - \frac{1}{d_o}.$$

Entering known values, we obtain a value for $\frac{1}{d_i}$:

$$\frac{1}{d_i} = \frac{1}{10.0\text{ cm}} - \frac{1}{7.50\text{ cm}} = \frac{-0.0333}{\text{cm}}.$$

This must be inverted to find d_{i}:

$$d_{\text{i}} = -\frac{\text{cm}}{0.0333} = -30.0\text{ cm}.$$

Now the thin lens equation can be used to find the magnification m, since both d_{i} and d_{o} are known. Entering their values gives

$$m = -\frac{d_{\text{i}}}{d_{\text{o}}} = -\frac{-30.0\text{ cm}}{10.0\text{ cm}} = 3.00$$

Discussion

A number of results in this example are true of all case 2 images, as well as being consistent with Figure 11. Magnification is indeed positive (as predicted), meaning the image is upright. The magnification is also greater than 1, meaning that the image is larger than the object—in this case, by a factor of 3. Note that the image distance is negative. This means the image is on the same side of the lens as the object. Thus the image cannot be projected and is virtual. (Negative values of d_{i} occur for virtual images.) The image is farther from the lens than the object, since the image distance is greater in magnitude than the object distance. The location of the image is not obvious when you look through a magnifier. In fact, since the image is bigger than the object, you may think the image is closer than the object. But the image is farther away, a fact that is useful in correcting farsightedness, as we shall see in a later section.

A third type of image is formed by a diverging or concave lens. Try looking through eyeglasses meant to correct nearsightedness. (See Figure 12.) You will see an image that is upright but smaller than the object. This means that the magnification is positive but less than 1. The ray diagram in Figure 13 shows that the image is on the same side of the lens as the object and, hence, cannot be projected—it is a virtual image. Note that the image is closer to the lens than the object. This is a *case* 3 image, formed for any object by a negative focal length or diverging lens.

A car when viewed through a concave lens looks upright.

Figure 12. A car viewed through a concave or diverging lens looks upright. This is a case 3 image. (credit: Daniel Oines, Flickr)

Figure (a) shows an upright object placed at d sub o equals seven point five cm and in front of a concave lens of on its left side. Parallel ray 1 falls on the lens and gets refracted and dotted backwards to pass through point F on the left side. Figure (b) shows ray 2 going straight through the center of the lens. Figure (c) combines both figures (a) and (b) and the dotted line and the solid line meet at a point on the left side of the lens forming a virtual image which is erect and diminished. Here h sub o is the height of the object above the axis and h sub i is the height of the image above the axis. The distance from the center to the image is d sub i equals 4.29 cm.

Figure 13. Ray tracing predicts the image location and size for a concave or diverging lens. Ray 1 enters parallel to the axis and is bent so that it appears to originate from the focal point. Ray 2 passes through the center of the lens without changing path. The two rays appear to come from a common point, locating the upright image. This is a case 3 image, which is closer to the lens than the object and smaller in height.

Example 4. Image Produced by a Concave Lens

Suppose an object such as a book page is held 7.50 cm from a concave lens of focal length –10.0 cm. Such a lens could be used in eyeglasses to correct pronounced nearsightedness. What magnification is produced?

Strategy and Concept

This example is identical to the preceding one, except that the focal length is negative for a concave or diverging lens. The method of solution is thus the same, but the results are different in important ways.

Solution

To find the magnification *m*, we must first find the image distance d_i using thin lens equation $\frac{1}{d_\mathrm{i}} = \frac{1}{f} - \frac{1}{d_\mathrm{o}}$, or its alternative rearrangement $d_\mathrm{i} = \frac{f d_\mathrm{o}}{d_\mathrm{o} - f}$.

We are given that f = –10.0 cm and d_o = 7.50 cm. Entering these yields a value for $\frac{1}{d_\mathrm{i}}$:

$$\frac{1}{d_\mathrm{i}} = \frac{1}{-10.0\text{ cm}} - \frac{1}{7.50\text{ cm}} = \frac{-0.2333}{\text{cm}}$$

This must be inverted to find d_i:

$$d_{\mathrm{i}} = -\frac{\mathrm{cm}}{0.2333} = -4.29\ \mathrm{cm}$$

Or

$$d_{\mathrm{i}} = \frac{(7.5)(-10)}{(7.5-(-10))} = -\frac{75}{17.5} = -4.29\ \mathrm{cm}$$

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

$$m = -\frac{d_{\mathrm{i}}}{d_{\mathrm{o}}} = -\frac{-4.29\ \mathrm{cm}}{7.50\ \mathrm{cm}} = 0.571$$

Discussion

A number of results in this example are true of all case 3 images, as well as being consistent with Figure 13. Magnification is positive (as predicted), meaning the image is upright. The magnification is also less than 1, meaning the image is smaller than the object—in this case, a little over half its size. The image distance is negative, meaning the image is on the same side of the lens as the object. (The image is virtual.) The image is closer to the lens than the object, since the image distance is smaller in magnitude than the object distance. The location of the image is not obvious when you look through a concave lens. In fact, since the image is smaller than the object, you may think it is farther away. But the image is closer than the object, a fact that is useful in correcting nearsightedness, as we shall see in a later section.

Table 1 summarizes the three types of images formed by single thin lenses. These are referred to as case 1, 2, and 3 images. Convex (converging) lenses can form either real or virtual images (cases 1 and 2, respectively), whereas concave (diverging) lenses can form only virtual images (always case 3). Real images are always inverted, but they can be either larger or smaller than the object. For example, a slide projector forms an image larger than the slide, whereas a camera makes an image smaller than the object being photographed. Virtual images are always upright and cannot be projected. Virtual images are larger than the object only in case 2, where a convex lens is used. The virtual image produced by a concave lens is always smaller than the object—a case 3 image. We can see and photograph virtual images only by using an additional lens to form a real image.

Table 1. Three Types of Images Formed By Thin Lenses

Type	Formed when	Image type	d_i	m
Case 1	f positive, $d_o > f$	real	positive	negative
Case 2	f positive, $d_o < f$	virtual	negative	positive $m > 1$
Case 3	f negative	virtual	negative	positive $m < 1$

In Image Formation by Mirrors, we shall see that mirrors can form exactly the same types of images as lenses.

Take-Home Experiment: Concentrating Sunlight

Find several lenses and determine whether they are converging or diverging. In general those that are thicker near the edges are diverging and those that are thicker near the center are converging. On a bright sunny day take the converging lenses outside and try focusing the sunlight onto a piece of paper. Determine the focal lengths of the lenses. Be careful because the paper may start to burn, depending on the type of lens you have selected.

Problem-Solving Strategies for Lenses

Step 1. Examine the situation to determine that image formation by a lens is involved.

Step 2. Determine whether ray tracing, the thin lens equations, or both are to be employed. A sketch is very useful even if ray tracing is not specifically required by the problem. Write symbols and values on the sketch.

Step 3. Identify exactly what needs to be determined in the problem (identify the unknowns).

Step 4. Make alist of what is given or can be inferred from the problem as stated (identify the knowns). It is helpful to determine whether the situation involves a case 1, 2, or 3 image. While these are just names for types of images, they have certain characteristics (given in Table 1) that can be of great use in solving problems.

Step 5. If ray tracing is required, use the ray tracing rules listed near the beginning of this section.

Step 6. Most quantitative problems require the use of the thin lens equations. These are solved in the usual manner by substituting knowns and solving for unknowns. Several worked examples serve as guides.

Step 7. Check to see if the answer is reasonable: Does it make sense? If you have identified the type of image (case 1, 2, or 3), you should assess whether your answer is consistent with the type of image, magnification, and so on.

Misconception Alert

We do not realize that light rays are coming from every part of the object, passing through every part of the lens, and all can be used to form the final image.

We generally feel the entire lens, or mirror, is needed to form an image. Actually, half a lens will form the same, though a fainter, image.

Section Summary

- Light rays entering a converging lens parallel to its axis cross one another at a single point on the opposite side.
- For a converging lens, the focal point is the point at which converging light rays cross; for a diverging lens, the focal point is the point from which diverging light rays appear to originate.
- The distance from the center of the lens to its focal point is called the focal length f.

  Power P of a lens is defined to be the inverse of its focal length, $P = \frac{1}{f}$.

  A lens that causes the light rays to bend away from its axis is

called a diverging lens.
Ray tracing is the technique of graphically determining the paths that light rays take.
The image in which light rays from one point on the object actually cross at the location of the image and can be projected onto a screen, a piece of film, or the retina of an eye is called a real image.

- Thin lens equations are

$$\frac{1}{d_o}+\frac{1}{d_i}=\frac{1}{f} \text{ and } \frac{h_i}{h_o}=-\frac{d_i}{d_o}=m \text{ (magnification).}$$

- The distance of the image from the center of the lens is called image distance.
An image that is on the same side of the lens as the object and cannot be projected on a screen is called a virtual image.

Conceptual Questions

1. It can be argued that a flat piece of glass, such as in a window, is like a lens with an infinite focal length. If so, where does it form an image? That is, how are d_i and d_o related?
2. You can often see a reflection when looking at a sheet of glass, particularly if it is darker on the other side. Explain why you can often see a double image in such circumstances.
3. When you focus a camera, you adjust the distance of the lens from the film. If the camera lens acts like a thin lens, why can it not be a fixed distance from the film for both near and distant objects?

4. A thin lens has two focal points, one on either side, at equal distances from its center, and should behave the same for light entering from either side. Look through your eyeglasses (or those of a friend) backward and forward and comment on whether they are thin lenses.
5. Will the focal length of a lens change when it is submerged in water? Explain.

Problems & Exercises

1. What is the power in diopters of a camera lens that has a 50.0 mm focal length?
2. Your camera's zoom lens has an adjustable focal length ranging from 80.0 to 200 mm. What is its range of powers?
3. What is the focal length of 1.75 D reading glasses found on the rack in a pharmacy?
4. You note that your prescription for new eyeglasses is −4.50 D. What will their focal length be?
5. How far from the lens must the film in a camera be, if the lens has a 35.0 mm focal length and is being used to photograph a flower 75.0 cm away? Explicitly show how you follow the steps in the Problem-Solving Strategy for lenses.
6. A certain slide projector has a 100 mm focal length lens. (a) How far away is the screen, if a slide is placed 103 mm from the lens and produces a sharp image?

(b) If the slide is 24.0 by 36.0 mm, what are the dimensions of the image? Explicitly show how you follow the steps in the *Problem-Solving Strategy for Lenses* (above).

7. A doctor examines a mole with a 15.0 cm focal length magnifying glass held 13.5 cm from the mole (a) Where is the image? (b) What is its magnification? (c) How big is the image of a 5.00 mm diameter mole?
8. How far from a piece of paper must you hold your father's 2.25 D reading glasses to try to burn a hole in the paper with sunlight?
9. A camera with a 50.0 mm focal length lens is being used to photograph a person standing 3.00 m away. (a) How far from the lens must the film be? (b) If the film is 36.0 mm high, what fraction of a 1.75 m tall person will fit on it? (c) Discuss how reasonable this seems, based on your experience in taking or posing for photographs.
10. A camera lens used for taking close-up photographs has a focal length of 22.0 mm. The farthest it can be placed from the film is 33.0 mm. (a) What is the closest object that can be photographed? (b) What is the magnification of this closest object?
11. Suppose your 50.0 mm focal length camera lens is 51.0 mm away from the film in the camera. (a) How far away is an object that is in focus? (b) What is the height of the object if its image is 2.00 cm high?
12. (a) What is the focal length of a magnifying glass that produces a magnification of 3.00 when held 5.00 cm from an object, such as a rare coin? (b) Calculate the power of the magnifier in diopters. (c) Discuss how this power compares to those for store-bought

reading glasses (typically 1.0 to 4.0 D). Is the magnifier's power greater, and should it be?

13. What magnification will be produced by a lens of power −4.00 D (such as might be used to correct myopia) if an object is held 25.0 cm away?
14. In Example 3, the magnification of a book held 7.50 cm from a 10.0 cm focal length lens was found to be 3.00. (a) Find the magnification for the book when it is held 8.50 cm from the magnifier. (b) Do the same for when it is held 9.50 cm from the magnifier. (c) Comment on the trend in m as the object distance increases as in these two calculations.
15. Suppose a 200 mm focal length telephoto lens is being used to photograph mountains 10.0 km away. (a) Where is the image? (b) What is the height of the image of a 1000 m high cliff on one of the mountains?
16. A camera with a 100 mm focal length lens is used to photograph the sun and moon. What is the height of the image of the sun on the film, given the sun is 1.40 $\times 10^6$ km in diameter and is 1.50 $\times 10^8$ km away?
17. Combine thin lens equations to show that the magnification for a thin lens is determined by its focal length and the object distance and is given by $$m = \frac{f}{(f - d_\text{o})}.$$

Glossary

converging lens: a convex lens in which light rays that enter it parallel to its axis converge at a single point on the opposite side

diverging lens: a concave lens in which light rays that enter it parallel to its axis bend away (diverge) from its axis

focal point: for a converging lens or mirror, the point at which converging light rays cross; for a diverging lens or mirror, the point from which diverging light rays appear to originate

focal length: distance from the center of a lens or curved mirror to its focal point

magnification: ratio of image height to object height

power: inverse of focal length

real image: image that can be projected

virtual image: image that cannot be projected

Selected Solutions to Problems & Exercises

2. 5.00 to 12.5 D

4. −0.222 m

6. (a) 3.43 m; (b) 0.800 by 1.20 m

7. (a) −1.35 m (on the object side of the lens); (b) +10.0; (c) 5.00 cm

8. 44.4 cm

10. (a) 6.60 cm; (b) −0.333

12. (a) +7.50 cm; (b) 13.3 D; (c) Much greater

14. (a) +6.67; (b) +20.0; (c) The magnification increases without limit (to infinity) as the object distance increases to the limit of the focal distance.

16. −0.933 mm

76. Image Formation by Mirrors

Learning Objectives

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

- Illustrate image formation in a flat mirror.
- Explain with ray diagrams the formation of an image using spherical mirrors.
- Determine focal length and magnification given radius of curvature, distance of object and image.

We only have to look as far as the nearest bathroom to find an example of an image formed by a mirror. Images in flat mirrors are the same size as the object and are located behind the mirror. Like lenses, mirrors can form a variety of images. For example, dental mirrors may produce a magnified image, just as makeup mirrors do. Security mirrors in shops, on the other hand, form images that are smaller than the object. We will use the law of reflection to understand how mirrors form images, and we will find that mirror images are analogous to those formed by lenses.

Figure 1 helps illustrate how a flat mirror forms an image. Two rays are shown emerging from the same point, striking the mirror, and being reflected into the observer's eye. The rays can diverge slightly, and both still get into the eye. If the rays are extrapolated backward, they seem to originate from a common point behind the mirror, locating the image. (The paths of the reflected rays into the eye are the same as if they had come directly from that point behind the mirror.) Using the law of reflection—the angle of reflection equals the angle of incidence—we can see that the image and object are the same distance from the mirror. This is a virtual image, since it cannot be projected—the rays only appear to originate from a common point behind the mirror. Obviously, if you walk behind the mirror, you cannot see the image, since the rays do not go there. But in front of the mirror, the rays behave exactly as if they had come from behind the mirror, so that is where the image is situated.

A bottle at a distance d sub o from a flat mirror. An observer's eye looks into the mirror and finds the image at d sub I behind the mirror. The incident rays fall onto the mirror and get reflected to the eye. The dotted lines represent reflected rays extrapolated backward and produce an image of the same size.

Figure 1. Two sets of rays from common points on an object are reflected by a flat mirror into the eye of an observer. The reflected rays seem to originate from behind the mirror, locating the virtual image.

Now let us consider the focal length of a mirror—for example, the concave spherical mirrors in Figure 2. Rays of light that strike the surface follow the law of reflection. For a mirror that is large compared with its radius of curvature, as in Figure 2a, we see that the reflected rays do not cross at the same point, and the mirror does not have a well-defined focal point. If the mirror had the shape of a parabola, the rays would all cross at a single point, and the mirror would have a well-defined focal point. But parabolic mirrors are much more expensive to make than spherical mirrors. The solution is to use a mirror that is small compared with its radius of curvature, as shown in Figure 2b. (This is the mirror equivalent of the thin lens approximation.) To a very good approximation, this mirror has a well-defined focal point at F that is the focal distance f from the center of the mirror. The focal length f of a concave mirror is positive, since it is a converging mirror.

Figure (a) shows a large concave spherical mirror. A beam of parallel rays is incident on the mirror; after reflection it converges at F. Figure (b) shows a concave mirror that is small when compared to its radius of curvature. A beam of parallel rays is incident on the mirror; after reflection it converges at F on the same side. The middle rays of the parallel beam are 1,2, and 3. The distance of F on ray 2 from the center of the mirror is its focal length small f.

Figure 2. (a) Parallel rays reflected from a large spherical mirror do not all cross at a common point. (b) If a spherical mirror is small compared with its radius of curvature, parallel rays are focused to a common point. The distance of the focal point from the center of the mirror is its focal length f. Since this mirror is converging, it has a positive focal length.

Just as for lenses, the shorter the focal length, the more powerful the mirror; thus, $P = \frac{1}{f}$ for a mirror, too. A more strongly curved mirror has a shorter focal length and a greater power. Using the law of reflection and some simple trigonometry, it can be shown that the focal length is half the radius of curvature, or $f = \frac{R}{2}$, where R is the radius of curvature of a spherical mirror. The smaller the radius of curvature, the smaller the focal length and, thus, the more powerful the mirror

The convex mirror shown in Figure 3 also has a focal point. Parallel rays of light reflected from the mirror seem to originate from the point F at the focal distance *f* behind the mirror. The focal length and power of a convex mirror are negative, since it is a diverging mirror.

Ray tracing is as useful for mirrors as for lenses. The rules for ray tracing for mirrors are based on the illustrations just discussed:

1. A ray approaching a concave converging mirror parallel to its axis is reflected through the focal point F of the mirror on the same side. (See rays 1 and 3 in Figure 2b.)
2. A ray approaching a convex diverging mirror parallel to its axis is reflected so that it seems to come from the focal point F behind the mirror. (See rays 1 and 3 in Figure 3.)

A convex spherical mirror. A beam of parallel rays incident on the mirror, after reflection, appear to come from F on ray 2 behind the mirror. Here the distance of

Figure 3. Parallel rays of light reflected from a convex spherical mirror (small in size compared with its radius of curvature) seem to originate from a well-defined focal point at the focal distance f behind the mirror. Convex mirrors diverge light rays and, thus, have a negative focal length.

3. Any ray striking the center of a mirror is followed by applying the law of reflection; it makes the same angle with the axis when leaving as when approaching. (See ray 2 in Figure 4.)
4. A ray approaching a concave converging mirror through its focal point is reflected parallel to its axis. (The reverse of rays 1 and 3 in Figure 2.)
5. A ray approaching a convex diverging mirror by heading toward its focal point on the opposite side is reflected parallel to the axis. (The reverse of rays 1 and 3 in Figure 3.)

We will use ray tracing to illustrate how images are formed by mirrors, and we can use ray tracing quantitatively to obtain numerical information. But since we assume each mirror is small compared with its radius of curvature, we can use the thin lens equations for mirrors just as we did for lenses.

Consider the situation shown in Figure 4, concave spherical mirror reflection, in which an object is placed farther from a concave (converging) mirror than its focal length. That is, f is positive and $d_o > f$, so that we may expect an image similar to the case 1 real image formed by a converging lens. Ray tracing in Figure 4 shows that the rays from a common point on the object all cross at a point on the same side of the mirror as the object. Thus a real image can be projected onto a screen placed at this location. The image distance is positive, and the image is inverted, so its magnification is negative. This is a *case 1 image for mirrors*. It differs from the case 1 image for lenses only in that the image is on the same side of the mirror as the object. It is otherwise identical.

Three incident rays, 1, 2, and 3, falling on a concave mirror. Ray 1 falls parallel, ray 2 falls making an angle with the axis and ray 3 passes through focal point F. These rays after reflection converge at a point below the axis. The image is inverted and enlarged and falls below the axis on the same side as the object. Here, the distance from the center of the mirror to F is the focal distance small f, distances of the object and the image from the mirror are d sub o and d sub I, respectively. The heights of the object and the image are h sub o and h sub I, respectively.

Figure 4. A case 1 image for a mirror. An object is farther from the converging mirror than its focal length. Rays from a common point on the object are traced using the rules in the text. Ray 1 approaches parallel to the axis, ray 2 strikes the center of the mirror, and ray 3 goes through the focal point on the way toward the mirror. All three rays cross at the same point after being reflected, locating the inverted real image. Although three rays are shown, only two of the three are needed to locate the image and determine its height.

Example 1. A Concave Reflector

Electric room heaters use a concave mirror to reflect infrared (IR) radiation from hot coils. Note that IR follows the same law of reflection as visible light. Given that the mirror has a radius of curvature of 50.0 cm and produces an image of the coils 3.00 m away from the mirror, where are the coils?

Strategy and Concept

We are given that the concave mirror projects a real image of the coils at an image distance d_i=3.00 m. The coils are the object, and we are asked to find their location—that is, to find the object distance d_o. We are also given the radius of curvature of the mirror, so that its focal length is $f = \frac{R}{2} = 25.0 \text{ cm}$ (positive since the mirror is concave or converging). Assuming the mirror is small compared with its radius of curvature, we can use the thin lens equations, to solve this problem.

Solution

Since d_i and f are known, thin lens equation can be used to find d_o: $\frac{1}{d_o} + \frac{1}{d_i} = \frac{1}{f}$.

Rearranging to isolate d_o gives $\frac{1}{d_o} = \frac{1}{f} - \frac{1}{d_i}$.

Entering known quantities gives a value for $\frac{1}{d_o}$:

$$\frac{1}{d_o} = \frac{1}{0.250 \text{ m}} - \frac{1}{3.00 \text{ m}} = \frac{3.667}{\text{m}}.$$

This must be inverted to find d_o:

$$d_o = \frac{1 \text{ m}}{3.667} = 27.3 \text{ cm}.$$

Discussion

Note that the object (the filament) is farther from the mirror than the mirror's focal length. This is a case 1 image ($d_o > f$ and f positive), consistent with the fact that a real image is formed. You will get the most concentrated thermal energy directly in front of the mirror and 3.00 m away from it. Generally, this is not desirable, since it could cause burns. Usually, you want the rays to emerge parallel, and this is accomplished by having the filament at the focal point of the mirror.

Note that the filament here is not much farther from the mirror than its focal length and that the image produced is considerably farther away. This is exactly analogous to a slide projector. Placing a slide only slightly farther away from the projector lens than its focal length produces an image significantly farther away. As the object gets closer to the focal distance, the image gets farther away. In fact, as the object distance approaches the focal length, the image distance approaches infinity and the rays are sent out parallel to one another.

Example 2. Solar Electric Generating System

One of the solar technologies used today for generating electricity is a device (called a parabolic trough or concentrating collector) that concentrates the sunlight onto a blackened pipe that contains a fluid. This heated fluid is pumped to a heat exchanger, where its heat energy is transferred to another system that is used to generate steam—and so generate electricity through a conventional steam cycle. Figure 5 shows such a working system in southern California. Concave mirrors are used to concentrate the sunlight onto the pipe. The mirror has the approximate shape of a section of a cylinder. For the problem, assume that the mirror is exactly one-quarter of a full cylinder.

1. If we wish to place the fluid-carrying pipe 40.0 cm from the concave mirror at the mirror's focal point, what will be the radius of curvature of the mirror?
2. Per meter of pipe, what will be the amount of sunlight concentrated onto the pipe, assuming the insolation (incident solar radiation) is 0.900 k W/m^2?
3. If the fluid-carrying pipe has a 2.00-cm diameter, what will be the temperature increase of the fluid per meter of pipe over a period of one minute? Assume all the solar radiation incident on the reflector is absorbed by the pipe, and that the fluid is mineral oil.

Strategy

To solve an *Integrated Concept Problem* we must first identify the physical principles involved. Part 1 is related to the current topic. Part 2 involves a little math, primarily geometry. Part 3 requires an understanding of heat and density.

Solution to Part 1

To a good approximation for a concave or semi-spherical surface, the point where the parallel rays from the sun converge will be at the focal point, so R = 2*f* = 80.0 cm.

Solution to Part 2

The insolation is 900 W $/\mathrm{m}^2$. We must find the cross-sectional area A of the concave mirror, since the power delivered is 900 W $/\mathrm{m}^2$ × A. The mirror in this case is a quarter-section of a cylinder, so the area for a length L of the mirror is $\mathrm{A} = \frac{1}{4}(2\pi\mathrm{R})\mathrm{L}$. The area for a length of 1.00 m is then

$$\mathrm{A} = \frac{\pi}{2}R\,(1.00\ \mathrm{m}) = \frac{(3.14)}{2}(0.800\ \mathrm{m})\,(1.00\ \mathrm{m}) = 1.26^2$$

The insolation on the 1.00-m length of pipe is then

$$\left(9.00 \times 10^2 \frac{\mathrm{W}}{\mathrm{m}^2}\right)\left(1.26\ \mathrm{m}^2\right) = 1130\ \mathrm{W}$$

Solution to Part 3

The increase in temperature is given by Q = *mc*ΔT. The mass *m* of the mineral oil in the one-meter section of pipe is

$$\begin{aligned} m &= \rho \mathrm{V} = \rho\pi\left(\tfrac{d}{2}\right)^2 (1.00 \text{ m}) \\ &= \left(8.00 \times 10^2 \text{ kg/m}^3\right)(3.14)(0.0100 \text{ m})^2 (1.00 \text{ m}) \\ &= 0.251 \text{ kg} \end{aligned}$$

Therefore, the increase in temperature in one minute is

$$\begin{aligned} \Delta T &= \frac{Q}{mc} \\ &= \frac{(1130 \text{ W})(60.0 \text{ s})}{(0.251 \text{ kg})(1670 \text{ J}\cdot \text{kg}/^\circ \text{C})} \\ &= 162^\circ\text{C} \end{aligned}$$

Discussion for Part 3

An array of such pipes in the California desert can provide a thermal output of 250 MW on a sunny day, with fluids reaching temperatures as high as 400°C. We are considering only one meter of pipe here, and ignoring heat losses along the pipe.

A parabolic trough solar thermal electric power plant located at Kramer Junction, California

Figure 5. Parabolic trough collectors are used to generate electricity in southern California. (credit: kjkolb, Wikimedia Commons)

What happens if an object is closer to a concave mirror than its focal length? This is analogous to a case 2 image for lenses ($d_o < f$ and f positive), which is a magnifier. In fact, this is how makeup mirrors act as magnifiers. Figure 6a uses ray tracing to locate the image of an object placed close to a concave mirror. Rays from a common point on the object are reflected in such a manner that they appear to be coming from behind the mirror, meaning that the image is virtual and cannot be projected. As with a magnifying glass, the image is upright and larger than the object. This is a *case 2 image for mirrors* and is exactly analogous to that for lenses.

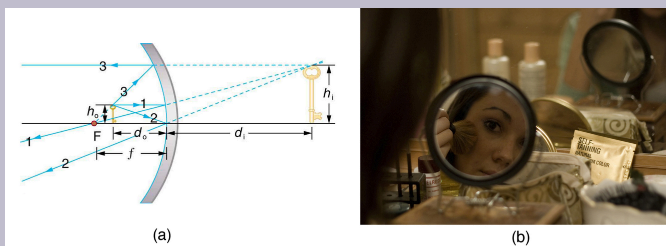


Figure 6. (a) Case 2 images for mirrors are formed when a converging mirror has an object closer to it than its focal length. Ray 1 approaches parallel to the axis, ray 2 strikes the center of the mirror, and ray 3 approaches the mirror as if it came from the focal point. (b) A magnifying mirror showing the reflection. (credit: Mike Melrose, Flickr)

All three rays appear to originate from the same point after being reflected, locating the upright virtual image behind the mirror and showing it to be larger than the object. (b) Makeup mirrors are perhaps the most common use of a concave mirror to produce a larger, upright image.

A convex mirror is a diverging mirror (f is negative) and forms only one type of image. It is a *case* 3 image—one that is upright and smaller than the object, just as for diverging lenses. Figure 7a uses ray tracing to illustrate the location and size of the case 3 image for mirrors. Since the image is behind the mirror, it cannot be projected and is thus a virtual image. It is also seen to be smaller than the object.

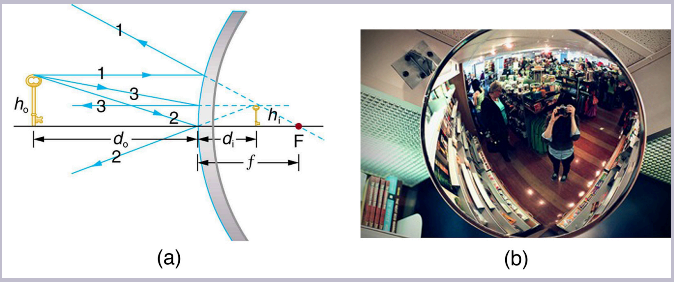


Figure 7. Case 3 images for mirrors are formed by any convex mirror. Ray 1 approaches parallel to the axis, ray 2 strikes the center of the mirror, and ray 3 approaches toward the focal point. All three rays appear to originate from the same point after being reflected, locating the upright virtual image behind the mirror and showing it to be smaller than the object. (b) Security mirrors are convex, producing a smaller, upright image. Because the image is smaller, a larger area is imaged compared to what would be observed for a flat mirror (and hence security is improved). (credit: Laura D'Alessandro, Flickr)

Example 3. Image in a Convex Mirror

A keratometer is a device used to measure the curvature of the cornea, particularly for fitting contact lenses. Light is reflected from the cornea, which acts like a convex mirror, and the keratometer measures the magnification of the image. The smaller the magnification, the smaller the radius of curvature of the cornea. If the light source is 12.0 cm from the cornea and the image's magnification is 0.0320, what is the cornea's radius of curvature?

Strategy

If we can find the focal length of the convex mirror formed by the cornea, we can find its radius of curvature (the radius of curvature is twice the focal length of a spherical mirror). We are given that the object distance is d_o = 12.0 cm and that m = 0.0320. We first solve for the image distance d_i, and then for f.

Solution

$m = -\frac{d_i}{d_o}$. Solving this expression for d_i gives d_i = $-md_o$.

Entering known values yields d_i = −(0.0320)(12.0 cm) = −0.384 cm.

$$\frac{1}{f} = \frac{1}{d_o} + \frac{1}{d_i}$$

Substituting known values,

$$\frac{1}{f} = \frac{1}{12.0 \text{ cm}} + \frac{1}{-0.384 \text{ cm}} = \frac{-2.52}{\text{cm}}.$$

This must be inverted to find f:

$$f = \frac{\text{cm}}{-2.52} = -0.400 \text{ cm}$$

The radius of curvature is twice the focal length, so that R = 2|f| = 0.800 cm.

Discussion

Although the focal length f of a convex mirror is defined to be negative, we take the absolute value to give us a positive value for R. The radius of curvature found here is reasonable for a cornea. The distance from cornea to retina in an adult eye is about 2.0 cm. In practice, many corneas are not spherical, complicating the job of fitting contact lenses. Note that the image distance here is negative, consistent with the fact that the image is behind the mirror, where it cannot be projected. In this section's Problems and Exercises, you will show that for a fixed object distance, the smaller the radius of curvature, the smaller the magnification.

The three types of images formed by mirrors (cases 1, 2, and 3) are exactly analogous to those formed by lenses, as summarized in the table at the end of Image Formation by Lenses. It is easiest to concentrate on only three types of images—then remember that concave mirrors act like convex lenses, whereas convex mirrors act like concave lenses.

Take-Home Experiment: Concave Mirrors Close to Home

Find a flashlight and identify the curved mirror used in it.

Find another flashlight and shine the first flashlight onto the second one, which is turned off. Estimate the focal length of the mirror. You might try shining a flashlight on the curved mirror behind the headlight of a car, keeping the headlight switched off, and determine its focal length.

Problem-Solving Strategy for Mirrors

Step 1. Examine the situation to determine that image formation by a mirror is involved.

Step 2. Refer to the Problem-Solving Strategies for Lenses. The same strategies are valid for mirrors as for lenses with one qualification—use the ray tracing rules for mirrors listed earlier in this section.

Section Summary

- The characteristics of an image formed by a flat mirror are: (a) The image and object are the same distance from the mirror, (b) The image is a virtual image, and (c) The image is situated behind the mirror.
- 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.

Conceptual Questions

1. What are the differences between real and virtual images? How can you tell (by looking) whether an image formed by a single lens or mirror is real or virtual?
2. Can you see a virtual image? Can you photograph one? Can one be projected onto a screen with additional lenses or mirrors? Explain your responses.
3. Is it necessary to project a real image onto a screen for it to exist?
4. At what distance is an image always located—at d_o, d_i, or f?
5. Under what circumstances will an image be located at the focal point of a lens or mirror?
6. What is meant by a negative magnification? What is meant by a magnification that is less than 1 in magnitude?
7. Can a case 1 image be larger than the object even though its magnification is always negative? Explain.
8. Figure 8 shows a light bulb between two mirrors. One mirror produces a beam of light with parallel rays; the other keeps light from escaping without being put into the beam. Where is the filament of the light in relation to the focal point or radius of curvature of each mirror?

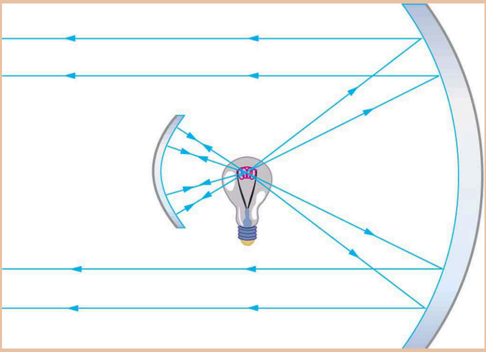


Figure 8. The two mirrors trap most of the bulb's light and form a directional beam as in a headlight.

9. The two mirrors trap most of the bulb's light and form a directional beam as in a headlight.
10. Two concave mirrors of different sizes are placed facing one another. A filament bulb is placed at the focus of the larger mirror. The rays after reflection from the larger mirror travel parallel to one another. The rays falling on the smaller mirror retrace their paths.
11. Devise an arrangement of mirrors allowing you to see the back of your head. What is the minimum number of mirrors needed for this task?
12. If you wish to see your entire body in a flat mirror (from head to toe), how tall should the mirror be? Does its size depend upon your distance away from the mirror? Provide a sketch.
13. It can be argued that a flat mirror has an infinite focal length. If so, where does it form an image? That is, how are d_i and d_o related?
14. Why are diverging mirrors often used for rear-view mirrors in vehicles? What is the main disadvantage of

using such a mirror compared with a flat one?

Problems & Exercises

1. What is the focal length of a makeup mirror that has a power of 1.50 D?
2. Some telephoto cameras use a mirror rather than a lens. What radius of curvature mirror is needed to replace a 800 mm focal length telephoto lens?
3. (a) Calculate the focal length of the mirror formed by the shiny back of a spoon that has a 3.00 cm radius of curvature. (b) What is its power in diopters?
4. Find the magnification of the heater element in Example 1. Note that its large magnitude helps spread out the reflected energy.
5. What is the focal length of a makeup mirror that produces a magnification of 1.50 when a person's face is 12.0 cm away?
6. A shopper standing 3.00 m from a convex security mirror sees his image with a magnification of 0.250. (a) Where is his image? (b) What is the focal length of the mirror? (c) What is its radius of curvature?
7. An object 1.50 cm high is held 3.00 cm from a person's cornea, and its reflected image is measured to be 0.167 cm high. (a) What is the magnification? (b) Where is the image? (c) Find the radius of curvature of the convex mirror formed by the cornea. (Note that this technique is used by optometrists to measure the

curvature of the cornea for contact lens fitting. The instrument used is called a keratometer, or curve measurer.)

8. Ray tracing for a flat mirror shows that the image is located a distance behind the mirror equal to the distance of the object from the mirror. This is stated $d_i = -d_o$, since this is a negative image distance (it is a virtual image). (a) What is the focal length of a flat mirror? (b) What is its power?
9. Show that for a flat mirror $h_i = h_o$, knowing that the image is a distance behind the mirror equal in magnitude to the distance of the object from the mirror.
10. Use the law of reflection to prove that the focal length of a mirror is half its radius of curvature. That is, prove that $f = \frac{R}{2}$. Note this is true for a spherical mirror only if its diameter is small compared with its radius of curvature.
11. Referring to the electric room heater considered in the first example in this section, calculate the intensity of IR radiation in W/m^2 projected by the concave mirror on a person 3.00 m away. Assume that the heating element radiates 1500 W and has an area of 100 cm^2, and that half of the radiated power is reflected and focused by the mirror.
12. Consider a 250-W heat lamp fixed to the ceiling in a bathroom. If the filament in one light burns out then the remaining three still work. Construct a problem in which you determine the resistance of each filament in order to obtain a certain intensity projected on the bathroom floor. The ceiling is 3.0 m high. The

problem will need to involve concave mirrors behind the filaments. Your instructor may wish to guide you on the level of complexity to consider in the electrical components.

Glossary

converging mirror: a concave mirror in which light rays that strike it parallel to its axis converge at one or more points along the axis

diverging mirror: a convex mirror in which light rays that strike it parallel to its axis bend away (diverge) from its axis

law of reflection: angle of reflection equals the angle of incidence

Selected Solutions to Problems & Exercises

1. +0.667 m

3. (a) -1.5×10^{-2} m; (b) −66.7 D

5. +0.360 m (concave)

7. (a) +0.111; (b) −0.334 cm (behind “mirror”); (c) 0.752cm

9.

$$m = \frac{h_\mathrm{i}}{h_\mathrm{o}} = -\frac{d_\mathrm{i}}{d_\mathrm{o}} = -\frac{-d_\mathrm{o}}{d_\mathrm{o}} = \frac{d_\mathrm{o}}{d_\mathrm{o}} = 1 \Rightarrow h_\mathrm{i} = h_\mathrm{o}$$

11. 6.82 $\mathrm{kW/m^2}$

PART X

VISION AND OPTICAL INSTRUMENTS

77. Introduction to Vision and Optical Instruments

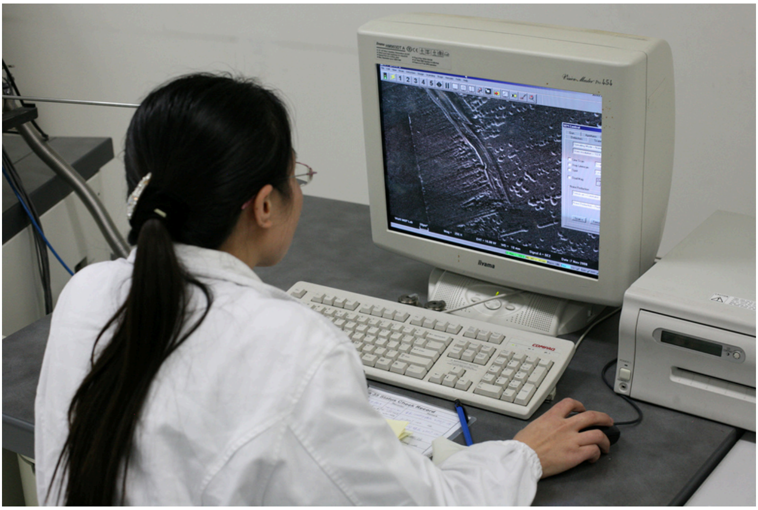


Figure 1. A scientist examines minute details on the surface of a disk drive at a magnification of 100,000 times. The image was produced using an electron microscope. (credit: Robert Scoble)

Explore how the image on the computer screen is formed. How is the image formation on the computer screen different from the image formation in your eye as you look down the microscope? How can videos of living cell processes be taken for viewing later on, and by many different people?

Seeing faces and objects we love and cherish is a delight—one's favorite teddy bear, a picture on the wall, or the sun rising over the mountains. Intricate images help us understand nature and are invaluable for developing techniques and technologies in order to improve the quality of life. The image of a red blood cell that almost fills the cross-sectional area of a tiny capillary makes us wonder how blood makes it through and not get stuck. We are able to see bacteria and viruses and understand their structure. It is the knowledge of physics that provides fundamental understanding and models required to develop new techniques and instruments. Therefore, physics is called an *enabling science*—a science that enables development and advancement in other areas. It is through optics and imaging that physics enables advancement in major areas of biosciences. This chapter illustrates the enabling nature of physics through an understanding of how a human eye is able to see and how we are able to use optical instruments to see beyond what is possible with the naked eye. It is convenient to categorize these instruments on the basis of geometric optics and wave optics.

78. Video: Refraction

Watch the following Physics Concept Trailer to see how LASIK eye surgery changes the shape of the cornea so that light can refract correctly on the cornea.

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://library.achievingthedream.org/austinccphysics2/?p=105#oembed-1

79. Physics of the Eye

Learning Objectives

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

- Explain the image formation by the eye.
- Explain why peripheral images lack detail and color.
- Define refractive indices.
- Analyze the accommodation of the eye for distant and near vision.

The eye is perhaps the most interesting of all optical instruments. The eye is remarkable in how it forms images and in the richness of detail and color it can detect. However, our eyes commonly need some correction, to reach what is called “normal” vision, but should be called ideal rather than normal. Image formation by our eyes and common vision correction are easy to analyze with the optics discussed in Geometric Optics.

Figure 1 shows the basic anatomy of the eye. The cornea and lens form a system that, to a good approximation, acts as a single thin lens. For clear vision, a real image must be projected onto the light-sensitive retina, which lies at a fixed distance from the lens. The lens of the eye adjusts its power to produce an image on the retina for objects at different distances. The center of the image falls on the fovea, which has the greatest density of light receptors and the greatest acuity (sharpness) in the visual field. The variable opening (or pupil) of the eye along with chemical adaptation allows the eye to detect light intensities from the lowest observable to 10^{10} times greater (without damage). This is an incredible range of detection. Our eyes perform a vast number of functions, such as sense direction, movement, sophisticated colors, and distance. Processing of visual nerve impulses begins with interconnections in the retina and continues in the brain. The optic nerve conveys signals received by the eye to the brain.

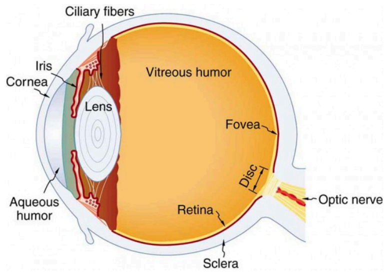


Figure 1. The cornea and lens of an eye act together to form a real image on the light-sensing retina, which has its densest concentration of receptors in the fovea and a blind spot over the optic nerve. The power of the lens of an eye is adjustable to provide an image on the retina for varying object distances. Layers of tissues with varying indices of refraction in the lens are shown here. However, they have been omitted from other pictures for clarity.

Refractive indices are crucial to image formation using lenses. Table 1 shows refractive indices relevant to the eye. The biggest change in the refractive index, and bending of rays, occurs at the cornea rather than the lens. The ray diagram in Figure 2 shows image formation by the cornea and lens of the eye. The rays bend according to the refractive indices provided in Table 1. The cornea provides about two-thirds of the power of the eye, owing to the fact that speed of light changes considerably while traveling from air into cornea. The lens provides the remaining power needed to produce an image on the retina. The cornea and lens can be treated as a single thin lens, even though the light rays pass through several layers of material (such as cornea, aqueous humor, several layers in the lens, and vitreous humor), changing direction at each interface. The image formed is much like the one produced by a single convex lens. This is a case 1 image. Images formed in the eye are inverted but the brain inverts them once more to make them seem upright.

Table 1. Refractive Indices Relevant to the Eye

Material	Index of Refraction
Water	1.33
Air	1.0
Cornea	1.38
Aqueous humor	1.34
Lens	1.41 average (varies throughout the lens, greatest in center)
Vitreous humor	1.34

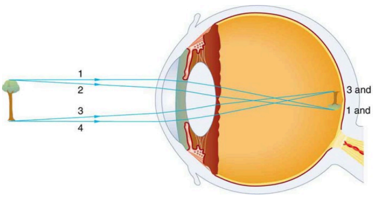


Figure 2. An image is formed on the retina with light rays converging most at the cornea and upon entering and exiting the lens. Rays from the top and bottom of the object are traced and produce an inverted real image on the retina. The distance to the object is drawn smaller than scale.

As noted, the image must fall precisely on the retina to produce clear vision—that is, the image distance d_i must equal the lens-to-retina distance. Because the lens-to-retina distance does not change, the image distance d_i must be the same for objects at all distances. The eye manages this by varying the power (and focal length) of the lens to accommodate for objects at various distances. The process of adjusting the eye's focal length is called *accommodation*. A person with normal (ideal) vision can see objects clearly at distances ranging from 25 cm to essentially infinity. However, although the near point (the shortest distance at which a sharp focus can be obtained) increases with age (becoming meters for some older people), we will consider it to be 25 cm in our treatment here.

Figure 3 shows the accommodation of the eye for distant and near vision. Since light rays from a nearby object can diverge and still enter the eye, the lens must be more converging (more powerful) for close vision than for distant vision. To be more converging, the lens is made thicker by the action of the ciliary muscle surrounding it. The eye is most relaxed when viewing distant objects, one reason that microscopes and telescopes are designed to produce distant images. Vision of very distant objects is called *totally relaxed*, while close vision is termed *accommodated*, with the closest vision being *fully accommodated*.

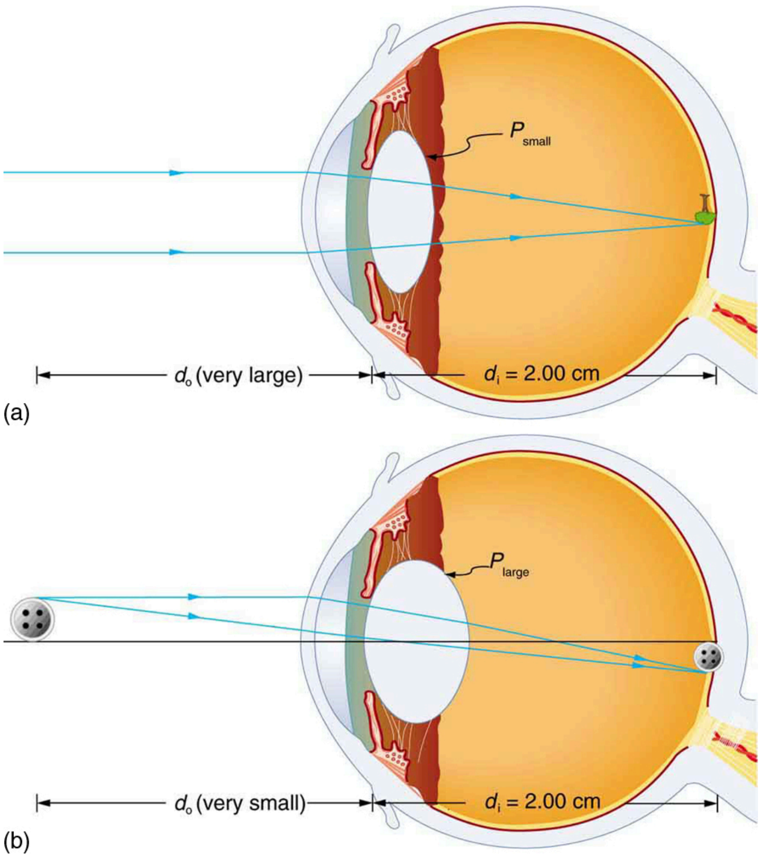


Figure 3. Relaxed and accommodated vision for distant and close objects. (a) Light rays from the same point on a distant object must be nearly parallel while entering the eye and more easily converge to produce an image on the retina. (b) Light rays from a nearby object can diverge more and still enter the eye. A more powerful lens is needed to converge them on the retina than if they were parallel.

We will use the thin lens equations to examine image formation by the eye quantitatively. First, note the power of a lens is given as $p = \frac{1}{f}$, so we rewrite the thin lens equations as $P = \frac{1}{d_o} + \frac{1}{d_i}$ and $\frac{h_i}{h_o} = -\frac{d_i}{d_o} = m$.

We understand that d_i must equal the lens-to-retina distance to obtain clear vision, and that normal vision is possible for objects at distances d_o = 25 cm to infinity.

Take-Home Experiment: The Pupil

Look at the central transparent area of someone's eye, the pupil, in normal room light. Estimate the diameter of the pupil. Now turn off the lights and darken the room. After a few minutes turn on the lights and promptly estimate the diameter of the pupil. What happens to the pupil as the eye adjusts to the room light? Explain your observations.

The eye can detect an impressive amount of detail, considering how small the image is on the retina. To get some idea of how small the image can be, consider the following example.

Example 1. Size of Image on Retina

What is the size of the image on the retina of a 1.20×10^{-2} cm diameter human hair, held at arm's length (60.0 cm) away? Take the lens-to-retina distance to be 2.00 cm.

Strategy

We want to find the height of the image h_i, given the height of the object is $h_o = 1.20 \times 10^{-2}$ cm. We also know that the object is 60.0 cm away, so that d_o=60.0 cm. For clear vision, the image distance must equal the lens-to-retina distance, and so $d_i = 2.00$ cm . The equation $\frac{h_i}{h_o} = -\frac{d_i}{d_o} = m$ can be used to find h_i with the known information.

Solution

The only unknown variable in the equation $\frac{h_i}{h_o} = -\frac{d_i}{d_o} = m$ is h_i:

$$\frac{h_i}{h_o} = -\frac{d_i}{d_o}$$

Rearranging to isolate h_i yields

$$h_i = -h_o \cdot \frac{d_i}{d_o}.$$

Substituting the known values gives

$$\begin{aligned} h_i &= -\left(1.20 \times 10^{-2}\text{ cm}\right)\frac{2.00\text{ cm}}{60.0\text{ cm}} \\ &= -4.00 \times 10^{-4}\text{ cm} \end{aligned}$$

Discussion

This truly small image is not the smallest discernible—that is, the limit to visual acuity is even smaller than this. Limitations on visual acuity have to do with the wave properties of light and will be discussed in the next chapter. Some limitation is also due to the inherent anatomy of the eye and processing that occurs in our brain.

Example 2. Power Range of the Eye

Calculate the power of the eye when viewing objects at the greatest and smallest distances possible with normal vision, assuming a lens-to-retina distance of 2.00 cm (a typical value).

Strategy

For clear vision, the image must be on the retina, and so d_i = 2.00 cm here. For distant vision, $d_\text{o} \approx \infty$, and for close vision, d_o = 25.0 cm, as discussed earlier. The equation $P=\frac{1}{d_\text{o}}+\frac{1}{d_\text{i}}$ as written just above, can be used directly to solve for P in both cases, since we know d_i and

d_o. Power has units of diopters, where $1\text{ D} = \frac{1}{\text{m}}$, and so we should express all distances in meters.

Solution

For distant vision,

$$P = \frac{1}{d_o} + \frac{1}{d_i} = \frac{1}{\infty} + \frac{1}{0.0200\text{ m}}$$

Since $\frac{1}{\infty} = 0$, this gives

$$P = 0 + \frac{50.0}{\text{m}} = 50.0\text{ D}$$ (distant vision).

Now, for close vision,

$$\begin{aligned} P &= \frac{1}{d_o} + \frac{1}{d_i} = \frac{1}{0.250\text{ m}} + \frac{1}{0.0200\text{ m}} \\ &= \frac{4.00}{\text{m}} + \frac{50.0}{\text{m}} = 4.00\text{ D} + 50.0\text{ D} \\ &= 54.0\text{ D (close vision)} \end{aligned}$$

Discussion

For an eye with this typical 2.00 cm lens-to-retina distance, the power of the eye ranges from 50.0 D (for distant totally relaxed vision) to 54.0 D (for close fully accommodated vision), which is an 8% increase. This increase in power for close vision is consistent with the preceding discussion and the ray tracing in Figure 3. An 8% ability to accommodate is considered normal but is typical for people who are about 40 years old. Younger people have greater accommodation ability, whereas older people gradually lose the ability to accommodate. When an optometrist identifies accommodation as a problem in elder people, it is most likely due to stiffening of the lens. The lens of the eye changes with age in ways that tend to preserve the ability to see distant objects clearly but do not allow the eye to accommodate for close vision, a condition called *presbyopia* (literally, elder eye). To correct this vision defect, we place a converging, positive power lens in front of the eye, such as found in reading glasses. Commonly available reading glasses are rated by their power in diopters, typically ranging from 1.0 to 3.5 D.

Section Summary

- Image formation by the eye is adequately described by the thin lens equations:
  $$P = \frac{1}{d_{\mathrm{o}}} + \frac{1}{d_{\mathrm{i}}} \text{ and } \frac{h_{\mathrm{i}}}{h_{\mathrm{o}}} = -\frac{d_{\mathrm{i}}}{d_{\mathrm{o}}} = m.$$
- The eye produces a real image on the retina by adjusting its focal length and power in a process called accommodation.
- For close vision, the eye is fully accommodated and has its greatest power, whereas for distant vision, it is totally relaxed and has its smallest power.
- The loss of the ability to accommodate with age is called presbyopia, which is corrected by the use of a converging lens to add power for close vision.

Conceptual Questions

1. If the lens of a person's eye is removed because of cataracts (as has been done since ancient times), why would you expect a spectacle lens of about 16 D to be prescribed?
2. A cataract is cloudiness in the lens of the eye. Is light dispersed or diffused by it?
3. When laser light is shone into a relaxed normal-vision eye to repair a tear by spot-welding the retina to the back of the eye, the rays entering the eye must be parallel. Why?
4. How does the power of a dry contact lens compare with its power when resting on the tear layer of the eye? Explain.
5. Why is your vision so blurry when you open your eyes while swimming under water? How does a face mask enable clear vision?

Problems & Exercises

Unless otherwise stated, the lens-to-retina distance is 2.00 *cm.*

1. What is the power of the eye when viewing an object 50.0 cm away?
2. Calculate the power of the eye when viewing an

object 3.00 m away.

3. (a) The print in many books averages 3.50 mm in height. How high is the image of the print on the retina when the book is held 30.0 cm from the eye? (b) Compare the size of the print to the sizes of rods and cones in the fovea and discuss the possible details observable in the letters. (The eye-brain system can perform better because of interconnections and higher order image processing.)
4. Suppose a certain person's visual acuity is such that he can see objects clearly that form an image 4.00 μm high on his retina. What is the maximum distance at which he can read the 75.0 cm high letters on the side of an airplane?
5. People who do very detailed work close up, such as jewellers, often can see objects clearly at much closer distance than the normal 25 cm. (a) What is the power of the eyes of a woman who can see an object clearly at a distance of only 8.00 cm? (b) What is the size of an image of a 1.00 mm object, such as lettering inside a ring, held at this distance? (c) What would the size of the image be if the object were held at the normal 25.0 cm distance?

Glossary

accommodation: the ability of the eye to adjust its focal length is known as accommodation

presbyopia: a condition in which the lens of the eye becomes progressively unable to focus on objects close to the viewer

Selected Solutions to Problems & Exercises

1. 52.0 D

3. (a) −0.233 mm; (b) The size of the rods and the cones is smaller than the image height, so we can distinguish letters on a page.

5. (a) +62.5 D; (b) −0.250 mm; (c) −0.0800 mm

80. Telescopes

Learning Objectives

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

- Outline the invention of a telescope.
- Describe the working of a telescope.

Telescopes are meant for viewing distant objects, producing an image that is larger than the image that can be seen with the unaided eye. Telescopes gather far more light than the eye, allowing dim objects to be observed with greater magnification and better resolution. Although Galileo is often credited with inventing the telescope, he actually did not. What he did was more important. He constructed several early telescopes, was the first to study the heavens with them, and made monumental discoveries using them. Among these are the moons of Jupiter, the craters and mountains on the Moon, the details of sunspots, and the fact that the Milky Way is composed of vast numbers of individual stars.

Figure 1a shows a telescope made of two lenses, the convex objective and the concave eyepiece, the same construction used by Galileo. Such an arrangement produces an upright image and is used in spyglasses and opera glasses.

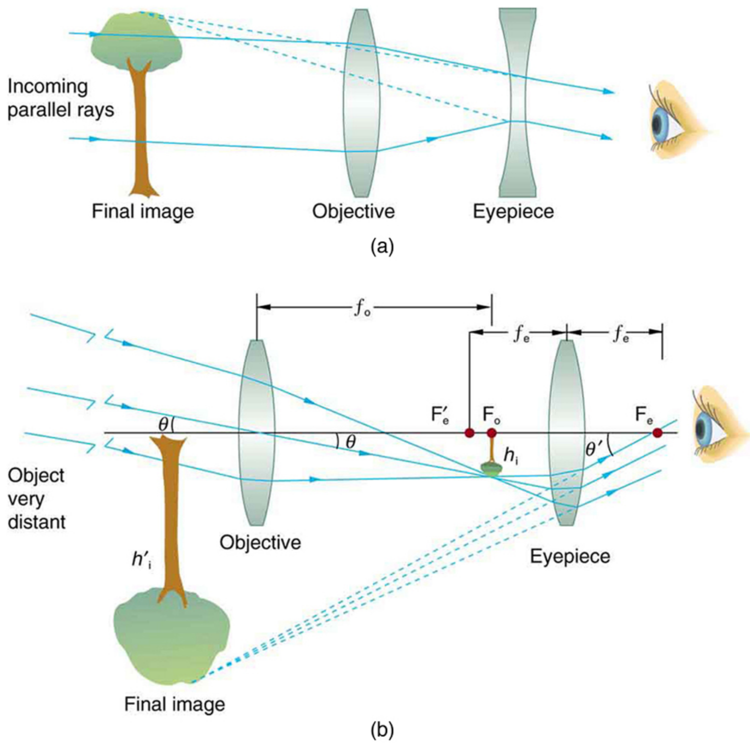


Figure 1. (a) Galileo made telescopes with a convex objective and a concave eyepiece. These produce an upright image and are used in spyglasses. (b) Most simple telescopes have two convex lenses. The objective forms a case 1 image that is the object for the eyepiece. The eyepiece forms a case 2 final image that is magnified.

The most common two-lens telescope, like the simple microscope, uses two convex lenses and is shown in Figure 1b. The object is so far away from the telescope that it is essentially at infinity compared with the focal lengths of the lenses ($d_o \approx \infty$). The first image is thus produced at $d_i = f_o$, as shown in the figure. To prove this, note that

$$\frac{1}{d_\mathrm{i}} = \frac{1}{f_\mathrm{o}} - \frac{1}{d_\mathrm{o}} = \frac{1}{f_\mathrm{o}} - \frac{1}{\infty}$$

Because $\frac{1}{\infty} = 0$, this simplifies to $\frac{1}{d_\mathrm{i}} = \frac{1}{f_\mathrm{o}}$, which implies that $d_\mathrm{i} = f_\mathrm{o}$, as claimed. It is true that for any distant object and any lens or mirror, the image is at the focal length.

The first image formed by a telescope objective as seen in Figure 1b will not be large compared with what you might see by looking at the object directly. For example, the spot formed by sunlight focused on a piece of paper by a magnifying glass is the image of the Sun, and it is small. The telescope eyepiece (like the microscope eyepiece) magnifies this first image. The distance between the eyepiece and the objective lens is made slightly less than the sum of their focal lengths so that the first image is closer to the eyepiece than its focal length. That is, d_o' is less than f_e, and so the eyepiece forms a case 2 image that is large and to the left for easy viewing. If the angle subtended by an object as viewed by the unaided eye is θ, and the angle subtended by the telescope image is θ', then the *angular magnification* M is defined to be their ratio. That is, $M = \frac{\theta'}{\theta}$. It can be shown that the angular magnification of a telescope is related to the focal lengths of the objective and eyepiece; and is given by

$$M = \frac{\theta'}{\theta} = -\frac{f_\mathrm{o}}{f_\mathrm{e}}$$

The minus sign indicates the image is inverted. To obtain the greatest angular magnification, it is best to have a long focal length objective and a short focal length eyepiece. The greater the angular magnification M, the larger an object will appear when viewed through a telescope, making more details visible. Limits to observable details are imposed by many factors, including lens quality and atmospheric disturbance.

The image in most telescopes is inverted, which is unimportant for observing the stars but a real problem for other applications, such as telescopes on ships or telescopic gun sights. If an upright image is needed, Galileo's arrangement in Figure 1a can be used. But a more common arrangement is to use a third convex lens as an eyepiece, increasing the distance between the first two and inverting the image once again as seen in Figure 2.

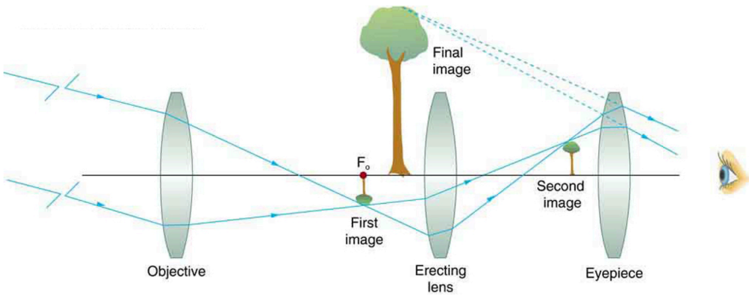


Figure 2. This arrangement of three lenses in a telescope produces an upright final image. The first two lenses are far enough apart that the second lens inverts the image of the first one more time. The third lens acts as a magnifier and keeps the image upright and in a location that is easy to view.

A telescope can also be made with a concave mirror as its first element or objective, since a concave mirror acts like a convex lens as seen in Figure 3. Flat mirrors are often employed in optical instruments to make them more compact or to send light to cameras and other sensing devices. There are many advantages to using mirrors rather than lenses for telescope objectives. Mirrors can be constructed much larger than lenses and can, thus, gather large amounts of light, as needed to view distant galaxies, for example. Large and relatively flat mirrors have very long focal lengths, so that great angular magnification is possible.

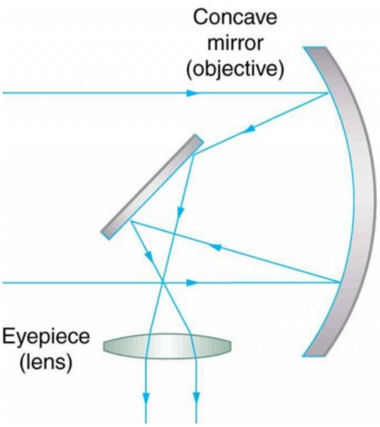


Figure 3. A two-element telescope composed of a mirror as the objective and a lens for the eyepiece is shown. This telescope forms an image in the same manner as the two-convex-lens telescope already discussed, but it does not suffer from chromatic aberrations. Such telescopes can gather more light, since larger mirrors than lenses can be constructed.

Telescopes, like microscopes, can utilize a range of frequencies from the electromagnetic spectrum. Figure 4a shows the Australia Telescope Compact Array, which uses six 22-m antennas for mapping the southern skies using radio waves. Figure 4b shows the focusing of x rays on the Chandra X-ray Observatory—a satellite orbiting earth since 1999 and looking at high temperature events as exploding stars, quasars, and black holes. X rays, with much more energy and shorter wavelengths than RF and light, are mainly absorbed and not reflected when incident perpendicular to the medium. But they can be reflected when incident at small glancing angles, much like a rock will skip on a lake if thrown at a small angle. The mirrors for the Chandra consist of a long barrelled pathway and 4 pairs of mirrors to focus the rays at a point 10 meters away from the entrance. The mirrors are extremely smooth and consist of a glass ceramic base with a thin coating of metal (iridium). Four pairs of precision manufactured mirrors are exquisitely shaped and aligned so that x rays ricochet off the mirrors like bullets off a wall, focusing on a spot.

A current exciting development is a collaborative effort involving 17 countries to construct a Square Kilometre Array (SKA) of telescopes capable of covering from 80 MHz to 2 GHz. The initial stage of the project is the construction of the Australian Square Kilometre Array Pathfinder in Western Australia (see Figure 5). The project will use cutting-edge technologies such as *adaptive optics* in which the lens or mirror is constructed from lots of carefully aligned tiny lenses and mirrors that can be manipulated using computers. A range of rapidly changing distortions can be minimized by deforming or tilting the tiny lenses and mirrors. The use of adaptive optics in vision correction is a current area of research.

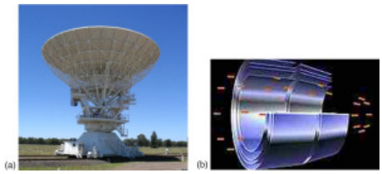


Figure 4. (a) The Australia Telescope Compact Array at Narrabri (500 km NW of Sydney). (credit: Ian Bailey) (b) The focusing of x rays on the Chandra Observatory, a satellite orbiting earth. X rays ricochet off 4 pairs of mirrors forming a barrelled pathway leading to the focus point. (credit: NASA)

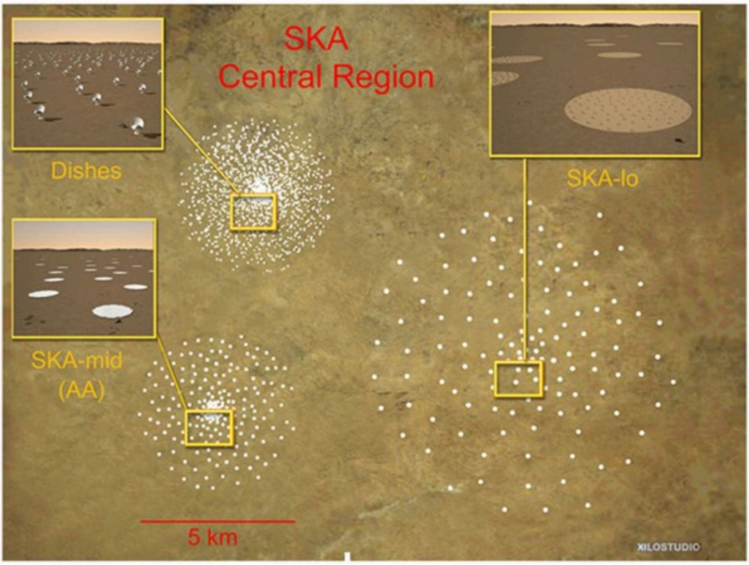


Figure 5. An artist's impression of the Australian Square Kilometre Array Pathfinder in Western Australia is displayed. (credit: SPDO, XILOSTUDIOS)

Section Summary

- Simple telescopes can be made with two lenses. They are used for viewing objects at large distances and utilize the entire range of the electromagnetic spectrum.
- The angular magnification M for a telescope is given by $M = \frac{\theta'}{\theta} = -\frac{f_o}{f_e}$, where θ is the angle subtended by an object viewed by the unaided eye, θ' is the angle subtended by a magnified image, and f_o and f_e are the focal lengths of the objective and the eyepiece.

Conceptual Questions

1. If you want your microscope or telescope to project a real image onto a screen, how would you change the placement of the eyepiece relative to the objective?

Problems & Exercises

Unless otherwise stated, the lens-to-retina distance is 2.00 cm.

1. What is the angular magnification of a telescope that has a 100 cm focal length objective and a 2.50 cm focal length eyepiece?
2. Find the distance between the objective and eyepiece lenses in the telescope in the above problem needed to produce a final image very far from the observer, where vision is most relaxed. Note that a telescope is normally used to view very distant objects.
3. A large reflecting telescope has an objective mirror with a 10.0 m radius of curvature. What angular magnification does it produce when a 3.00 m focal length eyepiece is used?
4. A small telescope has a concave mirror with a 2.00 m radius of curvature for its objective. Its eyepiece is

a 4.00 cm focal length lens. (a) What is the telescope's angular magnification? (b) What angle is subtended by a 25,000 km diameter sunspot? (c) What is the angle of its telescopic image?

5. A 7.5× binocular produces an angular magnification of −7.50, acting like a telescope. (Mirrors are used to make the image upright.) If the binoculars have objective lenses with a 75.0 cm focal length, what is the focal length of the eyepiece lenses?
6. **Construct Your Own Problem.** Consider a telescope of the type used by Galileo, having a convex objective and a concave eyepiece as illustrated in Figure 1a. Construct a problem in which you calculate the location and size of the image produced. Among the things to be considered are the focal lengths of the lenses and their relative placements as well as the size and location of the object. Verify that the angular magnification is greater than one. That is, the angle subtended at the eye by the image is greater than the angle subtended by the object.

Glossary

adaptive optics: optical technology in which computers adjust the lenses and mirrors in a device to correct for image distortions

angular magnification: a ratio related to the focal lengths of the objective and eyepiece and given as $M=-\frac{f_{\text{o}}}{f_{\text{e}}}$

Selected Solutions to Problems & Exercises

1. −40.0
3. −1.67
5. +10.0 cm

81. Vision Correction

Learning Objectives

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

- Identify and discuss common vision defects.
- Explain nearsightedness and farsightedness corrections.
- Explain laser vision correction.

The need for some type of vision correction is very common. Common vision defects are easy to understand, and some are simple to correct. Figure 1 illustrates two common vision defects. *Nearsightedness*, or *myopia*, is the inability to see distant objects clearly while close objects are clear. The eye overconverges the nearly parallel rays from a distant object, and the rays cross in front of the retina. More divergent rays from a close object are converged on the retina for a clear image. The distance to the farthest object that can be seen clearly is called the *far point* of the eye (normally infinity). *Farsightedness*, or *hyperopia*, is the inability to see close objects clearly while distant objects may be clear. A farsighted eye does not converge sufficient rays from a close object to make the rays meet on the retina. Less diverging rays from a distant object can be converged for a clear image. The distance to the closest object that can be seen clearly is called the *near point* of the eye (normally 25 cm).

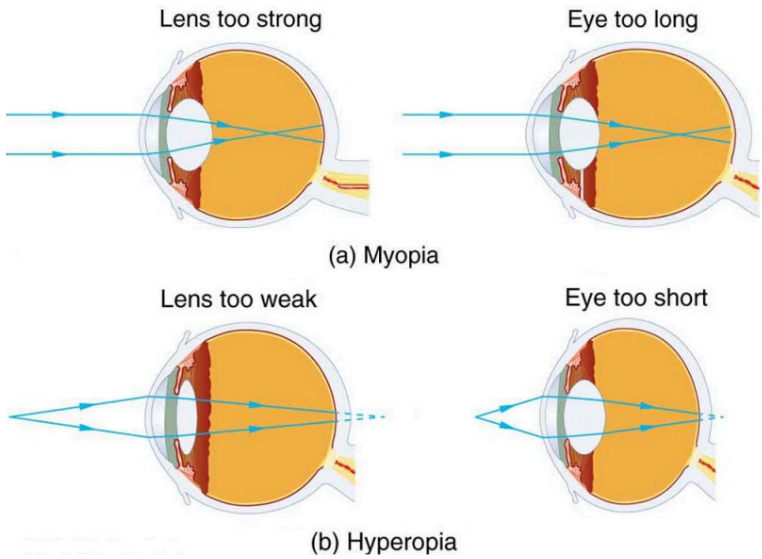


Figure 1. (a) The nearsighted (myopic) eye converges rays from a distant object in front of the retina; thus, they are diverging when they strike the retina, producing a blurry image. This can be caused by the lens of the eye being too powerful or the length of the eye being too great. (b) The farsighted (hyperopic) eye is unable to converge the rays from a close object by the time they strike the retina, producing blurry close vision. This can be caused by insufficient power in the lens or by the eye being too short.

Since the nearsighted eye over converges light rays, the correction for nearsightedness is to place a diverging spectacle lens in front of the eye. This reduces the power of an eye that is too powerful. Another way of thinking about this is that a diverging spectacle lens produces a case 3 image, which is closer to the eye than the object (see Figure 2). To determine the spectacle power needed for correction, you must know the person's far point—that is, you must know the greatest distance at which the person can see clearly. Then the image produced by a spectacle lens must be at this distance or closer for the nearsighted person to be able to see it clearly. It is worth noting that wearing glasses does not change the eye in any way. The eyeglass lens is simply used to create an image of the object at a distance where the nearsighted person can see it clearly. Whereas someone not wearing glasses can see clearly *objects* that fall between their near point and their far point, someone wearing glasses can see *images* that fall between their near point and their far point.

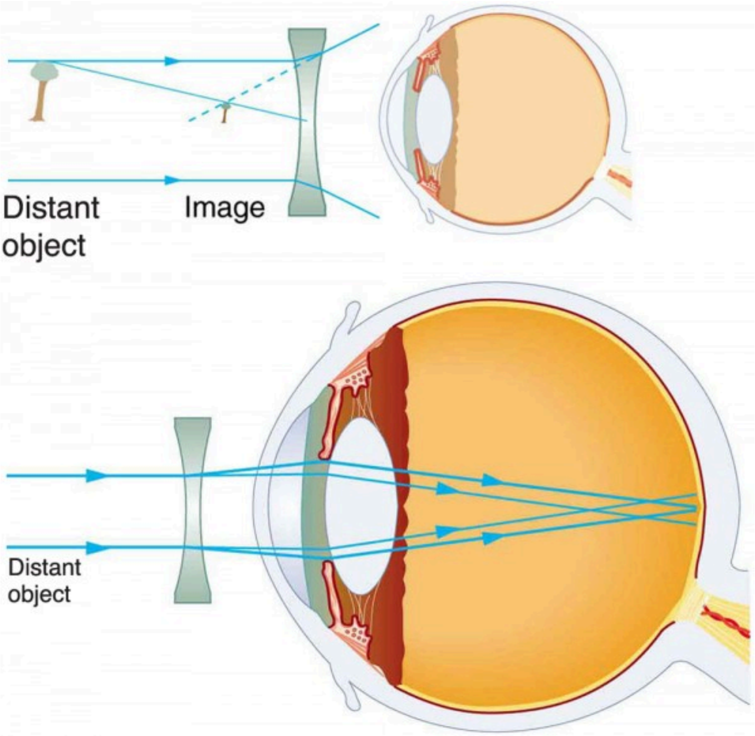


Figure 2. Correction of nearsightedness requires a diverging lens that compensates for the overconvergence by the eye. The diverging lens produces an image closer to the eye than the object, so that the nearsighted person can see it clearly.

Example 1. Correcting Nearsightedness

What power of spectacle lens is needed to correct the vision of a nearsighted person whose far point is 30.0 cm? Assume the spectacle (corrective) lens is held 1.50 cm away from the eye by eyeglass frames.

Strategy

You want this nearsighted person to be able to see very distant objects clearly. That means the spectacle lens must produce an image 30.0 cm from the eye for an object very far away. An image 30.0 cm from the eye will be 28.5 cm to the left of the spectacle lens (see Figure 2). Therefore, we must get $d_i = -28.5$ cm when $d_o \approx \infty$. The image distance is negative, because it is on the same side of the spectacle as the object.

Solution

Since d_i and d_o are known, the power of the spectacle lens can be found using $P = \frac{1}{d_o} + \frac{1}{d_i}$ as written earlier:

$$P = \frac{1}{d_o} + \frac{1}{d_i} = \frac{1}{\infty} + \frac{1}{-0.285\text{ m}}$$

Since $\frac{1}{\infty} = 0$, we obtain:

$$P = 0 - \frac{3.51}{\text{m}} = -3.51\ \text{D}.$$

Discussion

The negative power indicates a diverging (or concave) lens, as expected. The spectacle produces a case 3 image closer to the eye, where the person can see it. If you examine eyeglasses for nearsighted people, you will find the lenses are thinnest in the center. Additionally, if you examine a prescription for eyeglasses for nearsighted people, you will find that the prescribed power is negative and given in units of diopters.

Since the farsighted eye under converges light rays, the correction for farsightedness is to place a converging spectacle lens in front of the eye. This increases the power of an eye that is too weak. Another way of thinking about this is that a converging spectacle lens produces a case 2 image, which is farther from the eye than the object (see Figure 3). To determine the spectacle power needed for correction, you must know the person's near point—that is, you must know the smallest distance at which the person can see clearly. Then the image produced by a spectacle lens must be at this distance or farther for the farsighted person to be able to see it clearly.

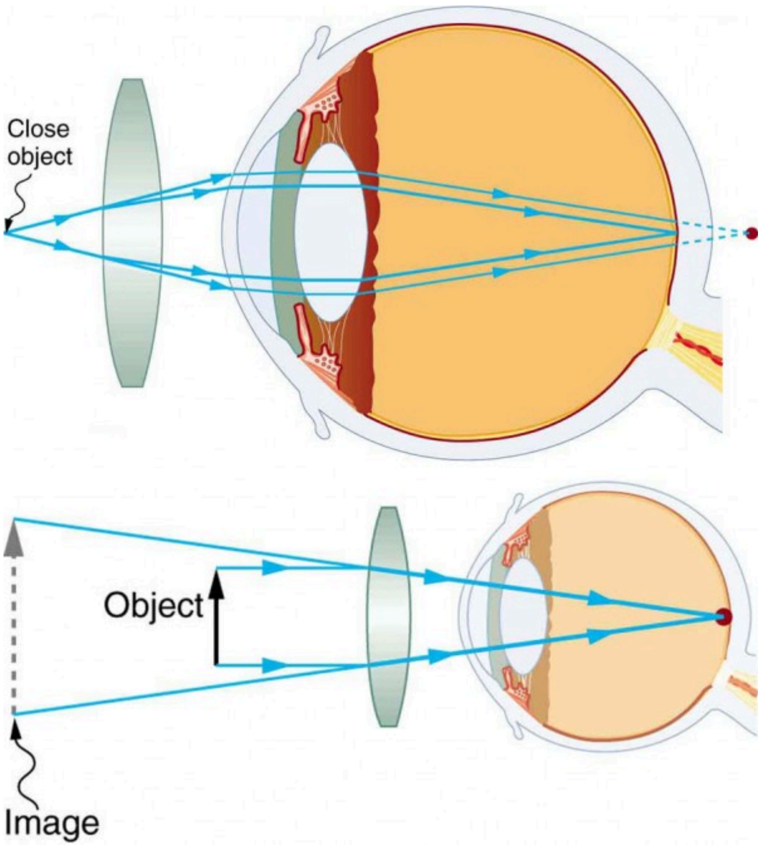


Figure 3. Correction of farsightedness uses a converging lens that compensates for the under convergence by the eye. The converging lens produces an image farther from the eye than the object, so that the farsighted person can see it clearly.

Example 2. Correcting Farsightedness

What power of spectacle lens is needed to allow a farsighted person, whose near point is 1.00 m, to see an object clearly that is 25.0 cm away? Assume the spectacle (corrective) lens is held 1.50 cm away from the eye by eyeglass frames.

Strategy

When an object is held 25.0 cm from the person's eyes, the spectacle lens must produce an image 1.00 m away (the near point). An image 1.00 m from the eye will be 98.5 cm to the left of the spectacle lens because the spectacle lens is 1.50 cm from the eye (see Figure 3). Therefore, d_i=−98.5 cm. The image distance is negative, because it is on the same side of the spectacle as the object. The object is 23.5 cm to the left of the spectacle, so that d_o=23.5 cm.

Solution

Since d_i and d_o are known, the power of the spectacle lens can be found using $P=\dfrac{1}{d_o}+\dfrac{1}{d_i}$:

$$\begin{aligned} P &= \tfrac{1}{d_o}+\tfrac{1}{d_i}=\frac{1}{0.235\text{ m}}+\frac{1}{-0.985\text{ m}} \\ &= 4.26\text{ D}-1.02\text{ D}=3.24\text{ D} \end{aligned}$$

Discussion

The positive power indicates a converging (convex) lens, as expected. The convex spectacle produces a case 2 image farther from the eye, where the person can see it. If you examine eyeglasses of farsighted people, you will find the lenses to be thickest in the center. In addition, a prescription of eyeglasses for farsighted people has a prescribed power that is positive.

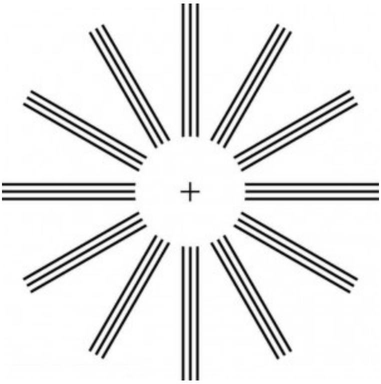


Figure 4. This chart can detect astigmatism, unevenness in the focus of the eye. Check each of your eyes separately by looking at the center cross (without spectacles if you wear them). If lines along some axes appear darker or clearer than others, you have an astigmatism.

Another common vision defect is *astigmatism*, an unevenness or asymmetry in the focus of the eye. For example, rays passing through a vertical region of the eye may focus closer than rays passing through a horizontal region, resulting in the image appearing elongated. This is mostly due to irregularities in the shape of the cornea but can also be due to lens irregularities or unevenness in the retina. Because of these irregularities, different parts of the lens system produce images at different locations. The eye-brain system can compensate for some of these irregularities, but they generally manifest themselves as less distinct vision or sharper images along certain axes. Figure

4 shows a chart used to detect astigmatism. Astigmatism can be at least partially corrected with a spectacle having the opposite irregularity of the eye. If an eyeglass prescription has a cylindrical correction, it is there to correct astigmatism. The normal corrections for short- or farsightedness are spherical corrections, uniform along all axes.

Contact lenses have advantages over glasses beyond their cosmetic aspects. One problem with glasses is that as the eye moves, it is not at a fixed distance from the spectacle lens. Contacts rest on and move with the eye, eliminating this problem. Because contacts cover a significant portion of the cornea, they provide superior peripheral vision compared with eyeglasses. Contacts also correct some corneal astigmatism caused by surface irregularities. The tear layer between the smooth contact and the cornea fills in the irregularities. Since the index of refraction of the tear layer and the cornea are very similar, you now have a regular optical surface in place of an irregular one. If the curvature of a contact lens is not the same as the cornea (as may be necessary with some individuals to obtain a comfortable fit), the tear layer between the contact and cornea acts as a lens. If the tear layer is thinner in the center than at the edges, it has a negative power, for example. Skilled optometrists will adjust the power of the contact to compensate.

Laser vision correction has progressed rapidly in the last few years. It is the latest and by far the most successful in a series of procedures that correct vision by reshaping the cornea. As noted at the beginning of this section, the cornea accounts for about two-thirds of the power of the eye. Thus, small adjustments of its curvature have the same effect as putting a lens in front of the eye. To a reasonable approximation, the power of multiple lenses placed close together equals the sum of their powers. For example, a concave spectacle lens (for nearsightedness) having $P = -3.00\text{ D}$ has the same effect on vision as reducing the power of the eye itself by 3.00 D. So to correct the eye for nearsightedness, the cornea is flattened to reduce its power. Similarly, to correct for farsightedness, the curvature of the cornea is enhanced to increase the power of the eye—the same effect as the positive power spectacle lens used for farsightedness. Laser vision correction uses high intensity electromagnetic radiation to ablate (to remove material from the surface) and reshape the corneal surfaces.

Today, the most commonly used laser vision correction procedure is *Laser in situ Keratomileusis* (LASIK). The top layer of the cornea is surgically peeled back and the underlying tissue ablated by multiple bursts of finely controlled ultraviolet radiation produced by an excimer laser. Lasers are used because they not only produce well-focused intense light, but they also emit very pure wavelength electromagnetic radiation that can be controlled more accurately than mixed wavelength light. The 193 nm wavelength UV commonly used is extremely and strongly absorbed by corneal tissue, allowing precise evaporation of very thin layers. A computer controlled program applies more bursts, usually at a rate of 10 per second, to the areas that require deeper removal. Typically a spot less than 1 mm in diameter and about 0.3 μm in thickness is removed by each burst. Nearsightedness, farsightedness, and astigmatism can be corrected with an accuracy that produces normal distant vision in more than 90% of the patients, in many cases right away. The corneal flap is

Figure 5. Laser vision correction is being performed using the LASIK procedure. Reshaping of the cornea by laser ablation is based on a careful assessment of the patient's vision and is computer controlled. The upper corneal layer is temporarily peeled back and minimally disturbed in LASIK, providing for more rapid and less painful healing of the less sensitive tissues below. (credit: U.S. Navy photo by Mass Communication Specialist 1st Class Brien Aho)

replaced; healing takes place rapidly and is nearly painless. More than 1 million Americans per year undergo LASIK (see Figure 5).

Section Summary

- Nearsightedness, or myopia, is the inability to see distant objects and is corrected with a diverging lens to reduce power.
- Farsightedness, or hyperopia, is the inability to see close objects and is corrected with a converging lens to increase power.
- In myopia and hyperopia, the corrective lenses produce images at a distance that the person can see clearly—the far point and near point, respectively.

Conceptual Questions

1. It has become common to replace the cataract-clouded lens of the eye with an internal lens. This intraocular lens can be chosen so that the person has perfect distant vision. Will the person be able to read without glasses? If the person was nearsighted, is the power of the intraocular lens greater or less than the removed lens?
2. If the cornea is to be reshaped (this can be done surgically or with contact lenses) to correct myopia, should its curvature be made greater or smaller? Explain. Also explain how hyperopia can be corrected.
3. If there is a fixed percent uncertainty in LASIK reshaping of the cornea, why would you expect those

people with the greatest correction to have a poorer chance of normal distant vision after the procedure?

4. A person with presbyopia has lost some or all of the ability to accommodate the power of the eye. If such a person's distant vision is corrected with LASIK, will she still need reading glasses? Explain.

Problems & Exercises

1. What is the far point of a person whose eyes have a relaxed power of 50.5 D?
2. What is the near point of a person whose eyes have an accommodated power of 53.5 D?
3. (a) A laser vision correction reshaping the cornea of a myopic patient reduces the power of his eye by 9.00 D, with a ±5.0% uncertainty in the final correction. What is the range of diopters for spectacle lenses that this person might need after LASIK procedure? (b) Was the person nearsighted or farsighted before the procedure? How do you know?
4. In a LASIK vision correction, the power of a patient's eye is increased by 3.00 D. Assuming this produces normal close vision, what was the patient's near point before the procedure?
5. What was the previous far point of a patient who had laser vision correction that reduced the power of her eye by 7.00 D, producing normal distant vision for her?

6. A severely myopic patient has a far point of 5.00 cm. By how many diopters should the power of his eye be reduced in laser vision correction to obtain normal distant vision for him?
7. A student's eyes, while reading the blackboard, have a power of 51.0 D. How far is the board from his eyes?
8. The power of a physician's eyes is 53.0 D while examining a patient. How far from her eyes is the feature being examined?
9. A young woman with normal distant vision has a 10.0% ability to accommodate (that is, increase) the power of her eyes. What is the closest object she can see clearly?
10. The far point of a myopic administrator is 50.0 cm. (a) What is the relaxed power of his eyes? (b) If he has the normal 8.00% ability to accommodate, what is the closest object he can see clearly?
11. A very myopic man has a far point of 20.0 cm. What power contact lens (when on the eye) will correct his distant vision?
12. Repeat the previous problem for eyeglasses held 1.50 cm from the eyes.
13. A myopic person sees that her contact lens prescription is –4.00 D. What is her far point?
14. Repeat the previous problem for glasses that are 1.75 cm from the eyes.
15. The contact lens prescription for a mildly farsighted person is 0.750 D, and the person has a near point of 29.0 cm. What is the power of the tear layer between the cornea and the lens if the correction is ideal, taking the tear layer into account?
16. A nearsighted man cannot see objects clearly

beyond 20 cm from his eyes. How close must he stand to a mirror in order to see what he is doing when he shaves?

17. A mother sees that her child's contact lens prescription is 0.750 D. What is the child's near point?
18. Repeat the previous problem for glasses that are 2.20 cm from the eyes.
19. The contact lens prescription for a nearsighted person is −4.00 D and the person has a far point of 22.5 cm. What is the power of the tear layer between the cornea and the lens if the correction is ideal, taking the tear layer into account?
20. **Unreasonable Results.** A boy has a near point of 50 cm and a far point of 500 cm. Will a −4.00 D lens correct his far point to infinity?

Glossary

nearsightedness: another term for myopia, a visual defect in which distant objects appear blurred because their images are focused in front of the retina rather than being focused on the retina

myopia: a visual defect in which distant objects appear blurred because their images are focused in front of the retina rather than being focused on the retina

far point: the object point imaged by the eye onto the retina in an unaccommodated eye

farsightedness: another term for hyperopia, the condition of an eye where incoming rays of light reach the retina before they converge into a focused image

hyperopia: the condition of an eye where incoming rays of light reach the retina before they converge into a focused image

near point: the point nearest the eye at which an object is accurately focused on the retina at full accommodation

astigmatism: the result of an inability of the cornea to properly focus an image onto the retina

laser vision correction: a medical procedure used to correct astigmatism and eyesight deficiencies such as myopia and hyperopia

Selected Solutions to Problems & Exercises

1. 2.00 m

3. (a) ±0.45 D; (b) The person was nearsighted because the patient was myopic and the power was reduced.

5. 0.143 m

7. 1.00 m

9. 20.0 cm

11. −5.00 D

13. 25.0 cm

15. −0.198 D

17. 30.8 cm

19. −0.444 D

82. Microscopes

Learning Objectives

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

- Investigate different types of microscopes.
- Learn how image is formed in a compound microscope.

Figure 1. Multiple lenses and mirrors are used in this microscope. (credit: U.S. Navy photo by Tom Watanabe)

Although the eye is marvelous in its ability to see objects large and small, it obviously has limitations to the smallest details it can detect. Human desire to see beyond what is possible with the naked eye led to the use of optical instruments. In this section we will examine microscopes, instruments for enlarging the detail that we cannot see with the unaided eye. The microscope is a multiple-element system having more than a single lens or mirror. (See Figure 1.) A microscope can be made from two convex lenses. The image formed by the first element becomes the object for the second element. The second element forms its own image, which is the object for the third element, and so on. Ray tracing helps to visualize the image formed. If the device is composed of thin lenses and mirrors that obey the thin lens equations, then it is not difficult to describe their behavior numerically.

Microscopes were first developed in the early 1600s by eyeglass makers in The Netherlands and Denmark. The simplest *compound microscope* is constructed from two convex lenses as shown schematically in Figure 2. The first lens is called the *objective lens*, and has typical magnification values from 5× to 100×. In standard microscopes, the objectives are mounted such that when you switch between objectives, the sample remains in focus. Objectives arranged in this way are described as parfocal. The second, the *eyepiece*, also referred to as the ocular, has several lenses which slide inside a cylindrical barrel. The focusing ability is provided by the movement of both the objective lens and the eyepiece. The purpose of a microscope is to magnify small objects, and both lenses contribute to the final magnification. Additionally, the final enlarged image is produced in a location far enough from the observer to be easily viewed, since the eye cannot focus on objects or images that are too close.

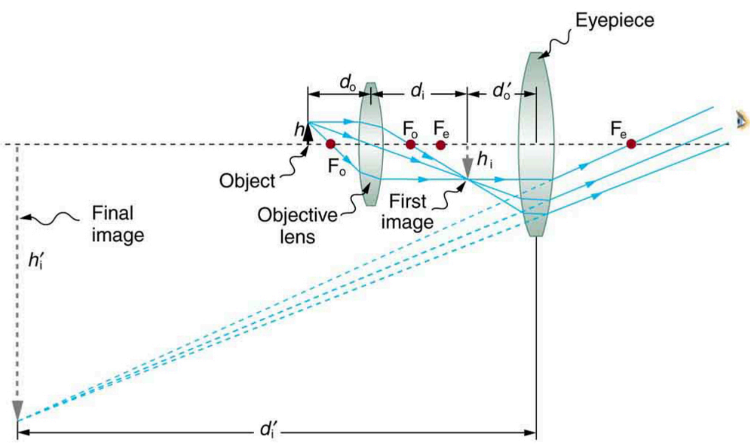


Figure 2. A compound microscope composed of two lenses, an objective and an eyepiece. The objective forms a case 1 image that is larger than the object. This first image is the object for the eyepiece. The eyepiece forms a case 2 final image that is further magnified.

To see how the microscope in Figure 2 forms an image, we consider its two lenses in succession. The object is slightly farther away from the objective lens than its focal length f_o, producing a case 1 image that is larger than the object. This first image is the object for the second lens, or eyepiece. The eyepiece is intentionally located so it can further magnify the image. The eyepiece is placed so that the first image is closer to it than its focal length f_e. Thus the eyepiece acts as a magnifying glass, and the final image is made even larger. The final image remains inverted, but it is farther from the observer, making it easy to view (the eye is most relaxed when viewing distant objects and normally cannot focus closer than 25 cm). Since each lens produces a magnification that multiplies the height of the image, it is apparent that the overall magnification m is the product of the individual magnifications: $m = m_o m_e$, where m_o is the magnification of the objective and m_e is the magnification of the eyepiece. This equation can be generalized for any combination of thin lenses and mirrors that obey the thin lens equations.

Overall Magnification

The overall magnification of a multiple-element system is the product of the individual magnifications of its elements.

Example 1. Microscope Magnification

Calculate the magnification of an object placed 6.20 mm from a compound microscope that has a 6.00 mm focal length objective and a 50.0 mm focal length eyepiece. The objective and eyepiece are separated by 23.0 cm.

Strategy and Concept

This situation is similar to that shown in Figure 2. To find the overall magnification, we must find the magnification of the objective, then the magnification of the eyepiece. This involves using the thin lens equation.

Solution

The magnification of the objective lens is given as

$$m_{\mathrm{o}} = -\frac{d_{\mathrm{i}}}{d_{\mathrm{o}}},$$

where d_{o} and d_{i} are the object and image distances, respectively, for the objective lens as labeled in Figure 2. The object distance is given to be d_{o}=6.20 mm, but the image distance d_{i} is not known. Isolating d_{i}, we have

$$\frac{1}{d_{\mathrm{i}}} = \frac{1}{f_{\mathrm{o}}} - \frac{1}{d_{\mathrm{o}}},$$

where f_{o} is the focal length of the objective lens. Substituting known values gives

$$\frac{1}{d_i} = \frac{1}{6.00 \text{ mm}} - \frac{1}{6.20 \text{ mm}} = \frac{0.00538}{\text{mm}}.$$

We invert this to find d_i: $d_i = 186$ mm.

Substituting this into the expression for m_o gives

$$m_o = -\frac{d_i}{d_o} = -\frac{186 \text{ mm}}{6.20 \text{ mm}} = -30.0.$$

Now we must find the magnification of the eyepiece, which is given by $m_e = -\frac{d_i\prime}{d_o\prime}$, where d_i' and d_o' are the image and object distances for the eyepiece (see Figure 2). The object distance is the distance of the first image from the eyepiece. Since the first image is 186 mm to the right of the objective and the eyepiece is 230 mm to the right of the objective, the object distance is d_o' = 230 mm − 186 mm = 44.0 mm. This places the first image closer to the eyepiece than its focal length, so that the eyepiece will form a case 2 image as shown in the figure. We still need to find the location of the final image d_i' in order to find the magnification. This is done as before to obtain a value for $\frac{1}{d_i\prime}$:

$$\frac{1}{d_i\prime} = \frac{1}{f_e} - \frac{1}{d_o\prime} = \frac{1}{50.0 \text{ mm}} - \frac{1}{44.0 \text{ mm}} = \frac{0.00273}{\text{mm}}$$

Inverting gives

$$d_i\prime = -\frac{\text{mm}}{0.00273} = -367 \text{ mm}.$$

The eyepiece's magnification is thus

$$m_e = -\frac{d_i\prime}{d_o\prime} = -\frac{-367 \text{ mm}}{44.0 \text{ mm}} = 8.33.$$

So the overall magnification is $m = m_o m_e = (-30.0)(8.33) = -250$.

Discussion

Both the objective and the eyepiece contribute to the overall magnification, which is large and negative, consistent with Figure 2, where the image is seen to be large and inverted. In this case, the image is virtual and inverted, which cannot happen for a single element (case 2 and case 3 images for single elements are virtual and upright). The final image is 367 mm (0.367 m) to the left of the eyepiece. Had the eyepiece been placed farther from the objective, it could have formed a case 1 image to the right. Such an image could be projected on a screen, but it would be behind the head of the person in the figure and not appropriate for direct viewing. The procedure used to solve this example is applicable in any multiple-element system. Each element is treated in turn, with each forming an image that becomes the object for the next element. The process is not more difficult than for single lenses or mirrors, only lengthier.

Normal optical microscopes can magnify up to 1500× with a theoretical resolution of −0.2 µm. The lenses can be quite complicated and are composed of multiple elements to reduce aberrations. Microscope objective lenses are particularly important as they primarily gather light from the specimen. Three parameters describe microscope objectives: the *numerical aperture* (NA), the magnification (m), and the working distance. The NA is related to the light gathering ability of a lens and is obtained using the angle of acceptance θ formed by the maximum cone of rays focusing on the specimen (see Figure 3a) and is given by NA = n sin α, where n is the refractive index of the medium between the lens and the specimen and $\alpha = \frac{\theta}{2}$. As the angle of acceptance given by θ increases, NA becomes larger and more light is gathered from a smaller focal region giving higher resolution. A 0.75 NA objective gives more detail than a 0.10 NA objective.

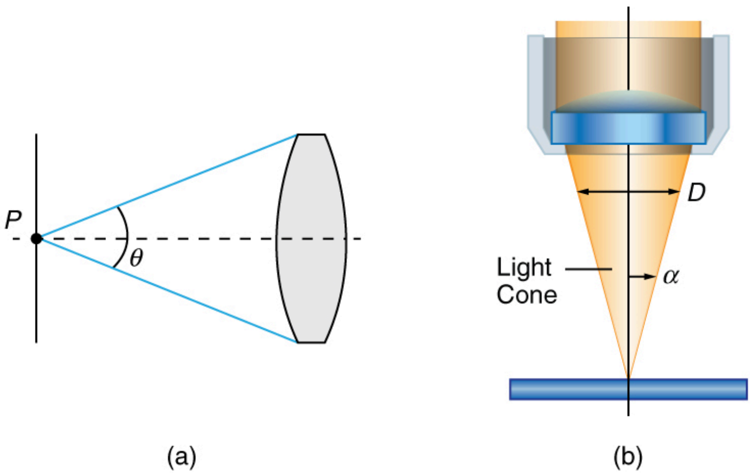


Figure 3. (a) The numerical aperture of a microscope objective lens refers to the light-gathering ability of the lens and is calculated using half the angle of acceptance . (b) Here, is half the acceptance angle for light rays from a specimen entering a camera lens, and is the diameter of the aperture that controls the light entering the lens.

While the numerical aperture can be used to compare resolutions of various objectives, it does not indicate how far the lens could be from the specimen. This is specified by the "working distance," which is the distance (in mm usually) from the front lens element of the objective to the specimen, or cover glass. The higher the NA the closer the lens will be to the specimen and the more chances there are of breaking the cover slip and damaging both the specimen and the lens. The focal length of an objective lens is different than the working distance. This is because objective lenses are made of a combination of lenses and the focal length is measured from inside the barrel. The working distance is a parameter that microscopists can use more readily as it is measured from the outermost lens. The working distance decreases as the NA and magnification both increase.

The term $f/\#$ in general is called the f-number and is used to denote the light per unit area reaching the image plane. In photography, an image of an object at infinity is formed at the focal point and the f-number is given by the ratio of the focal length f of the lens and the diameter D of the aperture controlling the light into the lens (see Figure 3b). If the acceptance angle is small the NA of the lens can also be used as given below.

$$f/\# = \frac{f}{D} \approx \frac{1}{2NA}$$

As the f-number decreases, the camera is able to gather light from a larger angle, giving wide-angle photography. As usual there is a trade-off. A greater $f/\#$ means less light reaches the image plane. A setting of $f/16$ usually allows one to take pictures in bright sunlight as the aperture diameter is small. In optical fibers, light needs to be focused into the fiber. Figure 4 shows the angle used in calculating the NA of an optical fiber.

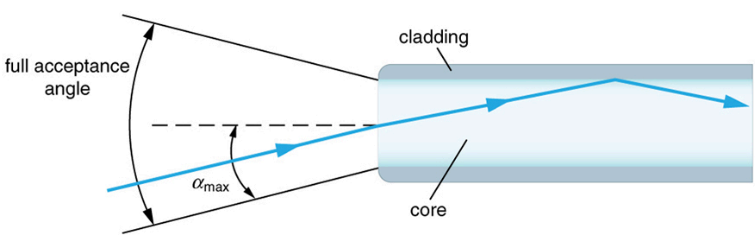


Figure 4. Light rays enter an optical fiber. The numerical aperture of the optical fiber can be determined by using the angle αmax.

Can the NA be larger than 1.00? The answer is 'yes' if we use immersion lenses in which a medium such as oil, glycerine or water is placed between the objective and the microscope cover slip. This minimizes the mismatch in refractive indices as light rays go through different media, generally providing a greater light-gathering ability and an increase in resolution. Figure 5 shows light rays when using air and immersion lenses.

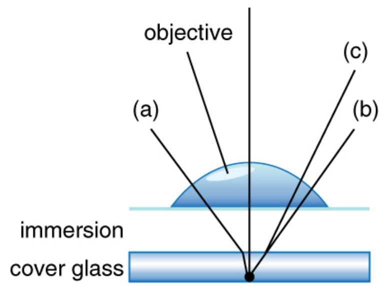


Figure 5. Light rays from a specimen entering the objective. Paths for immersion medium of air (a), water (b) (n = 1.33), and oil (c) (n = 1.51) are shown. The water and oil immersions allow more rays to enter the objective, increasing the resolution.

When using a microscope we do not see the entire extent of the sample. Depending on the eyepiece and objective lens we see a restricted region which we say is the field of view. The objective is then manipulated in two-dimensions above the sample to view other regions of the sample. Electronic scanning of either the objective or the sample is used in scanning microscopy. The image formed at each point during the scanning is combined using a computer to generate an image of a larger region of the sample at a selected magnification.

When using a microscope, we rely on gathering light to form an image. Hence most specimens need to be illuminated, particularly at higher magnifications, when observing details that are so small that they reflect only small amounts of light. To make such objects easily visible, the intensity of light falling on them needs to be increased. Special illuminating systems called condensers are used for this purpose. The type of condenser that is suitable for an application depends on how the specimen is examined, whether by transmission, scattering or reflecting. See Figure 6 for an example of each. White light sources are common and lasers are often used. Laser light illumination tends to be quite intense and it is important to ensure that the light does not result in the degradation of the specimen.

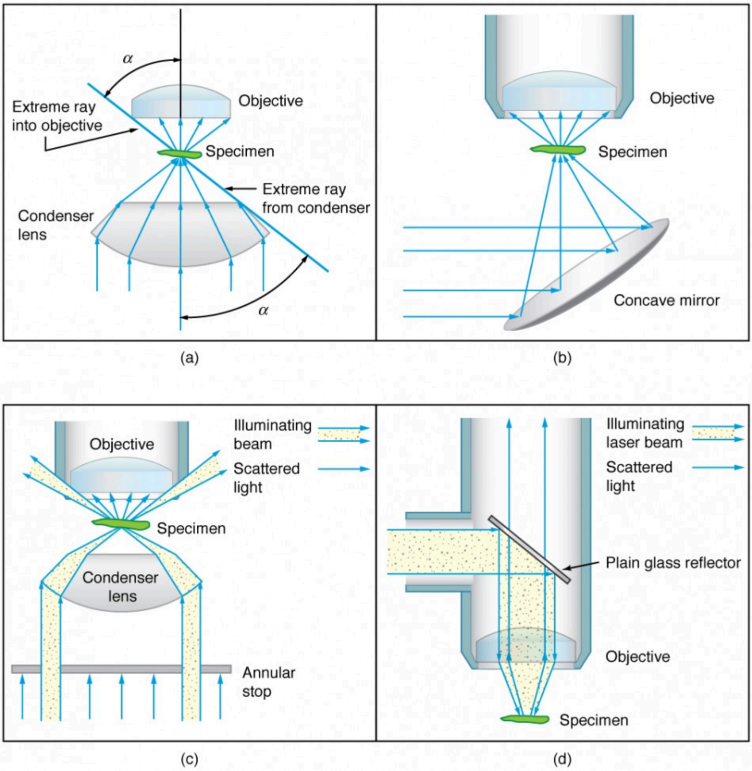


Figure 6. Illumination of a specimen in a microscope. (a) Transmitted light from a condenser lens. (b) Transmitted light from a mirror condenser. (c) Dark field illumination by scattering (the illuminating beam misses the objective lens). (d) High magnification illumination with reflected light – normally laser light.

We normally associate microscopes with visible light but x ray and electron microscopes provide greater resolution. The focusing and basic physics is the same as that just described, even though the lenses require different technology. The electron microscope requires vacuum chambers so that the electrons can proceed unheeded. Magnifications of 50 million times provide the ability to determine positions of individual atoms within materials. An electron microscope is shown in Figure 7.

Figure 7. An electron microscope has the capability to image individual atoms on a material. The microscope uses vacuum technology, sophisticated detectors and state of the art image processing software. (credit: Dave Pape)

We do not use our eyes to form images; rather images are recorded electronically and displayed on computers. In fact observing and saving images formed by optical microscopes on computers is now done routinely. Video recordings of what occurs in a microscope can be made for viewing by many people at later dates. Physics provides the science and tools needed to generate the sequence of time-lapse images of meiosis similar to the sequence sketched in Figure 8.

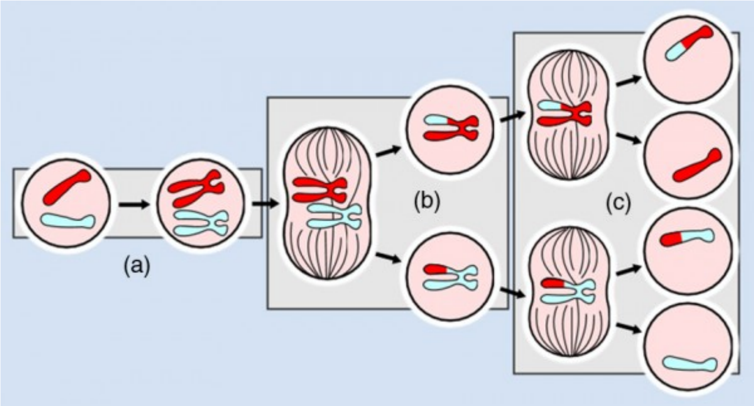


Figure 8. The image shows a sequence of events that takes place during meiosis. (credit: PatríciaR, Wikimedia Commons; National Center for Biotechnology Information)

Take-Home Experiment: Make a Lens

Look through a clear glass or plastic bottle and describe what you see. Now fill the bottle with water and describe what you see. Use the water bottle as a lens to produce the image of a bright object and estimate the focal length of the water bottle lens. How is the focal length a function of the depth of water in the bottle?

Section Summary

- The microscope is a multiple-element system having more than a single lens or mirror.

- Many optical devices contain more than a single lens or mirror. These are analysed by considering each element sequentially. The image formed by the first is the object for the second, and so on. The same ray tracing and thin lens techniques apply to each lens element.
- The overall magnification of a multiple-element system is the product of the magnifications of its individual elements. For a two-element system with an objective and an eyepiece, this is $m = m_o m_e$, where m_o is the magnification of the objective and m_e is the magnification of the eyepiece, such as for a microscope.
- Microscopes are instruments for allowing us to see detail we would not be able to see with the unaided eye and consist of a range of components.
- The eyepiece and objective contribute to the magnification. The numerical aperture NA of an objective is given by $\mathrm{NA} = n \sin \alpha$ where n is the refractive index and α the angle of acceptance.
- Immersion techniques are often used to improve the light gathering ability of microscopes. The specimen is illuminated by transmitted, scattered or reflected light though a condenser.
- The $f/\#$ describes the light gathering ability of a lens. It is given by $f/\# = \dfrac{f}{D} \approx \dfrac{1}{2NA}$.

Conceptual Questions

1. Geometric optics describes the interaction of light with macroscopic objects. Why, then, is it correct to use geometric optics to analyse a microscope's

image?
2. The image produced by the microscope in Figure 2 cannot be projected. Could extra lenses or mirrors project it? Explain.
3. Why not have the objective of a microscope form a case 2 image with a large magnification? (Hint: Consider the location of that image and the difficulty that would pose for using the eyepiece as a magnifier.)
4. What advantages do oil immersion objectives offer?
5. How does the NA of a microscope compare with the NA of an optical fiber?

Problems & Exercises

1. A microscope with an overall magnification of 800 has an objective that magnifies by 200. (a) What is the magnification of the eyepiece? (b) If there are two other objectives that can be used, having magnifications of 100 and 400, what other total magnifications are possible?
2. (a) What magnification is produced by a 0.150 cm focal length microscope objective that is 0.155 cm from the object being viewed? (b) What is the overall magnification if an 8× eyepiece (one that produces a magnification of 8.00) is used?
3. (a) Where does an object need to be placed relative to a microscope for its 0.500 cm focal length

objective to produce a magnification of −400? (b) Where should the 5.00 cm focal length eyepiece be placed to produce a further fourfold (4.00) magnification?

4. You switch from a 1.40 NA 60× oil immersion objective to a 1.40 NA 60× oil immersion objective. What are the acceptance angles for each? Compare and comment on the values. Which would you use first to locate the target area on your specimen?
5. An amoeba is 0.305 cm away from the 0.300 cm focal length objective lens of a microscope. (a) Where is the image formed by the objective lens? (b) What is this image's magnification? (c) An eyepiece with a 2.00 cm focal length is placed 20.0 cm from the objective. Where is the final image? (d) What magnification is produced by the eyepiece? (e) What is the overall magnification? (See Figure 2.)
6. You are using a standard microscope with a 0.10 NA 4× objective and switch to a 0.65 NA 40× objective. What are the acceptance angles for each? Compare and comment on the values. Which would you use first to locate the target area on of your specimen? (See Figure 3.)
7. **Unreasonable Results.** Your friends show you an image through a microscope. They tell you that the microscope has an objective with a 0.500 cm focal length and an eyepiece with a 5.00 cm focal length. The resulting overall magnification is 250,000. Are these viable values for a microscope?

Glossary

compound microscope: a microscope constructed from two convex lenses, the first serving as the ocular lens(close to the eye) and the second serving as the objective lens

objective lens: the lens nearest to the object being examined

eyepiece: the lens or combination of lenses in an optical instrument nearest to the eye of the observer

numerical aperture: a number or measure that expresses the ability of a lens to resolve fine detail in an object being observed. Derived by mathematical formula $NA = n \sin \alpha$, where n is the refractive index of the medium between the lens and the specimen and $\alpha = \frac{\theta}{2}$

Selected Solutions to Problems & Exercises

1. (a) 4.00; (b) 1600

3. (a) 0.501 cm; (b) Eyepiece should be 204 cm behind the objective lens.

5. (a) +18.3 cm (on the eyepiece side of the objective lens); (b) −60.0; (c) −11.3 cm (on the objective side of the eyepiece); (d) +6.67; (e) −400

83. Color and Color Vision

Learning Objectives

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

- Explain the simple theory of color vision.
- Outline the coloring properties of light sources.
- Describe the retinex theory of color vision.

The gift of vision is made richer by the existence of color. Objects and lights abound with thousands of hues that stimulate our eyes, brains, and emotions. Two basic questions are addressed in this brief treatment—what does color mean in scientific terms, and how do we, as humans, perceive it?

Simple Theory of Color Vision

We have already noted that color is associated with the wavelength of visible electromagnetic radiation. When our eyes receive pure-wavelength light, we tend to see only a few colors. Six of these (most often listed) are red, orange, yellow, green, blue, and violet. These are the rainbow of colors produced when white light is dispersed according to different wavelengths. There are thousands of other *hues* that we can perceive. These include brown, teal, gold, pink, and white. One simple theory of color vision implies that all these hues are our eye's response to different combinations of wavelengths.

This is true to an extent, but we find that color perception is even subtler than our eye's response for various wavelengths of light.

The two major types of light-sensing cells (photoreceptors) in the retina are *rods and cones*

Take-Home Experiment: Rods and Cones

1. Go into a darkened room from a brightly lit room, or from outside in the Sun. How long did it take to start seeing shapes more clearly? What about color? Return to the bright room. Did it take a few minutes before you could see things clearly?
2. Demonstrate the sensitivity of foveal vision. Look at the letter G in the word ROGERS. What about the clarity of the letters on either side of G?

Cones are most concentrated in the fovea, the central region of the retina. There are no rods here. The fovea is at the center of the macula, a 5 mm diameter region responsible for our central vision. The cones work best in bright light and are responsible for high resolution vision. There are about 6 million cones in the human retina. There are three types of cones, and each type is sensitive to different ranges of wavelengths, as illustrated in Figure 1.

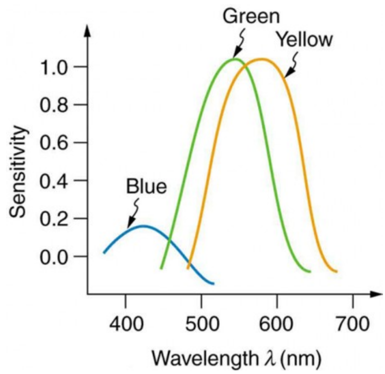


Figure 1. The image shows the relative sensitivity of the three types of cones, which are named according to wavelengths of greatest sensitivity. Rods are about 1000 times more sensitive, and their curve peaks at about 500 nm. Evidence for the three types of cones comes from direct measurements in animal and human eyes and testing of color blind people.

A *simplified theory of color vision* is that there are three *primary colors* corresponding to the three types of cones. The thousands of other hues that we can distinguish among are created by various combinations of stimulations of the three types of cones. Color television uses a three-color system in which the screen is covered with equal numbers of red, green, and blue phosphor dots. The broad range of hues a viewer sees is produced by various combinations of these three colors. For example, you will perceive yellow when red and green are illuminated with the correct ratio of intensities. White may be sensed when all three are illuminated. Then, it would seem that all hues can be produced by adding three primary colors in various proportions. But there is an indication that color vision is more sophisticated. There is no unique set of three primary colors. Another set that works is yellow, green, and blue. A further indication of the need for a more complex theory of color vision is that various different combinations can produce the same hue. Yellow can be sensed with yellow light, or with a combination of red and green, and also with white light from which violet has been removed. The three-primary-colors aspect of color vision is well established; more sophisticated theories expand on it rather than deny it.

Consider why various objects display color—that is, why are feathers blue and red in a crimson rosella? The *true color of an object* is defined by its absorptive or reflective characteristics. Figure 2 shows white light falling on three different objects, one pure blue, one pure red, and one black, as well as pure red light falling on a white object. Other hues are created by more complex absorption characteristics. Pink, for example on a galah cockatoo, can be due to weak absorption of all colors except red. An object can appear a different color under non-white illumination. For example, a pure blue object illuminated with pure red light will *appear* black, because it absorbs all the red light falling on it. But, the true color of the object is blue, which is independent of illumination.

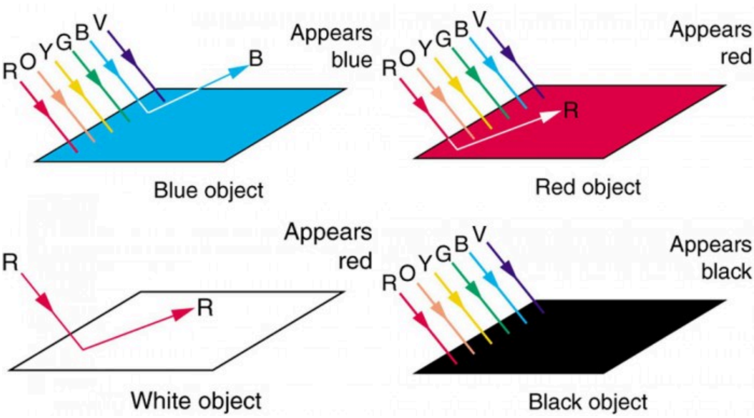


Figure 2. Absorption characteristics determine the true color of an object. Here, three objects are illuminated by white light, and one by pure red light. White is the equal mixture of all visible wavelengths; black is the absence of light.

Similarly, *light sources have colors* that are defined by the wavelengths they produce. A helium-neon laser emits pure red light. In fact, the phrase "pure red light" is defined by having a sharp constrained spectrum, a characteristic of laser light. The Sun produces a broad yellowish spectrum, fluorescent lights emit bluish-white light, and incandescent lights emit reddish-white hues as seen in Figure 3. As you would expect, you sense these colors when viewing the light source directly or when illuminating a white object with them. All of this fits neatly into the simplified theory that a combination of wavelengths produces various hues.

Take-Home Experiment: Exploring Color Addition

This activity is best done with plastic sheets of different colors as they allow more light to pass through to our eyes. However, thin sheets of paper and fabric can also be used. Overlay different colors of the material and hold them up to a white light. Using the theory described above, explain the colors you observe. You could also try mixing different crayon colors.

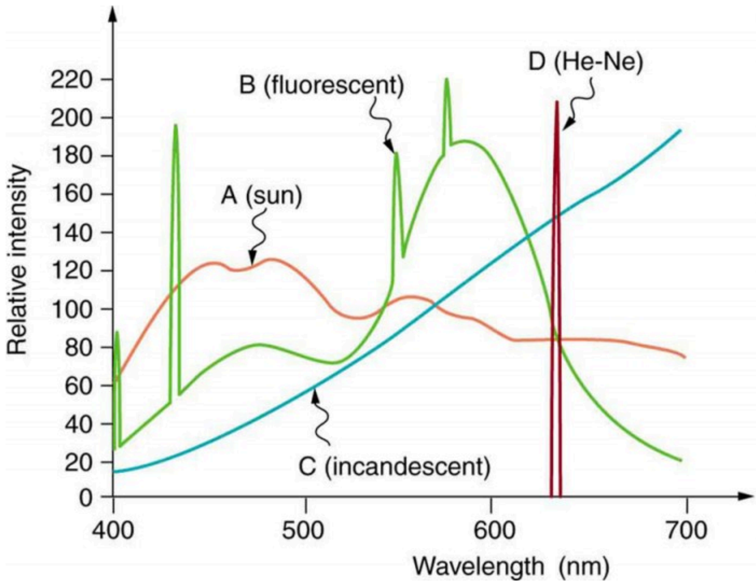


Figure 3. Emission spectra for various light sources are shown. Curve A is average sunlight at Earth's surface, curve B is light from a fluorescent lamp, and curve C is the output of an incandescent light. The spike for a helium-neon laser (curve D) is due to its pure wavelength emission. The spikes in the fluorescent output are due to atomic spectra—a topic that will be explored later.

Color Constancy and a Modified Theory of Color Vision

The eye-brain color-sensing system can, by comparing various objects in its view, perceive the true color of an object under varying lighting conditions—an ability that is called *color constancy*. We can sense that a white tablecloth, for example, is white whether it is illuminated by sunlight, fluorescent light, or candlelight. The wavelengths entering the eye are quite different in each case, as the graphs in Figure 3 imply, but our color vision can detect the true color by comparing the tablecloth with its surroundings.

Theories that take color constancy into account are based on a large body of anatomical evidence as well as perceptual studies. There are nerve connections among the light receptors on the retina, and there are far fewer nerve connections to the brain than there are rods and cones. This means that there is signal processing in the eye before information is sent to the brain. For example, the eye makes comparisons between adjacent light receptors and is very sensitive to edges as seen in Figure 4. Rather than responding simply to the light entering the eye, which is uniform in the various rectangles in this figure, the eye responds to the edges and senses false darkness variations.

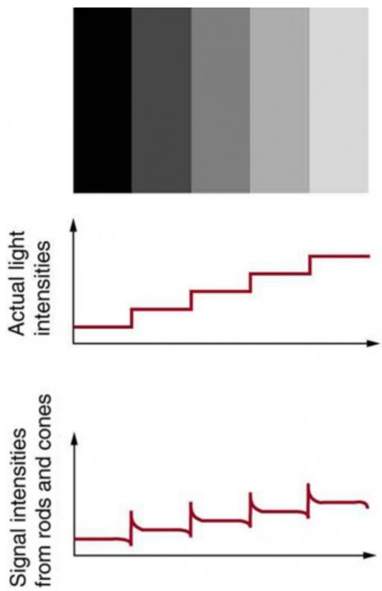
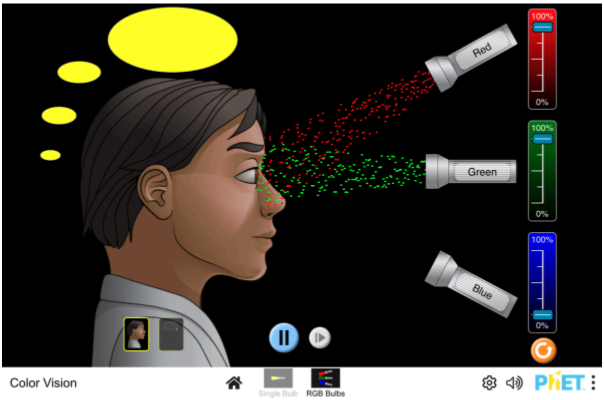


Figure 4. The importance of edges is shown. Although the grey strips are uniformly shaded, as indicated by the graph immediately below them, they do not appear uniform at all. Instead, they are perceived darker on the dark side and lighter on the light side of the edge, as shown in the bottom graph. This is due to nerve impulse processing in the eye.

One theory that takes various factors into account was advanced by Edwin Land (1909–1991), the creative founder of the Polaroid Corporation. Land proposed, based partly on his many elegant experiments, that the three types of cones are organized into systems called *retinexes*. Each retinex forms an image that is compared with the others, and the eye-brain system thus can compare a candle-illuminated white table cloth with its generally reddish surroundings and determine that it is actually white. This *retinex theory of color vision* is an example of modified theories of color vision that attempt to account for its subtleties. One striking experiment performed by Land demonstrates that some type of image comparison may produce color vision. Two pictures are taken of a scene on black-and-white film, one using a red filter, the other a blue filter. Resulting black-and-white slides are then projected and superimposed on a screen, producing a black-and-white image, as expected. Then a red filter is placed in front of the slide taken with a red filter, and the images are again superimposed on a screen. You would expect an image in various shades of pink, but instead, the image appears to humans in full color with all the hues of the original scene. This implies that color vision can be induced by comparison of the black-and-white and red images. Color vision is not completely understood or explained, and the retinex theory is not totally accepted. It is apparent that color vision is much subtler than what a first look might imply.

PhET Explorations: Color Vision

Make a whole rainbow by mixing red, green, and blue light. Change the wavelength of a monochromatic beam or filter white light. View the light as a solid beam, or see the individual photons.



Click to run the simulation.

Section Summary

- The eye has four types of light receptors—rods and three types of color-sensitive cones.
- The rods are good for night vision, peripheral vision, and motion changes, while the cones are responsible for central vision and color.
- We perceive many hues, from light having mixtures of wavelengths.
- A simplified theory of color vision states that there are three primary colors, which correspond to the three types of cones, and that various combinations of the primary colors produce all the hues.
- The true color of an object is related to its relative absorption

of various wavelengths of light. The color of a light source is related to the wavelengths it produces.

- Color constancy is the ability of the eye-brain system to discern the true color of an object illuminated by various light sources.
- The retinex theory of color vision explains color constancy by postulating the existence of three retinexes or image systems, associated with the three types of cones that are compared to obtain sophisticated information.

Conceptual Questions

1. A pure red object on a black background seems to disappear when illuminated with pure green light. Explain why.
2. What is color constancy, and what are its limitations?
3. There are different types of color blindness related to the malfunction of different types of cones. Why would it be particularly useful to study those rare individuals who are color blind only in one eye or who have a different type of color blindness in each eye?
4. Propose a way to study the function of the rods alone, given they can sense light about 1000 times dimmer than the cones.

Glossary

hues: identity of a color as it relates specifically to the spectrum

rods and cones: two types of photoreceptors in the human retina; rods are responsible for vision at low light levels, while cones are active at higher light levels

simplified theory of color vision: a theory that states that there are three primary colors, which correspond to the three types of cones

color constancy: a part of the visual perception system that allows people to perceive color in a variety of conditions and to see some consistency in the color

retinex: a theory proposed to explain color and brightness perception and constancies; is a combination of the words retina and cortex, which are the two areas responsible for the processing of visual information

retinex theory of color vision: the ability to perceive color in an ambient-colored environment

84. Aberrations

Learning Objective

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

- Describe optical aberration.

Real lenses behave somewhat differently from how they are modeled using the thin lens equations, producing *aberrations*. An aberration is a distortion in an image. There are a variety of aberrations due to a lens size, material, thickness, and position of the object. One common type of aberration is chromatic aberration, which is related to color. Since the index of refraction of lenses depends on color or wavelength, images are produced at different places and with different magnifications for different colors. (The law of reflection is independent of wavelength, and so mirrors do not have this problem. This is another advantage for mirrors in optical systems such as telescopes.)

Figure 1a shows chromatic aberration for a single convex lens and its partial correction with a two-lens system. Violet rays are bent more than red, since they have a higher index of refraction and are thus focused closer to the lens. The diverging lens partially corrects this, although it is usually not possible to do so completely. Lenses of different materials and having different dispersions may be used. For example an achromatic doublet consisting of a converging lens made of crown glass and a diverging lens made of flint glass in contact can dramatically reduce chromatic aberration (see Figure 1b).

Quite often in an imaging system the object is off-center. Consequently, different parts of a lens or mirror do not refract or reflect the image to the same point. This type of aberration is called a coma and is shown in Figure 2. The image in this case often appears pear-shaped. Another common aberration is spherical aberration where rays converging from the outer edges of a lens converge to a focus closer to the lens and rays closer to the axis focus further (see Figure 3). Aberrations due to astigmatism in the lenses of the eyes are discussed in [Vision Correction](), and a chart used to detect astigmatism is shown in Figure 4. Such aberrations and can also be an issue with manufactured lenses.

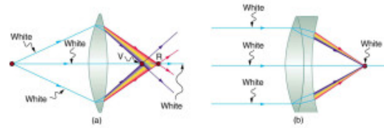


Figure 1. (a) Chromatic aberration is caused by the dependence of a lens's index of refraction on color (wavelength). The lens is more powerful for violet (V) than for red (R), producing images with different locations and magnifications. (b) Multiple-lens systems can partially correct chromatic aberrations, but they may require lenses of different materials and add to the expense of optical systems such as cameras.

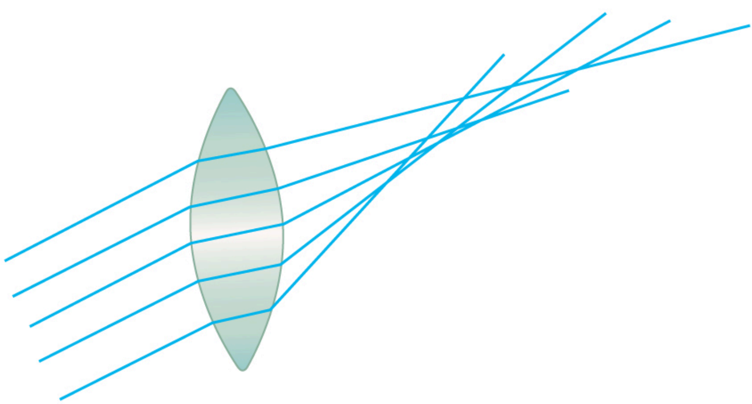


Figure 2. A coma is an aberration caused by an object that is off-center, often resulting in a pear-shaped image. The rays originate from points that are not on the optical axis and they do not converge at one common focal point.

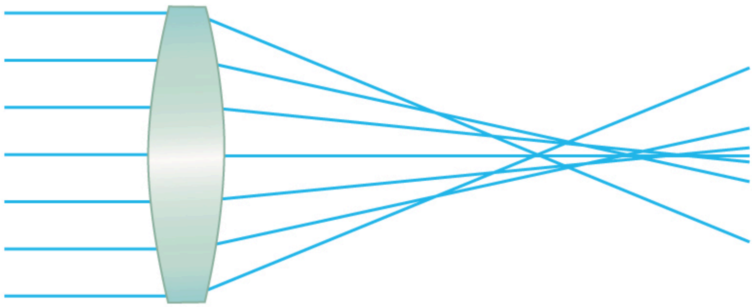


Figure 3. Spherical aberration is caused by rays focusing at different distances from the lens.

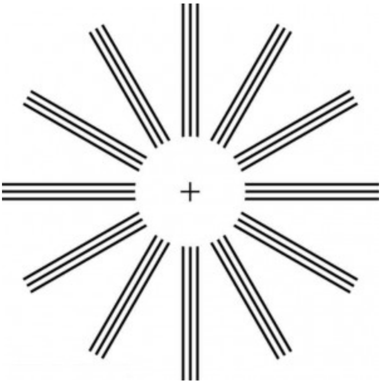


Figure 4. This chart can detect astigmatism, unevenness in the focus of the eye. Check each of your eyes separately by looking at the center cross (without spectacles if you wear them). If lines along some axes appear darker or clearer than others, you have an astigmatism.

The image produced by an optical system needs to be bright enough to be discerned. It is often a challenge to obtain a sufficiently bright image. The brightness is determined by the amount of light passing through the optical system. The optical components determining the brightness are the diameter of the lens and the diameter of pupils, diaphragms or aperture stops placed in front of lenses. Optical systems often have entrance and exit pupils to specifically reduce aberrations but they inevitably reduce brightness as well. Consequently, optical systems need to strike a balance between the various components used. The iris in the eye dilates and constricts, acting as an entrance pupil. You can see objects more clearly by looking through a small hole made with your hand in the shape of a fist. Squinting, or using a small hole in a piece of paper, also will make the object sharper.

So how are aberrations corrected? The lenses may also have specially shaped surfaces, as opposed to the simple spherical shape that is relatively easy to produce. Expensive camera lenses are large in diameter, so that they can gather more light, and need several elements to correct for various aberrations. Further, advances in materials science have resulted in lenses with a range of refractive indices—technically referred to as graded index (GRIN) lenses. Spectacles often have the ability to provide a range of focusing ability using similar techniques. GRIN lenses are particularly important at the end of optical fibers in endoscopes. Advanced computing techniques allow for a range of corrections on images after the image has been collected and certain characteristics of the optical system are known. Some of these techniques are sophisticated versions of what are available on commercial packages like Adobe Photoshop.

Section Summary

- Aberrations or image distortions can arise due to the finite thickness of optical instruments, imperfections in the optical components, and limitations on the ways in which the components are used.
- The means for correcting aberrations range from better components to computational techniques.

Conceptual Questions

1. List the various types of aberrations. What causes them and how can each be reduced?

Problems & Exercises

Integrated Concepts. (a) During laser vision correction, a brief burst of 193 nm ultraviolet light is projected onto the cornea of the patient. It makes a spot 1.00 mm in diameter and deposits 0.500 mJ of energy. Calculate the depth of the layer ablated, assuming the corneal tissue has the same properties as water and is initially at 34.0°C. The tissue's temperature is increased to 100°C and evaporated without further temperature increase.
(b) Does your answer imply that the shape of the cornea can be finely controlled?

Glossary

aberration: failure of rays to converge at one focus because of limitations or defects in a lens or mirror

Solutions to Problems & Exercises

(a) 0.251 µm; (b) Yes, this thickness implies that the shape of the cornea can be very finely controlled, producing normal distant vision in more than 90% of patients.

PART XI

WAVE OPTICS

85. Introduction to Wave Optics

Figure 1. The colors reflected by this compact disc vary with angle and are not caused by pigments. Colors such as these are direct evidence of the wave character of light. (credit: Infopro, Wikimedia Commons)

Examine a compact disc under white light, noting the colors observed and locations of the colors. Determine if the spectra are formed by diffraction from circular lines centered at the middle of the disc and, if so, what is their spacing. If not, determine the type of spacing. Also with the CD, explore the spectra of a few light sources, such as a candle flame, incandescent bulb, halogen light, and fluorescent light. Knowing the spacing of the rows of pits in the compact disc, estimate the maximum spacing that will allow the given number of megabytes of information to be stored.

If you have ever looked at the reds, blues, and greens in a sunlit soap bubble and wondered how straw-colored soapy water could produce them, you have hit upon one of the many phenomena that can only be explained by the wave character of light (see Figure 2). The same is true for the colors seen in an oil slick or in the light reflected from a compact disc. These and other interesting phenomena, such as the dispersion of white light into a rainbow of colors when passed through a narrow slit, cannot be explained fully by geometric optics. In these cases, light interacts with small objects and exhibits its wave characteristics. The branch of optics that considers the behavior of light when it exhibits wave characteristics (particularly when it interacts with small objects) is called wave optics (sometimes called physical optics). It is the topic of this chapter.

Figure 2. These soap bubbles exhibit brilliant colors when exposed to sunlight. How are the colors produced if they are not pigments in the soap? (credit: Scott Robinson, Flickr)

86. The Wave Aspect of Light: Interference

Learning Objectives

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

- Discuss the wave character of light.
- Identify the changes when light enters a medium.

We know that visible light is the type of electromagnetic wave to which our eyes respond. Like all other electromagnetic waves, it obeys the equation $c = f\lambda$, where $c = 3 \times 10^8$ m/s is the speed of light in vacuum, f is the frequency of the electromagnetic waves, and λ is its wavelength. The range of visible wavelengths is approximately 380 to 760 nm. As is true for all waves, light travels in straight lines and acts like a ray when it interacts with objects several times as large as its wavelength. However, when it interacts with smaller objects, it displays its wave characteristics prominently. Interference is the hallmark of a wave, and in Figure 1 both the ray and wave characteristics of light can be seen. The laser beam emitted by the observatory epitomizes a ray, traveling in a straight line. However, passing a pure-wavelength beam through vertical slits with a size close to the wavelength of the beam reveals the wave character of light, as the beam spreads out horizontally into a pattern of bright and dark regions caused by systematic constructive and destructive interference. Rather than spreading out, a ray would continue traveling straight ahead after passing through slits.

Making Connections: Waves

The most certain indication of a wave is interference. This wave characteristic is most prominent when the wave interacts with an object that is not large compared with the wavelength. Interference is observed for water waves, sound waves, light waves, and (as we will see in Special Relativity) for matter waves, such as electrons scattered from a crystal.

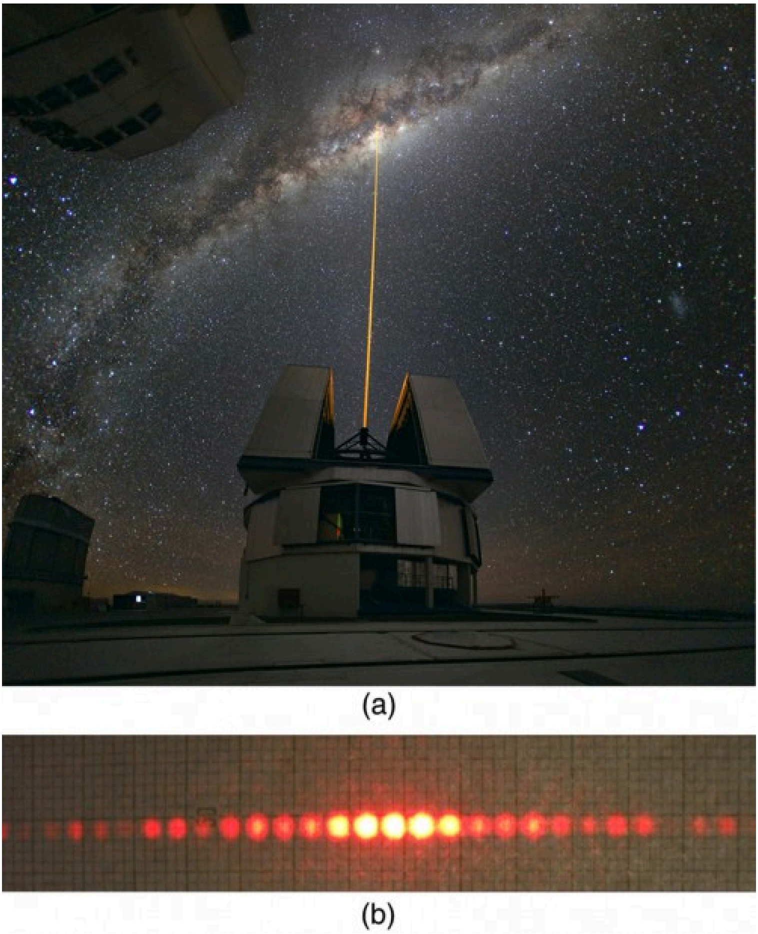


Figure 1. (a) The laser beam emitted by an observatory acts like a ray, traveling in a straight line. This laser beam is from the Paranal Observatory of the European Southern Observatory. (credit: Yuri Beletsky, European Southern Observatory) (b) A laser beam passing through a grid of vertical slits produces an interference pattern—characteristic of a wave. (credit: Shim'on and Slava Rybka, Wikimedia Commons)

Light has wave characteristics in various media as well as in a vacuum. When light goes from a vacuum to some medium, like water, its speed and wavelength change, but its frequency f remains the same. (We can think of light as a forced oscillation that must have the frequency of the original source.) The speed of light in a medium is $v = \frac{c}{n}$, where n is its index of refraction. If we divide both sides of equation $c = f\lambda$ by n, we get $\frac{c}{n} = v = \frac{f\lambda}{n}$. This implies that $v = f\lambda_n$, where λ_n is the *wavelength in a medium* and that $\lambda_n = \frac{\lambda}{n}$, where λ is the wavelength in vacuum and n is the medium's index of refraction. Therefore, the wavelength of light is smaller in any medium than it is in vacuum. In water, for example, which has $n = 1.333$, the range of visible wavelengths is $\frac{380\text{ nm}}{1.333}$ to $\frac{760\text{ nm}}{1.333}$, or $\lambda_n = 285$ to 570 nm. Although wavelengths change while traveling from one medium to another, colors do not, since colors are associated with frequency.

Section Summary

- Wave optics is the branch of optics that must be used when light interacts with small objects or whenever the wave characteristics of light are considered.
- Wave characteristics are those associated with interference and diffraction.
- Visible light is the type of electromagnetic wave to which our eyes respond and has a wavelength in the range of 380 to 760 nm.
- Like all EM waves, the following relationship is valid in vacuum: $c = f\lambda$, where $c = 3 \times 10^8$ m/s is the speed of light, f is the frequency of the electromagnetic wave, and λ is its wavelength in vacuum.
- The wavelength λ_n of light in a medium with index of refraction

n is $\lambda_n = \frac{\lambda}{n}$. Its frequency is the same as in vacuum.

Conceptual Questions

1. What type of experimental evidence indicates that light is a wave?
2. Give an example of a wave characteristic of light that is easily observed outside the laboratory.

Problems & Exercises

1. Show that when light passes from air to water, its wavelength decreases to 0.750 times its original value.
2. Find the range of visible wavelengths of light in crown glass.
3. What is the index of refraction of a material for which the wavelength of light is 0.671 times its value in a vacuum? Identify the likely substance.
4. Analysis of an interference effect in a clear solid shows that the wavelength of light in the solid is 329 nm. Knowing this light comes from a He-Ne laser and has a wavelength of 633 nm in air, is the substance zircon or diamond?
5. What is the ratio of thicknesses of crown glass and water that would contain the same number of

wavelengths of light?

Glossary

wavelength in a medium: {\lambda }_{\text{n}}=\text{\lambda}{n}\\, where λ is the wavelength in vacuum, and n is the index of refraction of the medium

Selected Solutions to Problems & Exercises

1. $$\frac{1}{1.333}=0.750$$

3. 1.49, Polystyrene

5. 0.877 glass to water

87. Huygens's Principle: Diffraction

Learning Objectives

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

- Discuss the propagation of transverse waves.
- Discuss Huygens's principle.
- Explain the bending of light.

Figure 1 shows how a transverse wave looks as viewed from above and from the side. A light wave can be imagined to propagate like this, although we do not actually see it wiggling through space. From above, we view the wavefronts (or wave crests) as we would by looking down on the ocean waves. The side view would be a graph of the electric or magnetic field. The view from above is perhaps the most useful in developing concepts about wave optics.

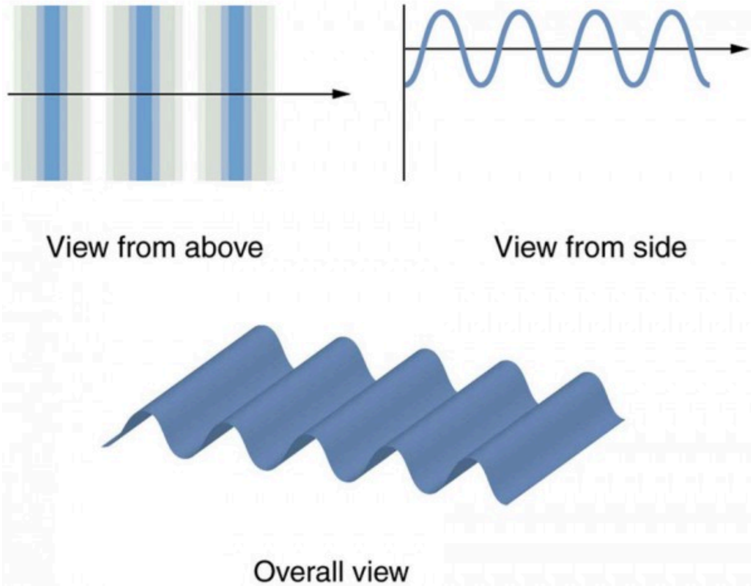


Figure 1. A transverse wave, such as an electromagnetic wave like light, as viewed from above and from the side. The direction of propagation is perpendicular to the wavefronts (or wave crests) and is represented by an arrow like a ray.

The Dutch scientist Christiaan Huygens (1629–1695) developed a useful technique for determining in detail how and where waves propagate. Starting from some known position, *Huygens's principle* states that:

Every point on a wavefront is a source of wavelets that spread out in the forward direction at the same speed as the wave itself. The new wavefront is a line tangent to all of the wavelets.

Figure 2 shows how Huygens's principle is applied. A wavefront is the long edge that moves, for example, the crest or the trough. Each point on the wavefront emits a semicircular wave that moves at the propagation speed v. These are drawn at a time t later, so that they have moved a distance $s = vt$. The new wavefront is a line tangent to the wavelets and is where we would expect the wave to be a time t later. Huygens's principle works for all types of waves, including water waves, sound waves, and light waves. We will find it useful not only in describing how light waves propagate, but also in explaining the laws of reflection and refraction. In addition, we will see that Huygens's principle tells us how and where light rays interfere.

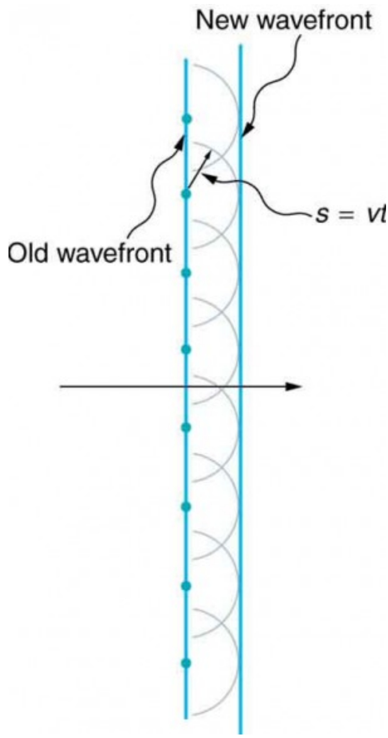


Figure 2. Huygens's principle applied to a straight wavefront. Each point on the wavefront emits a semicircular wavelet that moves a distance . The new wavefront is a line tangent to the wavelets.

Figure 3 shows how a mirror reflects an incoming wave at an angle equal to the incident angle, verifying the law of reflection. As the wavefront strikes the mirror, wavelets are first emitted from the left part of the mirror and then the right. The wavelets closer to the left have had time to travel farther, producing a wavefront traveling in the direction shown.

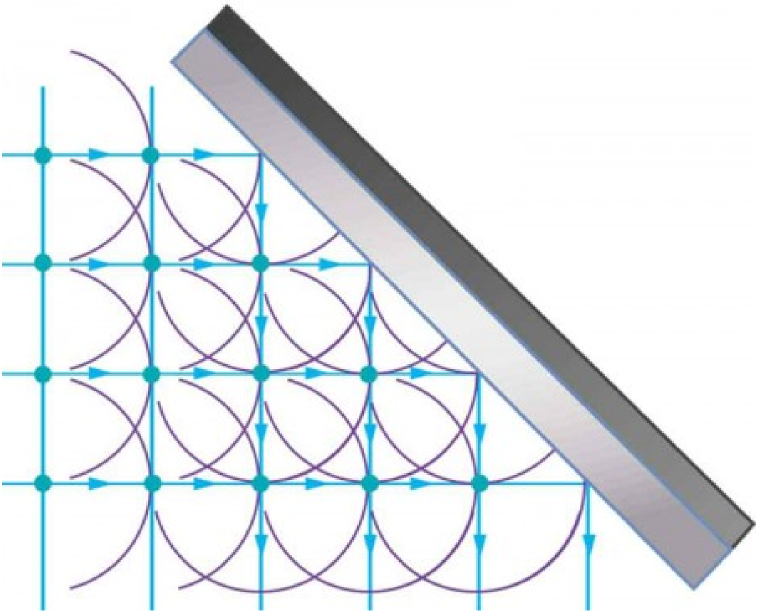


Figure 3. Huygens's principle applied to a straight wavefront striking a mirror. The wavelets shown were emitted as each point on the wavefront struck the mirror. The tangent to these wavelets shows that the new wavefront has been reflected at an angle equal to the incident angle. The direction of propagation is perpendicular to the wavefront, as shown by the downward-pointing arrows.

The law of refraction can be explained by applying Huygens's principle to a wavefront passing from one medium to another (see Figure 4). Each wavelet in the figure was emitted when the wavefront crossed the interface between the media. Since the speed of light is smaller in the second medium, the waves do not travel as far in a given time, and the new wavefront changes direction as shown. This explains why a ray changes direction to become closer to the perpendicular when light slows down. Snell's law can be derived from the geometry in Figure 4, but this is left as an exercise for ambitious readers.

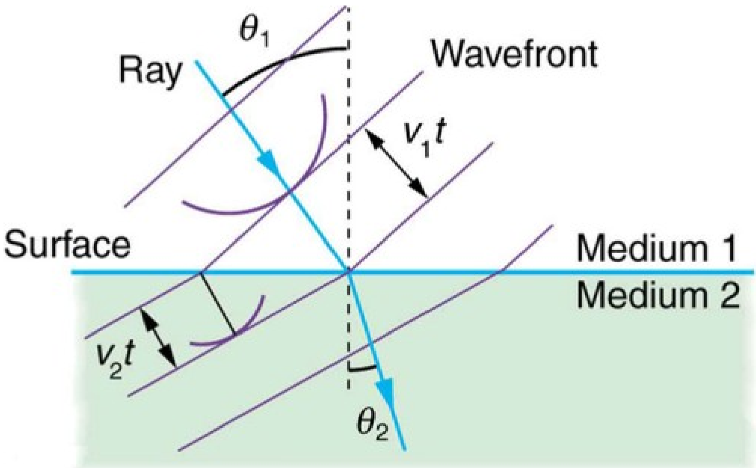


Figure 4. Huygens's principle applied to a straight wavefront traveling from one medium to another where its speed is less. The ray bends toward the perpendicular, since the wavelets have a lower speed in the second medium.

What happens when a wave passes through an opening, such as light shining through an open door into a dark room? For light, we expect to see a sharp shadow of the doorway on the floor of the room, and we expect no light to bend around corners into other parts of the room. When sound passes through a door, we expect to hear it everywhere in the room and, thus, expect that sound spreads out when passing through such an opening (see Figure 5). What is the difference between the behavior of sound waves and light waves in this case? The answer is that light has very short wavelengths and acts like a ray. Sound has wavelengths on the order of the size of the door and bends around corners (for frequency of 1000 Hz, $\lambda = \frac{c}{f} = \frac{330 \text{ m/s}}{1000 \text{ s}^{-1}} = 0.33 \text{ m}$, about three times smaller than the width of the doorway).

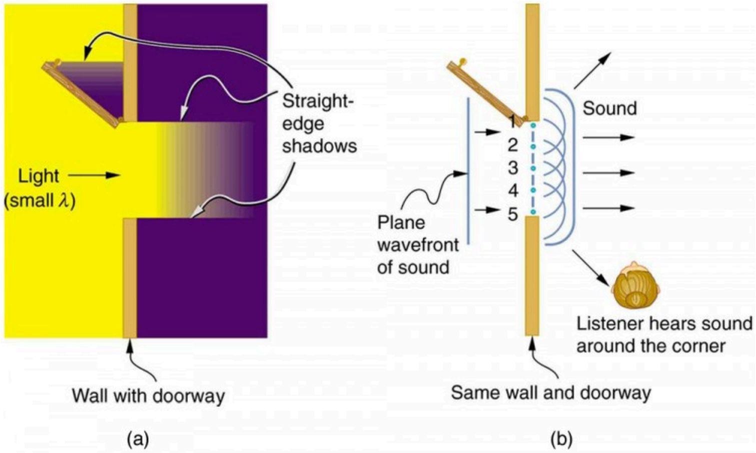


Figure 5. (a) Light passing through a doorway makes a sharp outline on the floor. Since light's wavelength is very small compared with the size of the door, it acts like a ray. (b) Sound waves bend into all parts of the room, a wave effect, because their wavelength is similar to the size of the door.

If we pass light through smaller openings, often called slits, we can use Huygens's principle to see that light bends as sound does (see Figure 6). The bending of a wave around the edges of an opening or an obstacle is called *diffraction*. Diffraction is a wave characteristic and occurs for all types of waves. If diffraction is observed for some phenomenon, it is evidence that the phenomenon is a wave. Thus the horizontal diffraction of the laser beam after it passes through slits in Figure 7 is evidence that light is a wave.

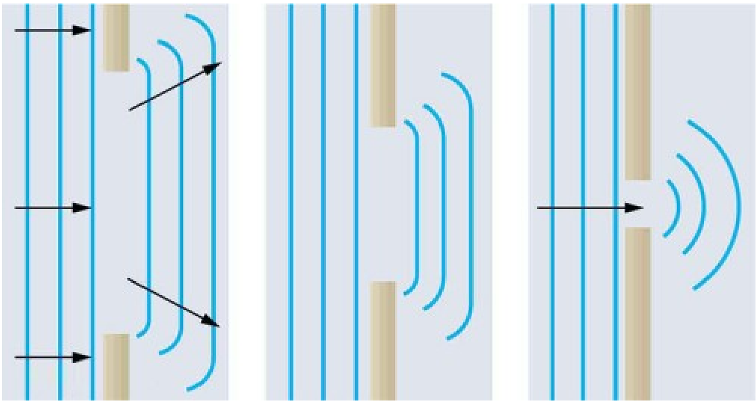


Figure 6. Huygens's principle applied to a straight wavefront striking an opening. The edges of the wavefront bend after passing through the opening, a process called diffraction. The amount of bending is more extreme for a small opening, consistent with the fact that wave characteristics are most noticeable for interactions with objects about the same size as the wavelength.

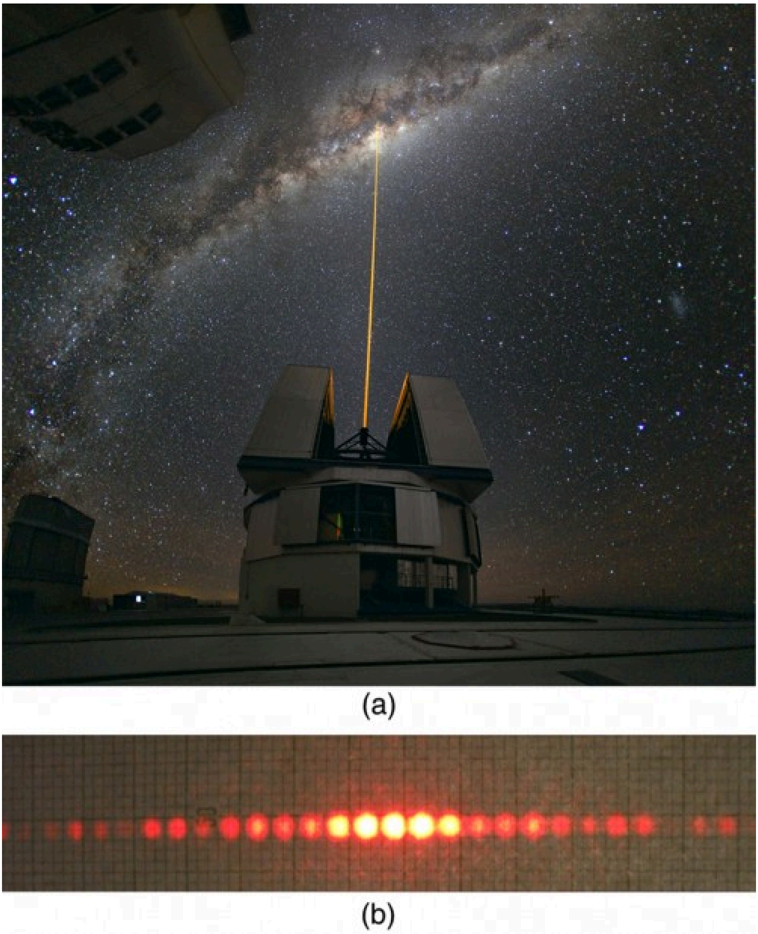


Figure 7. (a) The laser beam emitted by an observatory acts like a ray, traveling in a straight line. This laser beam is from the Paranal Observatory of the European Southern Observatory. (credit: Yuri Beletsky, European Southern Observatory) (b) A laser beam passing through a grid of vertical slits produces an interference pattern—characteristic of a wave. (credit: Shim'on and Slava Rybka, Wikimedia Commons)

Section Summary

- An accurate technique for determining how and where waves propagate is given by Huygens's principle: Every point on a wavefront is a source of wavelets that spread out in the forward direction at the same speed as the wave itself. The new wavefront is a line tangent to all of the wavelets.
- Diffraction is the bending of a wave around the edges of an opening or other obstacle.

Conceptual Questions

1. How do wave effects depend on the size of the object with which the wave interacts? For example, why does sound bend around the corner of a building while light does not?
2. Under what conditions can light be modeled like a ray? Like a wave?
3. Go outside in the sunlight and observe your shadow. It has fuzzy edges even if you do not. Is this a diffraction effect? Explain.
4. Why does the wavelength of light decrease when it passes from vacuum into a medium? State which attributes change and which stay the same and, thus, require the wavelength to decrease.
5. Does Huygens's principle apply to all types of waves?

Glossary

diffraction: the bending of a wave around the edges of an opening or an obstacle

Huygens's principle: every point on a wavefront is a source of wavelets that spread out in the forward direction at the same speed as the wave itself. The new wavefront is a line tangent to all of the wavelets

88. Young's Double Slit Experiment

Learning Objectives

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

- Explain the phenomena of interference.
- Define constructive interference for a double slit and destructive interference for a double slit.

Although Christiaan Huygens thought that light was a wave, Isaac Newton did not. Newton felt that there were other explanations for color, and for the interference and diffraction effects that were observable at the time. Owing to Newton's tremendous stature, his view generally prevailed. The fact that Huygens's principle worked was not considered evidence that was direct enough to prove that light is a wave. The acceptance of the wave character of light came many years later when, in 1801, the English physicist and physician Thomas Young (1773–1829) did his now-classic double slit experiment (see Figure 1).

Why do we not ordinarily observe wave behavior for light, such as observed in Young's double slit experiment? First, light must interact with something small, such as the closely spaced slits used by Young, to show pronounced wave effects. Furthermore, Young first passed light from a single source (the Sun) through a single slit to make the light somewhat coherent. By *coherent*, we mean waves are in phase or have a definite phase relationship. *Incoherent* means the waves have random phase relationships. Why did Young then pass the light through a double slit? The answer to this question is that two slits provide two coherent light sources that then interfere constructively or destructively. Young used sunlight, where each wavelength forms its own pattern, making the effect more difficult to see. We illustrate the double slit experiment with monochromatic (single λ) light to clarify the effect. Figure 2 shows the pure constructive and destructive interference of two waves having the same wavelength and amplitude.

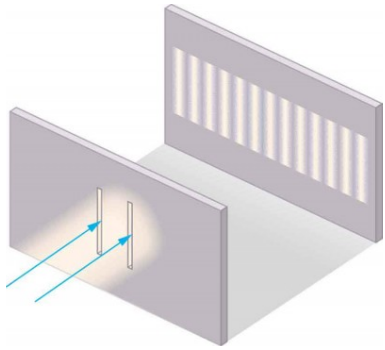


Figure 1. Young's double slit experiment. Here pure-wavelength light sent through a pair of vertical slits is diffracted into a pattern on the screen of numerous vertical lines spread out horizontally. Without diffraction and interference, the light would simply make two lines on the screen.

When light passes through narrow slits, it is diffracted into semicircular waves, as shown in Figure 3a. Pure constructive interference occurs where the waves are crest to crest or trough to trough. Pure destructive interference occurs where they are crest to trough.

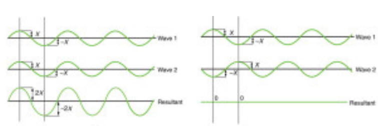


Figure 2. The amplitudes of waves add. (a) Pure constructive interference is obtained when identical waves are in phase. (b) Pure destructive interference occurs when identical waves are exactly out of phase, or shifted by half a wavelength.

The light must fall on a screen and be scattered into our eyes for us to see the pattern. An analogous pattern for water waves is shown in Figure 3b. Note that regions of constructive and destructive interference move out from the slits at well-defined angles to the original beam. These angles depend on wavelength and the distance between the slits, as we shall see below.

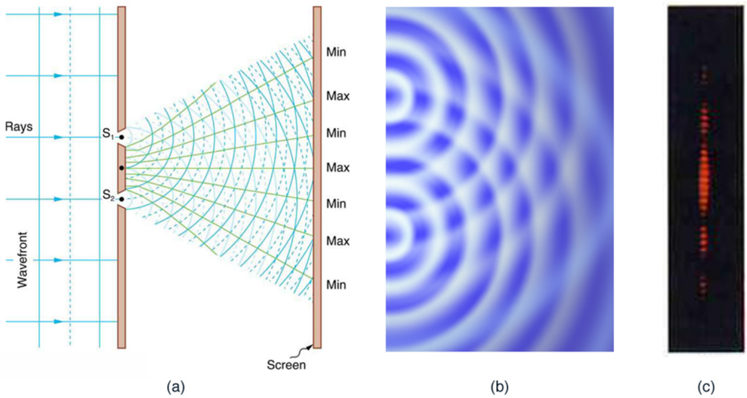


Figure 3. Double slits produce two coherent sources of waves that interfere. (a) Light spreads out (diffracts) from each slit, because the slits are narrow. These waves overlap and interfere constructively (bright lines) and destructively (dark regions). We can only see this if the light falls onto a screen and is scattered into our eyes. (b) Double slit interference pattern for water waves are nearly identical to that for light. Wave action is greatest in regions of constructive interference and least in regions of destructive interference. (c) When light that has passed through double slits falls on a screen, we see a pattern such as this. (credit: PASCO)

To understand the double slit interference pattern, we consider how two waves travel from the slits to the screen, as illustrated in Figure 4. Each slit is a different distance from a given point on the screen. Thus different numbers of wavelengths fit into each path. Waves start out from the slits in phase (crest to crest), but they may end up out of phase (crest to trough) at the screen if the paths differ in length by half a wavelength, interfering destructively as shown in Figure 4a. If the paths differ by a whole wavelength, then the waves arrive in phase (crest to crest) at the screen, interfering constructively as shown in Figure 4b. More generally, if the paths taken by the two waves differ by any half-integral number of wavelengths [$(1/2)\lambda$, $(3/2)\lambda$, $(5/2)\lambda$, etc.], then destructive interference occurs. Similarly, if the paths taken by the two waves differ by any integral number of wavelengths (λ, 2λ, 3λ, etc.), then constructive interference occurs.

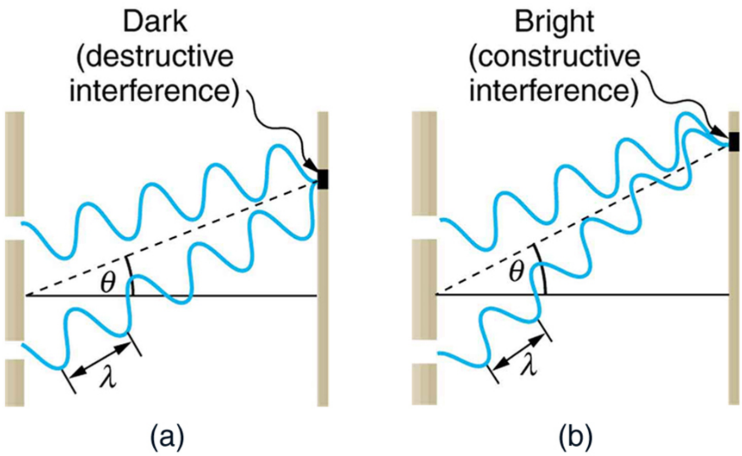


Figure 4. Waves follow different paths from the slits to a common point on a screen. (a) Destructive interference occurs here, because one path is a half wavelength longer than the other. The waves start in phase but arrive out of phase. (b) Constructive interference occurs here because one path is a whole wavelength longer than the other. The waves start out and arrive in phase.

Take-Home Experiment: Using Fingers as Slits

Look at a light, such as a street lamp or incandescent bulb, through the narrow gap between two fingers held close together. What type of pattern do you see? How does it change when you allow the fingers to move a little farther apart? Is it more distinct for a monochromatic source, such as the yellow light from a sodium vapor lamp, than for an incandescent bulb?

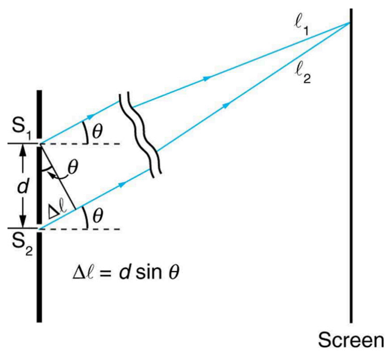


Figure 5. The paths from each slit to a common point on the screen differ by an amount dsinθ, assuming the distance to the screen is much greater than the distance between slits (not to scale here).

Figure 5 shows how to determine the path length difference for waves traveling from two slits to a common point on a screen. If the screen is a large distance away compared with the distance between the slits, then the angle θ between the path and a line from the slits to the screen (see the figure) is nearly the same for each path. The difference between the paths is shown in the figure; simple trigonometry shows it to be $d \sin\theta$, where d is the distance between the slits. To obtain *constructive interference for a double slit*, the path length difference must be an integral multiple of the wavelength, or $d \sin\theta = m\lambda$, for $m = 0, 1, -1, 2, -2, \ldots$ (constructive).

Similarly, to obtain *destructive interference for a double slit*, the path length difference must be a half-integral multiple of the wavelength, or

$$d\sin\theta = \left(m + \frac{1}{2}\right)\lambda, \text{ for } m = 0, 1, -1, 2, -2, \ldots \text{ (destructive)}$$

,

where λ is the wavelength of the light, d is the distance between slits, and θ is the angle from the original direction of the beam as discussed above. We call m the *order* of the interference. For example, $m = 4$ is fourth-order interference.

The equations for double slit interference imply that a series of bright and dark lines are formed. For vertical slits, the light spreads out horizontally on either side of the incident beam into a pattern called interference fringes, illustrated in Figure 6. The intensity of the bright fringes falls off on either side, being brightest at the center. The closer the slits are, the more is the spreading of the bright fringes. We can see this by examining the equation $d \sin \theta = m\lambda$, for $m = 0, 1, -1, 2, -2, \ldots$.

For fixed λ and m, the smaller d is, the larger θ must be, since $\sin\theta = \frac{m\lambda}{d}$. This is consistent with our contention that wave effects are most noticeable when the object the wave encounters (here, slits a distance d apart) is small. Small d gives large θ, hence a large effect.

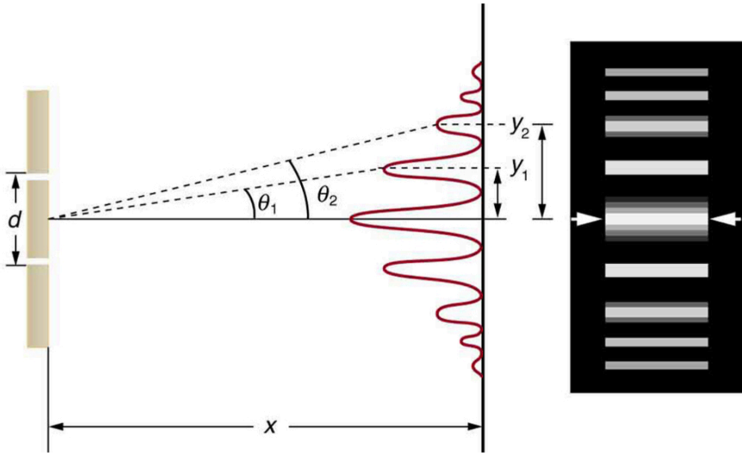


Figure 6. The interference pattern for a double slit has an intensity that falls off with angle. The photograph shows multiple bright and dark lines, or fringes, formed by light passing through a double slit.

Example 1. Finding a Wavelength from an Interference Pattern

Suppose you pass light from a He-Ne laser through two slits separated by 0.0100 mm and find that the third bright line on a screen is formed at an angle of 10.95° relative to the incident beam. What is the wavelength of the light?

Strategy

The third bright line is due to third-order constructive interference, which means that $m = 3$. We are given $d =$ 0.0100 mm and $\theta = 10.95°$. The wavelength can thus be found using the equation $d \sin \theta = m\lambda$ for constructive interference.

Solution

The equation is $d \sin \theta = m\lambda$. Solving for the wavelength λ gives $\lambda = \dfrac{d \sin \theta}{m}$.

Substituting known values yields

$$\begin{aligned} \lambda &= \frac{(0.0100 \text{ nm})(\sin 10.95^\circ)}{3} \\ &= 6.33 \times 10^{-4} \text{ nm} = 633 \text{ nm} \end{aligned}$$

Discussion

To three digits, this is the wavelength of light emitted by the common He-Ne laser. Not by coincidence, this red color is similar to that emitted by neon lights. More important, however, is the fact that interference patterns can be used to measure wavelength. Young did this for visible wavelengths. This analytical technique is still widely used to measure electromagnetic spectra. For a given order, the angle for constructive interference increases with λ, so that spectra (measurements of intensity versus wavelength) can be obtained.

Example 2. Calculating Highest Order Possible

Interference patterns do not have an infinite number of lines, since there is a limit to how big m can be. What is the highest-order constructive interference possible with the system described in the preceding example?

Strategy and Concept

The equation $d \sin \theta = m\lambda$ (for $m = 0, 1, -1, 2, -2, \ldots$) describes constructive interference. For fixed values of d and λ, the larger m is, the larger $\sin \theta$ is. However, the maximum value that $\sin\theta$ can have is 1, for an angle of 90°. (Larger angles imply that light goes backward and does not reach the screen at all.) Let us find which m corresponds to this maximum diffraction angle.

Solution

Solving the equation $d\sin\theta = m\lambda$ for m gives

$$\lambda = \frac{d\sin\theta}{m}.$$

Taking $\sin\theta = 1$ and substituting the values of d and λ from the preceding example gives

$$m = \frac{(0.0100\ \text{mm})\,(1)}{633\ \text{nm}} \approx 15.8$$

Therefore, the largest integer m can be is 15, or $m = 15$.

Discussion

The number of fringes depends on the wavelength and slit separation. The number of fringes will be very large for large slit separations. However, if the slit separation becomes much greater than the wavelength, the intensity of the interference pattern changes so that the screen has two bright lines cast by the slits, as expected when light behaves like a ray. We also note that the fringes get fainter further away from the center. Consequently, not all 15 fringes may be observable.

Section Summary

- Young's double slit experiment gave definitive proof of the wave character of light.
- An interference pattern is obtained by the superposition of light from two slits.
- There is constructive interference when $d \sin \theta = m\lambda$ (for $m = 0, 1, -1, 2, -2, \ldots$), where d is the distance between the slits, θ is the angle relative to the incident direction, and m is the order of the interference.
- There is destructive interference when $d \sin \theta = m\lambda$ (for $m = 0, 1, -1, 2, -2, \ldots$).

Conceptual Questions

1. Young's double slit experiment breaks a single light beam into two sources. Would the same pattern be obtained for two independent sources of light, such as the headlights of a distant car? Explain.
2. Suppose you use the same double slit to perform Young's double slit experiment in air and then repeat the experiment in water. Do the angles to the same parts of the interference pattern get larger or smaller? Does the color of the light change? Explain.
3. Is it possible to create a situation in which there is only destructive interference? Explain.
4. Figure 7 shows the central part of the interference pattern for a pure wavelength of red light projected onto a double slit. The pattern is actually a combination of single slit and double slit interference.

Note that the bright spots are evenly spaced. Is this a double slit or single slit characteristic? Note that some of the bright spots are dim on either side of the center. Is this a single slit or double slit characteristic? Which is smaller, the slit width or the separation between slits? Explain your responses.

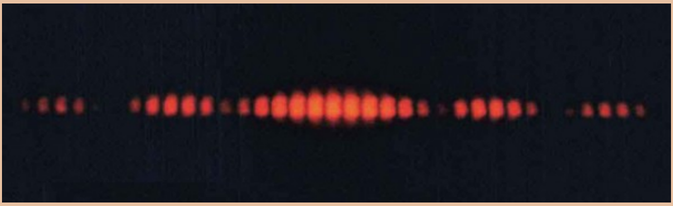


Figure 7. This double slit interference pattern also shows signs of single slit interference. (credit: PASCO)

Problems & Exercises

1. At what angle is the first-order maximum for 450-nm wavelength blue light falling on double slits separated by 0.0500 mm?
2. Calculate the angle for the third-order maximum of 580-nm wavelength yellow light falling on double slits separated by 0.100 mm.
3. What is the separation between two slits for which 610-nm orange light has its first maximum at an angle

of 30.0º?

4. Find the distance between two slits that produces the first minimum for 410-nm violet light at an angle of 45.0º.
5. Calculate the wavelength of light that has its third minimum at an angle of 30.0º when falling on double slits separated by 3.00 μm.
6. What is the wavelength of light falling on double slits separated by 2.00 μm if the third-order maximum is at an angle of 60.0º?
7. At what angle is the fourth-order maximum for the situation in Question 1?
8. What is the highest-order maximum for 400-nm light falling on double slits separated by 25.0 μm?
9. Find the largest wavelength of light falling on double slits separated by 1.20 μm for which there is a first-order maximum. Is this in the visible part of the spectrum?
10. What is the smallest separation between two slits that will produce a second-order maximum for 720-nm red light?
11. (a) What is the smallest separation between two slits that will produce a second-order maximum for any visible light? (b) For all visible light?
12. (a) If the first-order maximum for pure-wavelength light falling on a double slit is at an angle of 10.0º, at what angle is the second-order maximum? (b) What is the angle of the first minimum? (c) What is the highest-order maximum possible here?
13. Figure 8 shows a double slit located a distance x from a screen, with the distance from the center of the screen given by y. When the distance d between

the slits is relatively large, there will be numerous bright spots, called fringes. Show that, for small angles (where $\sin\theta \approx \theta$, with θ in radians), the distance between fringes is given by $\Delta y = \frac{x\lambda}{d}$.

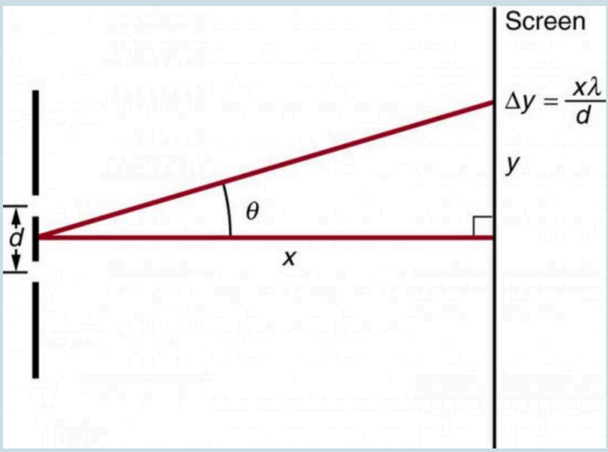


Figure 8. The distance between adjacent fringes is [latex]\Delta{y}=\frac{x\lambda}{d}\\[/latex], assuming the slit separation d is large compared with λ.

14. Using the result of the problem above, calculate the distance between fringes for 633-nm light falling on double slits separated by 0.0800 mm, located 3.00 m from a screen as in Figure 8.
15. Using the result of the problem two problems prior, find the wavelength of light that produces fringes 7.50 mm apart on a screen 2.00 m from double slits separated by 0.120 mm (see Figure 8).

Glossary

coherent: waves are in phase or have a definite phase relationship

constructive interference for a double slit: the path length difference must be an integral multiple of the wavelength

destructive interference for a double slit: the path length difference must be a half-integral multiple of the wavelength

incoherent: waves have random phase relationships

order: the integer m used in the equations for constructive and destructive interference for a double slit

Selected Solutions to Problems & Exercises

1. 0.516°

3. 1.22×10^{-6} m

5. 600 nm

7. 2.06°

9. 1200 nm (not visible)

11. (a) 760 nm; (b) 1520 nm

13. For small angles $\sin\theta - \tan\theta \approx \theta$ (in radians).

For two adjacent fringes we have, $d \sin \theta_{\text{m}} = m\lambda$ and $d \sin \theta_{\text{m}} + 1 = (m+1)\lambda$

Subtracting these equations gives

$$d\left(\sin\theta_{\mathrm{m}+1}-\sin\theta_{\mathrm{m}}\right)=[(m+1)-m]\lambda$$
$$d\left(\theta_{\mathrm{m}+1}-\theta_{\mathrm{m}}\right)=\lambda$$
$$\tan\theta_{\mathrm{m}}=\frac{y_{\mathrm{m}}}{x}\approx\theta_{\mathrm{m}}\Rightarrow d\left(\frac{y_{\mathrm{m}+1}}{x}-\frac{y_{\mathrm{m}}}{x}\right)=\lambda$$
$$d\frac{\Delta y}{x}=\lambda\Rightarrow\Delta y=\frac{\mathrm{x}\lambda}{d}$$

15. 450 nm

89. Multiple Slit Diffraction

Learning Objectives

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

- Discuss the pattern obtained from diffraction grating.
- Explain diffraction grating effects.

Figure 1. The colors reflected by this compact disc vary with angle and are not caused by pigments. Colors such as these are direct evidence of the wave character of light. (credit: Infopro, Wikimedia Commons)

An interesting thing happens if you pass light through a large number of evenly spaced parallel slits, called a *diffraction grating*. An interference pattern is created that is very similar to the one formed by a double slit (see Figure 2). A diffraction grating can be manufactured by scratching glass with a sharp tool in a number of precisely positioned parallel lines, with the untouched regions acting like slits. These can be photographically mass produced rather cheaply. Diffraction gratings work both for transmission of light, as in Figure 2, and for reflection of light, as on butterfly wings and the Australian opal in Figure 3 or the CD in Figure 1. In addition to their use as novelty items, diffraction gratings are commonly used for spectroscopic dispersion and analysis of light. What makes them particularly useful is the fact that they form a sharper pattern than double slits do. That is, their bright regions are narrower and brighter, while their dark regions are darker. Figure 4 shows idealized graphs demonstrating the sharper pattern. Natural diffraction gratings occur in the feathers of certain birds. Tiny, finger-like structures in regular patterns act as reflection gratings, producing constructive interference that gives the feathers colors not solely due to their pigmentation. This is called iridescence.

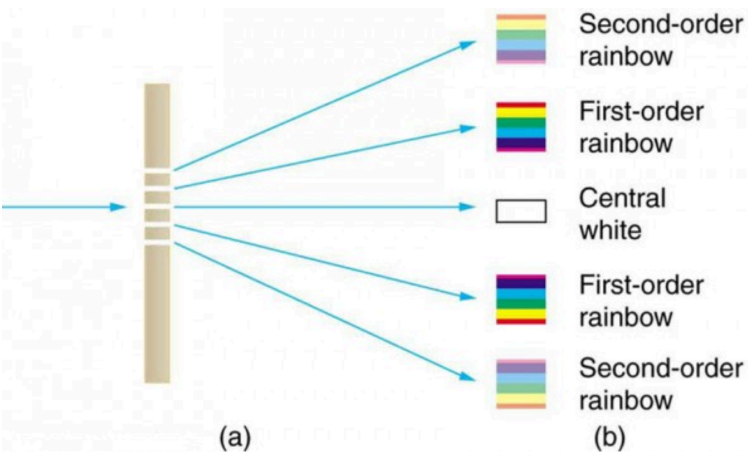


Figure 1. A diffraction grating is a large number of evenly spaced parallel slits. (a) Light passing through is diffracted in a pattern similar to a double slit, with bright regions at various angles. (b) The pattern obtained for white light incident on a grating. The central maximum is white, and the higher-order maxima disperse white light into a rainbow of colors.

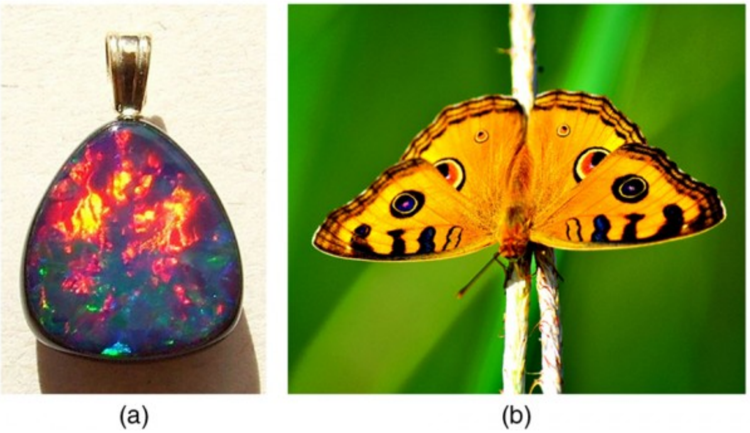


Figure 2. (a) This Australian opal and (b) the butterfly wings have rows of reflectors that act like reflection gratings, reflecting different colors at different angles. (credits: (a) Opals-On-Black.com, via Flickr (b) whologwhy, Flickr)

The analysis of a diffraction grating is very similar to that for a double slit (see Figure 5). As we know from our discussion of double slits in [Young's Double Slit Experiment](), light is diffracted by each slit and spreads out after passing through. Rays traveling in the same direction (at an angle θ relative to the incident direction) are shown in Figure 5. Each of these rays travels a different distance to a common point on a screen far away. The rays start in phase, and they can be in or out of phase when they reach a screen, depending on the difference in the path lengths traveled.

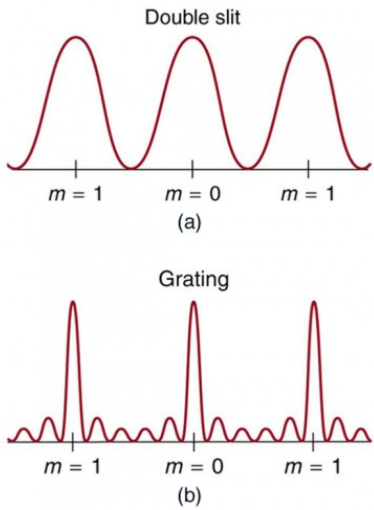


Figure 4. Idealized graphs of the intensity of light passing through a double slit (a) and a diffraction grating (b) for monochromatic light. Maxima can be produced at the same angles, but those for the diffraction grating are narrower and hence sharper. The maxima become narrower and the regions between darker as the number of slits is increased.

As seen in Figure 5, each ray travels a distance d sin θ different from that of its neighbor, where d is the distance between slits. If this distance equals an integral number of wavelengths, the rays all arrive in phase, and constructive interference (a maximum) is obtained. Thus, the condition necessary to obtain *constructive interference for a diffraction grating* is $d \sin\theta = m\lambda$, for $m = 0, 1, -1, 2, -2, \ldots$ (constructive) where d is the distance between slits in the grating, λ is the wavelength of light, and m is the order of the maximum. Note that this is exactly the same equation as for double slits separated by d. However, the slits are usually closer in diffraction gratings than in double slits, producing fewer maxima at larger angles.

In Figure 5, we see a diffraction grating showing light rays from each slit traveling in the same direction. Each ray travels a different distance to reach a common point on a screen (not shown). Each ray travels a distance $d \sin \theta$ different from that of its neighbor.

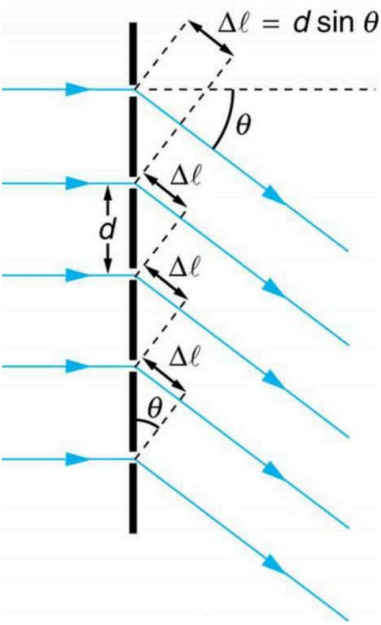


Figure 5.

Where are diffraction gratings used? Diffraction gratings are key components of monochromators used, for example, in optical imaging of particular wavelengths from biological or medical samples. A diffraction grating can be chosen to specifically analyze a wavelength emitted by molecules in diseased cells in a biopsy sample or to help excite strategic molecules in the sample with a selected frequency of light. Another vital use is in optical fiber technologies where fibers are designed to provide optimum performance at specific wavelengths. A range of diffraction gratings are available for selecting specific wavelengths for such use.

Take-Home Experiment: Rainbows on a CD

The spacing d of the grooves in a CD or DVD can be well determined by using a laser and the equation $d \sin \theta = m\lambda$, for $m = 0, 1, -1, 2, -2, \ldots$. However, we can still make a good estimate of this spacing by using white light and the rainbow of colors that comes from the interference. Reflect sunlight from a CD onto a wall and use your best judgment of the location of a strongly diffracted color to find the separation *d*.

Example 1. Calculating Typical Diffraction Grating Effects

Diffraction gratings with 10,000 lines per centimeter are readily available. Suppose you have one, and you send a beam of white light through it to a screen 2.00 m away.

1. Find the angles for the first-order diffraction of the shortest and longest wavelengths of visible light (380 and 760 nm).
2. What is the distance between the ends of the rainbow of visible light produced on the screen for first-order interference? (See Figure 6.)

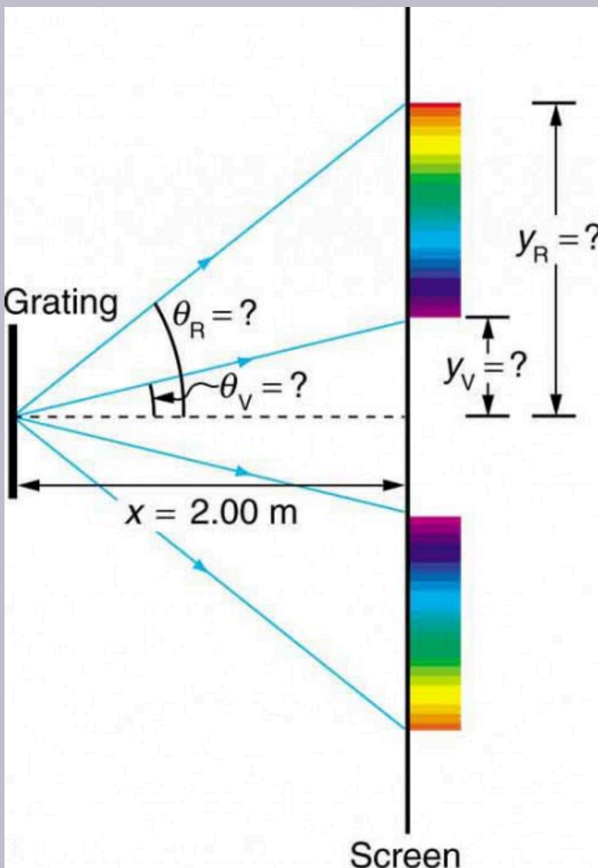


Figure 6. The diffraction grating considered in this example produces a rainbow of colors on a screen a distance from the grating. The distances along the screen are measured perpendicular to the x-direction. In other words, the rainbow pattern extends out of the page.

Strategy

The angles can be found using the equation $d \sin \theta = m\lambda$ (for $m = 0, 1, -1, 2, -2, \ldots$) once a value for the slit spacing d has been determined. Since there are 10,000 lines per centimeter, each line is separated by 1/10,000 of a centimeter. Once the angles are found, the distances along the screen can be found using simple trigonometry.

Solution for Part 1

The distance between slits is

$d = \frac{1\text{ cm}}{10,000} = 1.00 \times 10^{-4}\text{ cm}$ or 1.00×10^{-6} m.

Let us call the two angles θ_V for violet (380 nm) and θ_R for red (760 nm). Solving the equation $d \sin \theta_V = m\lambda$ for $\sin \theta_V$,

$\sin \theta_V = \frac{m\lambda_V}{d}$, where $m = 1$ for first order and $\lambda_V =$ 380 nm $= 3.80 \times 10^{-7}$ m. Substituting these values gives

$$\sin \theta_V = \frac{3.80 \times 10^{-7}\text{ m}}{1.00 \times 10^{-6}\text{ m}} = 0.380$$

Thus the angle θ_V is $\theta_V = \sin^{-1} 0.380 = 22.33°$.

Similarly,

$$\sin \theta_R = \frac{7.60 \times 10^{-7}\text{ m}}{1.00 \times 10^{-6}\text{ m}}$$

Thus the angle θ_R is $\theta_R = \sin^{-1} 0.760 = 49.46°$.

Notice that in both equations, we reported the results of these intermediate calculations to four significant figures to use with the calculation in Part 2.

Solution for Part 2

The distances on the screen are labeled y_V and y_R in Figure 6. Noting that $\tan\theta = \frac{y}{x}$, we can solve for y_V and y_R. That is, $y_V = x \tan\theta_V = (2.00 \text{ m})(\tan 22.33°) = 0.815$ m and $y_R = x \tan\theta_R = (2.00 \text{ m})(\tan 49.46°) = 2.338$ m.

The distance between them is therefore $y_R - y_V = 1.52$ m.

Discussion

The large distance between the red and violet ends of the rainbow produced from the white light indicates the potential this diffraction grating has as a spectroscopic tool. The more it can spread out the wavelengths (greater dispersion), the more detail can be seen in a spectrum. This depends on the quality of the diffraction grating—it must be very precisely made in addition to having closely spaced lines.

Section Summary

A diffraction grating is a large collection of evenly spaced parallel slits that produces an interference pattern similar to but sharper than that of a double slit.

There is constructive interference for a diffraction grating when $d \sin \theta = m\lambda$ (for $m = 0, 1, -1, 2, -2, \ldots$), where d is the distance between slits in the grating, λ is the wavelength of light, and m is the order of the maximum.

Conceptual Questions

1. What is the advantage of a diffraction grating over a double slit in dispersing light into a spectrum?
2. What are the advantages of a diffraction grating over a prism in dispersing light for spectral analysis?
3. Can the lines in a diffraction grating be too close together to be useful as a spectroscopic tool for visible light? If so, what type of EM radiation would the grating be suitable for? Explain.
4. If a beam of white light passes through a diffraction grating with vertical lines, the light is dispersed into rainbow colors on the right and left. If a glass prism disperses white light to the right into a rainbow, how does the sequence of colors compare with that produced on the right by a diffraction grating?
5. Suppose pure-wavelength light falls on a diffraction grating. What happens to the interference pattern if the same light falls on a grating that has more lines per centimeter? What happens to the interference pattern if a longer-wavelength light falls on the same grating? Explain how these two effects are consistent in terms of the relationship of wavelength to the distance between slits.
6. Suppose a feather appears green but has no green

pigment. Explain in terms of diffraction.

7. It is possible that there is no minimum in the interference pattern of a single slit. Explain why. Is the same true of double slits and diffraction gratings?

Problems & Exercises

1. A diffraction grating has 2000 lines per centimeter. At what angle will the first-order maximum be for 520-nm-wavelength green light?
2. Find the angle for the third-order maximum for 580-nm-wavelength yellow light falling on a diffraction grating having 1500 lines per centimeter.
3. How many lines per centimeter are there on a diffraction grating that gives a first-order maximum for 470-nm blue light at an angle of 25.0°?
4. What is the distance between lines on a diffraction grating that produces a second-order maximum for 760-nm red light at an angle of 60.0°?
5. Calculate the wavelength of light that has its second-order maximum at 45.0° when falling on a diffraction grating that has 5000 lines per centimeter.
6. An electric current through hydrogen gas produces several distinct wavelengths of visible light. What are the wavelengths of the hydrogen spectrum, if they form first-order maxima at angles of 24.2°, 25.7°, 29.1°, and 41.0° when projected on a diffraction grating having 10,000 lines per centimeter?

7. (a) What do the four angles in the above problem become if a 5000-line-per-centimeter diffraction grating is used? (b) Using this grating, what would the angles be for the second-order maxima? (c) Discuss the relationship between integral reductions in lines per centimeter and the new angles of various order maxima.
8. What is the maximum number of lines per centimeter a diffraction grating can have and produce a complete first-order spectrum for visible light?
9. The yellow light from a sodium vapor lamp seems to be of pure wavelength, but it produces two first-order maxima at 36.093° and 36.129° when projected on a 10,000 line per centimeter diffraction grating. What are the two wavelengths to an accuracy of 0.1 nm?
10. What is the spacing between structures in a feather that acts as a reflection grating, given that they produce a first-order maximum for 525-nm light at a 30.0° angle?
11. Structures on a bird feather act like a reflection grating having 8000 lines per centimeter. What is the angle of the first-order maximum for 600-nm light?
12. An opal such as that shown in Figure 2 acts like a reflection grating with rows separated by about 8 μm. If the opal is illuminated normally, (a) at what angle will red light be seen and (b) at what angle will blue light be seen?
13. At what angle does a diffraction grating produces a second-order maximum for light having a first-order maximum at 20.0°?
14. Show that a diffraction grating cannot produce a

second-order maximum for a given wavelength of light unless the first-order maximum is at an angle less than 30.0º.

15. If a diffraction grating produces a first-order maximum for the shortest wavelength of visible light at 30.0º, at what angle will the first-order maximum be for the longest wavelength of visible light?
16. (a) Find the maximum number of lines per centimeter a diffraction grating can have and produce a maximum for the smallest wavelength of visible light. (b) Would such a grating be useful for ultraviolet spectra? (c) For infrared spectra?
17. (a) Show that a 30,000-line-per-centimeter grating will not produce a maximum for visible light. (b) What is the longest wavelength for which it does produce a first-order maximum? (c) What is the greatest number of lines per centimeter a diffraction grating can have and produce a complete second-order spectrum for visible light?
18. A He–Ne laser beam is reflected from the surface of a CD onto a wall. The brightest spot is the reflected beam at an angle equal to the angle of incidence. However, fringes are also observed. If the wall is 1.50 m from the CD, and the first fringe is 0.600 m from the central maximum, what is the spacing of grooves on the CD?
19. The analysis shown in the figure below also applies to diffraction gratings with lines separated by a distance *d*. What is the distance between fringes produced by a diffraction grating having 125 lines per centimeter for 600-nm light, if the screen is 1.50 m away?

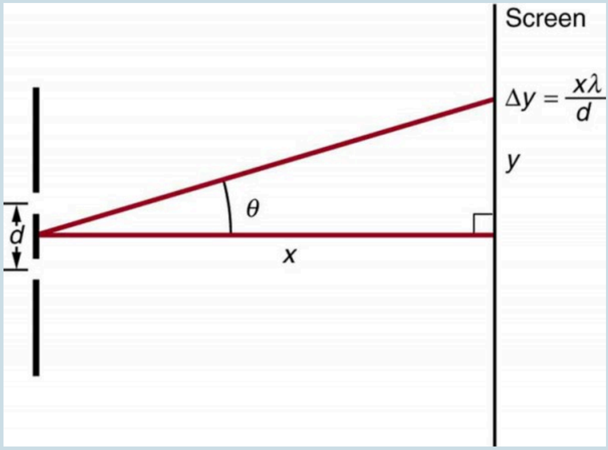


Figure 6. The distance between adjacent fringes is [latex]\Delta y=\frac{x\lambda}{d}\\[/latex], assuming the slit separation d is large compared with λ.

20. **Unreasonable Results.** Red light of wavelength of 700 nm falls on a double slit separated by 400 nm. (a) At what angle is the first-order maximum in the diffraction pattern? (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?
21. **Unreasonable Results.** (a) What visible wavelength has its fourth-order maximum at an angle of 25.0° when projected on a 25,000-line-per-centimeter diffraction grating? (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?
22. **Construct Your Own Problem.** Consider a spectrometer based on a diffraction grating.

Construct a problem in which you calculate the distance between two wavelengths of electromagnetic radiation in your spectrometer. Among the things to be considered are the wavelengths you wish to be able to distinguish, the number of lines per meter on the diffraction grating, and the distance from the grating to the screen or detector. Discuss the practicality of the device in terms of being able to discern between wavelengths of interest.

Glossary

constructive interference for a diffraction grating: occurs when the condition $d \sin \theta = m\lambda$ (form $= 0,1,-1,2,-2,\ldots$) is satisfied, where d is the distance between slits in the grating, λ is the wavelength of light, and m is the order of the maximum

diffraction grating: a large number of evenly spaced parallel slits

Selected Solution to Problems & Exercises

1. 5.97º

3. 8.99×10^{3}

5. 707 nm

7. (a) 11.8°,12.5°,14.1°,19.2°; (b) 24.2°,25.7°,29.1°,41.0°; (c)

Decreasing the number of lines per centimeter by a factor of x means that the angle for the x-order maximum is the same as the original angle for the first-order maximum.

9. 589.1 nm and 589.6 nm

11. 28.7°

13. 43.2°

15. 90.0°

17. (a) The longest wavelength is 333.3 nm, which is not visible; (b) 333 nm (UV); (c) 6.58×10^3 cm

19. 1.13×10^{-2} m

21. (a) 42.3 nm; (b) Not a visible wavelength. The number of slits in this diffraction grating is too large. Etching in integrated circuits can be done to a resolution of 50 nm, so slit separations of 400 nm are at the limit of what we can do today. This line spacing is too small to produce diffraction of light.

90. Single Slit Diffraction

Learning Objectives

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

- Discuss the single slit diffraction pattern.

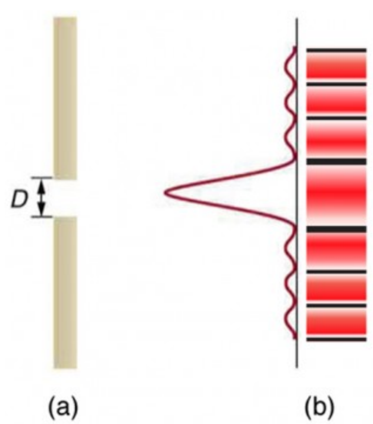
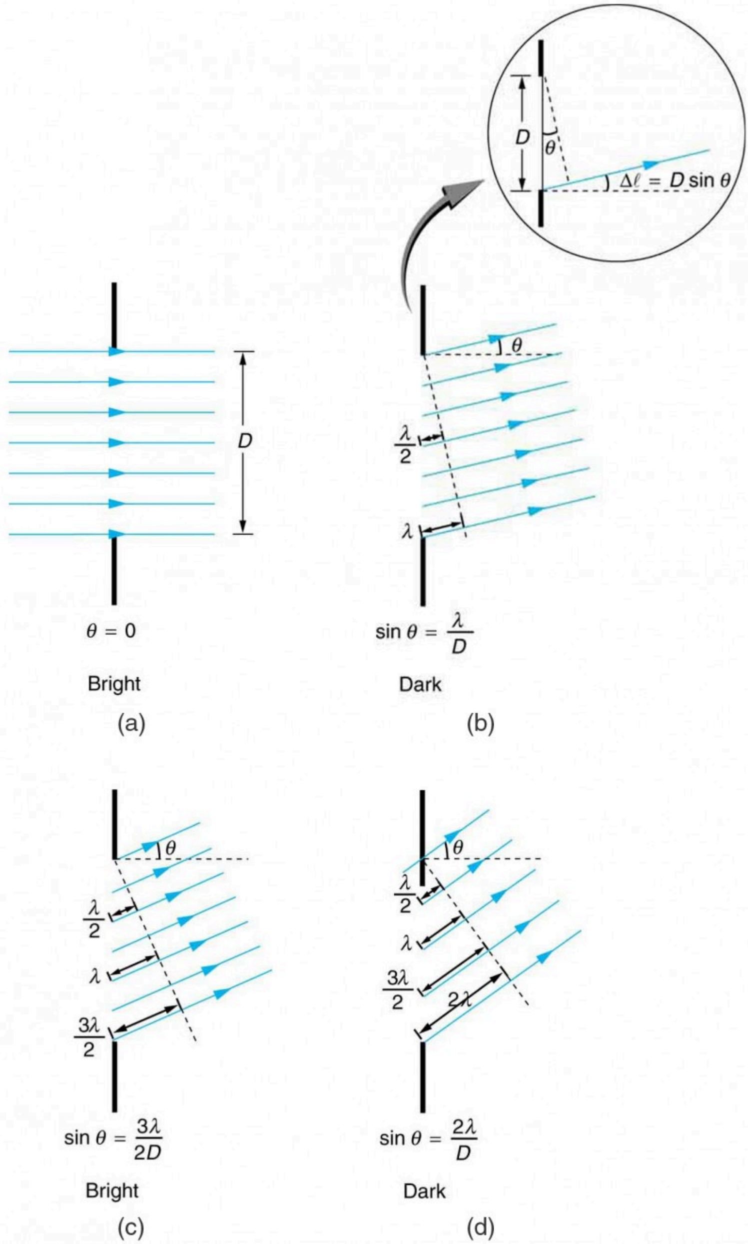


Figure 1. (a) Single slit diffraction pattern. Monochromatic light passing through a single slit has a central maximum and many smaller and dimmer maxima on either side. The central maximum is six times higher than shown. (b) The drawing shows the bright central maximum and dimmer and thinner maxima on either side.

Light passing through a single slit forms a diffraction pattern somewhat different from those formed by double slits or diffraction gratings. Figure 1 shows a single slit diffraction pattern. Note that the central maximum is larger than those on either side, and that the intensity decreases rapidly on either side. In contrast, a diffraction grating produces evenly spaced lines that dim slowly on either side of center.

The analysis of single slit diffraction is illustrated in Figure 2. Here we consider light coming from different parts of the *same* slit. According to Huygens's principle, every part of the wavefront in the slit emits wavelets. These are like rays that start out in phase and head in all directions. (Each ray is perpendicular to the wavefront of a wavelet.) Assuming the screen is very far away compared with the size of the slit, rays heading toward a common destination are nearly parallel. When they travel straight ahead, as in Figure 2a, they remain in phase, and a central maximum is obtained. However, when rays travel at an angle θ relative to the original direction of the beam, each travels a different distance to a common location, and they can arrive in or out of phase. In Figure 2b, the ray from the bottom travels a distance of one wavelength λ farther than the ray from the top. Thus a ray from the center travels a distance $\lambda/2$ farther than the one on the left, arrives out of phase, and interferes destructively. A ray from slightly above the center and one from slightly above the bottom will also cancel one another. In fact, each ray from the slit will have another to interfere destructively, and a minimum in intensity will occur at this angle. There will be another minimum at the same angle to the right of the incident direction of the light.



D
θ
Δℓ = D sin θ
θ
D
λ/2
λ
θ = 0
Bright
(a)
sin θ = λ/D
Dark
(b)
θ
λ/2
λ
3λ/2
sin θ = 3λ/2D
Bright
(c)
θ
λ/2
λ
3λ/2
2λ
sin θ = 2λ/D
Dark
(d)

Figure 2.

In Figure 2 we see that light passing through a single slit is diffracted in all directions and may interfere constructively or destructively, depending on the angle. The difference in path length for rays from either side of the slit is seen to be $D \sin \theta$.

At the larger angle shown in Figure 2c, the path lengths differ by $3\lambda/2$ for rays from the top and bottom of the slit. One ray travels a distance λ different from the ray from the bottom and arrives in phase, interfering constructively. Two rays, each from slightly above those two, will also add constructively. Most rays from the slit will have another to interfere with constructively, and a maximum in intensity will occur at this angle. However, all rays do not interfere constructively for this situation, and so the maximum is not as intense as the central maximum. Finally, in Figure 2d, the angle shown is large enough to produce a second minimum. As seen in the figure, the difference in path length for rays from either side of the slit is $D \sin \theta$, and we see that a destructive minimum is obtained when this distance is an integral multiple of the wavelength.

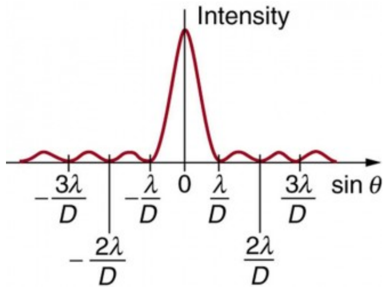


Figure 3. A graph of single slit diffraction intensity showing the central maximum to be wider and much more intense than those to the sides. In fact the central maximum is six times higher than shown here.

Thus, to obtain *destructive interference for a single slit*, $D \sin \theta = m\lambda$, for $m = 1, -1, 2, -2, 3, \ldots$ (destructive), where D is the slit width, λ is the light's wavelength, θ is the angle relative to the original direction of the light, and m is the order of the minimum. Figure 3 shows a graph of intensity for single slit interference, and it is apparent that the maxima on either side of the central maximum are much less intense and not as wide. This is consistent with the illustration in Figure 1b.

Example 1. Calculating Single Slit Diffraction

Visible light of wavelength 550 nm falls on a single slit and produces its second diffraction minimum at an angle of 45.0° relative to the incident direction of the light.

1. What is the width of the slit?
2. At what angle is the first minimum produced?

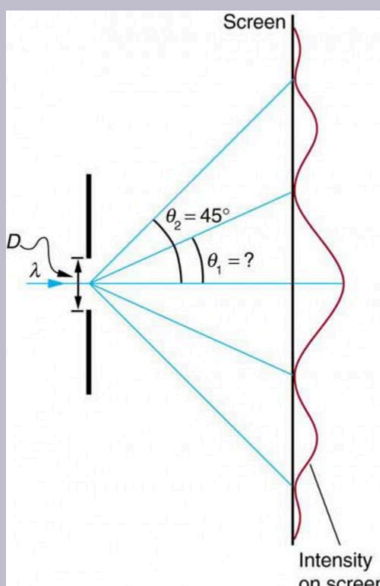


Figure 4.

A graph of the single slit diffraction pattern is analyzed in this example.

Strategy

From the given information, and assuming the screen is far away from the slit, we can use the equation $D \sin \theta = m\lambda$ first to find D, and again to find the angle for the first minimum θ_1.

Solution for Part 1

We are given that $\lambda = 550$ nm, $m = 2$, and $\theta_2 = 45.0°$. Solving the equation $D \sin \theta = m\lambda$ for D and substituting known values gives

$$\begin{aligned} D &= \frac{m\lambda}{\sin \theta_2} = \frac{2(550 \text{ nm})}{\sin 45.0^\circ} \\ &= \frac{1100 \times 10^{-9}}{0.707} \\ &= 1.56 \times 10^{-6} \end{aligned}$$

Solution for Part 2

Solving the equation $D \sin \theta = m\lambda$ for $\sin \theta_1$ and substituting the known values gives

$$\sin \theta_1 = \frac{m\lambda}{D} = \frac{1\left(550 \times 10^{-9} \text{ m}\right)}{1.56 \times 10^{-6} \text{ m}}$$

Thus the angle θ_1 is $\theta_1 = \sin^{-1} 0.354 = 20.7°$.

Discussion

We see that the slit is narrow (it is only a few times greater than the wavelength of light). This is consistent with the fact that light must interact with an object comparable in size to its wavelength in order to exhibit significant wave effects such as this single slit diffraction pattern. We also see that the central maximum extends 20.7º on either side of the original beam, for a width of about 41º. The angle between the first and second minima is only about 24º(45.0º − 20.7º). Thus the second maximum is only about half as wide as the central maximum.

Section Summary

- A single slit produces an interference pattern characterized by a broad central maximum with narrower and dimmer maxima to the sides.
- There is destructive interference for a single slit when $D \sin \theta = m\lambda$, (form = 1,−1,2,−2,3, . . .), where D is the slit width, λ is the light's wavelength, θ is the angle relative to the original direction of the light, and m is the order of the minimum. Note that there is no $m = 0$ minimum.

Conceptual Questions

1. As the width of the slit producing a single-slit diffraction pattern is reduced, how will the diffraction pattern produced change?

Problems & Exercises

1. (a) At what angle is the first minimum for 550-nm light falling on a single slit of width 1.00 μm? (b) Will there be a second minimum?
2. (a) Calculate the angle at which a 2.00-μm-wide slit produces its first minimum for 410-nm violet light. (b) Where is the first minimum for 700-nm red light?
3. (a) How wide is a single slit that produces its first minimum for 633-nm light at an angle of 28.0º? (b) At what angle will the second minimum be?
4. (a) What is the width of a single slit that produces its first minimum at 60.0º for 600-nm light? (b) Find the wavelength of light that has its first minimum at 62.0º.
5. Find the wavelength of light that has its third minimum at an angle of 48.6º when it falls on a single slit of width 3.00 μm.
6. Calculate the wavelength of light that produces its first minimum at an angle of 36.9º when falling on a single slit of width 1.00 μm.

7. (a) Sodium vapor light averaging 589 nm in wavelength falls on a single slit of width 7.50 μm. At what angle does it produces its second minimum? (b) What is the highest-order minimum produced?
8. (a) Find the angle of the third diffraction minimum for 633-nm light falling on a slit of width 20.0 μm. (b) What slit width would place this minimum at 85.0º?
9. (a) Find the angle between the first minima for the two sodium vapor lines, which have wavelengths of 589.1 and 589.6 nm, when they fall upon a single slit of width 2.00 μm. (b) What is the distance between these minima if the diffraction pattern falls on a screen 1.00 m from the slit? (c) Discuss the ease or difficulty of measuring such a distance.
10. (a) What is the minimum width of a single slit (in multiples of λ) that will produce a first minimum for a wavelength λ? (b) What is its minimum width if it produces 50 minima? (c) 1000 minima?
11. (a) If a single slit produces a first minimum at 14.5º, at what angle is the second-order minimum? (b) What is the angle of the third-order minimum? (c) Is there a fourth-order minimum? (d) Use your answers to illustrate how the angular width of the central maximum is about twice the angular width of the next maximum (which is the angle between the first and second minima).
12. A double slit produces a diffraction pattern that is a combination of single and double slit interference. Find the ratio of the width of the slits to the separation between them, if the first minimum of the single slit pattern falls on the fifth maximum of the double slit pattern. (This will greatly reduce the

intensity of the fifth maximum.)

13. **Integrated Concepts.** A water break at the entrance to a harbor consists of a rock barrier with a 50.0-m-wide opening. Ocean waves of 20.0-m wavelength approach the opening straight on. At what angle to the incident direction are the boats inside the harbor most protected against wave action?
14. **Integrated Concepts.** An aircraft maintenance technician walks past a tall hangar door that acts like a single slit for sound entering the hangar. Outside the door, on a line perpendicular to the opening in the door, a jet engine makes a 600-Hz sound. At what angle with the door will the technician observe the first minimum in sound intensity if the vertical opening is 0.800 m wide and the speed of sound is 340 m/s?

Glossary

destructive interference for a single slit: occurs when $D \sin \theta = m\lambda$, (form=1,−1,2,−2,3, . . .), where D is the slit width, λ is the light's wavelength, θ is the angle relative to the original direction of the light, and m is the order of the minimum

Selected Solutions to Problems & Exercises

1. (a) 33.4º; (b) No

3. (a) 1.35×10^{-6} m; (b) 69.9º

5. 750 nm

7. (a) 9.04º; (b) 12

9. (a) 0.0150º; (b) 0.262 mm; (c) This distance is not easily measured by human eye, but under a microscope or magnifying glass it is quite easily measurable.

11. (a) 30.1º; (b) 48.7º; (c) No; (d) $2\theta_1 = (2)(14.5º) = 29º$, $\theta_2 - \theta_1 = 30.05º - 14.5º = 15.56º$. Thus, $29º \approx (2)(15.56º) = 31.1º$.

13. 23.6º and 53.1º

91. Limits of Resolution: The Rayleigh Criterion

Learning Objectives

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

- Discuss the Rayleigh criterion.

Light diffracts as it moves through space, bending around obstacles, interfering constructively and destructively. While this can be used as a spectroscopic tool—a diffraction grating disperses light according to wavelength, for example, and is used to produce spectra—diffraction also limits the detail we can obtain in images. Figure 1a shows the effect of passing light through a small circular aperture. Instead of a bright spot with sharp edges, a spot with a fuzzy edge surrounded by circles of light is obtained. This pattern is caused by diffraction similar to that produced by a single slit. Light from different parts of the circular aperture interferes constructively and destructively. The effect is most noticeable when the aperture is small, but the effect is there for large apertures, too.

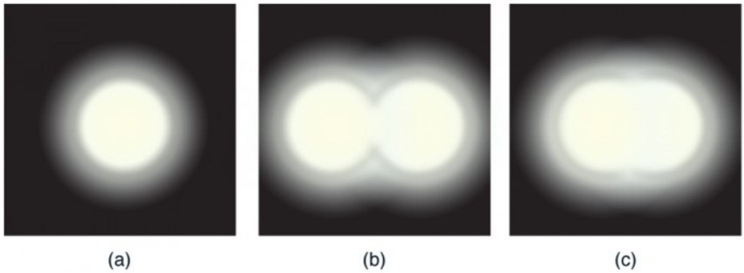


Figure 1. (a) Monochromatic light passed through a small circular aperture produces this diffraction pattern. (b) Two point light sources that are close to one another produce overlapping images because of diffraction. (c) If they are closer together, they cannot be resolved or distinguished.

How does diffraction affect the detail that can be observed when light passes through an aperture? Figure 1b shows the diffraction pattern produced by two point light sources that are close to one another. The pattern is similar to that for a single point source, and it is just barely possible to tell that there are two light sources rather than one. If they were closer together, as in Figure 1c, we could not distinguish them, thus limiting the detail or resolution we can obtain. This limit is an inescapable consequence of the wave nature of light.

There are many situations in which diffraction limits the resolution. The acuity of our vision is limited because light passes through the pupil, the circular aperture of our eye. Be aware that the diffraction-like spreading of light is due to the limited diameter of a light beam, not the interaction with an aperture. Thus light passing through a lens with a diameter *D* shows this effect and spreads, blurring the image, just as light passing through an aperture of diameter *D* does. So diffraction limits the resolution of any system having a lens or mirror. Telescopes are also limited by diffraction, because of the finite diameter *D* of their primary mirror.

Take-Home Experiment: Resolution of the Eye

Draw two lines on a white sheet of paper (several mm apart). How far away can you be and still distinguish the two lines? What does this tell you about the size of the eye's pupil? Can you be quantitative? (The size of an adult's pupil is discussed in Physics of the Eye.)

Just what is the limit? To answer that question, consider the diffraction pattern for a circular aperture, which has a central maximum that is wider and brighter than the maxima surrounding it (similar to a slit) (see Figure 2a). It can be shown that, for a circular aperture of diameter D, the first minimum in the diffraction pattern occurs at $\theta = 1.22\dfrac{\lambda}{D}$ (providing the aperture is large compared with the wavelength of light, which is the case for most optical instruments). The accepted criterion for determining the diffraction limit to resolution based on this angle was developed by Lord Rayleigh in the 19th century. The *Rayleigh criterion* for the diffraction limit to resolution states that *two images are just resolvable when the center of the diffraction pattern of one is directly over the first minimum of the diffraction pattern of the other*. See Figure 2b. The first minimum is at an angle of $\theta = 1.22\dfrac{\lambda}{D}$, so that two point objects are just resolvable if they are separated by the angle

$$\theta = 1.22\frac{\lambda}{D},$$

where λ is the wavelength of light (or other electromagnetic radiation) and D is the diameter of the aperture, lens, mirror, etc., with which the two objects are observed. In this expression, θ has units of radians.

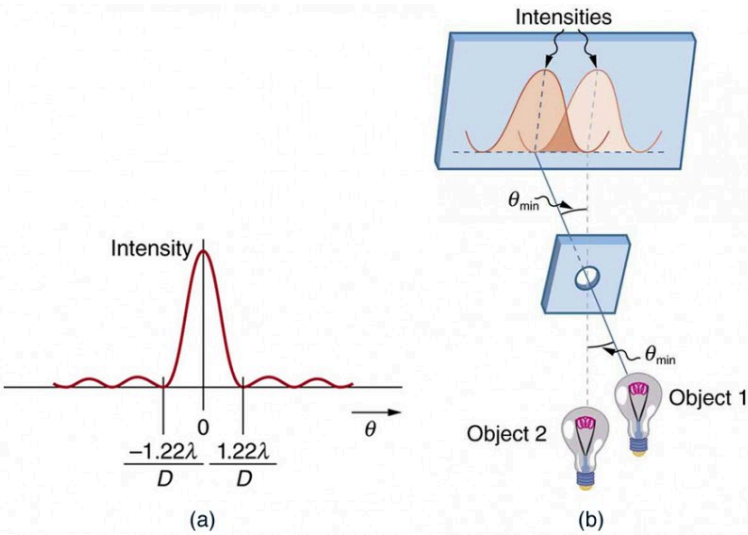


Figure 2. (a) Graph of intensity of the diffraction pattern for a circular aperture. Note that, similar to a single slit, the central maximum is wider and brighter than those to the sides. (b) Two point objects produce overlapping diffraction patterns. Shown here is the Rayleigh criterion for being just resolvable. The central maximum of one pattern lies on the first minimum of the other.

Making Connections: Limits to Knowledge

All attempts to observe the size and shape of objects are limited by the wavelength of the probe. Even the small wavelength of light prohibits exact precision. When extremely small wavelength probes as with an electron microscope are used, the system is disturbed, still limiting our knowledge, much as making an electrical measurement alters a circuit. Heisenberg's uncertainty principle asserts that this limit is fundamental and inescapable, as we shall see in quantum mechanics.

Example 1. Calculating Diffraction Limits of the Hubble Space Telescope

The primary mirror of the orbiting Hubble Space Telescope has a diameter of 2.40 m. Being in orbit, this telescope avoids the degrading effects of atmospheric distortion on its resolution.

1. What is the angle between two just-resolvable point light sources (perhaps two stars)? Assume an average light wavelength of 550 nm.
2. If these two stars are at the 2 million light year distance of the Andromeda galaxy, how close together can they be and still be resolved? (A light year, or ly, is the distance light travels in 1 year.)

Strategy

The Rayleigh criterion stated in the equation $\theta = 1.22\dfrac{\lambda}{D}$ gives the smallest possible angle θ between point sources, or the best obtainable resolution. Once this angle is found, the distance between stars can be calculated, since we are given how far away they are.

Solution for Part 1

The Rayleigh criterion for the minimum resolvable angle is $\theta = 1.22\dfrac{\lambda}{D}$.

Entering known values gives

$$\begin{aligned}\theta &= 1.22\frac{550\times10^{-9}\ \text{m}}{2.40\ \text{m}}\\ &= 2.80\times10^{-7}\ \text{rad}\end{aligned}$$

Solution for Part 2

The distance s between two objects a distance *r* away and separated by an angle θ is s = $r\theta$.

Substituting known values gives

$$\begin{aligned}s &= \left(2.0\times10^{6}\ \text{ly}\right)\left(2.80\times10^{-7}\ \text{rad}\right)\\ &= 0.56\ \text{ly}\end{aligned}$$

Discussion

The angle found in Part 1 is extraordinarily small (less than 1/50,000 of a degree), because the primary mirror is so large compared with the wavelength of light. As noticed, diffraction effects are most noticeable when light interacts with objects having sizes on the order of the wavelength of light. However, the effect is still there, and there is a diffraction limit to what is observable. The actual resolution of the Hubble Telescope is not quite as good as that found here. As with all instruments, there are other effects, such as non-uniformities in mirrors or aberrations in lenses that further limit resolution. However, Figure 3 gives an indication of the extent of the detail observable with the Hubble because of its size and quality and especially because it is above the Earth's atmosphere.

Figure 3. These two photographs of the M82 galaxy give an idea of the observable detail using the Hubble Space Telescope compared with that using a ground-based telescope. (a) On the left is a ground-based image. (credit: Ricnun, Wikimedia Commons) (b) The photo on the right was captured by Hubble. (credit: NASA, ESA, *and the Hubble Heritage Team* (STScI/AURA))

The answer in Part 2 indicates that two stars separated by about half a light year can be resolved. The average distance between stars in a galaxy is on the order of 5 light years in the outer parts and about 1 light year near the galactic center. Therefore, the Hubble can resolve most of the individual stars in Andromeda galaxy, even though it lies at such a huge distance that its light takes 2 million years for its light to reach us. Figure 4 shows another mirror used to observe radio waves from outer space.

Figure 4. A 305-m-diameter natural bowl at Arecibo in Puerto Rico is lined with reflective material, making it into a radio telescope. It is the largest curved focusing dish in the world. Although D for Arecibo is much larger than for the Hubble Telescope, it detects much longer wavelength radiation and its diffraction limit is significantly poorer than Hubble's. Arecibo is still very useful, because important information is carried by radio waves that is not carried by visible light. (credit: Tatyana Temirbulatova, Flickr)

Diffraction is not only a problem for optical instruments but also for the electromagnetic radiation itself. Any beam of light having a finite diameter D and a wavelength λ exhibits diffraction spreading. The beam spreads out with an angle θ given by the equation $\theta = 1.22\frac{\lambda}{D}$. Take, for example, a laser beam made of rays as parallel as possible (angles between rays as close to $\theta = 0°$ as possible) instead spreads out at an angle $\theta = 1.22\frac{\lambda}{D}$, where D is the diameter of the beam and λ is its wavelength. This spreading is impossible to observe for a flashlight, because its beam is not very parallel to start with. However, for long-distance transmission of laser beams or microwave signals, diffraction spreading can be significant (see Figure 5). To avoid this, we can increase D. This is done for laser light sent to the Moon to measure its distance from the Earth. The laser beam is expanded through a telescope to make D much larger and θ smaller.

In Figure 5 we see that the beam produced by this microwave transmission antenna will spread out at a minimum angle $\theta = 1.22\frac{\lambda}{D}$ due to diffraction. It is impossible to produce a near-parallel beam, because the beam has a limited diameter.

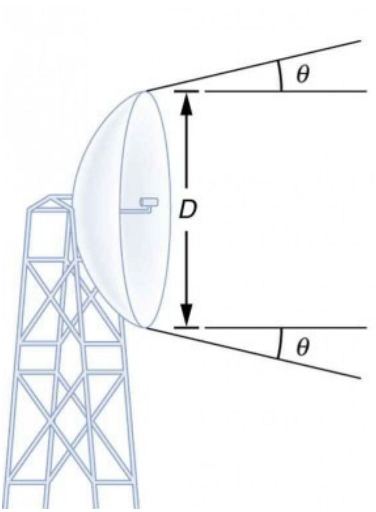


Figure 5.

In most biology laboratories, resolution is presented when the use of the microscope is introduced. The ability of a lens to produce sharp images of two closely spaced point objects is called resolution. The smaller the distance x by which two objects can be separated and still be seen as distinct, the greater the resolution. The resolving power of a lens is defined as that distance x. An expression for resolving power is obtained from the Rayleigh criterion. In Figure 6a we have two point objects separated by a distance x. According to the Rayleigh criterion, resolution is possible when the minimum angular separation is

$$\theta = 1.22\frac{\lambda}{D} = \frac{x}{d}$$

where d is the distance between the specimen and the objective lens, and we have used the small angle approximation (i.e., we have assumed that x is much smaller than d), so that $\tan\theta \approx \sin\theta \approx \theta$.

Therefore, the resolving power is

$$x = 1.22\frac{\lambda d}{D}$$

Another way to look at this is by re-examining the concept of Numerical Aperture (NA) discussed in Microscopes. There, NA is a measure of the maximum acceptance angle at which the fiber will take light and still contain it within the fiber. Figure 6b shows a lens and an object at point P. The NA here is a measure of the ability of the lens to gather light and resolve fine detail. The angle subtended by the lens at its focus is defined to be $\theta = 2\alpha$. From the Figure and again using the small angle approximation, we can write

$$\sin\alpha = \frac{\frac{D}{2}}{d} = \frac{D}{2d}$$

The NA for a lens is NA = $n \sin\alpha$, where n is the index of refraction of the medium between the objective lens and the object at point P.

From this definition for NA, we can see that

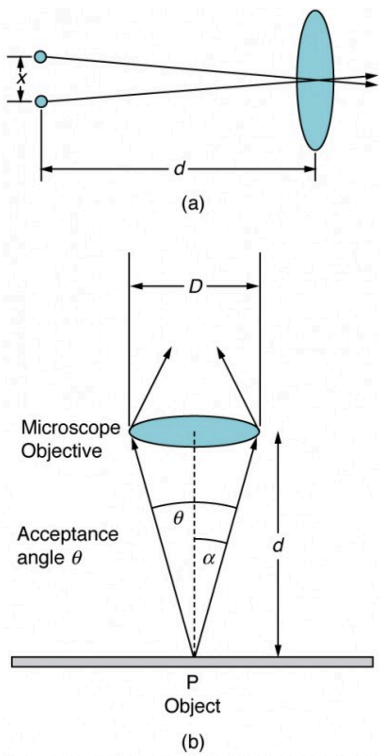


Figure 6. (a) Two points separated by at distance x and a positioned a distance d away from the objective. (credit: Infopro, Wikimedia Commons) (b) Terms and symbols used in discussion of resolving power for a lens and an object at point P. (credit: Infopro, Wikimedia Commons)

$$x = 1.22\frac{\lambda d}{D} = 1.22\frac{\lambda}{2\sin\alpha} = 0.61\frac{\lambda n}{NA}$$

In a microscope, NA is important because it relates to the resolving power of a lens. A lens with a large NA will be able to resolve finer details. Lenses with larger NA will also be able to collect more light and so give a brighter image. Another way to describe this situation is that the larger the NA, the larger the cone of light that can be brought into the lens, and so more of the diffraction modes will be collected. Thus the microscope has more information to form a clear image, and so its resolving power will be higher.

One of the consequences of diffraction is that the focal point of a beam has a finite width and intensity distribution. Consider focusing when only considering geometric optics, shown in Figure 7a. The focal point is infinitely small with a huge intensity and the capacity to incinerate most samples irrespective of the NA of the objective lens. For wave optics, due to diffraction, the focal point spreads to become a focal spot (see Figure 7b) with the size of the spot decreasing with increasing NA. Consequently, the intensity in the focal spot increases with increasing NA. The higher the NA, the greater the chances of photodegrading the specimen. However, the spot never becomes a true point.

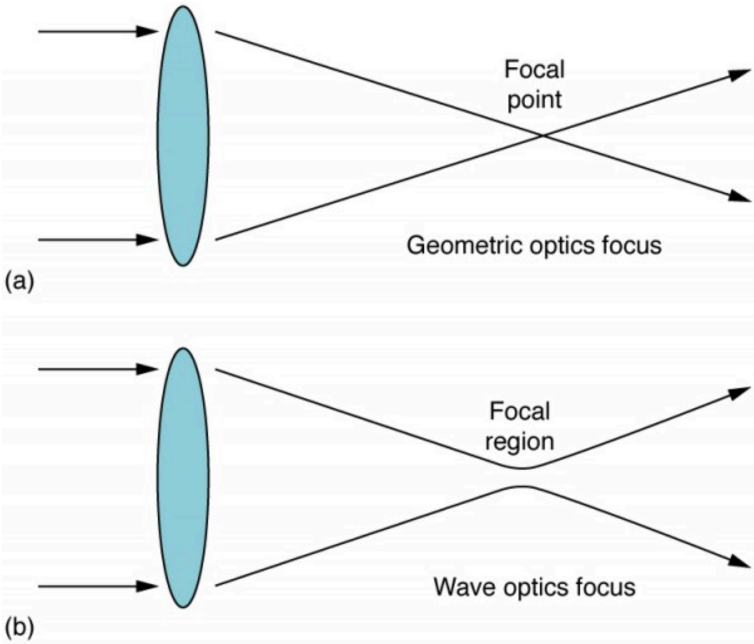


Figure 7. (a) In geometric optics, the focus is a point, but it is not physically possible to produce such a point because it implies infinite intensity. (b) In wave optics, the focus is an extended region.

Section Summary

- Diffraction limits resolution.
- For a circular aperture, lens, or mirror, the Rayleigh criterion states that two images are just resolvable when the center of the diffraction pattern of one is directly over the first minimum of the diffraction pattern of the other.
- This occurs for two point objects separated by the angle $\theta = 1.22\frac{\lambda}{D}$, where λ is the wavelength of light (or other electromagnetic radiation) and D is the diameter of the aperture, lens, mirror, etc. This equation also gives the angular

spreading of a source of light having a diameter *D*.

Conceptual Questions

1. A beam of light always spreads out. Why can a beam not be created with parallel rays to prevent spreading? Why can lenses, mirrors, or apertures not be used to correct the spreading?

Problems & Exercises

1. The 300-m-diameter Arecibo radio telescope pictured in Figure 4 detects radio waves with a 4.00 cm average wavelength. (a) What is the angle between two just-resolvable point sources for this telescope? (b) How close together could these point sources be at the 2 million light year distance of the Andromeda galaxy?
2. Assuming the angular resolution found for the Hubble Telescope in Example 1, what is the smallest detail that could be observed on the Moon?
3. Diffraction spreading for a flashlight is insignificant compared with other limitations in its optics, such as spherical aberrations in its mirror. To show this, calculate the minimum angular spreading of a flashlight beam that is originally 5.00 cm in diameter with an average wavelength of 600 nm.

4. (a) What is the minimum angular spread of a 633-nm wavelength He-Ne laser beam that is originally 1.00 mm in diameter? (b) If this laser is aimed at a mountain cliff 15.0 km away, how big will the illuminated spot be? (c) How big a spot would be illuminated on the Moon, neglecting atmospheric effects? (This might be done to hit a corner reflector to measure the round-trip time and, hence, distance.)
5. A telescope can be used to enlarge the diameter of a laser beam and limit diffraction spreading. The laser beam is sent through the telescope in opposite the normal direction and can then be projected onto a satellite or the Moon. (a) If this is done with the Mount Wilson telescope, producing a 2.54-m-diameter beam of 633-nm light, what is the minimum angular spread of the beam? (b) Neglecting atmospheric effects, what is the size of the spot this beam would make on the Moon, assuming a lunar distance of 3.84×10^8 m?
6. The limit to the eye's acuity is actually related to diffraction by the pupil. (a) What is the angle between two just-resolvable points of light for a 3.00-mm-diameter pupil, assuming an average wavelength of 550 nm? (b) Take your result to be the practical limit for the eye. What is the greatest possible distance a car can be from you if you can resolve its two headlights, given they are 1.30 m apart? (c) What is the distance between two just-resolvable points held at an arm's length (0.800 m) from your eye? (d) How does your answer to (c) compare to details you normally observe in everyday circumstances?
7. What is the minimum diameter mirror on a

telescope that would allow you to see details as small as 5.00 km on the Moon some 384,000 km away? Assume an average wavelength of 550 nm for the light received.

8. You are told not to shoot until you see the whites of their eyes. If the eyes are separated by 6.5 cm and the diameter of your pupil is 5.0 mm, at what distance can you resolve the two eyes using light of wavelength 555 nm?
9. (a) The planet Pluto and its Moon Charon are separated by 19,600 km. Neglecting atmospheric effects, should the 5.08-m-diameter Mount Palomar telescope be able to resolve these bodies when they are 4.50×10^9 km from Earth? Assume an average wavelength of 550 nm. (b) In actuality, it is just barely possible to discern that Pluto and Charon are separate bodies using an Earth-based telescope. What are the reasons for this?
10. The headlights of a car are 1.3 m apart. What is the maximum distance at which the eye can resolve these two headlights? Take the pupil diameter to be 0.40 cm.
11. When dots are placed on a page from a laser printer, they must be close enough so that you do not see the individual dots of ink. To do this, the separation of the dots must be less than Raleigh's criterion. Take the pupil of the eye to be 3.0 mm and the distance from the paper to the eye of 35 cm; find the minimum separation of two dots such that they cannot be resolved. How many dots per inch (dpi) does this correspond to?
12. **Unreasonable Results.** An amateur astronomer

wants to build a telescope with a diffraction limit that will allow him to see if there are people on the moons of Jupiter. (a) What diameter mirror is needed to be able to see 1.00 m detail on a Jovian Moon at a distance of 7.50×10^8 km from Earth? The wavelength of light averages 600 nm. (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

13. **Construct Your Own Problem.** Consider diffraction limits for an electromagnetic wave interacting with a circular object. Construct a problem in which you calculate the limit of angular resolution with a device, using this circular object (such as a lens, mirror, or antenna) to make observations. Also calculate the limit to spatial resolution (such as the size of features observable on the Moon) for observations at a specific distance from the device. Among the things to be considered are the wavelength of electromagnetic radiation used, the size of the circular object, and the distance to the system or phenomenon being observed.

Glossary

Rayleigh criterion: two images are just resolvable when the center of the diffraction pattern of one is directly over the first minimum of the diffraction pattern of the other

Selected Solutions to Problems & Exercises

1. (a) $1.63 \times 10{-}4$ rad; (b) 326 ly

3. 1.46×10^{-5} rad

5. (a) 3.04×10^{-7} rad; (b) Diameter of 235 m

7. 5.15 cm

9. (a) Yes. Should easily be able to discern; (b) The fact that it is just barely possible to discern that these are separate bodies indicates the severity of atmospheric aberrations.

92. Thin Film Interference

Learning Objectives

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

- Discuss the rainbow formation by thin films.

Figure 1. These soap bubbles exhibit brilliant colors when exposed to sunlight. (credit: Scott Robinson, Flickr)

The bright colors seen in an oil slick floating on water or in a sunlit soap bubble are caused by interference. The brightest colors are those that interfere constructively. This interference is between light reflected from different surfaces of a thin film; thus, the effect is known as *thin film interference*. As noticed before, interference effects are most prominent when light interacts with something having a size similar to its wavelength. A thin film is one having a thickness t smaller than a few times the wavelength of light, λ. Since color is associated indirectly with λ and since all interference depends in some way on the ratio of λ to the size of the object involved, we should expect to see different colors for different thicknesses of a film, as in Figure 1.

What causes thin film interference? Figure 2 shows how light reflected from the top and bottom surfaces of a film can interfere. Incident light is only partially reflected from the top surface of the film (ray 1). The remainder enters the film and is itself partially reflected from the bottom surface. Part of the light reflected from the bottom surface can emerge from the top of the film (ray 2) and interfere with light reflected from the top (ray 1). Since the ray that enters the film travels a greater distance, it may be in or out of phase with the ray reflected from the top. However, consider for a moment, again, the bubbles in Figure 1. The bubbles are darkest where they are thinnest. Furthermore, if you observe a soap bubble carefully, you will note it gets dark at the point where it breaks. For very thin films, the difference in path lengths of ray 1 and ray 2 in Figure 2 is negligible; so why should they interfere destructively and not constructively? The answer is that a phase change can occur upon reflection. The rule is as follows:

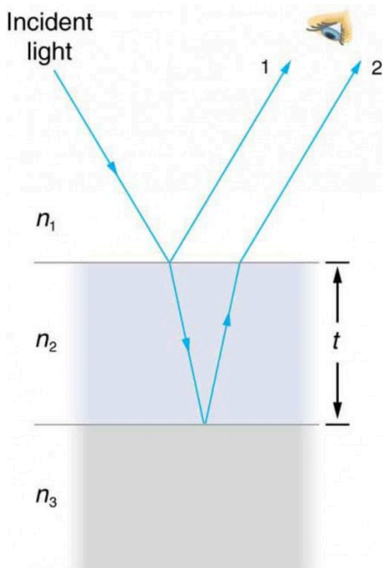


Figure 2. Light striking a thin film is partially reflected (ray 1) and partially refracted at the top surface. The refracted ray is partially reflected at the bottom surface and emerges as ray 2. These rays will interfere in a way that depends on the thickness of the film and the indices of refraction of the various media.

> When light reflects from a medium having an index of refraction greater than that of the medium in which it is traveling, a 180° phase change (or a λ / 2 shift) occurs.

If the film in Figure 2 is a soap bubble (essentially water with air on both sides), then there is a $\lambda/2$ shift for ray 1 and none for ray 2. Thus, when the film is very thin, the path length difference between the two rays is negligible, they are exactly out of phase, and destructive interference will occur at all wavelengths and so the soap bubble will be dark here.

The thickness of the film relative to the wavelength of light is the other crucial factor in thin film interference. Ray 2 in Figure 2 travels a greater distance than ray 1. For light incident perpendicular to the surface, ray 2 travels a distance approximately $2t$ farther than ray 1. When this distance is an integral or half-integral multiple of the wavelength in the medium ($\lambda_n = \lambda/n$, where λ is the wavelength in vacuum and n is the index of refraction), constructive or destructive interference occurs, depending also on whether there is a phase change in either ray.

Example 1. Calculating Non-reflective Lens Coating Using Thin Film Interference

Sophisticated cameras use a series of several lenses. Light can reflect from the surfaces of these various lenses and degrade image clarity. To limit these reflections, lenses are coated with a thin layer of magnesium fluoride that causes destructive thin film interference. What is the thinnest this film can be, if its index of refraction is 1.38 and it is designed to limit the reflection of 550-nm light, normally the most intense visible wavelength? The index of refraction of glass is 1.52.

Strategy

Refer to Figure 2 and use n_1=100 for air, n_2 = 1.38, and n_3 = 1.52. Both ray 1 and ray 2 will have a $\lambda/2$ shift upon reflection. Thus, to obtain destructive interference, ray 2 will need to travel a half wavelength farther than ray 1. For rays incident perpendicularly, the path length difference is $2t$.

Solution

To obtain destructive interference here,

$$2t = \frac{\lambda_{n2}}{2},$$

where λ_{n2} is the wavelength in the film and is given by $\lambda_{n2} = \dfrac{\lambda}{n_2}$.

Thus,

$$2t = \frac{\lambda / n_2}{2}$$

Solving for t and entering known values yields

$$\begin{aligned} t &= \frac{\lambda / n_2}{4} = \frac{550 \text{ nm}/1.38}{4} \\ &= 99.6 \text{ nm} \end{aligned}$$

Discussion

Films such as the one in this example are most effective in producing destructive interference when the thinnest layer is used, since light over a broader range of incident angles will be reduced in intensity. These films are called non-reflective coatings; this is only an approximately correct description, though, since other wavelengths will only be partially cancelled. Non-reflective coatings are used in car windows and sunglasses.

Thin film interference is most constructive or most destructive when the path length difference for the two rays is an integral or half-integral wavelength, respectively. That is, for rays incident perpendicularly, $2t = \lambda_n, 2\lambda_n, 3\lambda_n, \ldots$ or $2t = \frac{\lambda_n}{2}, \frac{3\lambda_n}{2}, \frac{5\lambda_n}{2}, \ldots$.

To know whether interference is constructive or destructive, you must also determine if there is a phase change upon reflection. Thin film interference thus depends on film thickness, the wavelength of light, and the refractive indices. For white light incident on a film that varies in thickness, you will observe rainbow colors of constructive interference for various wavelengths as the thickness varies.

Example 2. Soap Bubbles: More Than One Thickness can be Constructive

1. What are the three smallest thicknesses of a soap bubble that produce constructive interference for red light with a wavelength of 650 nm? The index of refraction of soap is taken to be the same as that of water.
2. What three smallest thicknesses will give destructive interference?

Strategy and Concept

Use Figure 2 to visualize the bubble. Note that $n_1 = n_3 = 1.00$ for air, and $n_2 = 1.333$ for soap (equivalent to water). There is a $\frac{\lambda}{2}$ shift for ray 1 reflected from the top surface of the bubble, and no shift for ray 2 reflected from the bottom surface. To get constructive interference, then, the path length difference ($2t$) must be a half-integral multiple of the wavelength—the first three being $\frac{\lambda_n}{2}, \frac{3\lambda_n}{2}, \text{ and } \frac{5\lambda_n}{2}$. To get destructive interference, the path length difference must be an integral multiple of the wavelength—the first three being 0, λ_n, and $2\lambda_n$.

Solution for Part 1

Constructive interference occurs here when

$$2t_c = \frac{\lambda_n}{2}, \frac{3\lambda_n}{2}, \text{ and } \frac{5\lambda_n}{2}, \ldots.$$

The smallest constructive thickness t_c thus is

$$\begin{aligned} t_c &= \frac{\lambda_n}{4} = \frac{\lambda/n}{4} = \frac{650 \text{ nm}/1.333}{4} \\ &= 122 \text{ nm} \end{aligned}$$

The next thickness that gives constructive interference is $t'_c = 3\lambda_n/4$, so that $t'_c = 366$ nm.

Finally, the third thickness producing constructive interference is $t''_c \leq 5\lambda_n/4$, so that $t''_c = 610$ nm.

Solution for Part 2

For *destructive interference*, the path length difference here is an integral multiple of the wavelength. The first occurs for zero thickness, since there is a phase change at the top surface. That is, $t_d = 0$.

The first non-zero thickness producing destructive interference is $2t'_d = \lambda_n$.

Substituting known values gives

$$\begin{aligned} t'_d &= \frac{\lambda}{2} = \frac{\lambda/n}{2} = \frac{650\text{ nm}/1.333}{2} \\ &= 244\text{ nm} \end{aligned}$$

Finally, the third destructive thickness is $2t''_d = 2\lambda_n$, so that

$$\begin{aligned} t''_d &= \lambda_n = \frac{\lambda}{n} = \frac{650\text{ nm}}{1.333} \\ &= 488\text{ nm} \end{aligned}$$

Discussion

If the bubble was illuminated with pure red light, we would see bright and dark bands at very uniform increases in thickness. First would be a dark band at 0 thickness, then bright at 122 nm thickness, then dark at 244 nm, bright at 366 nm, dark at 488 nm, and bright at 610 nm. If the bubble varied smoothly in thickness, like a smooth wedge, then the bands would be evenly spaced.

Another example of thin film interference can be seen when microscope slides are separated (see Figure 3). The slides are very flat, so that the wedge of air between them increases in thickness very uniformly. A phase change occurs at the second surface but not the first, and so there is a dark band where the slides touch. The rainbow colors of constructive interference repeat, going from violet to red again and again as the distance between the slides increases. As the layer of air increases, the bands become more difficult to see, because slight changes in incident angle have greater effects on path length differences. If pure-wavelength light instead of white light is used, then bright and dark bands are obtained rather than repeating rainbow colors.

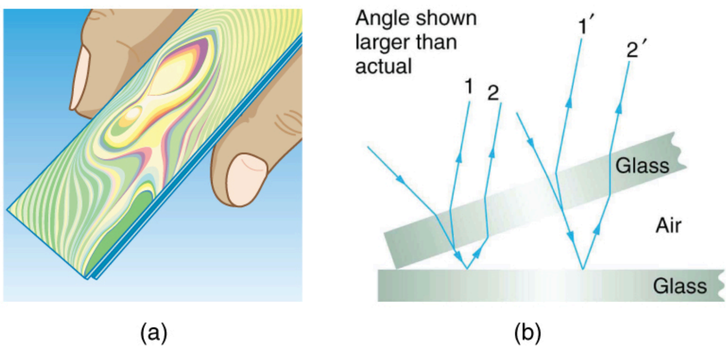


Figure 3. (a) The rainbow color bands are produced by thin film interference in the air between the two glass slides. (b) Schematic of the paths taken by rays in the wedge of air between the slides.

An important application of thin film interference is found in the manufacturing of optical instruments. A lens or mirror can be compared with a master as it is being ground, allowing it to be shaped to an accuracy of less than a wavelength over its entire surface. Figure 4 illustrates the phenomenon called Newton's rings, which occurs when the plane surfaces of two lenses are placed together. (The circular bands are called Newton's rings because Isaac Newton described them and their use in detail. Newton did not discover them; Robert Hooke did, and Newton did not believe they were due to the wave character of light.) Each successive ring of a given color indicates an increase of only one wavelength in the distance between the lens and the blank, so that great precision can be obtained. Once the lens is perfect, there will be no rings.

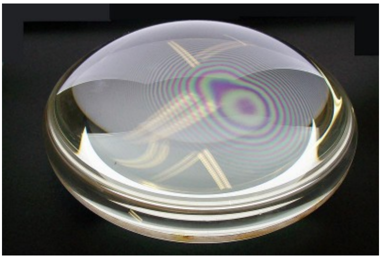


Figure 4. "Newton's rings" interference fringes are produced when two plano-convex lenses are placed together with their plane surfaces in contact. The rings are created by interference between the light reflected off the two surfaces as a result of a slight gap between them, indicating that these surfaces are not precisely plane but are slightly convex. (credit: Ulf Seifert, Wikimedia Commons)

The wings of certain moths and butterflies have nearly iridescent colors due to thin film interference. In addition to pigmentation, the wing's color is affected greatly by constructive interference of certain wavelengths reflected from its film-coated surface. Car manufacturers are offering special paint jobs that use thin film interference to produce colors that change with angle. This expensive option is based on variation of thin film path length differences with angle. Security features on credit cards, banknotes, driving licenses and similar items prone to forgery use thin film interference, diffraction gratings, or holograms. Australia led the way with dollar bills printed on polymer with a diffraction grating security feature making the currency difficult to forge. Other countries such as New Zealand and Taiwan are using similar technologies, while the United States currency includes a thin film interference effect.

Making Connections: Take-Home Experiment—Thin Film Interference

One feature of thin film interference and diffraction gratings is that the pattern shifts as you change the angle at which you look or move your head. Find examples of thin film interference and gratings around you. Explain how the patterns change for each specific example. Find examples where the thickness changes giving rise to changing colors. If you can find two microscope slides, then try observing the effect shown in Figure 3. Try separating one end of the two slides with a hair or maybe a thin piece of paper and observe the effect.

Problem-Solving Strategies for Wave Optics

Step 1. *Examine the situation to determine that interference is involved.* Identify whether slits or thin film interference are considered in the problem.

Step 2. *If slits are involved*, note that diffraction gratings and double slits produce very similar interference patterns, but that gratings have narrower (sharper) maxima. Single slit patterns are characterized by a large central maximum and smaller maxima to the sides.

Step 3. *If thin film interference is involved, take note of the path length difference between the two rays that interfere*. Be certain to use the wavelength in the medium involved, since it differs from the wavelength in vacuum. Note also that there is an additional $\lambda/2$ phase shift when light reflects from a medium with a greater index of refraction.

Step 4. *Identify exactly what needs to be determined in the problem (identify the unknowns)*. A written list is useful. Draw a diagram of the situation. Labeling the diagram is useful.

Step 5. *Make a list of what is given or can be inferred from the problem as stated (identify the knowns).*

Step 6. *Solve the appropriate equation for the quantity to be determined (the unknown), and enter the knowns*. Slits, gratings, and the Rayleigh limit involve equations.

Step 7. *For thin film interference, you will have constructive interference for a total shift that is an integral number of wavelengths. You will have destructive interference for a total shift of a half-integral number of wavelengths*. Always keep in mind that crest to crest is constructive whereas crest to trough is destructive.

Step 8. *Check to see if the answer is reasonable: Does it make sense*? Angles in interference patterns cannot be greater than 90°, for example.

Section Summary

- Thin film interference occurs between the light reflected from the top and bottom surfaces of a film. In addition to the path length difference, there can be a phase change.

- When light reflects from a medium having an index of refraction greater than that of the medium in which it is traveling, a 180° phase change (or a $\frac{\lambda}{2}$ shift) occurs.

Conceptual Questions

1. What effect does increasing the wedge angle have on the spacing of interference fringes? If the wedge angle is too large, fringes are not observed. Why?
2. How is the difference in paths taken by two originally in-phase light waves related to whether they interfere constructively or destructively? How can this be affected by reflection? By refraction?
3. Is there a phase change in the light reflected from either surface of a contact lens floating on a person's tear layer? The index of refraction of the lens is about 1.5, and its top surface is dry.
4. In placing a sample on a microscope slide, a glass cover is placed over a water drop on the glass slide. Light incident from above can reflect from the top and bottom of the glass cover and from the glass slide below the water drop. At which surfaces will there be a phase change in the reflected light?
5. Answer the above question if the fluid between the two pieces of crown glass is carbon disulfide.
6. While contemplating the food value of a slice of ham, you notice a rainbow of color reflected from its moist surface. Explain its origin.
7. An inventor notices that a soap bubble is dark at its

thinnest and realizes that destructive interference is taking place for all wavelengths. How could she use this knowledge to make a non-reflective coating for lenses that is effective at all wavelengths? That is, what limits would there be on the index of refraction and thickness of the coating? How might this be impractical?

8. A non-reflective coating like the one described in Example 1 works ideally for a single wavelength and for perpendicular incidence. What happens for other wavelengths and other incident directions? Be specific.
9. Why is it much more difficult to see interference fringes for light reflected from a thick piece of glass than from a thin film? Would it be easier if monochromatic light were used?

Problems & Exercises

1. A soap bubble is 100 nm thick and illuminated by white light incident perpendicular to its surface. What wavelength and color of visible light is most constructively reflected, assuming the same index of refraction as water?
2. An oil slick on water is 120 nm thick and illuminated by white light incident perpendicular to its surface. What color does the oil appear (what is the most constructively reflected wavelength), given its index

of refraction is 1.40?

3. Calculate the minimum thickness of an oil slick on water that appears blue when illuminated by white light perpendicular to its surface. Take the blue wavelength to be 470 nm and the index of refraction of oil to be 1.40.
4. Find the minimum thickness of a soap bubble that appears red when illuminated by white light perpendicular to its surface. Take the wavelength to be 680 nm, and assume the same index of refraction as water.
5. A film of soapy water (n = 1.33) on top of a plastic cutting board has a thickness of 233 nm. What color is most strongly reflected if it is illuminated perpendicular to its surface?
6. What are the three smallest non-zero thicknesses of soapy water (n = 1.33) on Plexiglas if it appears green (constructively reflecting 520-nm light) when illuminated perpendicularly by white light?
7. Suppose you have a lens system that is to be used primarily for 700-nm red light. What is the second thinnest coating of fluorite (magnesium fluoride) that would be non-reflective for this wavelength?
8. (a) As a soap bubble thins it becomes dark, because the path length difference becomes small compared with the wavelength of light and there is a phase shift at the top surface. If it becomes dark when the path length difference is less than one-fourth the wavelength, what is the thickest the bubble can be and appear dark at all visible wavelengths? Assume the same index of refraction as water. (b) Discuss the fragility of the film considering the thickness found.

9. A film of oil on water will appear dark when it is very thin, because the path length difference becomes small compared with the wavelength of light and there is a phase shift at the top surface. If it becomes dark when the path length difference is less than one-fourth the wavelength, what is the thickest the oil can be and appear dark at all visible wavelengths? Oil has an index of refraction of 1.40.
10. Figure 3 shows two glass slides illuminated by pure-wavelength light incident perpendicularly. The top slide touches the bottom slide at one end and rests on a 0.100-mm-diameter hair at the other end, forming a wedge of air. (a) How far apart are the dark bands, if the slides are 7.50 cm long and 589-nm light is used? (b) Is there any difference if the slides are made from crown or flint glass? Explain.
11. Figure 3 shows two 7.50-cm-long glass slides illuminated by pure 589-nm wavelength light incident perpendicularly. The top slide touches the bottom slide at one end and rests on some debris at the other end, forming a wedge of air. How thick is the debris, if the dark bands are 1.00 mm apart?
12. Repeat Question 1, but take the light to be incident at a 45° angle.
13. Repeat Question 2, but take the light to be incident at a 45° angle.
14. **Unreasonable Results.** To save money on making military aircraft invisible to radar, an inventor decides to coat them with a non-reflective material having an index of refraction of 1.20, which is between that of air and the surface of the plane. This, he reasons, should be much cheaper than designing Stealth

bombers. (a) What thickness should the coating be to inhibit the reflection of 4.00-cm wavelength radar? (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

Glossary

thin film interference: interference between light reflected from different surfaces of a thin film

Selected Solutions to Problems & Exercises

1. 532 nm (green)

3. 83.9 nm

5. 620 nm (orange)

7. 380 nm

9. 33.9 nm

11. 4.42×10^{-5} m

13. The oil film will appear black, since the reflected light is not in the visible part of the spectrum.

93. Polarization

Learning Objectives

By the end of this section, you will be able

- Discuss the meaning of polarization.
- Discuss the property of optical activity of certain materials.

Polaroid sunglasses are familiar to most of us. They have a special ability to cut the glare of light reflected from water or glass (see Figure 1). Polaroids have this ability because of a wave characteristic of light called polarization. What is polarization? How is it produced? What are some of its uses? The answers to these questions are related to the wave character of light.

Figure 1. These two photographs of a river show the effect of a polarizing filter in reducing glare in light reflected from the surface of water. Part (b) of this Figure was taken with a polarizing filter and part (a) was not. As a result, the reflection of clouds and sky observed in part (a) is not observed in part (b). Polarizing sunglasses are particularly useful on snow and water. (credit: Amithshs, Wikimedia Commons)

Light is one type of electromagnetic (EM) wave. As noted earlier, EM waves are *transverse waves* consisting of varying electric and magnetic fields that oscillate perpendicular to the direction of propagation (see Figure 2). There are specific directions for the oscillations of the electric and magnetic fields. *Polarization* is the attribute that a wave's oscillations have a definite direction relative to the direction of propagation of the wave. (This is not the same type of polarization as that discussed for the separation of charges.) Waves having such a direction are said to be *polarized*. For an EM wave, we define the *direction of polarization* to be the direction parallel to the electric field. Thus we can think of the electric field arrows as showing the direction of polarization, as in Figure 2.

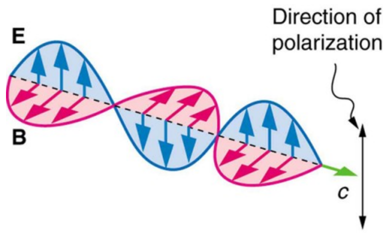


Figure 2. An EM wave, such as light, is a transverse wave. The electric and magnetic fields are perpendicular to the direction of propagation.

To examine this further, consider the transverse waves in the ropes shown in Figure 3. The oscillations in one rope are in a vertical plane and are said to be *vertically polarized*. Those in the other rope are in a horizontal plane and are *horizontally polarized*. If a vertical slit is placed on the first rope, the waves pass through. However, a vertical slit blocks the horizontally polarized waves. For EM waves, the direction of the electric field is analogous to the disturbances on the ropes.

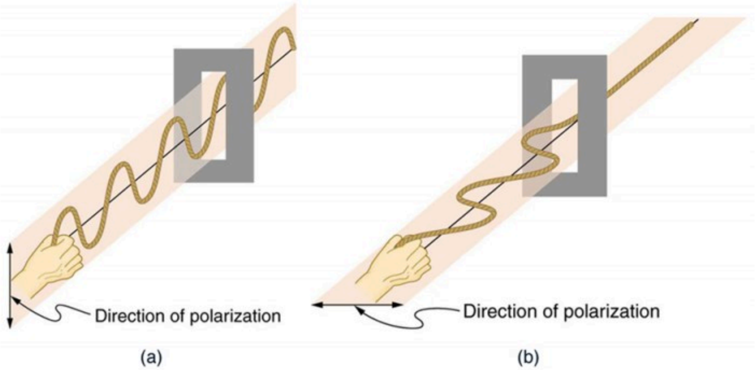


Figure 3. The transverse oscillations in one rope are in a vertical plane, and those in the other rope are in a horizontal plane. The first is said to be vertically polarized, and the other is said to be horizontally polarized. Vertical slits pass vertically polarized waves and block horizontally polarized waves.

The Sun and many other light sources produce waves that are randomly polarized (see Figure 4). Such light is said to be *unpolarized* because it is composed of many waves with all possible directions of polarization. Polaroid materials, invented by the founder of Polaroid Corporation, Edwin Land, act as a *polarizing* slit for light, allowing only polarization in one direction to pass through. Polarizing filters are composed of long molecules aligned in one direction. Thinking of the molecules as many slits, analogous to those for the oscillating ropes, we can understand why only light with a specific polarization can get through. The *axis of a polarizing filter* is the

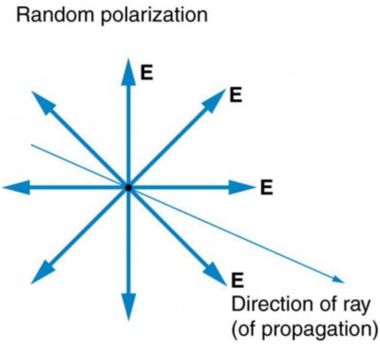


Figure 4. The slender arrow represents a ray of unpolarized light. The bold arrows represent the direction of polarization of the individual waves composing the ray. Since the light is unpolarized, the arrows point in all directions.

direction along which the filter passes the electric field of an EM wave (see Figure 5).

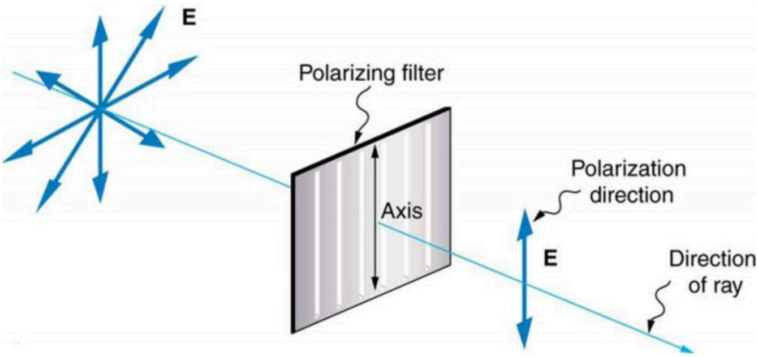


Figure 5. A polarizing filter has a polarization axis that acts as a slit passing through electric fields parallel to its direction. The direction of polarization of an EM wave is defined to be the direction of its electric field.

Figure 6 shows the effect of two polarizing filters on originally unpolarized light. The first filter polarizes the light along its axis. When the axes of the first and second filters are aligned (parallel), then all of the polarized light passed by the first filter is also passed by the second. If the second polarizing filter is rotated, only the component of the light parallel to the second filter's axis is passed. When the axes are perpendicular, no light is passed by the second.

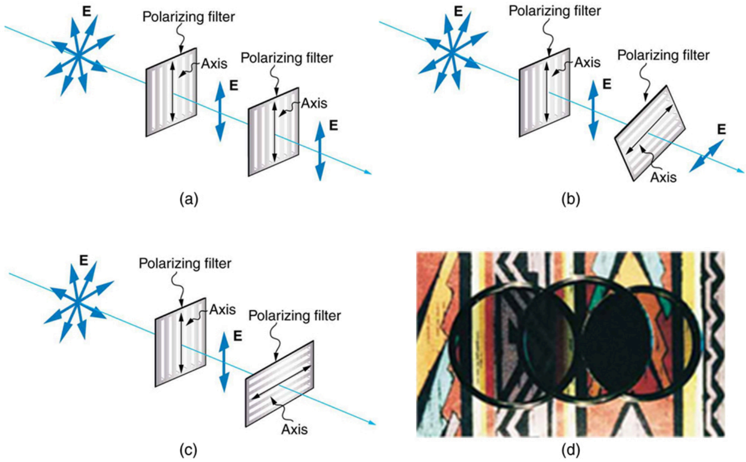


Figure 6. The effect of rotating two polarizing filters, where the first polarizes the light. (a) All of the polarized light is passed by the second polarizing filter, because its axis is parallel to the first. (b) As the second is rotated, only part of the light is passed. (c) When the second is perpendicular to the first, no light is passed. (d) In this photograph, a polarizing filter is placed above two others. Its axis is perpendicular to the filter on the right (dark area) and parallel to the filter on the left (lighter area). (credit: P.P. Urone)

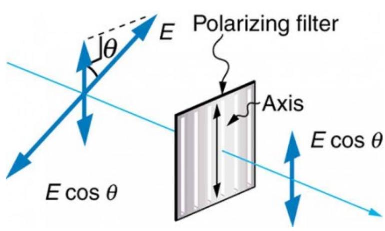


Figure 7. A polarizing filter transmits only the component of the wave parallel to its axis, , reducing the intensity of any light not polarized parallel to its axis.

Only the component of the EM wave parallel to the axis of a filter is passed. Let us call the angle between the direction of polarization and the axis of a filter θ. If the electric field has an amplitude E, then the transmitted part of the wave has an amplitude E cos θ (see Figure 7). Since the intensity of a wave is proportional to its amplitude squared, the intensity I of the transmitted wave is related to the incident wave by $I = I_0 \cos^2 \theta$, where I_0 is the intensity of the polarized wave before passing through the filter. (The above equation is known as Malus's law.)

Example 1. Calculating Intensity Reduction by a Polarizing Filter

What angle is needed between the direction of polarized light and the axis of a polarizing filter to reduce its intensity by 90.0%?

Strategy

When the intensity is reduced by 90.0%, it is 10.0% or 0.100 times its original value. That is, $I = 0.100 I_0$. Using this information, the equation $I = I_0 \cos^2\theta$ can be used to solve for the needed angle.

Solution

Solving the equation $I = I_0 \cos^2\theta$ for $\cos\theta$ and substituting with the relationship between I and I_0 gives

$$\cos\theta = \sqrt{\frac{I}{I_0}} = \sqrt{\frac{0.100 I_0}{I_0}} = 0.3162$$

Solving for θ yields $\theta = \cos^{-1} 0.3162 = 71.6^\circ$.

Discussion

A fairly large angle between the direction of polarization and the filter axis is needed to reduce the intensity to 10.0% of its original value. This seems reasonable based on experimenting with polarizing films. It is interesting that, at an angle of 45º, the intensity is reduced to 50% of its original value (as you will show in this section's Problems & Exercises). Note that 71.6º is 18.4º from reducing the intensity to zero, and that at an angle of 18.4º the intensity is reduced to 90.0% of its original value (as you will also show in Problems & Exercises), giving evidence of symmetry.

Polarization by Reflection

By now you can probably guess that Polaroid sunglasses cut the glare in reflected light because that light is polarized. You can check this for yourself by holding Polaroid sunglasses in front of you and rotating them while looking at light reflected from water or glass. As you rotate the sunglasses, you will notice the light gets bright and dim, but not completely black. This implies the reflected light is partially polarized and cannot be completely blocked by a polarizing filter.

Figure 8 illustrates what happens when unpolarized light is reflected from a surface. Vertically polarized light is preferentially refracted at the surface, so that *the reflected light is left more horizontally polarized*. The reasons for this phenomenon are beyond the scope of this text, but a convenient mnemonic for remembering this is to imagine the polarization direction to be like an arrow. Vertical polarization would be like an arrow perpendicular to the surface and would be more likely to stick and not be reflected. Horizontal polarization is like an arrow bouncing on its side and would be more likely to be reflected. Sunglasses with vertical axes would then block more reflected light than unpolarized light from other sources.

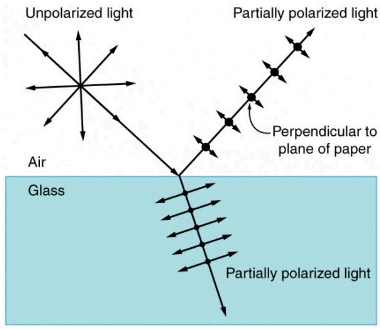


Figure 8. Polarization by reflection. Unpolarized light has equal amounts of vertical and horizontal polarization. After interaction with a surface, the vertical components are preferentially absorbed or refracted, leaving the reflected light more horizontally polarized. This is akin to arrows striking on their sides bouncing off, whereas arrows striking on their tips go into the surface.

Since the part of the light that is not reflected is refracted, the amount of polarization depends on the indices of refraction of the media involved. It can be shown that *reflected light is completely polarized* at a angle of reflection θ_b, given by $\tan\theta_b = \frac{n_2}{n_1}$, where n_1 is the medium in which the incident and reflected light travel and n_2 is the index of refraction of the medium that forms the interface that reflects the light. This equation is known as *Brewster's law*, and θ_b is known as *Brewster's angle*, named after the 19th-century Scottish physicist who discovered them.

Things Great and Small: Atomic Explanation of Polarizing Filters

Polarizing filters have a polarization axis that acts as a slit. This slit passes electromagnetic waves (often visible light) that have an electric field parallel to the axis. This is accomplished with long molecules aligned perpendicular to the axis as shown in Figure 9.

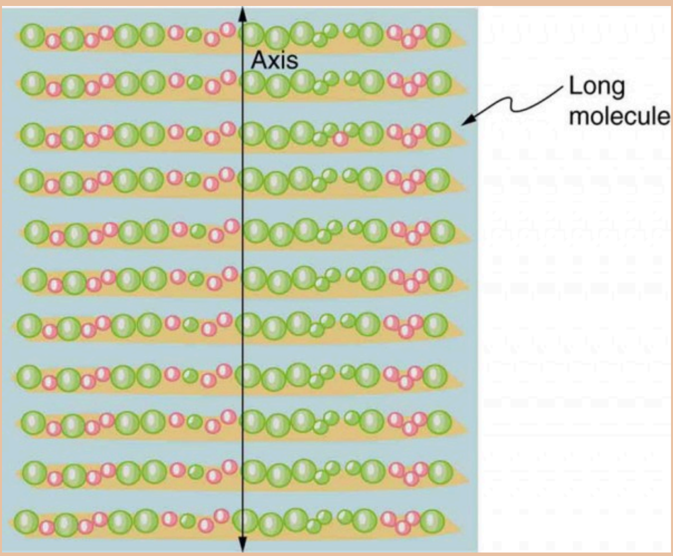


Figure 9. Long molecules are aligned perpendicular to the axis of a polarizing filter. The component of the electric field in an EM wave perpendicular to these molecules passes through the filter, while the component parallel to the molecules is absorbed.

Figure 10 illustrates how the component of the electric field parallel to the long molecules is absorbed. An electromagnetic wave is composed of oscillating electric and magnetic fields. The electric field is strong compared with the magnetic field and is more effective in exerting force on charges in the molecules. The most affected charged particles are the electrons in the molecules, since electron masses are small. If the electron is forced to oscillate, it can absorb energy from the EM wave. This reduces the fields in the wave and, hence, reduces its intensity. In long molecules, electrons can more easily oscillate parallel to the molecule than in the perpendicular direction. The electrons are bound to the molecule and are more restricted in their movement perpendicular to the molecule. Thus, the electrons can absorb EM waves that have a component of their electric field parallel to the molecule. The electrons are much less responsive to electric fields perpendicular to the molecule and will allow those fields to pass. Thus the axis of the polarizing filter is perpendicular to the length of the molecule.

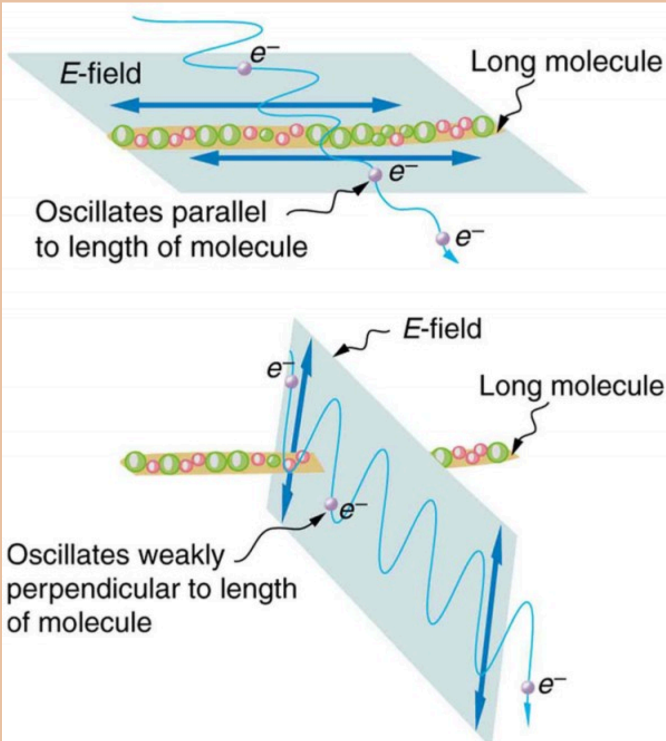


Figure 10. Artist's conception of an electron in a long molecule oscillating parallel to the molecule. The oscillation of the electron absorbs energy and reduces the intensity of the component of the EM wave that is parallel to the molecule.

Example 2. Calculating Polarization by

Reflection

1. At what angle will light traveling in air be completely polarized horizontally when reflected from water?
2. From glass?

Strategy

All we need to solve these problems are the indices of refraction. Air has $n_1 = 1.00$, water has $n_2 = 1.333$, and crown glass has $n'_2 = 1.520$. The equation $\tan\theta_\mathrm{b} = \frac{n_2}{n_1}$ can be directly applied to find θ_b in each case.

Solution for Part 1

Putting the known quantities into the equation

$$\tan\theta_\mathrm{b} = \frac{n_2}{n_1}$$ gives

$$\tan\theta_\mathrm{b} = \frac{n_2}{n_1} = \frac{1.333}{1.00} = 1.333.$$

Solving for the angle θ_b yields

$$\theta_\mathrm{b} = \tan^{-1} 1.333 = 53.1^\circ.$$

Solution for Part 2

Similarly, for crown glass and air,

$$\tan\theta'_\mathrm{b} = \frac{n'_2}{n_1} = \frac{1.520}{1.00} = 1.52.$$

Thus,

$$\theta'_\mathrm{b} = \tan^{-1} 1.52 = 56.7^\circ.$$

Discussion

Light reflected at these angles could be completely blocked by a good polarizing filter held with its *axis vertical*. Brewster's angle for water and air are similar to those for glass and air, so that sunglasses are equally effective for light reflected from either water or glass under similar circumstances. Light not reflected is refracted into these media. So at an incident angle equal to Brewster's angle, the refracted light will be slightly polarized vertically. It will not be completely polarized vertically, because only a small fraction of the incident light is reflected, and so a significant amount of horizontally polarized light is refracted.

Polarization by Scattering

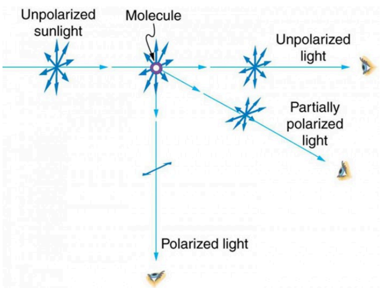


Figure 11. Polarization by scattering. Unpolarized light scattering from air molecules shakes their electrons perpendicular to the direction of the original ray. The scattered light therefore has a polarization perpendicular to the original direction and none parallel to the original direction.

If you hold your Polaroid sunglasses in front of you and rotate them while looking at blue sky, you will see the sky get bright and dim. This is a clear indication that light scattered by air is partially polarized. Figure 11 helps illustrate how this happens. Since light is a transverse EM wave, it vibrates the electrons of air molecules perpendicular to the direction it is traveling. The electrons then radiate like small antennae. Since they are oscillating perpendicular to the direction of the light ray, they produce EM radiation that is polarized perpendicular to the direction of the ray. When viewing the light along a line perpendicular to the original ray, as in Figure 11, there can be no polarization in the scattered light parallel to the original ray, because that would require the original ray to be a longitudinal wave. Along other directions, a component of the other polarization can be projected along the line of sight, and the scattered light will only be partially polarized. Furthermore, multiple scattering can bring light to your eyes from other directions and can contain different polarizations.

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 12. It is possible to manipulate this characteristic quickly and in small well-defined regions to create the contrast patterns we see in so many LCD devices.

In flat screen LCD televisions, there is a large light at the back of the TV. The light travels to the front screen through millions of tiny units called pixels (picture elements). One of these is shown in Figure 12 (a) and (b). Each unit has three cells, with red, blue, or green filters, each controlled independently. When the voltage across a liquid crystal is switched off, the liquid crystal passes the light through the particular filter. One can vary the picture contrast by varying the strength of the voltage applied to the liquid crystal.

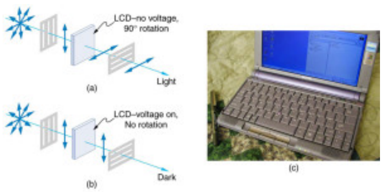


Figure 12. (a) Polarized light is rotated 90° by a liquid crystal and then passed by a polarizing filter that has its axis perpendicular to the original polarization direction. (b) When a voltage is applied to the liquid crystal, the polarized light is not rotated and is blocked by the filter, making 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 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.

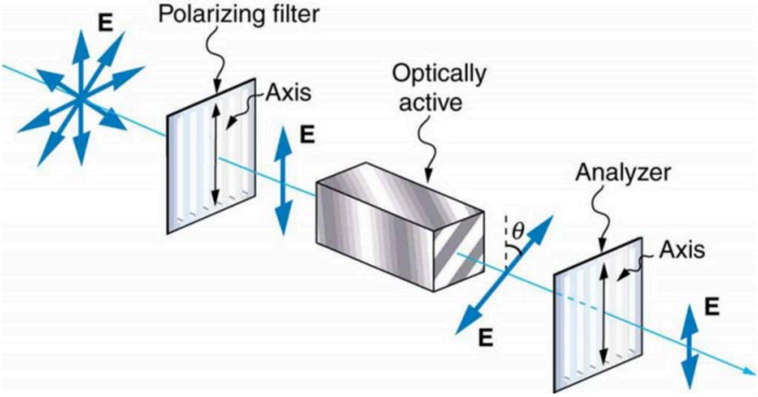


Figure 13. Optical activity is the ability of some substances to rotate the plane of polarization of light passing through them. The rotation is detected with a polarizing filter or analyzer.

Glass and plastic become optically active when stressed; the greater the stress, the greater the effect. Optical stress analysis on complicated shapes can be performed by making plastic models of them and observing them through crossed filters, as seen in Figure 14. It is apparent that the effect depends on wavelength as well as stress. The wavelength dependence is sometimes also used for artistic purposes.

Another interesting phenomenon associated with polarized light is the ability of some crystals to split an unpolarized beam of light into two. Such crystals are said to be *birefringent* (see Figure 15). Each of the separated rays has a specific polarization. One behaves normally and is called the ordinary ray, whereas the other does not obey Snell's law and is called the extraordinary ray. Birefringent crystals can be used to produce polarized beams from unpolarized light. Some birefringent materials preferentially absorb one of the polarizations. These materials are called dichroic and can produce polarization by this preferential absorption. This is fundamentally how polarizing filters and other polarizers work. The interested reader is invited to further pursue the numerous properties of materials related to polarization.

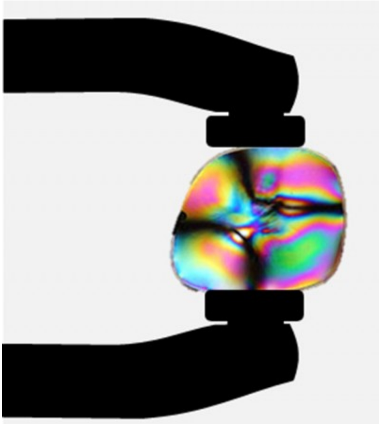


Figure 14. Optical stress analysis of a plastic lens placed between crossed polarizers. (credit: Infopro, Wikimedia Commons)

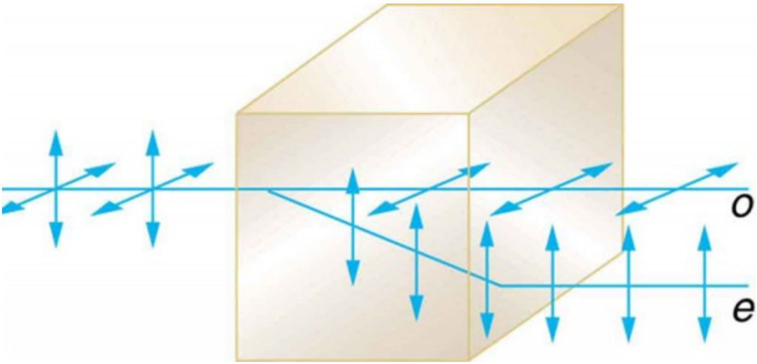


Figure 15. Birefringent materials, such as the common mineral calcite, split unpolarized beams of light into two. The ordinary ray behaves as expected, but the extraordinary ray does not obey Snell's law.

Section Summary

- Polarization is the attribute that wave oscillations have a definite direction relative to the direction of propagation of the wave.
- EM waves are transverse waves that may be polarized.
- The direction of polarization is defined to be the direction parallel to the electric field of the EM wave.
- Unpolarized light is composed of many rays having random polarization directions.
- Light can be polarized by passing it through a polarizing filter or other polarizing material. The intensity I of polarized light after passing through a polarizing filter is $I = I_0 \cos^2 \theta$, where I_0 is the original intensity and θ is the angle between the direction of polarization and the axis of the filter.
- Polarization is also produced by reflection.
- Brewster's law states that reflected light will be completely polarized at the angle of reflection θ_b, known as Brewster's angle, given by a statement known as Brewster's law: $\tan \theta_b = \frac{n_2}{n_1}$, where n_1 is the medium in which the incident and reflected light travel and n_2 is the index of refraction of the medium that forms the interface that reflects the light.
- Polarization can also be produced by scattering.
- There are a number of types of optically active substances that rotate the direction of polarization of light passing through them.

Conceptual Questions

1. Under what circumstances is the phase of light changed by reflection? Is the phase related to polarization?
2. Can a sound wave in air be polarized? Explain.
3. No light passes through two perfect polarizing filters with perpendicular axes. However, if a third polarizing filter is placed between the original two, some light can pass. Why is this? Under what circumstances does most of the light pass?
4. Explain what happens to the energy carried by light that it is dimmed by passing it through two crossed polarizing filters.
5. When particles scattering light are much smaller than its wavelength, the amount of scattering is proportional to $\frac{1}{\lambda^4}$. Does this mean there is more scattering for small λ than large λ? How does this relate to the fact that the sky is blue?
6. Using the information given in the preceding question, explain why sunsets are red.
7. When light is reflected at Brewster's angle from a smooth surface, it is 100% polarized parallel to the surface. Part of the light will be refracted into the surface. Describe how you would do an experiment to determine the polarization of the refracted light. What direction would you expect the polarization to have and would you expect it to be 100%?

Problems & Exercises

1. What angle is needed between the direction of polarized light and the axis of a polarizing filter to cut its intensity in half?
2. The angle between the axes of two polarizing filters is 45.0°. By how much does the second filter reduce the intensity of the light coming through the first?
3. If you have completely polarized light of intensity 150 W/m^2, what will its intensity be after passing through a polarizing filter with its axis at an 89.0° angle to the light's polarization direction?
4. What angle would the axis of a polarizing filter need to make with the direction of polarized light of intensity 1.00 kW/m^2 to reduce the intensity to 10.0 W/m^2?
5. At the end of Example 1, it was stated that the intensity of polarized light is reduced to 90.0% of its original value by passing through a polarizing filter with its axis at an angle of 18.4° to the direction of polarization. Verify this statement.
6. Show that if you have three polarizing filters, with the second at an angle of 45° to the first and the third at an angle of 90.0° to the first, the intensity of light passed by the first will be reduced to 25.0% of its value. (This is in contrast to having only the first and third, which reduces the intensity to zero, so that placing the second between them increases the intensity of the transmitted light.)
7. Prove that, if I is the intensity of light transmitted by two polarizing filters with axes at an angle θ and

I' is the intensity when the axes are at an angle $90.0° - \theta$, then $I + I' = I_0$ the original intensity. (Hint: Use the trigonometric identities $\cos(90.0° - \theta) = \sin\theta$ and $\cos^2\theta + \sin^2\theta = 1$.)

8. At what angle will light reflected from diamond be completely polarized?
9. What is Brewster's angle for light traveling in water that is reflected from crown glass?
10. A scuba diver sees light reflected from the water's surface. At what angle will this light be completely polarized?
11. At what angle is light inside crown glass completely polarized when reflected from water, as in a fish tank?
12. Light reflected at 55.6° from a window is completely polarized. What is the window's index of refraction and the likely substance of which it is made?
13. (a) Light reflected at 62.5° from a gemstone in a ring is completely polarized. Can the gem be a diamond? (b) At what angle would the light be completely polarized if the gem was in water?
14. If θ_b is Brewster's angle for light reflected from the top of an interface between two substances, and θ'_b is Brewster's angle for light reflected from below, prove that $\theta_b + \theta'_b = 90.0°$.
15. **Integrated Concepts.** If a polarizing filter reduces the intensity of polarized light to 50.0% of its original value, by how much are the electric and magnetic fields reduced?
16. **Integrated Concepts.** Suppose you put on two pairs of Polaroid sunglasses with their axes at an angle of

15.0°. How much longer will it take the light to deposit a given amount of energy in your eye compared with a single pair of sunglasses? Assume the lenses are clear except for their polarizing characteristics.

17. **Integrated Concepts.** (a) On a day when the intensity of sunlight is 1.00 kW/m^2, a circular lens 0.200 m in diameter focuses light onto water in a black beaker. Two polarizing sheets of plastic are placed in front of the lens with their axes at an angle of 20.0°. Assuming the sunlight is unpolarized and the polarizers are 100% efficient, what is the initial rate of heating of the water in °C/s, assuming it is 80.0% absorbed? The aluminum beaker has a mass of 30.0 grams and contains 250 grams of water. (b) Do the polarizing filters get hot? Explain.

Glossary

axis of a polarizing filter: the direction along which the filter passes the electric field of an EM wave

birefringent: crystals that split an unpolarized beam of light into two beams

Brewster's angle: $\theta_{\text{b}} = \tan^{-1}\left(\frac{n_2}{n_1}\right)$, where n_2 is the index of refraction of the medium from which the light is reflected and n_1 is the index of refraction of the medium in which the reflected light travels

Brewster's law: $\tan\theta_{\text{b}} = \frac{n_2}{n_1}$, where n_1 is the medium in which the incident and reflected light travel and n_2 is the index of refraction of the medium that forms the interface that reflects the light

direction of polarization: the direction parallel to the electric field for EM waves

horizontally polarized: the oscillations are in a horizontal plane

optically active: substances that rotate the plane of polarization of light passing through them

polarization: the attribute that wave oscillations have a definite direction relative to the direction of propagation of the wave

polarized: waves having the electric and magnetic field oscillations in a definite direction

reflected light that is completely polarized: light reflected at the angle of reflection θ_b, known as Brewster's angle

unpolarized: waves that are randomly polarized

vertically polarized: the oscillations are in a vertical plane

Selected Solutions to Problems & Exercises

1. 45.0º

3. 45.7 mW/m^2

5. 90.0%

7. I_0

9. 48.8º

11. 41.2º

13. (a) 1.92, not diamond (Zircon); (b) 55.2º

15. $B_2 = 0.707\ B_1$

17. (a) 2.07×10^{-2} °C/s; (b) Yes, the polarizing filters get hot

because they absorb some of the lost energy from the sunlight.

94. *Extended Topic* Microscopy Enhanced by the Wave Characteristics of Light

Learning Objective

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

- Discuss the different types of microscopes.

Physics research underpins the advancement of developments in microscopy. As we gain knowledge of the wave nature of electromagnetic waves and methods to analyze and interpret signals, new microscopes that enable us to "see" more are being developed. It is the evolution and newer generation of microscopes that are described in this section.

The use of microscopes (microscopy) to observe small details is limited by the wave nature of light. Owing to the fact that light diffracts significantly around small objects, it becomes impossible to observe details significantly smaller than the wavelength of light. One rule of thumb has it that all details smaller than about λ are difficult to observe. Radar, for example, can detect the size of an aircraft, but not its individual rivets, since the wavelength of most radar is several centimeters or greater. Similarly, visible light cannot detect individual atoms, since atoms are about 0.1 nm in size and visible wavelengths range from 380 to 760 nm. Ironically, special techniques used to obtain the best possible resolution with microscopes take advantage of the same wave characteristics of light that ultimately limit the detail.

Making Connections: Waves

All attempts to observe the size and shape of objects are limited by the wavelength of the probe. Sonar and medical ultrasound are limited by the wavelength of sound they employ. We shall see that this is also true in electron microscopy, since electrons have a wavelength. Heisenberg's uncertainty principle asserts that this limit is fundamental and inescapable, as we shall see in quantum mechanics.

The most obvious method of obtaining better detail is to utilize shorter wavelengths. *Ultraviolet* (UV) *microscopes* have been constructed with special lenses that transmit UV rays and utilize photographic or electronic techniques to record images. The shorter UV wavelengths allow somewhat greater detail to be observed, but drawbacks, such as the hazard of UV to living tissue and the need for special detection devices and lenses (which tend to be dispersive in the UV), severely limit the use of UV microscopes. Elsewhere, we will explore practical uses of very short wavelength EM waves, such as x rays, and other short-wavelength probes, such as electrons in electron microscopes, to detect small details.

Another difficulty in microscopy is the fact that many microscopic objects do not absorb much of the light passing through them. The lack of contrast makes image interpretation very difficult. *Contrast* is the difference in intensity between objects and the background on which they are observed. Stains (such as dyes, fluorophores, etc.) are commonly employed to enhance contrast, but these tend to be application specific. More general wave interference techniques can be used to produce contrast. Figure 1 shows the passage of light through a sample. Since the indices of refraction differ, the number of wavelengths in the paths differs. Light emerging from the object is thus out of phase with light from the background and will interfere differently, producing enhanced contrast, especially if the light is coherent and monochromatic—as in laser light.

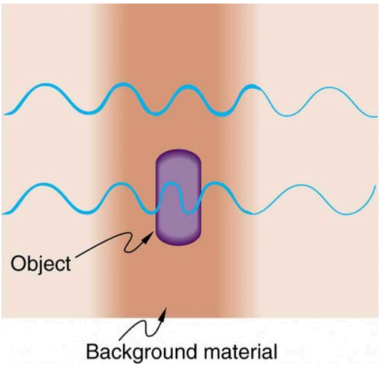


Figure 1. Light rays passing through a sample under a microscope will emerge with different phases depending on their paths. The object shown has a greater index of refraction than the background, and so the wavelength decreases as the ray passes through it. Superimposing these rays produces interference that varies with path, enhancing contrast between the object and background.

Interference microscopes enhance contrast between objects and background by superimposing a reference beam of light upon the light emerging from the sample. Since light from the background and objects differ in phase, there will be different amounts of constructive and destructive interference, producing the desired contrast in final intensity. Figure 2 shows schematically how this is done. Parallel rays of light from a source are split into two beams by a half-silvered mirror. These beams are called the object and reference beams. Each beam passes through identical optical elements, except that the object beam passes through the object we wish to observe microscopically. The light beams are recombined by another half-silvered mirror and interfere. Since the light rays passing through different parts of the object have different phases, interference will be significantly different and, hence, have greater contrast between them.

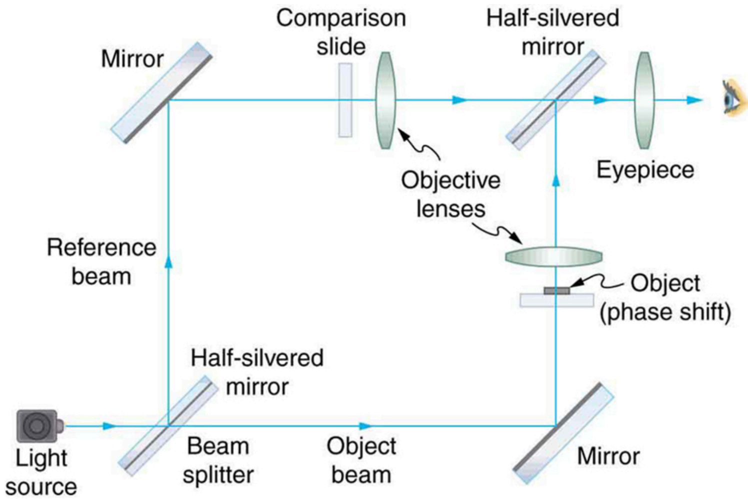


Figure 2. An interference microscope utilizes interference between the reference and object beam to enhance contrast. The two beams are split by a half-silvered mirror; the object beam is sent through the object, and the reference beam is sent through otherwise identical optical elements. The beams are recombined by another half-silvered mirror, and the interference depends on the various phases emerging from different parts of the object, enhancing contrast.

Another type of microscope utilizing wave interference and differences in phases to enhance contrast is called the *phase-contrast microscope*. While its principle is the same as the interference microscope, the phase-contrast microscope is simpler to use and construct. Its impact (and the principle upon which it is based) was so important that its developer, the Dutch physicist Frits Zernike (1888–1966), was awarded the Nobel Prize in 1953. Figure 3 shows the basic construction of a phase-contrast microscope. Phase differences between light passing through the object and background are produced by passing the rays through different parts of a phase plate (so called because it shifts the phase of the light passing through it). These two light rays are superimposed in the image plane, producing contrast due to their interference.

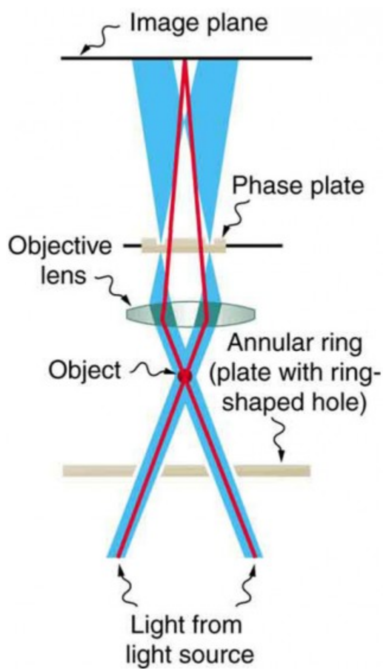


Figure 3. Simplified construction of a phase-contrast microscope. Phase differences between light passing through the object and background are produced by passing the rays through different parts of a phase plate. The light rays are superimposed in the image plane, producing contrast due to their interference.

A *polarization microscope* also enhances contrast by utilizing a wave characteristic of light. Polarization microscopes are useful for objects that are optically active or birefringent, particularly if those characteristics vary from place to place in the object. Polarized light is sent through the object and then observed through a polarizing filter that is perpendicular to the original polarization direction. Nearly transparent objects can then appear with strong color and in high contrast. Many polarization effects are wavelength dependent, producing color in the processed image. Contrast results from the action of the polarizing filter in passing only components parallel to its axis.

Apart from the UV microscope, the variations of microscopy discussed so far in this section are available as attachments to fairly standard microscopes or as slight variations. The next level of sophistication is provided by commercial *confocal microscopes*, which use the extended focal region shown in Figure 4b to obtain three-dimensional images rather than two-dimensional images.

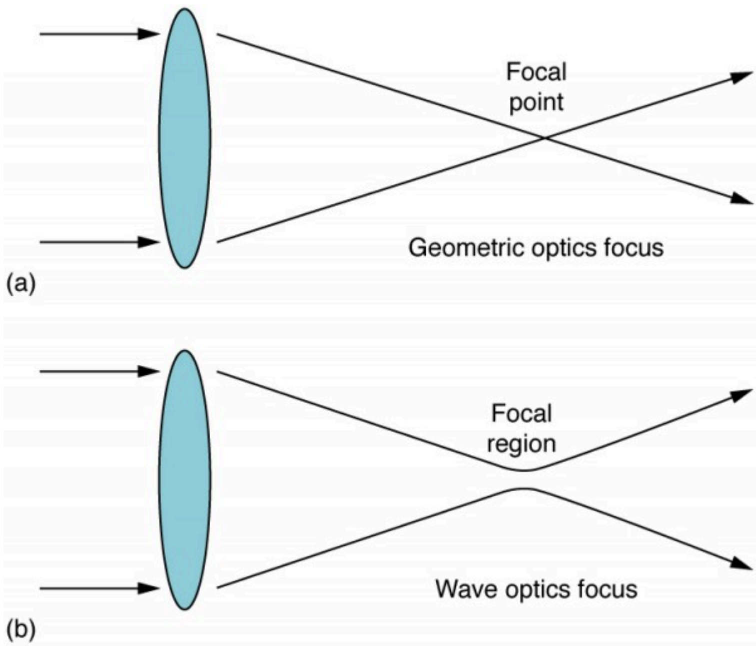


Figure 4. (a) In geometric optics, the focus is a point, but it is not physically possible to produce such a point because it implies infinite intensity. (b) In wave optics, the focus is an extended region.

Here, only a single plane or region of focus is identified; out-of-focus regions above and below this plane are subtracted out by a computer so the image quality is much better. This type of microscope makes use of fluorescence, where a laser provides the excitation light. Laser light passing through a tiny aperture called a pinhole forms an extended focal region within the specimen. The reflected light passes through the objective lens to a second pinhole and the photomultiplier detector, see Figure 5. The second pinhole is the key here and serves to block much of the light from points that are not at the focal point of the objective lens. The pinhole is conjugate (coupled) to the focal point of the lens. The second pinhole and detector are scanned, allowing reflected light from a small region or section of the extended focal region to be imaged at any one time. The out-of-focus light is excluded. Each image is stored in a computer, and a full scanned image is generated in a short time. Live cell processes can also be imaged at adequate scanning speeds allowing the imaging of three-dimensional microscopic movement. Confocal microscopy enhances images over conventional optical microscopy, especially for thicker specimens, and so has become quite popular.

The next level of sophistication is provided by microscopes attached to instruments that isolate and detect only a small wavelength band of light—monochromators and spectral analyzers. Here, the monochromatic light from a laser is scattered from the specimen. This scattered light shifts up or down as it excites particular energy levels in the sample. The uniqueness of the observed scattered light can give detailed information about the chemical composition of a given spot on the sample with high contrast—like molecular fingerprints. Applications are in materials science, nanotechnology, and the biomedical field. Fine details in biochemical processes over time can even be detected. The ultimate in microscopy is the electron microscope—to be discussed later. Research is being conducted into the development of new prototype microscopes that can become commercially available, providing better diagnostic and research capacities.

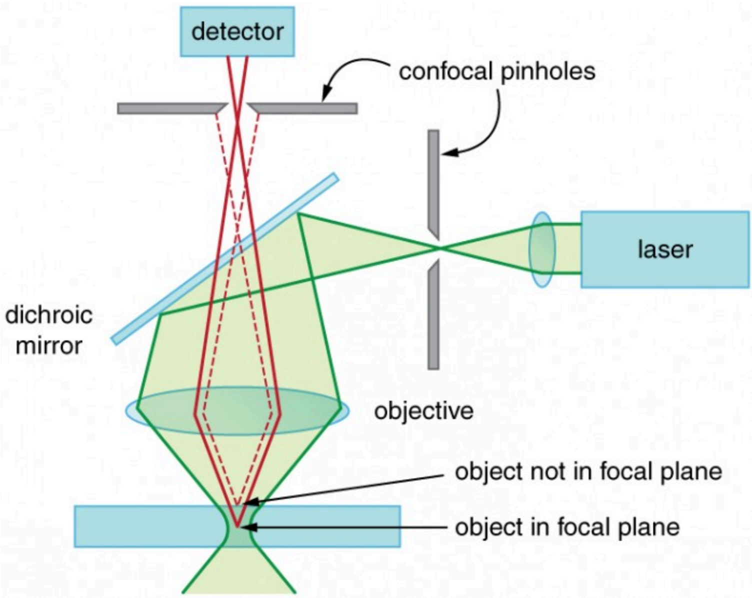


Figure 5. A confocal microscope provides three-dimensional images using pinholes and the extended depth of focus as described by wave optics. The right pinhole illuminates a tiny region of the sample in the focal plane. In-focus light rays from this tiny region pass through the dichroic mirror and the second pinhole to a detector and a computer. Out-of-focus light rays are blocked. The pinhole is scanned sideways to form an image of the entire focal plane. The pinhole can then be scanned up and down to gather images from different focal planes. The result is a three-dimensional image of the specimen.

Section Summary

- To improve microscope images, various techniques utilizing the wave characteristics of light have been developed. Many of these enhance contrast with interference effects.

Conceptual Questions

1. Explain how microscopes can use wave optics to improve contrast and why this is important.
2. A bright white light under water is collimated and directed upon a prism. What range of colors does one see emerging?

Glossary

confocal microscopes: microscopes that use the extended focal region to obtain three-dimensional images rather than two-dimensional images

contrast: the difference in intensity between objects and the background on which they are observed

interference microscopes: microscopes that enhance contrast between objects and background by superimposing a reference beam of light upon the light emerging from the sample

phase-contrast microscope: microscope utilizing wave interference and differences in phases to enhance contrast

polarization microscope: microscope that enhances contrast by utilizing a wave characteristic of light, useful for objects that are optically active

ultraviolet (UV) microscopes: microscopes constructed with special lenses that transmit UV rays and utilize photographic or electronic techniques to record images

PART XII

SPECIAL RELATIVITY

95. Introduction to Special Relativity

Figure 1. Special relativity explains why traveling to other star systems, such as these in the Orion Nebula, is unreasonable using our current level of technology. (credit: s58y, Flickr)

Have you ever looked up at the night sky and dreamed of traveling to other planets in faraway star systems? Would there be other life forms? What would other worlds look like? You might imagine that such an amazing trip would be possible if we could just travel fast enough, but you will read in this chapter why this is not true. In 1905 Albert Einstein developed the theory of special relativity. This theory explains the limit on an object's speed and describes the consequences.

Relativity. The word *relativity* might conjure an image of Einstein, but the idea did not begin with him. People have been exploring relativity for many centuries. Relativity is the study of how different observers measure the same event. Galileo and Newton developed the first correct version of classical relativity. Einstein developed the modern theory of relativity. Modern relativity is divided into two parts. *Special relativity* deals with observers who are moving at constant velocity. *General relativity* deals with observers who are undergoing acceleration. Einstein is famous because his theories of relativity made revolutionary predictions. Most importantly, his theories have been verified to great precision in a vast range of experiments, altering forever our concept of space and time.

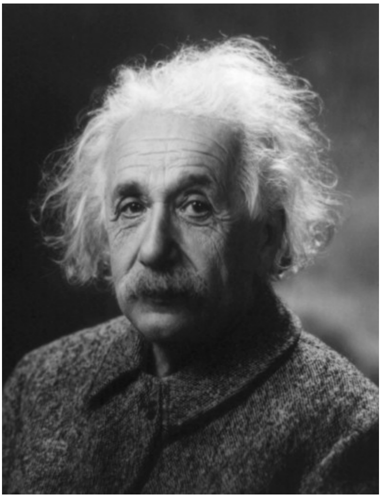


Figure 2. Many people think that Albert Einstein (1879–1955) was the greatest physicist of the 20th century. Not only did he develop modern relativity, thus revolutionizing our concept of the universe, he also made fundamental contributions to the foundations of quantum mechanics. (credit: The Library of Congress)

It is important to note that although classical mechanic, in general, and classical relativity, in particular, are limited, they are extremely good approximations for large, slow-moving objects. Otherwise, we could not use classical physics to launch satellites or build bridges. In the classical limit (objects larger than submicroscopic and moving slower than about 1% of the speed of light), relativistic mechanics becomes the same as classical mechanics. This fact will be noted at appropriate places throughout this chapter.

96. Einstein's Postulates

Learning Objectives

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

- State and explain both of Einstein's postulates.
- Explain what an inertial frame of reference is.
- Describe one way the speed of light can be changed

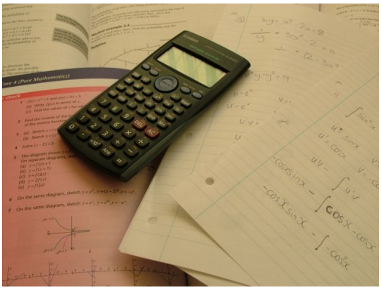


Figure 1. Special relativity resembles trigonometry in that both are reliable because they are based on postulates that flow one from another in a logical way. (credit: Jon Oakley, Flickr)

Have you ever used the Pythagorean Theorem and gotten a wrong answer? Probably not, unless you made a mistake in either your algebra or your arithmetic. Each time you perform the same calculation, you know that the answer will be the same. Trigonometry is reliable because of the certainty that one part always flows from another in a logical way. Each part is based on a set of postulates, and you can always connect the parts by applying those postulates. Physics is the same way with the exception that *all* parts must describe nature. If we are careful to choose the correct postulates, then our theory will follow and will be verified by experiment.

Einstein essentially did the theoretical aspect of this method for *relativity*. With two deceptively simple postulates and a careful consideration of how measurements are made, he produced the theory of *special relativity.*

Einstein's First Postulate

The first postulate upon which Einstein based the theory of special relativity relates to reference frames. All velocities are measured relative to some frame of reference. For example, a car's motion is measured relative to its starting point or the road it is moving over, a projectile's motion is measured relative to the surface it was launched from, and a planet's orbit is measured relative to the star it is orbiting around. The simplest frames of reference are those that are not accelerated and are not rotating. Newton's first law, the law of inertia, holds exactly in such a frame.

Inertial Reference Frame

An *inertial frame of reference* is a reference frame in which a body at rest remains at rest and a body in motion moves at a constant speed in a straight line unless acted on by an outside force.

The laws of physics seem to be simplest in inertial frames. For example, when you are in a plane flying at a constant altitude and speed, physics seems to work exactly the same as if you were standing on the surface of the Earth. However, in a plane that is taking off, matters are somewhat more complicated. In these cases, the net force on an object, F, is not equal to the product of mass and acceleration, ma. Instead, F is equal to ma plus a fictitious force. This situation is not as simple as in an inertial frame. Not only are laws of physics simplest in inertial frames, but they should be the same in all inertial frames, since there is no preferred frame and no absolute motion. Einstein incorporated these ideas into his *first postulate of special relativity*.

First Postulate of Special Relativity

The laws of physics are the same and can be stated in their simplest form in all inertial frames of reference.

As with many fundamental statements, there is more to this postulate than meets the eye. The laws of physics include only those that satisfy this postulate. We shall find that the definitions of relativistic momentum and energy must be altered to fit. Another outcome of this postulate is the famous equation $E = mc^2$.

Einstein’s Second Postulate

The second postulate upon which Einstein based his theory of special relativity deals with the speed of light. Late in the 19th century, the major tenets of classical physics were well established. Two of the most important were the laws of electricity and magnetism and Newton’s laws. In particular, the laws of electricity and magnetism predict that light travels at $c = 3.00 \times 10^8$ m/s in a vacuum, but they do not specify the frame of reference in which light has this speed.

There was a contradiction between this prediction and Newton's laws, in which velocities add like simple vectors. If the latter were true, then two observers moving at different speeds would see light traveling at different speeds. Imagine what a light wave would look like to a person traveling along with it at a speed c. If such a motion were possible then the wave would be stationary relative to the observer. It would have electric and magnetic fields that varied in strength at various distances from the observer but were constant in time. This is not allowed by Maxwell's equations. So either Maxwell's equations are wrong, or an object with mass cannot travel at speed c. Einstein concluded that the latter is true. An object with mass cannot travel at speed c. This conclusion implies that light in a vacuum must always travel at speed c relative to any observer. Maxwell's equations are correct, and Newton's addition of velocities is not correct for light.

Investigations such as Young's double slit experiment in the early-1800s had convincingly demonstrated that light is a wave. Many types of waves were known, and all travelled in some medium. Scientists therefore assumed that a medium carried light, even in a vacuum, and light travelled at a speed c relative to that medium. Starting in the mid-1880s, the American physicist A. A. Michelson, later aided by E. W. Morley, made a series of direct measurements of the speed of light. The results of their measurements were startling.

Michelson-Morley Experiment

The *Michelson-Morley experiment* demonstrated that the speed of light in a vacuum is independent of the motion of the Earth about the Sun.

The eventual conclusion derived from this result is that light, unlike mechanical waves such as sound, does not need a medium to carry it. Furthermore, the Michelson-Morley results implied that the speed of light c is independent of the motion of the source relative to the observer. That is, everyone observes light to move at speed c regardless of how they move relative to the source or one another. For a number of years, many scientists tried unsuccessfully to explain these results and still retain the general applicability of Newton's laws.

It was not until 1905, when Einstein published his first paper on special relativity, that the currently accepted conclusion was reached. Based mostly on his analysis that the laws of electricity and magnetism would not allow another speed for light, and only slightly aware of the Michelson-Morley experiment, Einstein detailed his *second postulate of special relativity*.

Second Postulate of Special Relativity

The speed of light c is a constant, independent of the relative motion of the source.

Deceptively simple and counterintuitive, this and the first postulate leave all else open for change. Some fundamental concepts do change. Among the changes are the loss of agreement on the elapsed time for an event, the variation of distance with speed, and the realization that matter and energy can be converted into one another. You will read about these concepts in the following sections.

Misconception Alert: Constancy of the Speed of Light

The speed of light is a constant $c = 3.00 \times 10^8$ m/s *in a vacuum*. If you remember the effect of the index of refraction from The Law of Refraction, the speed of light is lower in matter.

Section Summary

- Relativity is the study of how different observers measure the same event.
- Modern relativity is divided into two parts. Special relativity deals with observers who are in uniform (unaccelerated) motion, whereas general relativity includes accelerated relative motion and gravity. Modern relativity is correct in all circumstances and, in the limit of low velocity and weak gravitation, gives the same predictions as classical relativity.
- An inertial frame of reference is a reference frame in which a body at rest remains at rest and a body in motion moves at a constant speed in a straight line unless acted on by an outside force.
- Modern relativity is based on Einstein's two postulates. The

first postulate of special relativity is the idea that the laws of physics are the same and can be stated in their simplest form in all inertial frames of reference. The second postulate of special relativity is the idea that the speed of light *c* is a constant, independent of the relative motion of the source.
- The Michelson-Morley experiment demonstrated that the speed of light in a vacuum is independent of the motion of the Earth about the Sun.

Conceptual Questions

1. Which of Einstein's postulates of special relativity includes a concept that does not fit with the ideas of classical physics? Explain.
2. Is Earth an inertial frame of reference? Is the Sun? Justify your response.
3. When you are flying in a commercial jet, it may appear to you that the airplane is stationary and the Earth is moving beneath you. Is this point of view valid? Discuss briefly.

Glossary

relativity: the study of how different observers measure the same event

special relativity: the theory that, in an inertial frame of reference, the motion of an object is relative to the frame from which it is viewed or measured

inertial frame of reference: a reference frame in which a body at rest remains at rest and a body in motion moves at a constant speed in a straight line unless acted on by an outside force

first postulate of special relativity: the idea that the laws of physics are the same and can be stated in their simplest form in all inertial frames of reference

second postulate of special relativity: the idea that the speed of light *c* is a constant, independent of the source

Michelson-Morley experiment: an investigation performed in 1887 that proved that the speed of light in a vacuum is the same in all frames of reference from which it is viewed

97. Relativistic Energy

Learning Objectives

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

- Compute total energy of a relativistic object.
- Compute the kinetic energy of a relativistic object.
- Describe rest energy, and explain how it can be converted to other forms.
- Explain why massive particles cannot reach C.

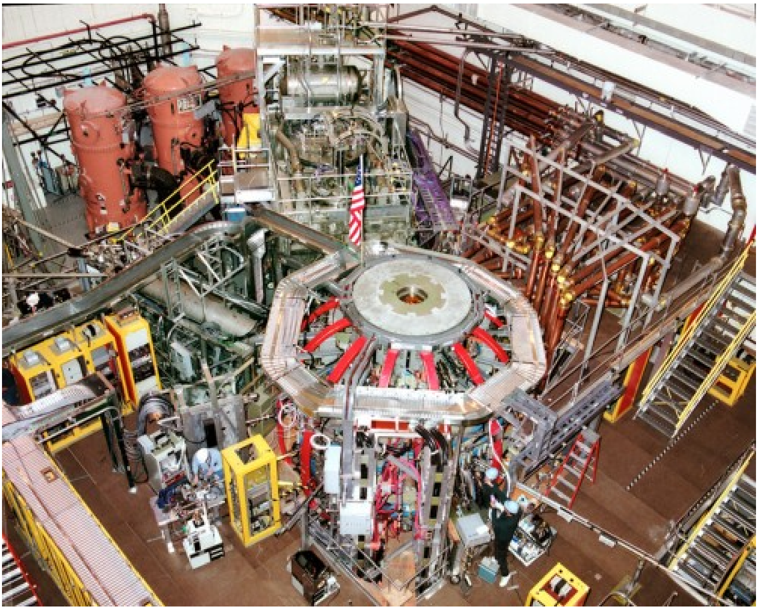


Figure 1. The National Spherical Torus Experiment (NSTX) has a fusion reactor in which hydrogen isotopes undergo fusion to produce helium. In this process, a relatively small mass of fuel is converted into a large amount of energy. (credit: Princeton Plasma Physics Laboratory)

A tokamak is a form of experimental fusion reactor, which can change mass to energy. Accomplishing this requires an understanding of relativistic energy. Nuclear reactors are proof of the conservation of relativistic energy.

Conservation of energy is one of the most important laws in physics. Not only does energy have many important forms, but each form can be converted to any other. We know that classically the total amount of energy in a system remains constant. Relativistically, energy is still conserved, provided its definition is altered to include the possibility of mass changing to energy, as in the reactions that occur within a nuclear reactor. Relativistic energy is intentionally defined so that it will be conserved in all inertial frames, just as is the case for relativistic momentum. As a consequence, we learn that several fundamental quantities are related in ways not known in classical physics. All of these relationships are verified by experiment and have fundamental consequences. The altered definition of energy contains some of the most fundamental and spectacular new insights into nature found in recent history.

Total Energy and Rest Energy

The first postulate of relativity states that the laws of physics are the same in all inertial frames. Einstein showed that the law of conservation of energy is valid relativistically, if we define energy to include a relativistic factor.

Total Energy

Total energy E is defined to be $E = \gamma mc^2$, where m is mass, c is the speed of light,

$$\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}}$$

and v is the velocity of the mass relative to an observer. There are many aspects of the total energy E that we will discuss—among them are how kinetic and potential energies are included in E, and how E is related to relativistic momentum. But first, note that at rest, total energy is not zero. Rather, when $v = 0$, we have $\gamma = 1$, and an object has rest energy.

Rest Energy

Rest energy is $E_0 = mc^2$.

This is the correct form of Einstein's most famous equation, which for the first time showed that energy is related to the mass of an object at rest. For example, if energy is stored in the object, its rest mass increases. This also implies that mass can be destroyed to release energy. The implications of these first two equations regarding relativistic energy are so broad that they were not completely recognized for some years after Einstein published them in 1907, nor was the experimental proof that they are correct widely recognized at first. Einstein, it should be noted, did understand and describe the meanings and implications of his theory.

Example 1. Calculating Rest Energy: Rest Energy is Very Large

Calculate the rest energy of a 1.00-g mass.

Strategy

One gram is a small mass—less than half the mass of a penny. We can multiply this mass, in SI units, by the speed of light squared to find the equivalent rest energy.

Solution

Identify the knowns: $m = 1.00 \times 10^{-3}$ kg; $c = 3.00 \times 10^{8}$ m/s

Identify the unknown: E_0

Choose the appropriate equation. $E_0 = mc^2$

Plug the knowns into the equation.

$$\begin{aligned} E_0 &= mc^2 = \left(1.00 \times 10^{-3}\ \text{kg}\right)\left(3.00 \times 10^{8}\ \text{m/s}\right)^2 \\ &= 9.00 \times 10^{13}\ \text{kg} \cdot\ \text{m}^2/\text{s}^2 \end{aligned}$$

Convert units.Noting that $1\ \text{kg} \cdot \text{m}^2/\text{s}^2 = 1$ J, we see the rest mass energy is $E_0 = 9.00 \times 10^{13}$ J.

Discussion

This is an enormous amount of energy for a 1.00-g mass. We do not notice this energy, because it is generally not available. Rest energy is large because the speed of light c is a large number and c^2 is a very large number, so that mc^2 is huge for any macroscopic mass. The 9.00×10^{13} J rest mass energy for 1.00 g is about twice the energy released by the

Hiroshima atomic bomb and about 10,000 times the kinetic energy of a large aircraft carrier. If a way can be found to convert rest mass energy into some other form (and all forms of energy can be converted into one another), then huge amounts of energy can be obtained from the destruction of mass.

Today, the practical applications of *the conversion of mass into another form of energy*, such as in nuclear weapons and nuclear power plants, are well known. But examples also existed when Einstein first proposed the correct form of relativistic energy, and he did describe some of them. Nuclear radiation had been discovered in the previous decade, and it had been a mystery as to where its energy originated. The explanation was that, in certain nuclear processes, a small amount of mass is destroyed and energy is released and carried by nuclear radiation. But the amount of mass destroyed is so small that it is difficult to detect that any is missing. Although Einstein proposed this as the source of energy in the radioactive salts then being studied, it was many years before there was broad recognition that mass could be and, in fact, commonly is converted to energy. (See Figure 2.)

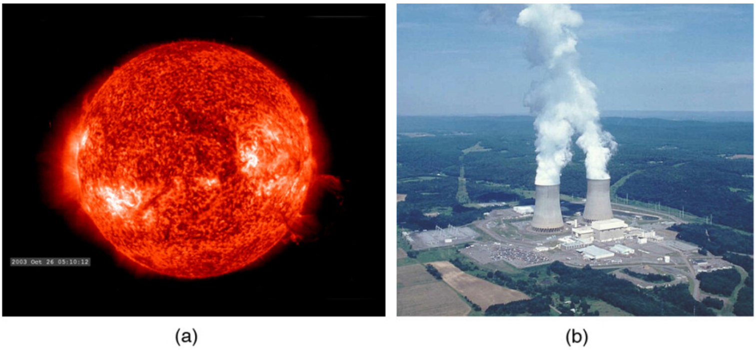


Figure 2. The Sun (a) and the Susquehanna Steam Electric Station (b) both convert mass into energy—the Sun via nuclear fusion, the electric station via nuclear fission. (credits: (a) NASA/Goddard Space Flight Center, Scientific Visualization Studio; *(b)* U.S. government*)*

Because of the relationship of rest energy to mass, we now consider mass to be a form of energy rather than something separate. There had not even been a hint of this prior to Einstein's work. Such conversion is now known to be the source of the Sun's energy, the energy of nuclear decay, and even the source of energy keeping Earth's interior hot.

Stored Energy and Potential Energy

What happens to energy stored in an object at rest, such as the energy put into a battery by charging it, or the energy stored in a toy gun's compressed spring? The energy input becomes part of the total energy of the object and, thus, increases its rest mass. All stored and potential energy becomes mass in a system. Why is it we don't ordinarily notice this? In fact, conservation of mass (meaning total mass is constant) was one of the great laws verified by 19th-century science. Why was it not noticed to be incorrect? Example 2 helps answer these questions.

Example 2. Calculating Rest Mass: A Small Mass Increase due to Energy Input

A car battery is rated to be able to move 600 ampere-hours ($A \cdot h$) of charge at 12.0 V.

1. Calculate the increase in rest mass of such a battery when it is taken from being fully depleted to being fully charged.
2. What percent increase is this, given the battery's mass is 20.0 kg?

Strategy

In Part 1, we first must find the energy stored in the battery, which equals what the battery can supply in the form of electrical potential energy. Since $PE_{elec} = qV$, we have to calculate the charge q in $600\ A \cdot h$, which is the product of the current I and the time t. We then multiply the result by 12.0 V. We can then calculate the battery's increase in mass using $\Delta E = PE_{elec} = (\Delta m)c^2$. Part 2 is a simple ratio converted to a percentage.

Solution for Part 1

Identify the knowns: $I \cdot t = 600\ A \cdot h$; $V = 12.0\ V$; $c = 3.00 \times 10^8\ m/s$

Identify the unknown: Δm

Choose the appropriate equation. $PE_{elec}=(\Delta m)c^2$

Rearrange the equation to solve for the unknown.

$$\Delta m = \frac{\text{PE}_{\text{elec}}}{c^2}.$$

Plug the knowns into the equation.

$$\begin{aligned}\Delta m &= \frac{\text{PE}_{\text{elec}}}{c^2}\\ &= \frac{qV}{c^2}\\ &= \frac{(It)V}{c^2}\\ &= \frac{(600\ \text{A}\cdot\text{h})(12.0\ \text{V})}{\left(3.00\times10^8\right)^2}\end{aligned}$$

Write amperes A as coulombs per second (C/s), and convert hours to seconds.

$$\begin{aligned}\Delta m &= \frac{\left(600\ \text{C/s}\cdot\text{h}\left(\frac{3600\ \text{s}}{1\ \text{h}}\right)\right)(12.0\ \text{J/C})}{\left(3.00\times10^8\ \text{m/s}\right)^2}\\ &= \frac{\left(2.16\times10^6\ \text{C}\right)(12.0\ \text{J/C})}{\left(3.00\times10^8\ \text{m/s}\right)^2}\end{aligned}$$

Using the conversion $1\ \text{kg}\cdot\text{m}^2/\text{s}^2 = 1\ \text{J}$, we can write the mass as $\Delta m = 2.88\times10^{-10}$ kg.

Solution for Part 2

Identify the knowns: $\Delta m = 2.88\times10^{-10}$ kg; m=20.0 kg

Identify the unknown: % change

Choose the appropriate equation.

$$\%\ \text{increase} = \frac{\Delta m}{m}\times100\%$$

Plug the knowns into the equation.

$$\begin{aligned}\% \text{ increase} &= \frac{\Delta m}{m} \times 100\% \\ &= \frac{2.88 \times 10^{-10}\text{ kg}}{20.0\text{ kg}} \times 100\% \\ &= 1.44 \times 10^{-9}\%\end{aligned}$$

Discussion

Both the actual increase in mass and the percent increase are very small, since energy is divided by c^2, a very large number. We would have to be able to measure the mass of the battery to a precision of a billionth of a percent, or 1 part in 10^{11}, to notice this increase. It is no wonder that the mass variation is not readily observed. In fact, this change in mass is so small that we may question how you could verify it is real. The answer is found in nuclear processes in which the percentage of mass destroyed is large enough to be measured. The mass of the fuel of a nuclear reactor, for example, is measurably smaller when its energy has been used. In that case, stored energy has been released (converted mostly to heat and electricity) and the rest mass has decreased. This is also the case when you use the energy stored in a battery, except that the stored energy is much greater in nuclear processes, making the change in mass measurable in practice as well as in theory.

Kinetic Energy and the Ultimate Speed Limit

Kinetic energy is energy of motion. Classically, kinetic energy has the familiar expression $\frac{1}{2}mv^2$. The relativistic expression for kinetic energy is obtained from the work-energy theorem. This theorem states that the net work on a system goes into kinetic energy. If our system starts from rest, then the work-energy theorem is $W_{net} = KE$.

Relativistically, at rest we have rest energy $E_0 = mc^2$. The work increases this to the total energy $E = \gamma mc^2$. Thus, $W_{net} = E - E_0 = \gamma mc^2 - mc^2 = (\gamma - 1)mc^2$.

Relativistically, we have $W_{net} = KE_{rel}$.

Relativistic Kinetic Energy

Relativistic kinetic energy is $KE_{rel} = (\gamma - 1)mc^2$.

When motionless, we have $v = 0$ and

$$\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} = 1,$$

so that $KE_{rel} = 0$ at rest, as expected. But the expression for relativistic kinetic energy (such as total energy and rest energy) does not look much like the classical $\frac{1}{2}mv^2$. To show that the classical expression for kinetic energy is obtained at low velocities, we note that the binomial expansion for γ at low velocities gives

$$\gamma = 1 + \frac{1v^2}{2c^2}.$$

A binomial expansion is a way of expressing an algebraic quantity as a sum of an infinite series of terms. In some cases, as in the limit of small velocity here, most terms are very small. Thus the expression derived for γ here is not exact, but it is a very accurate approximation. Thus, at low velocities,

$$\gamma - 1 = \frac{1v^2}{2c^2}.$$

Entering this into the expression for relativistic kinetic energy gives

$$\mathrm{KE}_{\mathrm{rel}} = \left(\frac{1v^2}{2c^2}\right)mc^2 = \frac{1}{2}mv^2 = \mathrm{KE}_{\mathrm{class}}.$$

So, in fact, relativistic kinetic energy does become the same as classical kinetic energy when $v << c$.

It is even more interesting to investigate what happens to kinetic energy when the velocity of an object approaches the speed of light. We know that γ becomes infinite as v approaches c, so that $\mathrm{KE}_{\mathrm{rel}}$ also becomes infinite as the velocity approaches the speed of light. (See Figure 3.) An infinite amount of work (and, hence, an infinite amount of energy input) is required to accelerate a mass to the speed of light.

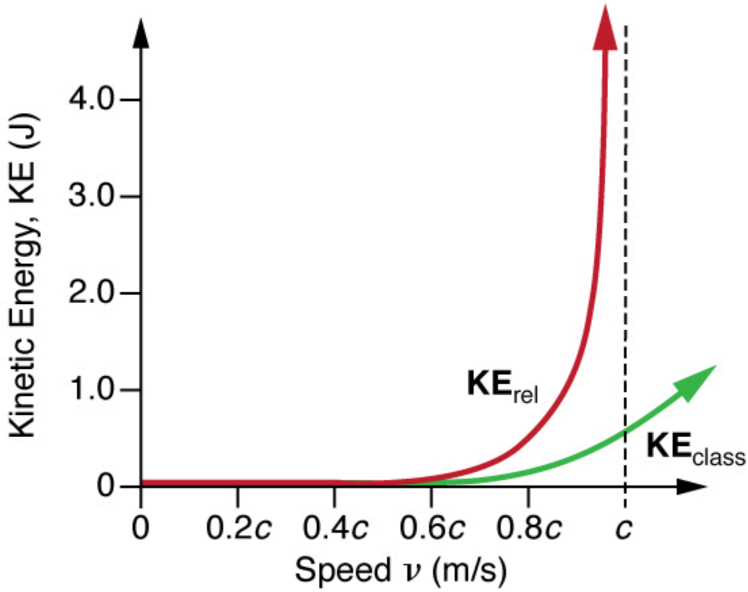


Figure 3. This graph of KE_{rel} *versus velocity shows how kinetic energy approaches infinity as velocity approaches the speed of light. It is thus not possible for an object having mass to reach the speed of light. Also shown is* KE_{class}*, the classical kinetic energy, which is similar to relativistic kinetic energy at low velocities. Note that much more energy is required to reach high velocities than predicted classically.*

The Speed of Light

No object with mass can attain the speed of light.

So the speed of light is the ultimate speed limit for any particle having mass. All of this is consistent with the fact that velocities less than c always add to less than c. Both the relativistic form for kinetic energy and the ultimate speed limit being c have been confirmed in detail in numerous experiments. No matter how much energy is put into accelerating a mass, its velocity can only approach—not reach—the speed of light.

Example 3. Comparing Kinetic Energy: Relativistic Energy Versus Classical Kinetic Energy

An electron has a velocity $v = 0.990c$.

1. Calculate the kinetic energy in MeV of the electron.
2. Compare this with the classical value for kinetic energy at this velocity. (The mass of an electron is 9.11×10^{-31} kg.)

Strategy

The expression for relativistic kinetic energy is always correct, but for Part 1 it must be used since the velocity is highly relativistic (close to c). First, we will calculate the relativistic factor γ, and then use it to determine the relativistic kinetic energy. For Part 2, we will calculate the classical kinetic energy (which would be close to the relativistic value if v were less than a few percent of c) and see that it is not the same.

Solution for Part 1

Identify the knowns: $v = 0.990c$; $m = 9.11 \times 10^{-31}$ kg

Identify the unknown: KE_{rel}

Choose the appropriate equation. $KE_{rel} = (\gamma - 1)mc^2$

Plug the knowns into the equation.First calculate γ. We will carry extra digits because this is an intermediate calculation.

$$\begin{aligned}\gamma &= \frac{1}{\sqrt{1-\frac{v^2}{c^2}}}\\ &= \frac{1}{\sqrt{1-\frac{(0.990c)^2}{c^2}}}\\ &= \frac{1}{\sqrt{1-(0.990)^2}}\\ &= 7.0888\end{aligned}$$

Next, we use this value to calculate the kinetic energy.

$$\begin{aligned}\mathrm{KE_{rel}} &= (\gamma - 1)mc^2\\ &= (7.0888 - 1)\left(9.11 \times 10^{-31}\ \mathrm{kg}\right)\left(3.00 \times 10^8\ \mathrm{m/s}\right)^2\\ &= 4.99 \times 10^{-13}\ \mathrm{J}\end{aligned}$$

Convert units.

$$\begin{aligned}\mathrm{KE_{rel}} &= \left(4.99 \times 10^{-13}\ \mathrm{J}\right)\left(\frac{1\ \mathrm{MeV}}{1.60\times10^{-13}\ \mathrm{J}}\right)\\ &= 3.12\ \mathrm{MeV}\end{aligned}$$

Solution for Part 2

List the knowns: $v = 0.990c$; $m = 9.11 \times 10^{-31}$ kg

List the unknown: KE_{class}

Choose the appropriate equation. $\text{KE}_{\text{class}} = \dfrac{1}{2}mv^2$

Plug the knowns into the equation.

$$\begin{aligned} KE_{\text{class}} &= \tfrac{1}{2}mv^2 \\ &= \tfrac{1}{2}\left(9.00 \times 10^{-31}\text{ kg}\right)(0.990)^2\left(3.00 \times 10^8\text{ m/s}\right)^2 \\ &= 4.02 \times 10^{-14}\text{ J} \end{aligned}$$

Convert units.

$$\begin{aligned} \text{KE}_{\text{class}} &= 4.02 \times 10^{-14}\text{ J}\left(\frac{1\text{ MeV}}{1.60 \times 10^{-13}\text{ J}}\right) \\ &= 0.251\text{ MeV} \end{aligned}$$

Discussion

As might be expected, since the velocity is 99.0% of the speed of light, the classical kinetic energy is significantly off from the correct relativistic value. Note also that the classical value is much smaller than the relativistic value. In fact, $\dfrac{\text{KE}_{\text{rel}}}{\text{KE}_{\text{class}}} = 12.4$ here. This is some indication of how difficult it is to get a mass moving close to the speed of light. Much more energy is required than predicted classically.

Some people interpret this extra energy as going into increasing the mass of the system, but, as discussed in Relativistic Momentum, this cannot be verified unambiguously. What is certain is that ever-increasing amounts of energy are needed to get the velocity of a mass a little closer to that of light. An energy of 3 MeV is a very small amount for an electron, and it can be achieved with present-day particle accelerators. SLAC, for example, can accelerate electrons to over 50×10^9 eV = 50,000 MeV.

Figure 4. The Fermi National Accelerator Laboratory, near Batavia, Illinois, was a subatomic particle collider that accelerated protons and antiprotons to attain energies up to 1 Tev (a trillion electronvolts). The circular ponds near the rings were built to dissipate waste heat. This accelerator was shut down in September 2011. (credit: Fermilab, Reidar Hahn)

Is there any point in getting v a little closer to c than 99.0% or 99.9%? The answer is yes. We learn a great deal by doing this. The energy that goes into a high-velocity mass can be converted to any other form, including into entirely new masses. (See Figure 4.) Most of what we know about the substructure of matter and the collection of exotic short-lived particles in nature has been learned this way. Particles are accelerated to extremely relativistic energies and made to collide with other particles, producing totally new species of particles. Patterns in the characteristics of these previously unknown particles hint at a basic substructure for all matter. These particles and some of their characteristics will be covered in Particle Physics.

Relativistic Energy and Momentum

We know classically that kinetic energy and momentum are related to each other, since

$$\mathrm{KE}_{\mathrm{class}} = \frac{p^2}{2m} = \frac{(mv)^2}{2m} = \frac{1}{2}mv^2.$$

Relativistically, we can obtain a relationship between energy and momentum by algebraically manipulating their definitions. This produces $E^2 = (pc)^2 + (mc^2)^2$, where E is the relativistic total energy and p is the relativistic momentum. This relationship between relativistic energy and relativistic momentum is more complicated than the classical, but we can gain some interesting new insights by examining it. First, total energy is related to momentum and rest mass. At rest, momentum is zero, and the equation gives the total energy to be the rest energy mc^2 (so this equation is consistent with the discussion of rest energy above). However, as the mass is accelerated, its momentum p increases, thus increasing the total energy. At sufficiently high velocities, the rest energy term $(mc^2)^2$ becomes negligible compared with the momentum term $(pc)^2$; thus, $E = pc$ at extremely relativistic velocities.

If we consider momentum p to be distinct from mass, we can determine the implications of the equation $E^2 = (pc)^2 + (mc^2)^2$, for a particle that has no mass. If we take m to be zero in this equation, then $E = pc, \text{ or } p = \dfrac{E}{c}$. Massless particles have this momentum. There are several massless particles found in nature, including photons (these are quanta of electromagnetic radiation). Another implication is that a massless particle must travel at speed c and only at speed c. While it is beyond the scope of this text to examine the relationship in the equation $E^2 = (pc)^2 + (mc^2)^2$, in detail, we can see that the relationship has important implications in special relativity.

Problem-Solving Strategies for Relativity

1. *Examine the situation to determine that it is necessary to use relativity*. Relativistic effects are related to
$$\gamma = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}},$$
the quantitative relativistic factor. If γ is very close to 1, then relativistic effects are small and differ very little from the usually easier classical calculations.
2. *Identify exactly what needs to be determined in the problem (identify the unknowns).*
3. *Make a list of what is given or can be inferred from the problem as stated (identify the knowns).* Look in particular for information on relative velocity v.
4. *Make certain you understand the conceptual aspects of the problem before making any calculations.* Decide, for example, which observer sees time dilated or length contracted before plugging into equations. If you have thought about who sees what, who is moving with the event being observed, who sees proper time, and so on, you will find it much easier to determine if your calculation is reasonable.
5. *Determine the primary type of calculation to be done to find the unknowns identified above.* You will find the section summary helpful in determining whether a length contraction, relativistic kinetic energy, or some other concept is involved.
6. *Do not round off during the calculation.* As noted in the text, you must often perform your calculations to

many digits to see the desired effect. You may round off at the very end of the problem, but do not use a rounded number in a subsequent calculation.

7. *Check the answer to see if it is reasonable: Does it make sense?* This may be more difficult for relativity, since we do not encounter it directly. But you can look for velocities greater than c or relativistic effects that are in the wrong direction (such as a time contraction where a dilation was expected).

Check Your Understanding

A photon decays into an electron-positron pair. What is the kinetic energy of the electron if its speed is 0.992c?

Solution

$$\mathrm{KE_{rel}} = (\gamma - 1)mc^2 = \left(\frac{1}{\sqrt{1-\frac{v^2}{c^2}}} - 1\right)mc^2$$

$$= \left(\frac{1}{\sqrt{1-\frac{(0.992c)^2}{c^2}}} - 1\right)(9.11 \times 10^{-31}\ \mathrm{kg})(3.00 \times 10^{8}\ \mathrm{m/s})^2 = 5.67 \times 10^{-13}\ \mathrm{J}$$

Section Summary

- Relativistic energy is conserved as long as we define it to include the possibility of mass changing to energy.
- Total Energy is defined as: $E = \gamma mc^2$, where $\gamma = \dfrac{1}{\sqrt{1 - \frac{v^2}{c^2}}}$.
- Rest energy is $E_0 = mc^2$, meaning that mass is a form of energy. If energy is stored in an object, its mass increases. Mass can be destroyed to release energy.
- We do not ordinarily notice the increase or decrease in mass of an object because the change in mass is so small for a large increase in energy.
- The relativistic work-energy theorem is $W_{net} = E - E_0 = \gamma mc^2 - mc^2 = (\gamma - 1)mc^2$.
- Relativistically, $W_{net} = KE_{rel}$, where KE_{rel} is the relativistic kinetic energy.
- Relativistic kinetic energy is $KE_{rel} = (\gamma - 1)mc^2$, where $\gamma = \dfrac{1}{\sqrt{1 - \frac{v^2}{c^2}}}$. At low velocities, relativistic kinetic energy reduces to classical kinetic energy.
- No object with mass can attain the speed of light because an infinite amount of work and an infinite amount of energy input is required to accelerate a mass to the speed of light.
- The equation $E^2 = (pc)^2 + (mc^2)^2$ relates the relativistic total energy E and the relativistic momentum p. At extremely high velocities, the rest energy mc^2 becomes negligible, and $E = pc$.

Conceptual Questions

1. How are the classical laws of conservation of energy and conservation of mass modified by modern relativity?
2. What happens to the mass of water in a pot when it cools, assuming no molecules escape or are added? Is this observable in practice? Explain.
3. Consider a thought experiment. You place an expanded balloon of air on weighing scales outside in the early morning. The balloon stays on the scales and you are able to measure changes in its mass. Does the mass of the balloon change as the day progresses? Discuss the difficulties in carrying out this experiment.
4. The mass of the fuel in a nuclear reactor decreases by an observable amount as it puts out energy. Is the same true for the coal and oxygen combined in a conventional power plant? If so, is this observable in practice for the coal and oxygen? Explain.
5. We know that the velocity of an object with mass has an upper limit of c. Is there an upper limit on its momentum? Its energy? Explain.
6. Given the fact that light travels at c, can it have mass? Explain.
7. If you use an Earth-based telescope to project a laser beam onto the Moon, you can move the spot across the Moon's surface at a velocity greater than the speed of light. Does this violate modern relativity? (Note that light is being sent from the Earth to the Moon, not across the surface of the Moon.)

Problems & Exercises

1. What is the rest energy of an electron, given its mass is 9.11×10^{-31} kg? Give your answer in joules and MeV.
2. Find the rest energy in joules and MeV of a proton, given its mass is 1.67×10^{-27} kg.
3. If the rest energies of a proton and a neutron (the two constituents of nuclei) are 938.3 and 939.6 MeV respectively, what is the difference in their masses in kilograms?
4. The Big Bang that began the universe is estimated to have released 10^{68} J of energy. How many stars could half this energy create, assuming the average star's mass is 4.00×10^{30} kg?
5. A supernova explosion of a 2.00×10^{31} kg star produces 1.00×10^{44} kg of energy. (a) How many kilograms of mass are converted to energy in the explosion? (b) What is the ratio $\frac{\Delta m}{m}$ of mass destroyed to the original mass of the star?
6. (a) Using data from [Table 1 in Conservation of Energy](), calculate the mass converted to energy by the fission of 1.00 kg of uranium. (b) What is the ratio of mass destroyed to the original mass, $\frac{\Delta m}{m}$?
7. (a) Using data from [Table 1 in Conservation of Energy](), calculate the amount of mass converted to energy by the fusion of 1.00 kg of hydrogen. (b) What is the ratio of mass destroyed to the original mass,

$\frac{\Delta m}{m}$? (c) How does this compare with $\frac{\Delta m}{m}$ for the fission of 1.00 kg of uranium?

8. There is approximately 10^{34} J of energy available from fusion of hydrogen in the world's oceans. (a) If 10^{33} J of this energy were utilized, what would be the decrease in mass of the oceans? (b) How great a volume of water does this correspond to? (c) Comment on whether this is a significant fraction of the total mass of the oceans.
9. A muon has a rest mass energy of 105.7 MeV, and it decays into an electron and a massless particle. (a) If all the lost mass is converted into the electron's kinetic energy, find γ for the electron. (b) What is the electron's velocity?
10. A π-meson is a particle that decays into a muon and a massless particle. The π-meson has a rest mass energy of 139.6 MeV, and the muon has a rest mass energy of 105.7 MeV. Suppose the π-meson is at rest and all of the missing mass goes into the muon's kinetic energy. How fast will the muon move?
11. (a) Calculate the relativistic kinetic energy of a 1000-kg car moving at 30.0 m/s if the speed of light were only 45.0 m/s. (b) Find the ratio of the relativistic kinetic energy to classical.
12. Alpha decay is nuclear decay in which a helium nucleus is emitted. If the helium nucleus has a mass of 6.80×10^{-27} kg and is given 5.00 MeV of kinetic energy, what is its velocity?
13. (a) Beta decay is nuclear decay in which an electron is emitted. If the electron is given 0.750 MeV of kinetic energy, what is its velocity? (b) Comment on

how the high velocity is consistent with the kinetic energy as it compares to the rest mass energy of the electron.

14. A positron is an antimatter version of the electron, having exactly the same mass. When a positron and an electron meet, they annihilate, converting all of their mass into energy. (a) Find the energy released, assuming negligible kinetic energy before the annihilation. (b) If this energy is given to a proton in the form of kinetic energy, what is its velocity? (c) If this energy is given to another electron in the form of kinetic energy, what is its velocity?

15. What is the kinetic energy in MeV of a π-meson that lives 1.40×10^{-16} s as measured in the laboratory, and 0.840×10^{-16} s when at rest relative to an observer, given that its rest energy is 135 MeV?

16. Find the kinetic energy in MeV of a neutron with a measured life span of 2065 s, given its rest energy is 939.6 MeV, and rest life span is 900s.

17. (a) Show that $\frac{(pc)^2}{(mc^2)^2} = \gamma^2 - 1$. This means that at large velocities $pc >> mc^2$. (b) Is $E \approx pc$ when γ = 30.0, as for the astronaut discussed in the twin paradox?

18. One cosmic ray neutron has a velocity of $0.250c$ relative to the Earth. (a) What is the neutron's total energy in MeV? (b) Find its momentum. (c) Is $E \approx pc$ in this situation? Discuss in terms of the equation given in part (a) of the previous problem.

19. What is γ for a proton having a mass energy of 938.3 MeV accelerated through an effective potential

of 1.0 TV (teravolt) at Fermilab outside Chicago?

20. (a) What is the effective accelerating potential for electrons at the Stanford Linear Accelerator, if $\gamma = 1.00 \times 10^5$ for them? (b) What is their total energy (nearly the same as kinetic in this case) in GeV?

21. (a) Using data from [Table 1 in Conservation of Energy](), find the mass destroyed when the energy in a barrel of crude oil is released. (b) Given these barrels contain 200 liters and assuming the density of crude oil is 750 kg/m^3, what is the ratio of mass destroyed to original mass, $\frac{\Delta m}{m}$?

22. (a) Calculate the energy released by the destruction of 1.00 kg of mass. (b) How many kilograms could be lifted to a 10.0 km height by this amount of energy?

23. A Van de Graaff accelerator utilizes a 50.0 MV potential difference to accelerate charged particles such as protons. (a) What is the velocity of a proton accelerated by such a potential? (b) An electron?

24. Suppose you use an average of 500 kW · h of electric energy per month in your home. (a) How long would 1.00 g of mass converted to electric energy with an efficiency of 38.0% last you? (b) How many homes could be supplied at the 500 kW · h per month rate for one year by the energy from the described mass conversion?

25. (a) A nuclear power plant converts energy from nuclear fission into electricity with an efficiency of 35.0%. How much mass is destroyed in one year to produce a continuous 1000 MW of electric power? (b) Do you think it would be possible to observe this mass loss if the total mass of the fuel is 10^4 kg?

26. Nuclear-powered rockets were researched for some years before safety concerns became paramount. (a) What fraction of a rocket's mass would have to be destroyed to get it into a low Earth orbit, neglecting the decrease in gravity? (Assume an orbital altitude of 250 km, and calculate both the kinetic energy (classical) and the gravitational potential energy needed.) (b) If the ship has a mass of 1.00×10^5 kg (100 tons), what total yield nuclear explosion in tons of TNT is needed?
27. The Sun produces energy at a rate of 4.00×10^{26} W by the fusion of hydrogen. (a) How many kilograms of hydrogen undergo fusion each second? (b) If the Sun is 90.0% hydrogen and half of this can undergo fusion before the Sun changes character, how long could it produce energy at its current rate? (c) How many kilograms of mass is the Sun losing per second? (d) What fraction of its mass will it have lost in the time found in part (b)?
28. **Unreasonable Results.** A proton has a mass of 1.67×10^{-27} kg. A physicist measures the proton's total energy to be 50.0 MeV. (a) What is the proton's kinetic energy? (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?
29. **Construct Your Own Problem.** Consider a highly relativistic particle. Discuss what is meant by the term "highly relativistic." (Note that, in part, it means that the particle cannot be massless.) Construct a problem in which you calculate the wavelength of such a particle and show that it is very nearly the same as the wavelength of a massless particle, such as

a photon, with the same energy. Among the things to be considered are the rest energy of the particle (it should be a known particle) and its total energy, which should be large compared to its rest energy.

30. **Construct Your Own Problem.** Consider an astronaut traveling to another star at a relativistic velocity. Construct a problem in which you calculate the time for the trip as observed on the Earth and as observed by the astronaut. Also calculate the amount of mass that must be converted to energy to get the astronaut and ship to the velocity travelled. Among the things to be considered are the distance to the star, the velocity, and the mass of the astronaut and ship. Unless your instructor directs you otherwise, do not include any energy given to other masses, such as rocket propellants.

Glossary

total energy: defined as $E = \gamma \mathrm{mc}^2$, where $\gamma = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}}$

rest energy: the energy stored in an object at rest: $E_0 = \mathrm{mc}^2$

relativistic kinetic energy: the kinetic energy of an object moving at relativistic speeds: $\mathrm{KE}_{\mathrm{rel}} = (\gamma - 1)\mathrm{mc}^2$, where $\gamma = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}}$

Selected Solutions to Problems & Exercises

1. 8.20×10^{-14} J; 0.512 MeV

3. 2.3×10^{-30} kg

5. (a) 1.11×10^{27} kg; (b) 5.56×10^{-5}

7. 7.1×10^{-3} kg

7.1×10^{-3}

The ratio is greater for hydrogen.

9. (a) 208; (b) 0.999988*c*

11. (a) 6.92×10^{5} J; (b) 1.54

13. (a) 0.914*c*; (b) The rest mass energy of an electron is 0.511 MeV, so the kinetic energy is approximately 150% of the rest mass energy. The electron should be traveling close to the speed of light.

15. 90.0 MeV

17. (a) $E^2 = p^2c^2 + m^2c^4 = \gamma^2 m^2 c^4$, so that $p^2c^2 = \left(\gamma^2 - 1\right)m^2c^4$, and therefore $\dfrac{(\mathrm{pc})^2}{(\mathrm{mc}^2)^2} = \gamma^2 - 1$

(b) yes

19. 1.07×10^{3}

21. (a) 6.56×10^{-8} kg; (b) 4.37×10^{-10}

23. (a) 0.314*c*; (b) 0.99995*c*

25. (a) 1.00 kg; (b) This much mass would be measurable,

but probably not observable just by looking because it is 0.01% of the total mass.

27. (a) 6.3×10^{11} kg/s; (b) 4.5×10^{10} y; (c) 4.44×10^{9} kg; (d) 0.32%

PART XIII

INTRODUCTION TO QUANTUM PHYSICS

98. Introduction to Quantum Physics

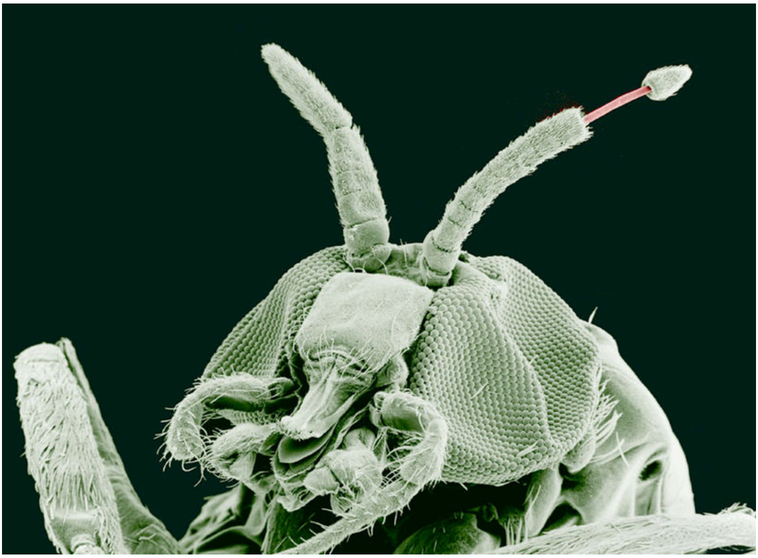


Figure 1. A black fly imaged by an electron microscope is as monstrous as any science-fiction creature. (credit: U.S. Department of Agriculture via Wikimedia Commons)

Quantum mechanics is the branch of physics needed to deal with submicroscopic objects. Because these objects are smaller than we can observe directly with our senses and generally must be observed with the aid of instruments, parts of quantum mechanics seem as foreign and bizarre as parts of relativity. But, like relativity, quantum mechanics has been shown to be valid—truth is often stranger than fiction.

Certain aspects of quantum mechanics are familiar to us. We accept as fact that matter is composed of atoms, the smallest unit of an element, and that these atoms combine to form molecules, the smallest unit of a compound. (See Figure 2.) While we cannot see the individual water molecules in a stream, for example, we are aware that this is because molecules are so small and so numerous in that stream. When introducing atoms, we commonly say that electrons orbit atoms in discrete shells around a tiny nucleus, itself composed of smaller particles called protons and neutrons. We are also aware that electric charge comes in tiny units carried almost entirely by electrons and protons. As with water molecules in a stream, we do not notice individual charges in the current through a lightbulb, because the charges are so small and so numerous in the macroscopic situations we sense directly.

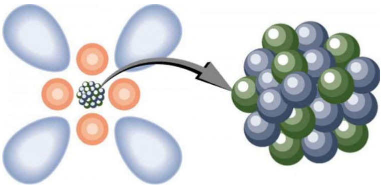


Figure 2. Atoms and their substructure are familiar examples of objects that require quantum mechanics to be fully explained. Certain of their characteristics, such as the discrete electron shells, are classical physics explanations. In quantum mechanics we conceptualize discrete "electron clouds" around the nucleus.

Making Connections: Realms of Physics

Classical physics is a good approximation of modern physics under conditions first discussed in the The Nature of Science and Physics. Quantum mechanics is valid in general, and it must be used rather than classical physics to describe small objects, such as atoms.

Atoms, molecules, and fundamental electron and proton charges are all examples of physical entities that are *quantized*—that is, they appear only in certain discrete values and do not have every conceivable value. Quantized is the opposite of continuous. We cannot have a fraction of an atom, or part of an electron's charge, or 14-1/3 cents, for example. Rather, everything is built of integral multiples of these substructures. Quantum physics is the branch of physics that deals with small objects and the quantization of various entities, including energy and angular momentum. Just as with classical physics, quantum physics has several subfields, such as mechanics and the study of electromagnetic forces. The *correspondence principle* states that in the classical limit (large, slow-moving objects), *quantum mechanics* becomes the same as classical physics. In this chapter, we begin the development of quantum mechanics and its description of the strange submicroscopic world. In later chapters, we will examine many areas, such as atomic and nuclear physics, in which quantum mechanics is crucial.

99. Quantization of Energy

Learning Objectives

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

- Explain Max Planck's contribution to the development of quantum mechanics.
- Explain why atomic spectra indicate quantization.

Planck's Contribution

Energy is quantized in some systems, meaning that the system can have only certain energies and not a continuum of energies, unlike the classical case. This would be like having only certain speeds at which a car can travel because its kinetic energy can have only certain values. We also find that some forms of energy transfer take place with discrete lumps of energy. While most of us are familiar with the quantization of matter into lumps called atoms, molecules, and the like, we are less aware that energy, too, can be quantized. Some of the earliest clues about the necessity of quantum mechanics over classical physics came from the quantization of energy.

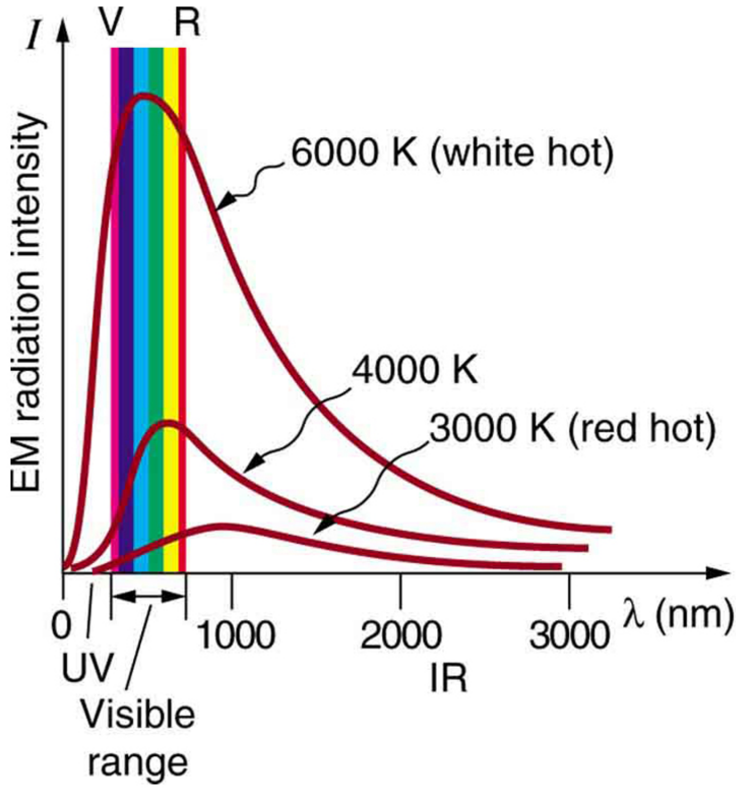


Figure 1. Graphs of blackbody radiation (from an ideal radiator) at three different radiator temperatures. The intensity or rate of radiation emission increases dramatically with temperature, and the peak of the spectrum shifts toward the visible and ultraviolet parts of the spectrum. The shape of the spectrum cannot be described with classical physics.

Where is the quantization of energy observed? Let us begin by considering the emission and absorption of electromagnetic (EM) radiation. The EM spectrum radiated by a hot solid is linked directly to the solid's temperature. (See Figure 1.) An ideal radiator is one that has an emissivity of 1 at all wavelengths and, thus, is jet black. Ideal radiators are therefore called *blackbodies*, and their EM radiation is called *blackbody radiation*. It was discussed that the total intensity of the radiation varies as T^4, the fourth power of the absolute temperature of the body, and that the peak of the spectrum shifts to shorter wavelengths at higher temperatures. All of this seems quite continuous, but it was the curve of the spectrum of intensity versus wavelength that gave a clue that the energies of the atoms in the solid are quantized. In fact, providing a theoretical explanation for the experimentally measured shape of the spectrum was a mystery at the turn of the century. When this "ultraviolet catastrophe" was eventually solved, the answers led to new technologies such as computers and the sophisticated imaging techniques described in earlier chapters. Once again, physics as an enabling science changed the way we live.

The German physicist Max Planck (1858–1947) used the idea that atoms and molecules in a body act like oscillators to absorb and emit radiation. The energies of the oscillating atoms and molecules had to be quantized to correctly describe the shape of the blackbody spectrum. Planck deduced that the energy of an oscillator having a frequency f is given by

$$E = \left(n + \frac{1}{2}\right)hf.$$

Here n is any nonnegative integer (0, 1, 2, 3, …). The symbol h stands for *Planck's constant*, given by $h = 6.626 \times 10^{-34}\ \text{J} \cdot \text{s}$.

The equation $E = \left(n + \frac{1}{2}\right)hf$ means that an oscillator having a frequency f (emitting and absorbing EM radiation of frequency f) can have its energy increase or decrease only in *discrete* steps of size $\Delta E = hf$.

It might be helpful to mention some macroscopic analogies of this quantization of energy phenomena. This is like a pendulum that has a characteristic oscillation frequency but can swing with only certain amplitudes. Quantization of energy also resembles a standing wave on a string that allows only particular harmonics described by integers. It is also similar to going up and down a hill using discrete stair steps rather than being able to move up and down a continuous slope. Your potential energy takes on discrete values as you move from step to step.

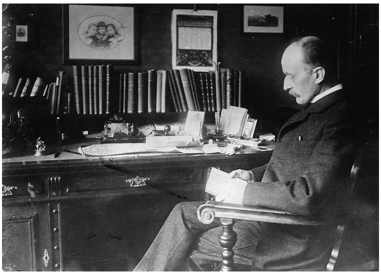


Figure 2. The German physicist Max Planck had a major influence on the early development of quantum mechanics, being the first to recognize that energy is sometimes quantized. Planck also made important contributions to special relativity and classical physics. (credit: Library of Congress, Prints and Photographs Division via Wikimedia Commons)

Using the quantization of oscillators, Planck was able to correctly describe the experimentally known shape of the blackbody spectrum. This was the first indication that energy is sometimes quantized on a small scale and earned him the Nobel Prize in Physics in 1918. Although Planck's theory comes from observations of a macroscopic object, its analysis is based on atoms and molecules. It was such a revolutionary departure from classical physics that Planck himself was reluctant to accept his own idea that energy states are not continuous. The general acceptance of Planck's energy quantization was greatly enhanced by Einstein's explanation of the photoelectric effect (discussed in the next section), which took energy quantization a step further. Planck was fully involved in the development of both early quantum mechanics and relativity. He quickly embraced Einstein's special relativity, published in 1905, and in 1906 Planck was the first to suggest the correct formula for relativistic momentum, $p = \gamma mu$.

Note that Planck's constant h is a very small number. So for an infrared frequency of 10^{14} Hz being emitted by a blackbody, for example, the difference between energy levels is only $\Delta E = hf = (6.63 \times 10^{-34}\ \text{J}\cdot\text{s})(10^{14}\ \text{Hz}) = 6.63 \times 10^{-20}\ \text{J}$, or about 0.4 eV. This 0.4 eV of energy is significant compared with typical atomic energies, which are on the order of an electron volt, or thermal energies, which are typically fractions of an electron volt. But on a macroscopic or classical scale, energies are typically on the order of joules. Even if macroscopic energies are quantized, the quantum steps are too small to be noticed. This is an example of the correspondence principle. For a large object, quantum mechanics produces results indistinguishable from those of classical physics.

Atomic Spectra

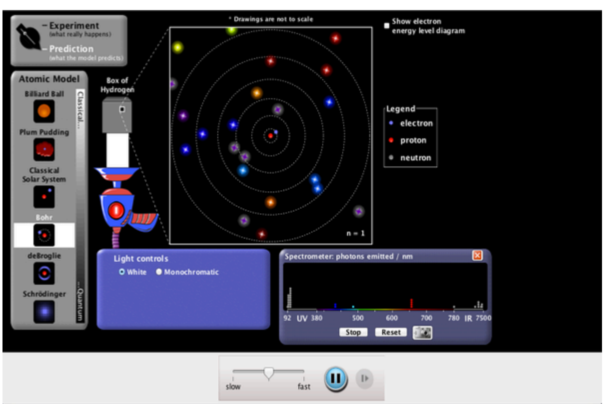
Now let us turn our attention to the *emission and absorption of EM radiation by gases*. The Sun is the most common example of a body containing gases emitting an EM spectrum that includes visible light. We also see examples in neon signs and candle flames. Studies of emissions of hot gases began more than two centuries ago, and it was soon recognized that these emission spectra contained huge amounts of information. The type of gas and its temperature, for example, could be determined. We now know that these EM emissions come from electrons transitioning between energy levels in individual atoms and molecules; thus, they are called *atomic spectra*. Atomic spectra remain an important analytical tool today. Figure 3 shows an example of an emission spectrum obtained by passing an electric discharge through a material. One of the most important characteristics of these spectra is that they are discrete. By this we mean that only certain wavelengths, and hence frequencies, are emitted. This is called a line spectrum. If frequency and energy are associated as $\Delta E = hf$, the energies of the electrons in the emitting atoms and molecules are quantized. This is discussed in more detail later in this chapter.

Figure 3. Emission spectrum of oxygen. When an electrical discharge is passed through a substance, its atoms and molecules absorb energy, which is reemitted as EM radiation. The discrete nature of these emissions implies that the energy states of the atoms and molecules are quantized. Such atomic spectra were used as analytical tools for many decades before it was understood why they are quantized. (credit: Teravolt, Wikimedia Commons)

It was a major puzzle that atomic spectra are quantized. Some of the best minds of 19th-century science failed to explain why this might be. Not until the second decade of the 20th century did an answer based on quantum mechanics begin to emerge. Again a macroscopic or classical body of gas was involved in the studies, but the effect, as we shall see, is due to individual atoms and molecules.

PhET Explorations: Models of the Hydrogen Atom

How did scientists Figure out the structure of atoms without looking at them? Try out different models by shooting light at the atom. Check how the prediction of the model matches the experimental results.



Click to download the simulation. Run using Java.

Section Summary

- The first indication that energy is sometimes quantized came from blackbody radiation, which is the emission of EM radiation by an object with an emissivity of 1.
- Planck recognized that the energy levels of the emitting atoms and molecules were quantized, with only the allowed values of $E = \left(n + \frac{1}{2}\right)\mathrm{hf}$, where n is any non-negative integer (0, 1, 2, 3, …).
- h is Planck's constant, whose value is $h = 6.626 \times 10^{-34}\ \mathrm{J \cdot s}$.
- Thus, the oscillatory absorption and emission energies of atoms and molecules in a blackbody could increase or decrease only in steps of size $\Delta E = hf$ where f is the frequency of the oscillatory nature of the absorption and emission of EM radiation.
- Another indication of energy levels being quantized in atoms and molecules comes from the lines in atomic spectra, which are the EM emissions of individual atoms and molecules.

Conceptual Questions

1. Give an example of a physical entity that is quantized. State specifically what the entity is and what the limits are on its values.
2. Give an example of a physical entity that is not quantized, in that it is continuous and may have a continuous range of values.
3. What aspect of the blackbody spectrum forced Planck to propose quantization of energy levels in its

atoms and molecules?

4. If Planck's constant were large, say 10^{34} times greater than it is, we would observe macroscopic entities to be quantized. Describe the motions of a child's swing under such circumstances.
5. Why don't we notice quantization in everyday events?

Problems & Exercises

1. A LiBr molecule oscillates with a frequency of 1.7×10^{13} Hz. (a) What is the difference in energy in eV between allowed oscillator states? (b) What is the approximate value of n for a state having an energy of 1.0 eV?
2. The difference in energy between allowed oscillator states in HBr molecules is 0.330 eV. What is the oscillation frequency of this molecule?
3. A physicist is watching a 15-kg orangutan at a zoo swing lazily in a tire at the end of a rope. He (the physicist) notices that each oscillation takes 3.00 s and hypothesizes that the energy is quantized. (a) What is the difference in energy in joules between allowed oscillator states? (b) What is the value of n for a state where the energy is 5.00 J? (c) Can the quantization be observed?

Glossary

blackbody: an ideal radiator, which can radiate equally well at all wavelengths

blackbody radiation: the electromagnetic radiation from a blackbody

Planck's constant: $h = 6.626 \times 10^{-34}\ \text{J} \cdot \text{s}$

atomic spectra: the electromagnetic emission from atoms and molecules

Selected Solutions to Problems & Exercises

1. (a) 0.070 eV; (b) 14

3. (a) 2.21×10^{34} J; (b) 2.26×10^{34}; (c) No

100. The Photoelectric Effect

Learning Objectives

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

- Describe a typical photoelectric-effect experiment.
- Determine the maximum kinetic energy of photoelectrons ejected by photons of one energy or wavelength, when given the maximum kinetic energy of photoelectrons for a different photon energy or wavelength.

When light strikes materials, it can eject electrons from them. This is called the *photoelectric effect*, meaning that light (*photo*) produces electricity. One common use of the photoelectric effect is in light meters, such as those that adjust the automatic iris on various types of cameras. In a similar way, another use is in solar cells, as you probably have in your calculator or have seen on a roof top or a roadside sign. These make use of the photoelectric effect to convert light into electricity for running different devices.

This effect has been known for more than a century and can be studied using a device such as that shown in Figure 1. This figure shows an evacuated tube with a metal plate and a collector wire that are connected by a variable voltage source, with the collector more negative than the plate. When light (or other EM radiation) strikes the plate in the evacuated tube, it may eject electrons. If the electrons have energy in electron volts (eV) greater than the potential difference between the plate and the wire in volts, some electrons will be collected on the wire. Since the electron energy in eV is eV, where q is the electron charge and V is the potential difference, the electron energy can be measured by adjusting the retarding voltage between the wire and the plate. The voltage that stops the electrons from reaching the wire equals the energy in eV. For example, if −3.00 V barely stops the electrons, their energy is 3.00 eV. The number of electrons ejected can be determined by measuring the current between the wire and plate. The more light, the more electrons; a little circuitry allows this device to be used as a light meter.

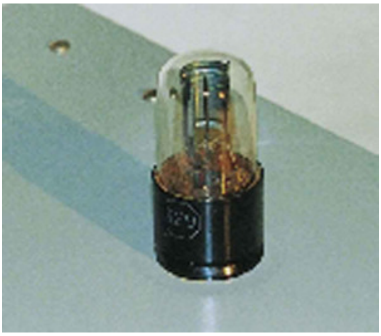


Figure 1. The photoelectric effect can be observed by allowing light to fall on the metal plate in this evacuated tube. Electrons ejected by the light are collected on the collector wire and measured as a current. A retarding voltage between the collector wire and plate can then be adjusted so as to determine the energy of the ejected electrons. For example, if it is sufficiently negative, no electrons will reach the wire. (credit: P.P. Urone)

What is really important about the photoelectric effect is what Albert Einstein deduced from it. Einstein realized that there were several characteristics of the photoelectric effect that could be explained only if *EM radiation is itself quantized*: the apparently continuous stream of energy in an EM wave is actually composed of energy quanta called photons. In his explanation of the photoelectric effect, Einstein defined a quantized unit or quantum of EM energy, which we now call a *photon*, with an energy proportional to the frequency of EM radiation. In equation form, the *photon energy*is $E = hf$, where E is the energy of a photon of frequency f and h is Planck's constant. This revolutionary idea looks similar to Planck's quantization of energy states in blackbody oscillators, but it is quite different. It is the quantization of EM radiation itself. EM waves are composed of photons and are not continuous smooth waves as described in previous chapters on optics. Their energy is absorbed and emitted in lumps, not continuously. This is exactly consistent with Planck's quantization of energy levels in blackbody oscillators, since these oscillators increase and decrease their energy in steps of hf by absorbing and emitting photons having $E = hf$. We do not observe this with our eyes, because there are so many photons in common light sources that individual photons go unnoticed. (See Figure 2.) The next section of the text (Photon Energies and the Electromagnetic Spectrum) is devoted to a discussion of photons and some of their characteristics and implications. For now, we will use the photon concept to explain the photoelectric effect, much as Einstein did.

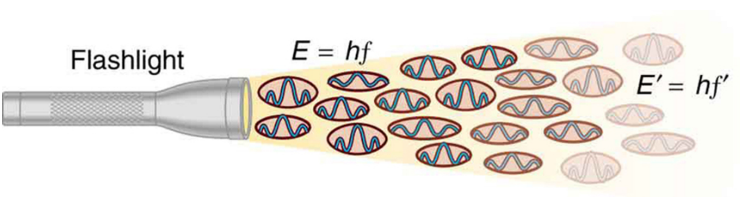


Figure 2. An EM wave of frequency f is composed of photons, or individual quanta of EM radiation. The energy of each photon is $E = hf$, where h is Planck's constant and f is the frequency of the EM radiation. Higher intensity means more photons per unit area. The flashlight emits large numbers of photons of many different frequencies, hence others have energy $E' = hf'$, and so on.

The photoelectric effect has the properties discussed below. All these properties are consistent with the idea that individual photons of EM radiation are absorbed by individual electrons in a material, with the electron gaining the photon's energy. Some of these properties are inconsistent with the idea that EM radiation is a simple wave. For simplicity, let us consider what happens with monochromatic EM radiation in which all photons have the same energy hf.

1. If we vary the frequency of the EM radiation falling on a material, we find the following: For a given material, there is a threshold frequency f_0 for the EM radiation below which no electrons are ejected, regardless of intensity. Individual photons interact with individual electrons. Thus if the photon energy is too small to break an electron away, no electrons will be ejected. If EM radiation was a simple wave, sufficient energy could be obtained by increasing the intensity.
2. *Once EM radiation falls on a material, electrons are ejected without delay*. As soon as an individual photon of a sufficiently high frequency is absorbed by an individual electron, the electron is ejected. If the EM radiation were a simple wave, several minutes would be required for sufficient energy to be deposited to the metal surface to eject an electron.
3. The number of electrons ejected per unit time is proportional to the intensity of the EM radiation and to no other characteristic. High-intensity EM radiation consists of large numbers of photons per unit area, with all photons having the same characteristic energy hf.
4. If we vary the intensity of the EM radiation and measure the energy of ejected electrons, we find the following: *The maximum kinetic energy of ejected electrons is independent of the intensity of the EM radiation*. Since there are so many electrons in a material, it is extremely unlikely that two photons will interact with the same electron at the same time, thereby increasing the energy given it. Instead (as noted in 3 above), increased intensity results in more electrons of the same energy being ejected. If EM radiation were a simple wave, a higher intensity could give more energy, and higher-energy

electrons would be ejected.

5. The kinetic energy of an ejected electron equals the photon energy minus the binding energy of the electron in the specific material. An individual photon can give all of its energy to an electron. The photon's energy is partly used to break the electron away from the material. The remainder goes into the ejected electron's kinetic energy. In equation form, this is given by $KE_e = hf - BE$, where KE_e is the maximum kinetic energy of the ejected electron, hf is the photon's energy, and BE is the *binding energy* of the electron to the particular material. (BE is sometimes called the *work function* of the material.) This equation, due to Einstein in 1905, explains the properties of the photoelectric effect quantitatively. An individual photon of EM radiation (it does not come any other way) interacts with an individual electron, supplying enough energy, BE, to break it away, with the remainder going to kinetic energy. The binding energy is $BE = hf_0$, where f_0 is the threshold frequency for the particular material. Figure 3 shows a graph of maximum KE_e versus the frequency of incident EM radiation falling on a particular material.

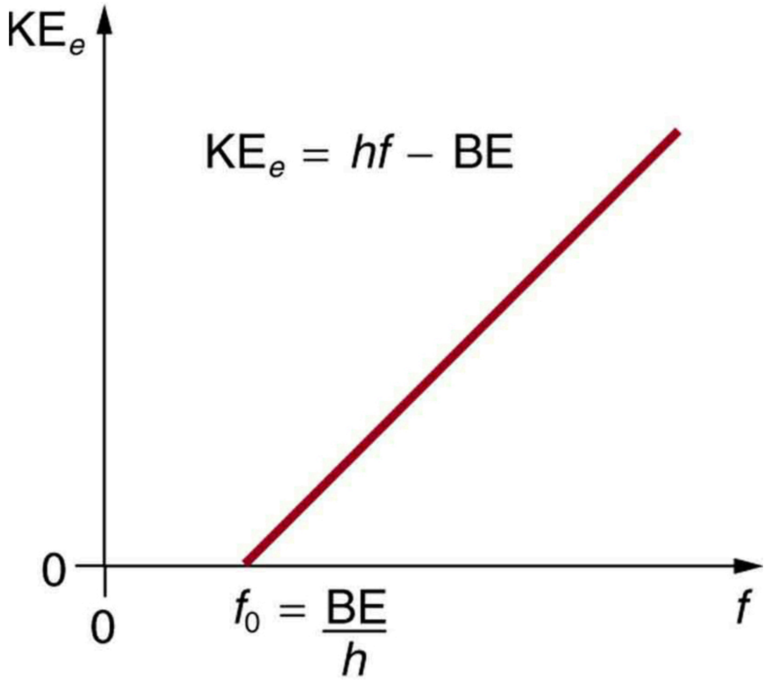


Figure 3. Photoelectric effect. A graph of the kinetic energy of an ejected electron, KE_e, versus the frequency of EM radiation impinging on a certain material. There is a threshold frequency below which no electrons are ejected, because the individual photon interacting with an individual electron has insufficient energy to break it away. Above the threshold energy, KE_e increases linearly with f, consistent with $KE_e = hf - BE$. The slope of this line is h—the data can be used to determine Planck's constant experimentally. Einstein gave the first successful explanation of such data by proposing the idea of photons—quanta of EM radiation.

Einstein's idea that EM radiation is quantized was crucial to the beginnings of quantum mechanics. It is a far more general concept than its explanation of the photoelectric effect might imply. All EM radiation can also be modeled in the form of photons, and the characteristics of EM radiation are entirely consistent with this fact. (As we will see in the next section, many aspects of EM radiation, such as the hazards of ultraviolet (UV) radiation, can be explained *only* by photon properties.) More famous for modern relativity, Einstein planted an important seed for quantum mechanics in 1905, the same year he published his first paper on special relativity. His explanation of the photoelectric effect was the basis for the Nobel Prize awarded to him in 1921. Although his other contributions to theoretical physics were also noted in that award, special and general relativity were not fully recognized in spite of having been partially verified by experiment by 1921. Although hero-worshipped, this great man never received Nobel recognition for his most famous work—relativity.

Example 1. Calculating Photon Energy and the Photoelectric Effect: A Violet Light

1. What is the energy in joules and electron volts of a photon of 420-nm violet light?

2. What is the maximum kinetic energy of electrons ejected from calcium by 420-nm violet light, given that the binding energy (or work function) of electrons for calcium metal is 2.71 eV?

Strategy

To solve Part 1, note that the energy of a photon is given by $E = hf$. For Part 2, once the energy of the photon is calculated, it is a straightforward application of $KE_e = hf - BE$ to find the ejected electron's maximum kinetic energy, since BE is given.

Solution for Part 1

Photon energy is given by $E = hf$.

Since we are given the wavelength rather than the frequency, we solve the familiar relationship $c = f\lambda$ for the frequency, yielding $f = \frac{c}{\lambda}$.

Combining these two equations gives the useful relationship $E = \frac{hc}{\lambda}$.

Now substituting known values yields

$$E = \frac{\left(6.63 \times 10^{-34}\ \text{J} \cdot\ \text{s}\right)\left(3.00 \times 10^{8}\ \text{m/s}\right)}{420 \times 10^{-9}\ \text{m}} = 4.74 \times 10^{-19}\ \text{J}$$

Converting to eV, the energy of the photon is

$$E = \left(4.47 \times 10^{-19}\ \text{J}\right)\frac{1\ \text{eV}}{1.6 \times 10^{-19}\ \text{J}} = 2.96\ \text{eV}$$

Solution for Part 2

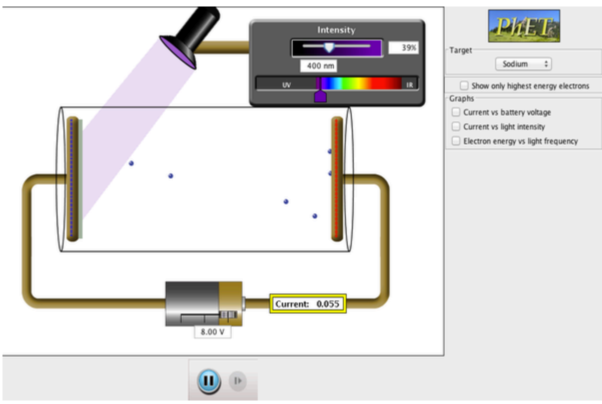
Finding the kinetic energy of the ejected electron is now a simple application of the equation $KE_e = hf - BE$. Substituting the photon energy and binding energy yields $KE_e = hf - BE = 2.96\ \text{eV} - 2.71\ \text{eV} = 0.246\ \text{eV}$.

Discussion

The energy of this 420-nm photon of violet light is a tiny fraction of a joule, and so it is no wonder that a single photon would be difficult for us to sense directly—humans are more attuned to energies on the order of joules. But looking at the energy in electron volts, we can see that this photon has enough energy to affect atoms and molecules. A DNA molecule can be broken with about 1 eV of energy, for example, and typical atomic and molecular energies are on the order of eV, so that the UV photon in this example could have biological effects. The ejected electron (called a *photoelectron*) has a rather low energy, and it would not travel far, except in a vacuum. The electron would be stopped by a retarding potential of but 0.26 eV. In fact, if the photon wavelength were longer and its energy less than 2.71 eV, then the formula would give a negative kinetic energy, an impossibility. This simply means that the 420-nm photons with their 2.96-eV energy are not much above the frequency threshold. You can show for yourself that the threshold wavelength is 459 nm (blue light). This means that if calcium metal is used in a light meter, the meter will be insensitive to wavelengths longer than those of blue light. Such a light meter would be completely insensitive to red light, for example.

PhET Explorations: Photoelectric Effect

See how light knocks electrons off a metal target, and recreate the experiment that spawned the field of quantum mechanics.



Click to download the simulation. Run using Java.

Section Summary

- The photoelectric effect is the process in which EM radiation ejects electrons from a material.
- Einstein proposed photons to be quanta of EM radiation having

energy E = hf, where f is the frequency of the radiation.

- All EM radiation is composed of photons. As Einstein explained, all characteristics of the photoelectric effect are due to the interaction of individual photons with individual electrons.
- The maximum kinetic energy KE_e of ejected electrons (photoelectrons) is given by $KE_e = hf - BE$, where hf is the photon energy and BE is the binding energy (or work function) of the electron to the particular material.

Conceptual Questions

1. Is visible light the only type of EM radiation that can cause the photoelectric effect?
2. Which aspects of the photoelectric effect cannot be explained without photons? Which can be explained without photons? Are the latter inconsistent with the existence of photons?
3. Is the photoelectric effect a direct consequence of the wave character of EM radiation or of the particle character of EM radiation? Explain briefly.
4. Insulators (nonmetals) have a higher BE than metals, and it is more difficult for photons to eject electrons from insulators. Discuss how this relates to the free charges in metals that make them good conductors.
5. If you pick up and shake a piece of metal that has electrons in it free to move as a current, no electrons fall out. Yet if you heat the metal, electrons can be boiled off. Explain both of these facts as they relate to the amount and distribution of energy involved with shaking the object as compared with heating it.

Problems & Exercises

1. What is the longest-wavelength EM radiation that can eject a photoelectron from silver, given that the binding energy is 4.73 eV? Is this in the visible range?
2. Find the longest-wavelength photon that can eject an electron from potassium, given that the binding energy is 2.24 eV. Is this visible EM radiation?
3. What is the binding energy in eV of electrons in magnesium, if the longest-wavelength photon that can eject electrons is 337 nm?
4. Calculate the binding energy in eV of electrons in aluminum, if the longest-wavelength photon that can eject them is 304 nm.
5. What is the maximum kinetic energy in eV of electrons ejected from sodium metal by 450-nm EM radiation, given that the binding energy is 2.28 eV?
6. UV radiation having a wavelength of 120 nm falls on gold metal, to which electrons are bound by 4.82 eV. What is the maximum kinetic energy of the ejected photoelectrons?
7. Violet light of wavelength 400 nm ejects electrons with a maximum kinetic energy of 0.860 eV from sodium metal. What is the binding energy of electrons to sodium metal?
8. UV radiation having a 300-nm wavelength falls on uranium metal, ejecting 0.500-eV electrons. What is the binding energy of electrons to uranium metal?
9. (a) What is the wavelength of EM radiation that ejects 2.00-eV electrons from calcium metal, given that the binding energy is 2.71 eV? (b) What type of

EM radiation is this?

10. Find the wavelength of photons that eject 0.100-eV electrons from potassium, given that the binding energy is 2.24 eV. Are these photons visible?
11. What is the maximum velocity of electrons ejected from a material by 80-nm photons, if they are bound to the material by 4.73 eV?
12. Photoelectrons from a material with a binding energy of 2.71 eV are ejected by 420-nm photons. Once ejected, how long does it take these electrons to travel 2.50 cm to a detection device?
13. A laser with a power output of 2.00 mW at a wavelength of 400 nm is projected onto calcium metal. (a) How many electrons per second are ejected? (b) What power is carried away by the electrons, given that the binding energy is 2.71 eV?
14. (a) Calculate the number of photoelectrons per second ejected from a 1.00-mm 2 area of sodium metal by 500-nm EM radiation having an intensity of 1.30 kW/m^2 (the intensity of sunlight above the Earth's atmosphere). (b) Given that the binding energy is 2.28 eV, what power is carried away by the electrons? (c) The electrons carry away less power than brought in by the photons. Where does the other power go? How can it be recovered?
15. **Unreasonable Results.** Red light having a wavelength of 700 nm is projected onto magnesium metal to which electrons are bound by 3.68 eV. (a) Use $KE_e = hf - BE$ to calculate the kinetic energy of the ejected electrons. (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

16. **Unreasonable Results.** (a) What is the binding energy of electrons to a material from which 4.00-eV electrons are ejected by 400-nm EM radiation? (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

Glossary

photoelectric effect: the phenomenon whereby some materials eject electrons when light is shined on them

photon: a quantum, or particle, of electromagnetic radiation

photon energy: the amount of energy a photon has; $E = hf$

binding energy: also called the work function; the amount of energy necessary to eject an electron from a material

Selected Solutions to Problems & Exercises

1. 263 nm

3. 3.69 eV

5. 0.483 eV

7. 2.25 eV

9. (a) 264 nm; (b) Ultraviolet

11. 1.95×10^{6} m/s

13. (a) 4.02×10^{15} s; (b) 0.256 mW

15. (a) −1.90 eV; (b) Negative kinetic energy; (c) That the electrons would be knocked free.

101. Photon Energies and the Electromagnetic Spectrum

Learning Objectives

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

- Explain the relationship between the energy of a photon in joules or electron volts and its wavelength or frequency.
- Calculate the number of photons per second emitted by a monochromatic source of specific wavelength and power.

Ionizing Radiation

A photon is a quantum of EM radiation. Its energy is given by E = hf and is related to the frequency f and wavelength λ of the radiation by

$$E = hf = \frac{hc}{\lambda}(\text{energy of a photon}),$$

where E is the energy of a single photon and c is the speed of light. When working with small systems, energy in eV is often useful. Note that Planck's constant in these units is $h = 4.14 \times 10^{-15}\ \text{eV}\cdot\text{s}$.

Since many wavelengths are stated in nanometers (nm), it is also useful to know that $hc = 1240\ \text{eV}\cdot\text{nm}$.

These will make many calculations a little easier.

All EM radiation is composed of photons. Figure 1 shows various divisions of the EM spectrum plotted against wavelength, frequency, and photon energy. Previously in this book, photon characteristics were alluded to in the discussion of some of the characteristics of UV, x rays, and γ rays, the first of which start with frequencies just above violet in the visible spectrum. It was noted that these types of EM radiation have characteristics much different than visible light. We can now see that such properties arise because photon energy is larger at high frequencies.

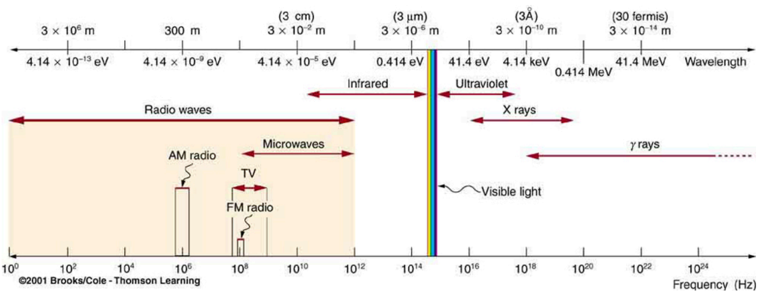


Figure 1. The EM spectrum, showing major categories as a function of photon energy in eV, as well as wavelength and frequency. Certain characteristics of EM radiation are directly attributable to photon energy alone.

Table 1. Representative Energies for Submicroscopic Effects (Order of Magnitude Only)

Rotational energies of molecules	10^{-5} eV
Vibrational energies of molecules	0.1 eV
Energy between outer electron shells in atoms	1 eV
Binding energy of a weakly bound molecule	1 eV
Energy of red light	2 eV
Binding energy of a tightly bound molecule	10 eV
Energy to ionize atom or molecule	10 to 1000 eV

Photons act as individual quanta and interact with individual electrons, atoms, molecules, and so on. The energy a photon carries is, thus, crucial to the effects it has. Table 1 lists representative submicroscopic energies in eV. When we compare photon energies from the EM spectrum in Figure 1 with energies in the table, we can see how effects vary with the type of EM radiation.

Gamma rays, a form of nuclear and cosmic EM radiation, can have the highest frequencies and, hence, the highest photon energies in the EM spectrum. For example, a γ-ray photon with $f = 10^{21}$ Hz has an energy $E = hf = 6.63 \times 10^{-13}$ J = 4.14 MeV. This is sufficient energy to ionize thousands of atoms and molecules, since only 10 to 1000 eV are needed per ionization. In fact, γ rays are one type of *ionizing radiation*, as are x rays and UV, because they produce ionization in materials that absorb them. Because so much ionization can be produced, a single γ-ray photon can cause significant damage to biological tissue, killing cells or damaging their ability to properly reproduce. When cell reproduction is disrupted, the result can be cancer, one of the known effects of exposure to ionizing radiation. Since cancer cells are rapidly reproducing, they are exceptionally sensitive to the disruption produced by ionizing radiation. This means that ionizing radiation has positive uses in cancer treatment as well as risks in producing cancer.

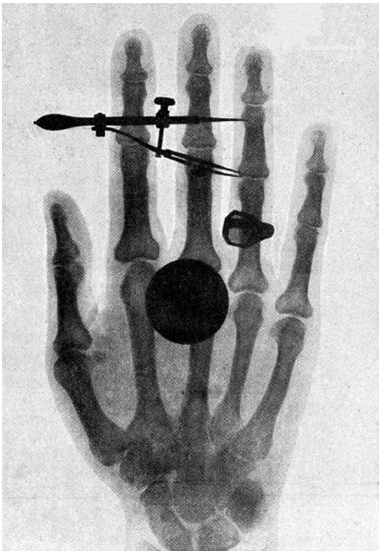


Figure 2. One of the first x-ray images, taken by Röentgen himself. The hand belongs to Bertha Röentgen, his wife. (credit: Wilhelm Conrad Röntgen, via Wikimedia Commons)

High photon energy also enables γ rays to penetrate materials, since a collision with a single atom or molecule is unlikely to absorb all the γ ray's energy. This can make γ rays useful as a probe, and they are sometimes used in medical imaging. *x rays*, as you can see in Figure 1, overlap with the low-frequency end of the γ ray range. Since x rays have energies of keV and up, individual x-ray photons also can produce large amounts of ionization. At lower photon energies, x rays are not as penetrating as γ rays and are slightly less hazardous. X rays are ideal for medical imaging, their most common use, and a fact that was recognized immediately upon their discovery in 1895 by the German physicist W. C. Roentgen (1845–1923). (See Figure 2.) Within one year of their discovery, x rays (for a time called Roentgen rays) were used for medical diagnostics. Roentgen received the 1901 Nobel Prize for the discovery of x rays.

Making Connections: Conservation of Energy

Once again, we find that conservation of energy allows us to consider the initial and final forms that energy takes, without having to make detailed calculations of the intermediate steps. Example 1 is solved by considering only the initial and final forms of energy.

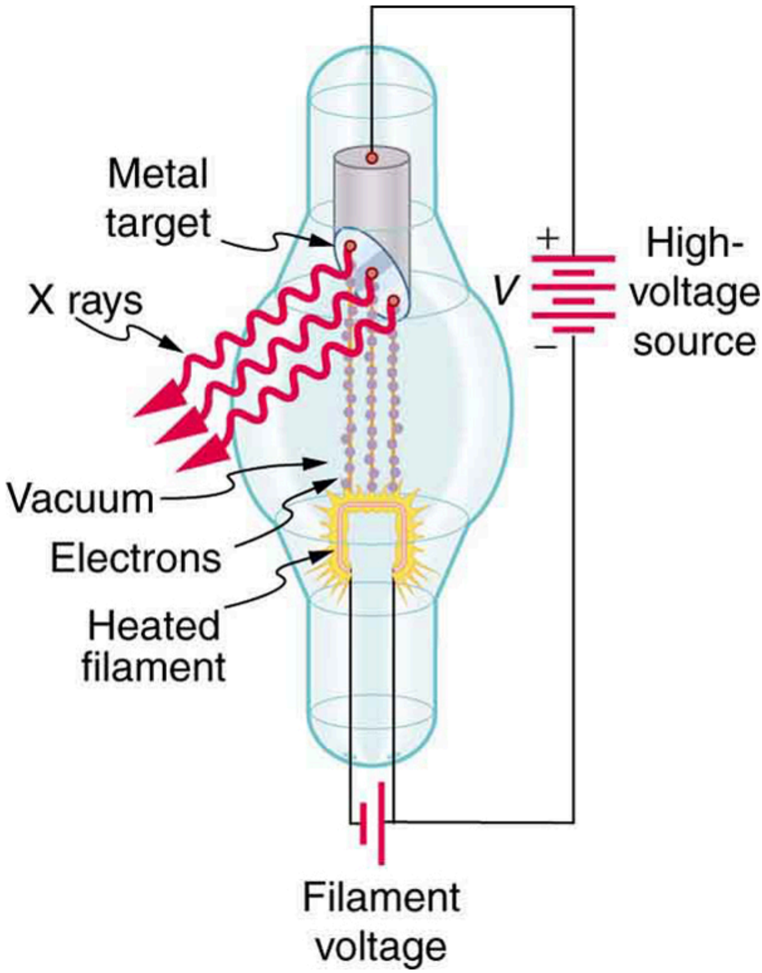


Figure 3. X rays are produced when energetic electrons strike the copper anode of this cathode ray tube (CRT). Electrons (shown here as separate particles) interact individually with the material they strike, sometimes producing photons of EM radiation.

While γ rays originate in nuclear decay, x rays are produced by the process shown in Figure 3. Electrons ejected by thermal agitation from a hot filament in a vacuum tube are accelerated through a high voltage, gaining kinetic energy from the electrical potential energy. When they strike the anode, the electrons convert their kinetic energy to a variety of forms, including thermal energy. But since an accelerated charge radiates EM waves, and since the electrons act individually, photons are also produced. Some of these x-ray photons obtain the kinetic energy of the electron. The accelerated electrons originate at the cathode, so such a tube is called a cathode ray tube (CRT), and various versions of them are found in older TV and computer screens as well as in x-ray machines.

Example 1. X-ray Photon Energy and X-ray Tube Voltage

Find the maximum energy in eV of an x-ray photon produced by electrons accelerated through a potential difference of 50.0 kV in a CRT like the one in Figure 3.

Strategy

Electrons can give all of their kinetic energy to a single photon when they strike the anode of a CRT. (This is something like the photoelectric effect in reverse.) The kinetic energy of the electron comes from electrical potential energy. Thus we can simply equate the maximum photon energy to the electrical potential energy—that is, $hf=qV$. (We do not have to calculate each step from beginning to end if we know that all of the starting energy qV is converted to the final form hf.)

Solution

The maximum photon energy is $hf = qV$, where q is the charge of the electron and V is the accelerating voltage. Thus, $hf = (1.60 \times 10^{-19}\text{ C})(50.0 \times 10^{3}\text{ V})$.

From the definition of the electron volt, we know $1\text{ eV} = 1.60 \times 10^{-19}\text{ J}$, where $1\text{ J} = 1\text{ C} \cdot \text{V}$. Gathering factors and converting energy to eV yields

$$hf = \left(50.0 \times 10^{3}\right)\left(1.60 \times 10^{-19}\text{ C} \cdot \text{ V}\right)\left(\frac{1\text{ eV}}{1.60 \times^{-19}\text{ C} \cdot \text{ V}}\right)$$

$$= \left(50.0 \times 10^{3}\right)\left(1\text{ eV}\right) = 50.0\text{ keV}$$

Discussion

This example produces a result that can be applied to many similar situations. If you accelerate a single elementary charge, like that of an electron, through a potential given in volts, then its energy in eV has the same numerical value. Thus a 50.0-kV potential generates 50.0 keV electrons, which in turn can produce photons with a maximum energy of 50 keV. Similarly, a 100-kV potential in an x-ray tube can generate up to 100-keV x-ray photons. Many x-ray tubes have adjustable voltages so that various energy x rays with differing energies, and therefore differing abilities to penetrate, can be generated.

Figure 4 shows the spectrum of x rays obtained from an x-ray tube. There are two distinct features to the spectrum. First, the smooth distribution results from electrons being decelerated in the anode material. A curve like this is obtained by detecting many photons, and it is apparent that the maximum energy is unlikely. This decelerating process produces radiation that is called *bremsstrahlung* (German for *braking radiation*). The second feature is the existence of sharp peaks in the spectrum; these are called *characteristic x rays*, since they are characteristic of the anode material. Characteristic x rays come from atomic excitations unique to a given type of anode material. They are akin to lines in atomic spectra, implying the energy levels of atoms are quantized. Phenomena such as discrete atomic spectra and characteristic x rays are explored further in Atomic Physics.

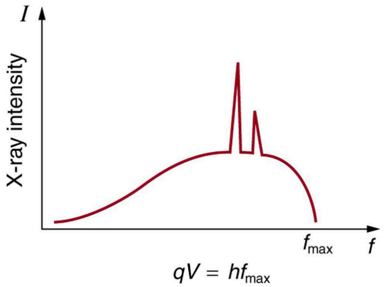


Figure 4. X-ray spectrum obtained when energetic electrons strike a material. The smooth part of the spectrum is bremsstrahlung, while the peaks are characteristic of the anode material. Both are atomic processes that produce energetic photons known as x-ray photons.

Ultraviolet radiation (approximately 4 eV to 300 eV) overlaps with the low end of the energy range of x rays, but UV is typically lower in energy. UV comes from the de-excitation of atoms that may be part of a hot solid or gas. These atoms can be given energy that they later release as UV by numerous processes, including electric discharge, nuclear explosion, thermal agitation, and exposure to x rays. A UV photon has sufficient energy to ionize atoms and molecules, which makes its effects different from those of visible light. UV thus has some of the same biological effects as γ rays and x rays. For example, it can cause skin cancer and is used as a sterilizer. The major difference is that several UV photons are required to disrupt cell reproduction or kill a bacterium, whereas single γ-ray and X-ray photons can do the same damage. But since UV does have the energy to alter molecules, it can do what visible light cannot. One of the beneficial aspects of UV is that it triggers the production of vitamin D in the skin, whereas visible light has insufficient energy per photon to alter the molecules that trigger this production. Infantile jaundice is treated by exposing the baby to UV (with eye protection), called phototherapy, the beneficial effects of which are thought to be related to its ability to help prevent the buildup of potentially toxic bilirubin in the blood.

Example 2. Photon Energy and Effects for UV

Short-wavelength UV is sometimes called vacuum UV, because it is strongly absorbed by air and must be studied in a vacuum. Calculate the photon energy in eV for 100-nm vacuum UV, and estimate the number of molecules it could ionize or break apart.

Strategy

Using the equation E = *hf* and appropriate constants, we can find the photon energy and compare it with energy information in Table 1.

Solution

The energy of a photon is given by

$$E = hf = \frac{hc}{\lambda}.$$

Using $hc = 1240\ \text{eV}\cdot\text{nm}$, we find that

$$E = hf = \frac{hc}{\lambda} = \frac{1240\ \text{eV}\ \cdot\ \text{nm}}{100\ \text{nm}} = 12.4\ \text{eV}$$

.

Discussion

According to Table 1, this photon energy might be able to ionize an atom or molecule, and it is about what is needed to break up a tightly bound molecule, since they are bound by approximately 10 eV. This photon energy could destroy about a dozen weakly bound molecules. Because of its high photon energy, UV disrupts atoms and molecules it interacts with. One good consequence is that all but the longest-wavelength UV is strongly absorbed and is easily blocked by sunglasses. In fact, most of the Sun's UV is absorbed by a thin layer of ozone in the upper atmosphere, protecting sensitive organisms on Earth. Damage to our ozone layer by the addition of such chemicals as CFCs has reduced this protection for us.

Visible Light

The range of photon energies for *visible light* from red to violet is 1.63 to 3.26 eV, respectively (left for this chapter's Problems and Exercises to verify). These energies are on the order of those between outer electron shells in atoms and molecules. This means that these photons can be absorbed by atoms and molecules. A *single* photon can actually stimulate the retina, for example, by altering a receptor molecule that then triggers a nerve impulse. Photons can be absorbed or emitted only by atoms and molecules that have precisely the correct quantized energy step to do so. For example, if a red photon of frequency f encounters a molecule that has an energy step, ΔE, equal to hf, then the photon can be absorbed. Violet flowers absorb red and reflect violet; this implies there is no energy step between levels in the receptor molecule equal to the violet photon's energy, but there is an energy step for the red.

There are some noticeable differences in the characteristics of light between the two ends of the visible spectrum that are due to photon energies. Red light has insufficient photon energy to expose most black-and-white film, and it is thus used to illuminate darkrooms where such film is developed. Since violet light has a higher photon energy, dyes that absorb violet tend to fade more quickly than those that do not. (See Figure 5.) Take a look at some faded color posters in a storefront some time, and you will notice that the blues and violets are the last to fade. This is because other dyes, such as red and green dyes, absorb blue and violet photons, the higher energies of which break up their weakly bound molecules. (Complex molecules such as those in dyes and DNA tend to be weakly bound.) Blue and violet dyes reflect those colors and, therefore, do not absorb these more energetic photons, thus suffering less molecular damage.

Figure 5. Why do the reds, yellows, and greens fade before the blues and violets when exposed to the Sun, as with this poster? The answer is related to photon energy. (credit: Deb Collins, Flickr)

Transparent materials, such as some glasses, do not absorb any visible light, because there is no energy step in the atoms or molecules that could absorb the light. Since individual photons interact with individual atoms, it is nearly impossible to have two photons absorbed simultaneously to reach a large energy step. Because of its lower photon energy, visible light can sometimes pass through many kilometers of a substance, while higher frequencies like UV, x ray, and γ rays are absorbed, because they have sufficient photon energy to ionize the material.

Example 3. How Many Photons per Second Does a Typical Light Bulb Produce?

Assuming that 10.0% of a 100-W light bulb's energy output is in the visible range (typical for incandescent bulbs) with an average wavelength of 580 nm, calculate the number of visible photons emitted per second.

Strategy

Power is energy per unit time, and so if we can find the energy per photon, we can determine the number of photons per second. This will best be done in joules, since power is given in watts, which are joules per second.

Solution

The power in visible light production is 10.0% of 100 W, or 10.0 J/s. The energy of the average visible photon is found by substituting the given average wavelength into the formula $E = \frac{hc}{\lambda}$.

This produces

$$E = \frac{\left(6.63 \times 10^{-34}\ \text{J} \cdot \text{s}\right)\left(3.00 \times 10^{8}\ \text{m/s}\right)}{580 \times 10^{-9}\ \text{m}} = 3.43 \times 10^{-19}\ \text{J}$$

The number of visible photons per second is thus

$$\text{photon/s} = \frac{10.0\ \text{J/s}}{3.42 \times 10^{-19}\ \text{J/photon}} = 2.92 \times 10^{19}\ \text{photon/s}$$

Discussion

This incredible number of photons per second is verification that individual photons are insignificant in ordinary human experience. It is also a verification of the correspondence principle—on the macroscopic scale, quantization becomes essentially continuous or classical. Finally, there are so many photons emitted by a 100-W lightbulb that it can be seen by the unaided eye many kilometers away.

Lower-Energy Photons

Infrared radiation (*IR*) has even lower photon energies than visible light and cannot significantly alter atoms and molecules. IR can be absorbed and emitted by atoms and molecules, particularly between closely spaced states. IR is extremely strongly absorbed by water, for example, because water molecules have many states separated by energies on the order of 10^{-5} eV to 10^{-2} eV, well within the IR and microwave energy ranges. This is why in the IR range, skin is almost jet black, with an emissivity near 1—there are many states in water molecules in the skin that can absorb a large range of IR photon energies. Not all molecules have this property. Air, for example, is nearly transparent to many IR frequencies.

Microwaves are the highest frequencies that can be produced by electronic circuits, although they are also produced naturally. Thus microwaves are similar to IR but do not extend to as high frequencies. There are states in water and other molecules that have the same frequency and energy as microwaves, typically about 10^{-5} eV. This is one reason why food absorbs microwaves more strongly than many other materials, making microwave ovens an efficient way of putting energy directly into food.

Photon energies for both IR and microwaves are so low that huge numbers of photons are involved in any significant energy transfer by IR or microwaves (such as warming yourself with a heat lamp or cooking pizza in the microwave). Visible light, IR, microwaves, and all lower frequencies cannot produce ionization with single photons and do not ordinarily have the hazards of higher frequencies. When visible, IR, or microwave radiation *is* hazardous, such as the inducement of cataracts by microwaves, the hazard is due to huge numbers of photons acting together (not to an accumulation of photons, such as sterilization by weak UV). The negative effects of visible, IR, or microwave radiation can be thermal effects, which could be produced by any heat source. But one difference is that at very high intensity, strong electric and magnetic fields can be produced by photons acting together. Such electromagnetic fields (EMF) can actually ionize materials.

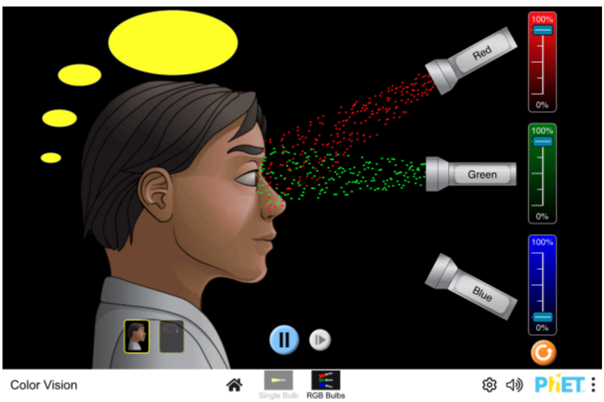
Misconception Alert: High-Voltage Power Lines

Although some people think that living near high-voltage power lines is hazardous to one's health, ongoing studies of the transient field effects produced by these lines show their strengths to be insufficient to cause damage. Demographic studies also fail to show significant correlation of ill effects with high-voltage power lines. The American Physical Society issued a report over 10 years ago on power-line fields, which concluded that the scientific literature and reviews of panels show no consistent, significant link between cancer and power-line fields. They also felt that the "diversion of resources to eliminate a threat which has no persuasive scientific basis is disturbing."

It is virtually impossible to detect individual photons having frequencies below microwave frequencies, because of their low photon energy. But the photons are there. A continuous EM wave can be modeled as photons. At low frequencies, EM waves are generally treated as time- and position-varying electric and magnetic fields with no discernible quantization. This is another example of the correspondence principle in situations involving huge numbers of photons.

PhET Explorations: Color Vision

Make a whole rainbow by mixing red, green, and blue light. Change the wavelength of a monochromatic beam or filter white light. View the light as a solid beam, or see the individual photons.



Click to run the simulation.

Section Summary

- Photon energy is responsible for many characteristics of EM radiation, being particularly noticeable at high frequencies.
- Photons have both wave and particle characteristics.

Conceptual Questions

1. Why are UV, x rays, and γ rays called ionizing radiation?
2. How can treating food with ionizing radiation help keep it from spoiling? UV is not very penetrating. What else could be used?
3. Some television tubes are CRTs. They use an approximately 30-kV accelerating potential to send electrons to the screen, where the electrons stimulate phosphors to emit the light that forms the pictures we watch. Would you expect x rays also to be created?
4. Tanning salons use “safe” UV with a longer wavelength than some of the UV in sunlight. This “safe” UV has enough photon energy to trigger the tanning mechanism. Is it likely to be able to cause cell damage and induce cancer with prolonged exposure?
5. Your pupils dilate when visible light intensity is reduced. Does wearing sunglasses that lack UV blockers increase or decrease the UV hazard to your eyes? Explain.
6. One could feel heat transfer in the form of infrared radiation from a large nuclear bomb detonated in the atmosphere 75 km from you. However, none of the profusely emitted x rays or γ rays reaches you. Explain.
7. Can a single microwave photon cause cell damage? Explain.
8. In an x-ray tube, the maximum photon energy is given by $hf = qV$. Would it be technically more correct

to say $hf = qV + BE$, where BE is the binding energy of electrons in the target anode? Why isn't the energy stated the latter way?

Problems & Exercises

1. What is the energy in joules and eV of a photon in a radio wave from an AM station that has a 1530-kHz broadcast frequency?
2. (a) Find the energy in joules and eV of photons in radio waves from an FM station that has a 90.0-MHz broadcast frequency. (b) What does this imply about the number of photons per second that the radio station must broadcast?
3. Calculate the frequency in hertz of a 1.00-MeV γ-ray photon.
4. (a) What is the wavelength of a 1.00-eV photon? (b) Find its frequency in hertz. (c) Identify the type of EM radiation.
5. Do the unit conversions necessary to show that $hc = 1240 \text{ eV} \cdot \text{nm}$, as stated in the text.
6. Confirm the statement in the text that the range of photon energies for visible light is 1.63 to 3.26 eV, given that the range of visible wavelengths is 380 to 760 nm.
7. (a) Calculate the energy in eV of an IR photon of frequency 2.00×10^{13} Hz. (b) How many of these photons would need to be absorbed simultaneously

by a tightly bound molecule to break it apart? (c) What is the energy in eV of a γ ray of frequency 3.00×10^{20} Hz? (d) How many tightly bound molecules could a single such γ ray break apart?

8. Prove that, to three-digit accuracy, $h = 4.14 \times 10^{-15}$ eV · s, as stated in the text.
9. (a) What is the maximum energy in eV of photons produced in a CRT using a 25.0-kV accelerating potential, such as a color TV? (b) What is their frequency?
10. What is the accelerating voltage of an x-ray tube that produces x rays with a shortest wavelength of 0.0103 nm?
11. (a) What is the ratio of power outputs by two microwave ovens having frequencies of 950 and 2560 MHz, if they emit the same number of photons per second? (b) What is the ratio of photons per second if they have the same power output?
12. How many photons per second are emitted by the antenna of a microwave oven, if its power output is 1.00 kW at a frequency of 2560 MHz?
13. Some satellites use nuclear power. (a) If such a satellite emits a 1.00-W flux of γ rays having an average energy of 0.500 MeV, how many are emitted per second? (b) These γ rays affect other satellites. How far away must another satellite be to only receive one γ ray per second per square meter?
14. (a) If the power output of a 650-kHz radio station is 50.0 kW, how many photons per second are produced? (b) If the radio waves are broadcast uniformly in all directions, find the number of photons per second per square meter at a distance of

100 km. Assume no reflection from the ground or absorption by the air.

15. How many x-ray photons per second are created by an x-ray tube that produces a flux of x rays having a power of 1.00 W? Assume the average energy per photon is 75.0 keV.

16. (a) How far away must you be from a 650-kHz radio station with power 50.0 kW for there to be only one photon per second per square meter? Assume no reflections or absorption, as if you were in deep outer space. (b) Discuss the implications for detecting intelligent life in other solar systems by detecting their radio broadcasts.

17. Assuming that 10.0% of a 100-W light bulb's energy output is in the visible range (typical for incandescent bulbs) with an average wavelength of 580 nm, and that the photons spread out uniformly and are not absorbed by the atmosphere, how far away would you be if 500 photons per second enter the 3.00-mm diameter pupil of your eye? (This number easily stimulates the retina.)

18. **Construct Your Own Problem.** Consider a laser pen. Construct a problem in which you calculate the number of photons per second emitted by the pen. Among the things to be considered are the laser pen's wavelength and power output. Your instructor may also wish for you to determine the minimum diffraction spreading in the beam and the number of photons per square centimeter the pen can project at some large distance. In this latter case, you will also need to consider the output size of the laser beam, the distance to the object being illuminated, and any

absorption or scattering along the way.

Glossary

gamma ray: also γ-ray; highest-energy photon in the EM spectrum

ionizing radiation: radiation that ionizes materials that absorb it

x ray: EM photon between γ-ray and UV in energy

bremsstrahlung: German for *braking radiation*; produced when electrons are decelerated

characteristic x rays: x rays whose energy depends on the material they were produced in

ultraviolet radiation: UV; ionizing photons slightly more energetic than violet light

visible light: the range of photon energies the human eye can detect

infrared radiation: photons with energies slightly less than red light

microwaves: photons with wavelengths on the order of a micron (µm)

Selected Solutions to Problems & Exercises

1. 6.34×10^{-9} eV, 1.01×10^{-27} J

3. 2.42×10^{20} Hz

5.

$$\begin{aligned} hc &= \left(6.62607 \times 10^{-34}\ \mathrm{J}\cdot\ \mathrm{s}\right)\left(2.99792 \times 10^{8}\ \mathrm{m/s}\right)\left(\frac{10^{9}\ \mathrm{nm}}{1\ \mathrm{m}}\right)\left(\frac{1.00000\ \mathrm{eV}}{1.60218\times10^{-19}\ \mathrm{J}}\right) \\ &= 1239.84\ \mathrm{eV}\cdot\mathrm{nm} \\ &\approx 1240\ \mathrm{eV}\cdot\ \mathrm{nm} \end{aligned}$$

7. (a) 0.0829 eV; (b) 121; (c) 1.24 MeV; (d) 1.24×10^{5}

9. (a) 25.0×10^{3} eV; (b) 6.04×10^{18} Hz

11. (a) 2.69; (b) 0.371

13. (a) 1.25×10^{13} photons/s; (b) 997 km

15. 8.33×10^{13} photons/s

17. 181 km

102. Photon Momentum

Learning Objectives

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

- Relate the linear momentum of a photon to its energy or wavelength, and apply linear momentum conservation to simple processes involving the emission, absorption, or reflection of photons.
- Account qualitatively for the increase of photon wavelength that is observed, and explain the significance of the Compton wavelength.

Measuring Photon Momentum

The quantum of EM radiation we call a *photon* has properties analogous to those of particles we can see, such as grains of sand. A photon interacts as a unit in collisions or when absorbed, rather than as an extensive wave. Massive quanta, like electrons, also act like macroscopic particles—something we expect, because they are the smallest units of matter. Particles carry momentum as well as energy. Despite photons having no mass, there has long been evidence that EM radiation carries momentum. (Maxwell and others who studied EM waves predicted that they would carry momentum.) It is now a well-established fact that photons *do* have momentum. In fact, photon momentum is suggested by the photoelectric effect, where photons knock electrons out of a substance. Figure 1 shows macroscopic evidence of photon momentum.

Figure 1 shows a comet with two prominent tails. What most people do not know about the tails is that they always point *away* from the Sun rather than trailing behind the comet (like the tail of Bo Peep's sheep). Comet tails are composed of gases and dust evaporated from the body of the comet and ionized gas. The dust particles recoil away from the Sun when photons scatter from them. Evidently, photons carry momentum in the direction of their motion (away from the Sun), and some of this momentum is transferred to dust particles in collisions. Gas atoms and molecules in the blue tail are most affected by other particles of radiation, such as protons and electrons emanating from the Sun, rather than by the momentum of photons.

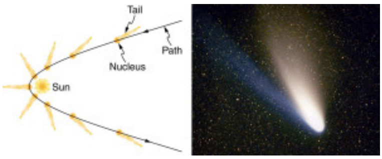


Figure 1. The tails of the Hale-Bopp comet point away from the Sun, evidence that light has momentum. Dust emanating from the body of the comet forms this tail. Particles of dust are pushed away from the Sun by light reflecting from them. The blue ionized gas tail is also produced by photons interacting with atoms in the comet material. (credit: Geoff Chester, U.S. Navy, via Wikimedia Commons)

Making Connections: Conservation of Momentum

Not only is momentum conserved in all realms of physics, but all types of particles are found to have momentum. We expect particles with mass to have momentum, but now we see that massless particles including photons also carry momentum.

Momentum is conserved in quantum mechanics just as it is in relativity and classical physics. Some of the earliest direct experimental evidence of this came from scattering of x-ray photons by electrons in substances, named Compton scattering after the American physicist Arthur H. Compton (1892–1962). Around 1923, Compton observed that x rays scattered from materials had a decreased energy and correctly analyzed this as being due to the scattering of photons from electrons. This phenomenon could be handled as a collision between two particles—a photon and an electron at rest in the material. Energy and momentum are conserved in the collision. (See Figure 2) He won a Nobel Prize in 1929 for the discovery of this scattering, now called the *Compton effect*, because it helped prove that *photon momentum* is given by $p = \frac{h}{\lambda}$, where h is Planck's constant and λ is the photon wavelength. (Note that relativistic momentum given as $p = \gamma m u$ is valid only for particles having mass.)

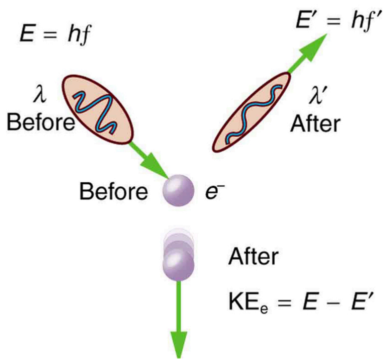


Figure 2. The Compton effect is the name given to the scattering of a photon by an electron. Energy and momentum are conserved, resulting in a reduction of both for the scattered photon. Studying this effect, Compton verified that photons have momentum.

We can see that photon momentum is small, since $p = \frac{h}{\lambda}$ and h is very small. It is for this reason that we do not ordinarily observe photon momentum. Our mirrors do not recoil when light reflects from them (except perhaps in cartoons). Compton saw the effects of photon momentum because he was observing x rays, which have a small wavelength and a relatively large momentum, interacting with the lightest of particles, the electron.

Example 1. Electron and Photon Momentum Compared

1. Calculate the momentum of a visible photon that has a wavelength of 500 nm.
2. Find the velocity of an electron having the same momentum.
3. What is the energy of the electron, and how does it compare with the energy of the photon?

Strategy

Finding the photon momentum is a straightforward application of its definition: $p = \dfrac{h}{\lambda}$. If we find the photon momentum is small, then we can assume that an electron with the same momentum will be nonrelativistic, making it easy to find its velocity and kinetic energy from the classical formulas.

Solution for Part 1

Photon momentum is given by the equation: $p = \dfrac{h}{\lambda}$.

Entering the given photon wavelength yields

$$p = \frac{6.63 \times 10^{-34}\ \mathrm{J} \cdot \mathrm{s}}{500 \times 10^{-9}\ \mathrm{m}} = 1.33 \times 10^{-27}\ \mathrm{kg} \cdot \mathrm{m/s}$$

Solution for Part 2

Since this momentum is indeed small, we will use the classical expression $p = mv$ to find the velocity of an electron with this momentum. Solving for v and using the known value for the mass of an electron gives

$$v = \frac{p}{m} = \frac{1.33 \times 10^{-27} \text{ kg} \cdot \text{ m/s}}{9.11 \times 10^{-31} \text{ kg}} = 1460 \text{ m/s} \approx 1460 \text{ m/s}$$

Solution for Part 3

The electron has kinetic energy, which is classically given by $\text{KE}_e = \frac{1}{2}mv^2$.

Thus,

$$\text{KE}_e = \frac{1}{2}\left(9.11 \times 10^{-3} \text{ kg}\right)\left(1455 \text{ m/s}\right)^2 = 9.64 \times 10^{-25} \text{ J}$$

.

Converting this to eV by multiplying by $\frac{1 \text{ eV}}{1.602 \times 10^{-19} \text{ J}}$ yields $\text{KE}_e = 6.02 \times 10^{-6}$ eV.

The photon energy E is

$$E = \frac{hc}{\lambda} = \frac{1240 \text{ eV} \cdot \text{ nm}}{500 \text{ nm}} = 2.48 \text{ eV},$$

which is about five orders of magnitude greater.

Discussion

Photon momentum is indeed small. Even if we have huge numbers of them, the total momentum they carry is small. An electron with the same momentum has a 1460 m/s velocity, which is clearly nonrelativistic. A more massive particle with the same momentum would have an even smaller velocity. This is borne out by the fact that it takes far less energy to give an electron the same momentum as a photon. But on a quantum-mechanical scale, especially for high-energy photons interacting with small masses, photon momentum is significant. Even on a large scale, photon momentum can have an effect if there are enough of them and if there is nothing to prevent the slow recoil of matter. Comet tails are one example, but there are also proposals to build space sails that use huge low-mass mirrors (made of aluminized Mylar) to reflect sunlight. In the vacuum of space, the mirrors would gradually recoil and could actually take spacecraft from place to place in the solar system. (See Figure 3.)

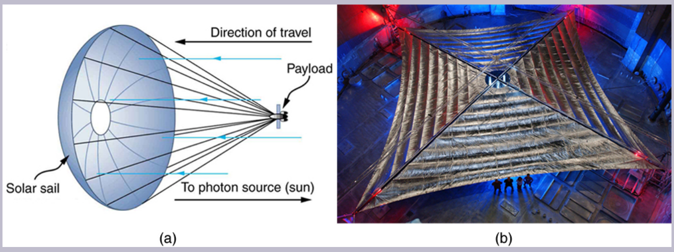


Figure 3. (a) Space sails have been proposed that use the momentum of sunlight reflecting from gigantic low-mass sails to propel spacecraft about the solar system. A Russian test model of this (the Cosmos 1) was launched in 2005, but did not make it into orbit due to a rocket failure. (b) A U.S. version of this, labeled LightSail-1, is scheduled for trial launches in the first part of this decade. It will have a 40-m2 sail. (credit: Kim Newton/NASA)

Relativistic Photon Momentum

There is a relationship between photon momentum *p* and photon energy E that is consistent with the relation given previously for the relativistic total energy of a particle as $E^2 = (pc)^2 + (mc)^2$. We know *m* is zero for a photon, but *p* is not, so that $E^2 = (pc)^2 + (mc)^2$ becomes E = *pc*, or $p = \frac{E}{c}$ (photons).

To check the validity of this relation, note that $E = \frac{hc}{\lambda}$ for a photon. Substituting this into $p = \frac{E}{c}$ yields

$$p = \frac{\frac{hc}{\lambda}}{c} = \frac{h}{\lambda},$$

as determined experimentally and discussed above. Thus, *p*=E/c is equivalent to Compton's result $p=h/\lambda$. For a further verification of the relationship between photon energy and momentum, see Example 2.

Photon Detectors

Almost all detection systems talked about thus far—eyes, photographic plates, photomultiplier tubes in microscopes, and CCD cameras—rely on particle-like properties of photons interacting with a sensitive area. A change is caused and either the change is cascaded or zillions of points are recorded to form an image we detect. These detectors are used in biomedical imaging systems, and there is ongoing research into improving the efficiency of receiving photons, particularly by cooling detection systems and reducing thermal effects.

Example 2. Photon Energy and Momentum

Show that $p = \dfrac{E}{c}$ for the photon considered in the Example 1.

Strategy

We will take the energy E found in Example 1, divide it by the speed of light, and see if the same momentum is obtained as before.

Solution

Given that the energy of the photon is 2.48 eV and converting this to joules, we get

$$p = \frac{E}{c} = \frac{(2.48 \text{ eV})\left(1.60 \times 10^{-16} \text{ J/eV}\right)}{3.00 \times 10^{8} \text{ m/s}} = 1.33 \times 10^{-27} \text{ kg} \cdot \text{m/s}$$

Discussion

This value for momentum is the same as found before (note that unrounded values are used in all calculations to avoid even small rounding errors), an expected verification of the relationship $p = \frac{E}{c}$. This also means the relationship between energy, momentum, and mass given by $E^2 = (pc)^2 + (mc)^2$ applies to both matter and photons. Once again, note that p is not zero, even when m is.

Problem-Solving Suggestion

Note that the forms of the constants h = 4.14 × 10^{-15} eV · s and hc = 1240 eV · nm may be particularly useful for this section's Problems and Exercises.

Section Summary

- Photons have momentum, given by $p = \frac{h}{\lambda}$, where λ is the photon wavelength.
- Photon energy and momentum are related by $p = \frac{E}{c}$, where $E = hf = \frac{hc}{\lambda}$ for a photon.

Conceptual Questions

1. Which formula may be used for the momentum of all particles, with or without mass?
2. Is there any measurable difference between the momentum of a photon and the momentum of

matter?

3. Why don't we feel the momentum of sunlight when we are on the beach?

Problems & Exercises

1. (a) Find the momentum of a 4.00-cm-wavelength microwave photon. (b) Discuss why you expect the answer to (a) to be very small.
2. (a) What is the momentum of a 0.0100-nm-wavelength photon that could detect details of an atom? (b) What is its energy in MeV?
3. (a) What is the wavelength of a photon that has a momentum of $5.00 \times 10^{-29}\ \text{kg} \cdot \text{m/s}$? (b) Find its energy in eV.
4. (a) A γ-ray photon has a momentum of 8.00×10^{-21} $\text{kg} \cdot \text{m/s}$. What is its wavelength? (b) Calculate its energy in MeV.
5. (a) Calculate the momentum of a photon having a wavelength of 2.50 µm. (b) Find the velocity of an electron having the same momentum. (c) What is the kinetic energy of the electron, and how does it compare with that of the photon?
6. Repeat the previous problem for a 10.0-nm-wavelength photon.
7. (a) Calculate the wavelength of a photon that has the same momentum as a proton moving at 1.00% of the speed of light. (b) What is the energy of the

photon in MeV? (c) What is the kinetic energy of the proton in MeV?

8. (a) Find the momentum of a 100-keV x-ray photon. (b) Find the equivalent velocity of a neutron with the same momentum. (c) What is the neutron's kinetic energy in keV?
9. Take the ratio of relativistic rest energy, $E = \gamma mc^2$, to relativistic momentum, $p = \gamma mu$, and show that in the limit that mass approaches zero, you find
$$\frac{E}{p} = c.$$
10. **Construct Your Own Problem.** Consider a space sail such as mentioned in Example 1. Construct a problem in which you calculate the light pressure on the sail in N/m^2 produced by reflecting sunlight. Also calculate the force that could be produced and how much effect that would have on a spacecraft. Among the things to be considered are the intensity of sunlight, its average wavelength, the number of photons per square meter this implies, the area of the space sail, and the mass of the system being accelerated.
11. **Unreasonable Results.** A car feels a small force due to the light it sends out from its headlights, equal to the momentum of the light divided by the time in which it is emitted. (a) Calculate the power of each headlight, if they exert a total force of 2.00×10^{-2} N backward on the car. (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

Glossary

photon momentum: the amount of momentum a photon has, calculated by $p = \frac{h}{\lambda} = \frac{E}{c}$

Compton effect: the phenomenon whereby x rays scattered from materials have decreased energy

Selected Solutions to Problems & Exercises

1. (a) $1.66 \times 10^{-32}\ \text{kg} \cdot \text{m/s}$; (b) The wavelength of microwave photons is large, so the momentum they carry is very small.

3. (a) 13.3 μm; (b) 9.38×10^{-2} eV

5. (a) $2.65 \times 10^{-28}\ \text{kg} \cdot \text{m/s}$; (b) 291 m/s; (c) electron 3.86×10^{-26} J, photon 7.96×10^{-20} J, ratio 2.06×10^{6}

7. (a) 1.32×10^{-13} m; (b) 9.39 MeV; (c) 4.70×10^{-2} MeV

9. $E = \gamma mc^2$ and $P = \gamma mu$, so

$$\frac{E}{P} = \frac{\gamma mc^2}{\gamma mu} = \frac{c^2}{u}$$

As the mass of particle approaches zero, its velocity u will approach c, so that the ratio of energy to momentum in this limit is

$$\lim_{m \to 0} \frac{E}{P} = \frac{c^2}{c} = c$$

which is consistent with the equation for photon energy.

11. (a) 3.00×10^6 W; (b) Headlights are way too bright; (c) Force is too large.

103. The Particle-Wave Duality

Learning Objectives

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

- Explain what the term particle-wave duality means, and why it is applied to EM radiation.

We have long known that EM radiation is a wave, capable of interference and diffraction. We now see that light can be modeled as photons, which are massless particles. This may seem contradictory, since we ordinarily deal with large objects that never act like both wave and particle. An ocean wave, for example, looks nothing like a rock. To understand small-scale phenomena, we make analogies with the large-scale phenomena we observe directly. When we say something behaves like a wave, we mean it shows interference effects analogous to those seen in overlapping water waves. (See Figure 1.) Two examples of waves are sound and EM radiation. When we say something behaves like a particle, we mean that it interacts as a discrete unit with no interference effects. Examples of particles include electrons, atoms, and photons of EM radiation. How do we talk about a phenomenon that acts like both a particle and a wave?

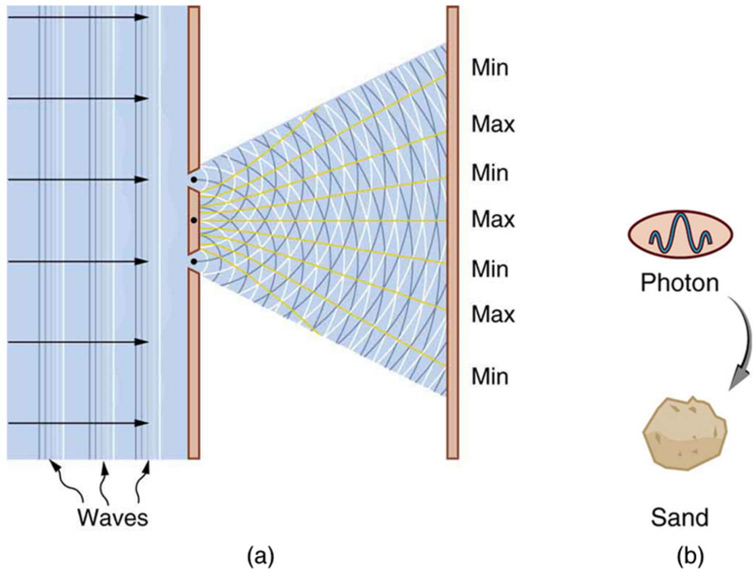


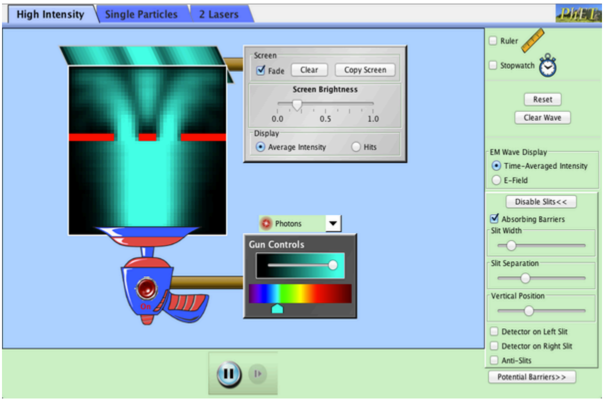
Figure 1. (a) The interference pattern for light through a double slit is a wave property understood by analogy to water waves. (b) The properties of photons having quantized energy and momentum and acting as a concentrated unit are understood by analogy to macroscopic particles.

There is no doubt that EM radiation interferes and has the properties of wavelength and frequency. There is also no doubt that it behaves as particles—photons with discrete energy. We call this twofold nature the *particle-wave duality*, meaning that EM radiation has both particle and wave properties. This so-called duality is simply a term for properties of the photon analogous to phenomena we can observe directly, on a macroscopic scale. If this term seems strange, it is because we do not ordinarily observe details on the quantum level directly, and our observations yield either particle *or* wavelike properties, but never both simultaneously.

Since we have a particle-wave duality for photons, and since we have seen connections between photons and matter in that both have momentum, it is reasonable to ask whether there is a particle-wave duality for matter as well. If the EM radiation we once thought to be a pure wave has particle properties, is it possible that matter has wave properties? The answer is yes. The consequences are tremendous, as we will begin to see in the next section.

PhET Explorations: Quantum Wave Interference

When do photons, electrons, and atoms behave like particles and when do they behave like waves? Watch waves spread out and interfere as they pass through a double slit, then get detected on a screen as tiny dots. Use quantum detectors to explore how measurements change the waves and the patterns they produce on the screen.



Click to download the simulation. Run using Java.

Section Summary

- EM radiation can behave like either a particle or a wave.
- This is termed particle-wave duality.

Glossary

particle-wave duality: the property of behaving like either a particle or a wave; the term for the phenomenon that all particles have wave characteristics

104. The Wave Nature of Matter

Learning Objective

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

- Describe the Davisson-Germer experiment, and explain how it provides evidence for the wave nature of electrons.

De Broglie Wavelength

In 1923 a French physics graduate student named Prince Louis-Victor de Broglie (1892–1987) made a radical proposal based on the hope that nature is symmetric. If EM radiation has both particle and wave properties, then nature would be symmetric if matter also had both particle and wave properties. If what we once thought of as an unequivocal wave (EM radiation) is also a particle, then what we think of as an unequivocal particle (matter) may also be a wave. De Broglie's suggestion, made as part of his doctoral thesis, was so radical that it was greeted with some skepticism. A copy of his thesis was sent to Einstein, who said it was not only probably correct, but that it might be of fundamental importance. With the support of Einstein and a few other prominent physicists, de Broglie was awarded his doctorate.

De Broglie took both relativity and quantum mechanics into account to develop the proposal that *all particles have a wavelength*, given by

$$\lambda = \frac{h}{p} \text{ (matter and photons)},$$

where h is Planck's constant and p is momentum. This is defined to be the *de Broglie wavelength*. (Note that we already have this for photons, from the equation $p = \frac{h}{\lambda}$.) The hallmark of a wave is interference. If matter is a wave, then it must exhibit constructive and destructive interference. Why isn't this ordinarily observed? The answer is that in order to see significant interference effects, a wave must interact with an object about the same size as its wavelength. Since h is very small, λ is also small, especially for macroscopic objects. A 3-kg bowling ball moving at 10 m/s, for example, has

$$\lambda = \frac{h}{p} = \frac{6.63 \times 10^{-34}\ \text{J} \cdot \text{s}}{(3\ \text{kg})\,(10\ \text{m/s})} = 2 \times 10^{-35}\ \text{m}.$$

This means that to see its wave characteristics, the bowling ball would have to interact with something about 10^{-35} m in size—far smaller than anything known. When waves interact with objects much larger than their wavelength, they show negligible interference effects and move in straight lines (such as light rays in geometric optics). To get easily observed interference effects from particles of matter, the longest wavelength and hence smallest mass possible would be useful. Therefore, this effect was first observed with electrons.

American physicists Clinton J. Davisson and Lester H. Germer in 1925 and, independently, British physicist G. P. Thomson (son of J. J. Thomson, discoverer of the electron) in 1926 scattered electrons from crystals and found diffraction patterns. These patterns are exactly consistent with interference of electrons having the de Broglie wavelength and are somewhat analogous to light interacting with a diffraction grating. (See Figure 1.)

Making Connections: Waves

All microscopic particles, whether massless, like photons, or having mass, like electrons, have wave properties. The relationship between momentum and wavelength is fundamental for all particles.

De Broglie's proposal of a wave nature for all particles initiated a remarkably productive era in which the foundations for quantum mechanics were laid. In 1926, the Austrian physicist Erwin Schrödinger (1887–1961) published four papers in which the wave nature of particles was treated explicitly with wave equations. At the same time, many others began important work.

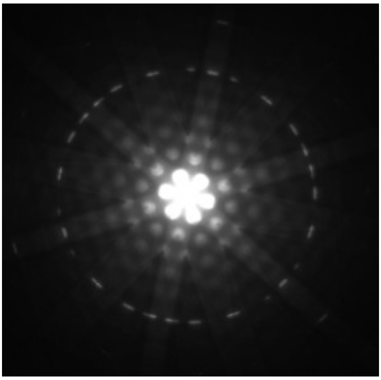


Figure 1. This diffraction pattern was obtained for electrons diffracted by crystalline silicon. Bright regions are those of constructive interference, while dark regions are those of destructive interference. (credit: Ndthe, Wikimedia Commons)

Among them was German physicist Werner Heisenberg (1901–1976) who, among many other contributions to quantum mechanics, formulated a mathematical treatment of the wave nature of matter that used matrices rather than wave equations. We will deal with some specifics in later sections, but it is worth noting that de Broglie's work was a watershed for the development of quantum mechanics. De Broglie was awarded the Nobel Prize in 1929 for his vision, as were Davisson and G. P. Thomson in 1937 for their experimental verification of de Broglie's hypothesis.

Example 1. Electron Wavelength versus Velocity and Energy

For an electron having a de Broglie wavelength of 0.167 nm (appropriate for interacting with crystal lattice structures that are about this size):

1. Calculate the electron's velocity, assuming it is nonrelativistic.
2. Calculate the electron's kinetic energy in eV.

Strategy

For Part 1, since the de Broglie wavelength is given, the electron's velocity can be obtained from $\lambda = \frac{h}{p}$ by using the nonrelativistic formula for momentum, $p = mv$.

For Part 2, once v is obtained (and it has been verified that v is nonrelativistic), the classical kinetic energy is simply $\frac{1}{2}mv^2$

Solution for Part 1

Substituting the nonrelativistic formula for momentum ($p = mv$) into the de Broglie wavelength gives

$$\lambda = \frac{h}{p} = \frac{h}{mv}.$$

Solving for v gives $v = \dfrac{h}{m\lambda}$.

Substituting known values yields

$$v = \frac{6.63 \times 10^{-34}\ \text{J} \cdot \text{s}}{\left(9.11 \times 10^{-31}\ \text{kg}\right)\left(0.167 \times 10^{-9}\ \text{m}\right)} = 4.36 \times 10^{6}\ \text{m/s}$$

Solution for Part 2

While fast compared with a car, this electron's speed is not highly relativistic, and so we can comfortably use the classical formula to find the electron's kinetic energy and convert it to eV as requested.

$$\begin{aligned} \text{KE} &= \tfrac{1}{2}mv^2 \\ &= \tfrac{1}{2}\left(9.11 \times 10^{-31}\ \text{kg}\right)\left(4.36 \times 10^{6}\ \text{m/s}\right)^2 \\ &= \left(86.4 \times 10^{-18}\ \text{J}\right)\left(\frac{1\ \text{eV}}{1.602 \times 10^{-19}\ \text{J}}\right) \\ &= 54.0\ \text{eV} \end{aligned}$$

Discussion

This low energy means that these 0.167-nm electrons could be obtained by accelerating them through a 54.0-V electrostatic potential, an easy task. The results also confirm the assumption that the electrons are nonrelativistic, since their velocity is just over 1% of the speed of light and the kinetic energy is about 0.01% of the rest energy of an electron (0.511 MeV). If the electrons had turned out to be relativistic, we would have had to use more involved calculations employing relativistic formulas.

Electron Microscopes

One consequence or use of the wave nature of matter is found in the electron microscope. As we have discussed, there is a limit to the detail observed with any probe having a wavelength. Resolution, or observable detail, is limited to about one wavelength. Since a potential of only 54 V can produce electrons with sub-nanometer wavelengths, it is easy to get electrons with much smaller wavelengths than those of visible light (hundreds of nanometers). Electron microscopes can, thus, be constructed to detect much smaller details than optical microscopes. (See Figure 2.)

There are basically two types of electron microscopes. The transmission electron microscope (TEM) accelerates electrons that are emitted from a hot filament (the cathode). The beam is broadened and then passes through the sample. A magnetic lens focuses the beam image onto a fluorescent screen, a photographic plate, or (most probably) a CCD (light sensitive camera), from which it is transferred to a computer. The TEM is similar to the optical microscope, but it requires a thin sample examined in a vacuum. However it can resolve details as small as 0.1 nm (10^{-10} m), providing magnifications of 100 million times the size of the original object. The TEM has allowed us to see individual atoms and structure of cell nuclei.

The scanning electron microscope (SEM) provides images by using secondary electrons produced by the primary beam interacting with the surface of the sample (see Figure 2). The SEM also uses magnetic lenses to focus the beam onto the sample. However, it moves the beam around electrically to "scan" the sample in the x and y directions. A CCD detector is used to process the data for each electron position, producing images like the one at the beginning of this chapter. The SEM has the advantage of not requiring a thin sample and of providing a 3-D view. However, its resolution is about ten times less than a TEM.

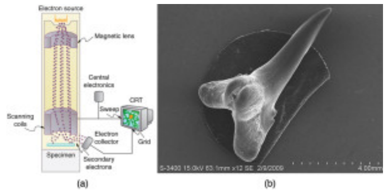


Figure 2. Schematic of a scanning electron microscope (SEM) (a) used to observe small details, such as those seen in this image of a tooth of a Himipristis, a type of shark (b). (credit: Dallas Krentzel, Flickr)

Electrons were the first particles with mass to be directly confirmed to have the wavelength proposed by de Broglie. Subsequently, protons, helium nuclei, neutrons, and many others have been observed to exhibit interference when they interact with objects having sizes similar to their de Broglie wavelength. The de Broglie wavelength for massless particles was well established in the 1920s for photons, and it has since been observed that all massless particles have a de Broglie wavelength $\lambda = \frac{h}{p}$. The wave nature of all particles is a universal characteristic of nature. We shall see in following sections that implications of the de Broglie wavelength include the quantization of energy in atoms and molecules, and an alteration of our basic view of nature on the microscopic scale. The next section, for example, shows that there are limits to the precision with which we may make predictions, regardless of how hard we try. There are even limits to the precision with which we may measure an object's location or energy.

Making Connections: A Submicroscopic Diffraction Grating

The wave nature of matter allows it to exhibit all the characteristics of other, more familiar, waves. Diffraction gratings, for example, produce diffraction patterns for light that depend on grating spacing and the wavelength of the light. This effect, as with most wave phenomena, is most pronounced when the wave interacts with objects having a size similar to its wavelength. For gratings, this is the spacing between multiple slits.) When electrons interact with a system having a spacing similar to the electron wavelength, they show the same types of interference patterns as light does for diffraction gratings, as shown at top left in Figure 3.

Atoms are spaced at regular intervals in a crystal as parallel planes, as shown in the bottom part of Figure 3. The spacings between these planes act like the openings in a diffraction grating. At certain incident angles, the paths of electrons scattering from successive planes differ by one wavelength and, thus, interfere constructively. At other angles, the path length differences are not an integral wavelength, and there is partial to total destructive interference. This type of scattering from a large crystal with well-defined lattice planes can produce dramatic interference patterns. It is called *Bragg reflection*, for the father-and-son team who first explored and analyzed it in some detail. The expanded view also shows the path-length differences and indicates how these depend on incident angle θ in a manner similar to the diffraction patterns for x rays reflecting from a crystal.

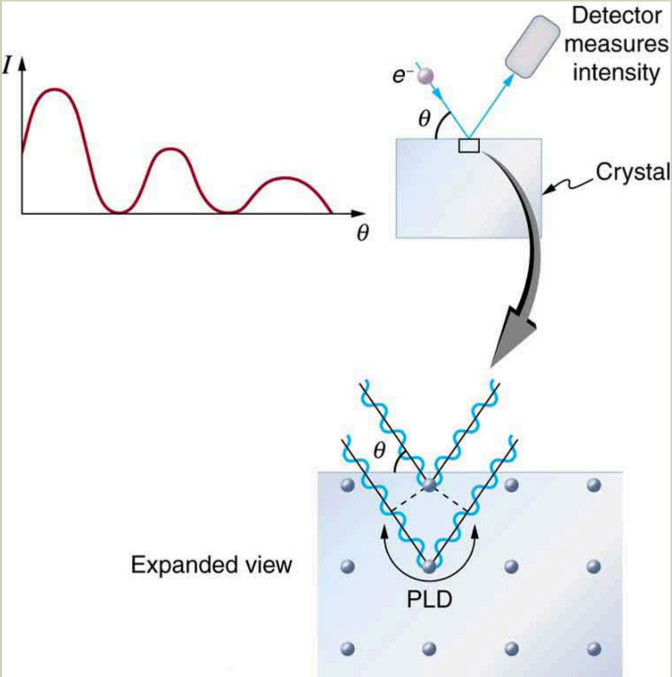


Figure 3. The diffraction pattern at top left is produced by scattering electrons from a crystal and is graphed as a function of incident angle relative to the regular array of atoms in a crystal, as shown at bottom. Electrons scattering from the second layer of atoms travel farther than those scattered from the top layer. If the path length difference (PLD) is an integral wavelength, there is constructive interference.

Let us take the spacing between parallel planes of atoms in the crystal to be *d*. As mentioned, if the path length difference (PLD) for the electrons is a whole number of wavelengths, there will be constructive interference—that is, PLD = $n\lambda (n = 1, 2, 3,\ldots)$. Because AB = BC = $d \sin \theta$, we have constructive interference when $n\lambda = 2d \sin \theta$. This relationship is called the *Bragg equation* and applies not only to electrons but also to x rays.

The wavelength of matter is a submicroscopic characteristic that explains a macroscopic phenomenon such as Bragg reflection. Similarly, the wavelength of light is a submicroscopic characteristic that explains the macroscopic phenomenon of diffraction patterns.

Section Summary

- Particles of matter also have a wavelength, called the de Broglie wavelength, given by $\lambda = \frac{h}{p}$, where p is momentum.
- Matter is found to have the same interference characteristics as any other wave.

Conceptual Questions

1. How does the interference of water waves differ from the interference of electrons? How are they analogous?
2. Describe one type of evidence for the wave nature of matter.
3. Describe one type of evidence for the particle

nature of EM radiation.

Problems & Exercises

1. At what velocity will an electron have a wavelength of 1.00 m?
2. What is the wavelength of an electron moving at 3.00% of the speed of light?
3. At what velocity does a proton have a 6.00-fm wavelength (about the size of a nucleus)? Assume the proton is nonrelativistic. (1 femtometer = 10^{-15} m.)
4. What is the velocity of a 0.400-kg billiard ball if its wavelength is 7.50 cm (large enough for it to interfere with other billiard balls)?
5. Find the wavelength of a proton moving at 1.00% of the speed of light.
6. Experiments are performed with ultracold neutrons having velocities as small as 1.00 m/s. (a) What is the wavelength of such a neutron? (b) What is its kinetic energy in eV?
7. (a) Find the velocity of a neutron that has a 6.00-fm wavelength (about the size of a nucleus). Assume the neutron is nonrelativistic. (b) What is the neutron's kinetic energy in MeV?
8. What is the wavelength of an electron accelerated through a 30.0-kV potential, as in a TV tube?
9. What is the kinetic energy of an electron in a TEM having a 0.0100-nm wavelength?

10. (a) Calculate the velocity of an electron that has a wavelength of 1.00 μm. (b) Through what voltage must the electron be accelerated to have this velocity?
11. The velocity of a proton emerging from a Van de Graaff accelerator is 25.0% of the speed of light. (a) What is the proton's wavelength? (b) What is its kinetic energy, assuming it is nonrelativistic? (c) What was the equivalent voltage through which it was accelerated?
12. The kinetic energy of an electron accelerated in an x-ray tube is 100 keV. Assuming it is nonrelativistic, what is its wavelength?
13. **Unreasonable Results.** (a) Assuming it is nonrelativistic, calculate the velocity of an electron with a 0.100-fm wavelength (small enough to detect details of a nucleus). (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

Glossary

de Broglie wavelength: the wavelength possessed by a particle of matter, calculated by $\lambda = \dfrac{h}{p}$

Selected Solutions to Problems & Exercises

1. $7.28 \times 10{-}4$ m

3. 6.62×10^{7} m/s

5. 1.32×10^{-13} m

7. (a) 6.62×10^{7} m/s; (b) 22.9 MeV

9. 15.1 keV

11. (a) 5.29 fm; (b) 4.70×10^{-12} J; (c) 29.4 MV

13. (a) 7.28×10^{12} m/s; (b) This is thousands of times the speed of light (an impossibility); (c) The assumption that the electron is non-relativistic is unreasonable at this wavelength.

PART XIV

ATOMIC PHYSICS

105. Introduction to Atomic Physics

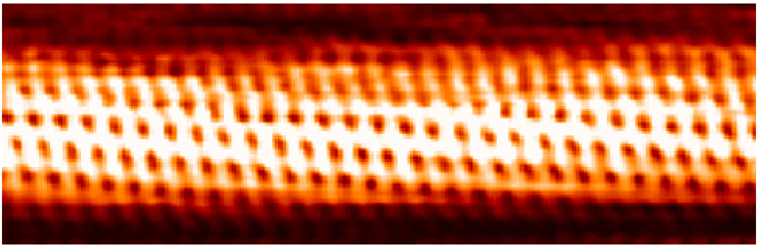


Figure 1. Individual carbon atoms are visible in this image of a carbon nanotube made by a scanning tunneling electron microscope. (credit: Taner Yildirim, National Institute of Standards and Technology, via Wikimedia Commons)

From childhood on, we learn that atoms are a substructure of all things around us, from the air we breathe to the autumn leaves that blanket a forest trail. Invisible to the eye, the existence and properties of atoms are used to explain many phenomena—a theme found throughout this text. In this chapter, we discuss the discovery of atoms and their own substructures; we then apply quantum mechanics to the description of atoms, and their properties and interactions. Along the way, we will find, much like the scientists who made the original discoveries, that new concepts emerge with applications far beyond the boundaries of atomic physics.

106. Discovery of the Atom

Learning Objective

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

- Describe the basic structure of the atom, the substructure of all matter.

How do we know that atoms are really there if we cannot see them with our eyes? A brief account of the progression from the proposal of atoms by the Greeks to the first direct evidence of their existence follows.

People have long speculated about the structure of matter and the existence of atoms. The earliest significant ideas to survive are due to the ancient Greeks in the fifth century BCE, especially those of the philosophers Leucippus and Democritus. (There is some evidence that philosophers in both India and China made similar speculations, at about the same time.) They considered the question of whether a substance can be divided without limit into ever smaller pieces. There are only a few possible answers to this question. One is that infinitesimally small subdivision is possible. Another is what Democritus in particular believed—that there is a smallest unit that cannot be further subdivided. Democritus called this the *atom*. We now know that atoms themselves can be subdivided, but their identity is destroyed in the process, so the Greeks were correct in a respect. The Greeks also felt that atoms were in constant motion, another correct notion.

The Greeks and others speculated about the properties of atoms, proposing that only a few types existed and that all matter was formed as various combinations of these types. The famous proposal that the basic elements were earth, air, fire, and water was brilliant, but incorrect. The Greeks had identified the most common examples of the four states of matter (solid, gas, plasma, and liquid), rather than the basic elements. More than 2000 years passed before observations could be made with equipment capable of revealing the true nature of atoms.

Over the centuries, discoveries were made regarding the properties of substances and their chemical reactions. Certain systematic features were recognized, but similarities between common and rare elements resulted in efforts to transmute them (lead into gold, in particular) for financial gain. Secrecy was endemic. Alchemists discovered and rediscovered many facts but did not make them broadly available. As the Middle Ages ended, alchemy gradually faded, and the science of chemistry arose. It was no longer possible, nor considered desirable, to keep discoveries secret. Collective knowledge grew, and by the beginning of the 19th century, an important fact was well established—the masses of reactants in specific chemical reactions always have a particular mass ratio. This is very strong indirect evidence that there are basic units (atoms and molecules) that have these same mass ratios. The English chemist John Dalton (1766–1844) did much of this work, with significant contributions by the Italian physicist Amedeo Avogadro (1776–1856). It was Avogadro who developed the idea of a fixed number of atoms and molecules in a mole, and this special number is called Avogadro's number in his honor. The Austrian physicist Johann Josef Loschmidt was the first to measure the value of the constant in 1865 using the kinetic theory of gases.

Patterns and Systematics

The recognition and appreciation of patterns has enabled us to make many discoveries. The periodic table of elements was proposed as an organized summary of the known elements long before all elements had been discovered, and it led to many other discoveries. We shall see in later chapters that patterns in the properties of subatomic particles led to the proposal of quarks as their underlying structure, an idea that is still bearing fruit.

Knowledge of the properties of elements and compounds grew, culminating in the mid-19th-century development of the periodic table of the elements by Dmitri Mendeleev (1834–1907), the great Russian chemist. Mendeleev proposed an ingenious array that highlighted the periodic nature of the properties of elements. Believing in the systematics of the periodic table, he also predicted the existence of then-unknown elements to complete it. Once these elements were discovered and determined to have properties predicted by Mendeleev, his periodic table became universally accepted.

Also during the 19th century, the kinetic theory of gases was developed. Kinetic theory is based on the existence of atoms and molecules in random thermal motion and provides a microscopic explanation of the gas laws, heat transfer, and thermodynamics (see Introduction to Temperature, Kinetic Theory, and the Gas Laws and Introduction to Laws of Thermodynamics). Kinetic theory works so well that it is another strong indication of the existence of atoms.

But it is still indirect evidence—individual atoms and molecules had not been observed. There were heated debates about the validity of kinetic theory until direct evidence of atoms was obtained.

The first truly direct evidence of atoms is credited to Robert Brown, a Scottish botanist. In 1827, he noticed that tiny pollen grains suspended in still water moved about in complex paths. This can be observed with a microscope for any small particles in a fluid. The motion is caused by the random thermal motions of fluid molecules colliding with particles in the fluid, and it is now called *Brownian motion*. (See Figure 1.) Statistical fluctuations in the numbers of molecules striking the sides of a visible particle cause it to move first this way, then that. Although the molecules cannot be directly observed, their effects on the particle can be. By examining Brownian motion, the size of molecules can be calculated. The smaller and more numerous they are, the smaller the fluctuations in the numbers striking different sides.

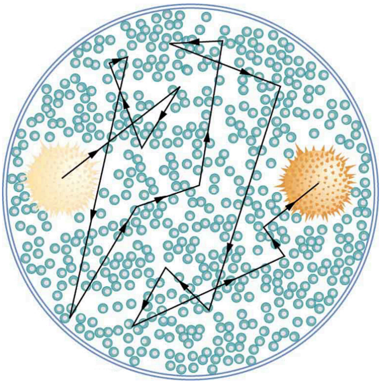


Figure 1. The position of a pollen grain in water, measured every few seconds under a microscope, exhibits Brownian motion. Brownian motion is due to fluctuations in the number of atoms and molecules colliding with a small mass, causing it to move about in complex paths. This is nearly direct evidence for the existence of atoms, providing a satisfactory alternative explanation cannot be found.

It was Albert Einstein who, starting in his epochal year of 1905, published several papers that explained precisely how Brownian motion could be used to measure the size of atoms and molecules. (In 1905 Einstein created special relativity, proposed photons as quanta of EM radiation, and produced a theory of Brownian motion that allowed the size of atoms to be determined. All of this was done in his spare time, since he worked days as a patent examiner. Any one of these very basic works could have been the crowning achievement of an entire career—yet Einstein did even more in later years.) Their sizes were only approximately known to be 10^{-10} m, based on a comparison of latent heat of vaporization and surface tension made in about 1805 by Thomas Young of double-slit fame and the famous astronomer and mathematician Simon Laplace.

Using Einstein's ideas, the French physicist Jean-Baptiste Perrin (1870–1942) carefully observed Brownian motion; not only did he confirm Einstein's theory, he also produced accurate sizes for atoms and molecules. Since molecular weights and densities of materials were well established, knowing atomic and molecular sizes allowed a precise value for Avogadro's number to be obtained. (If we know how big an atom is, we know how many fit into a certain volume.) Perrin also used these ideas to explain atomic and molecular agitation effects in sedimentation, and he received the 1926 Nobel Prize for his achievements. Most scientists were already convinced of the existence of atoms, but the accurate observation and analysis of Brownian motion was conclusive—it was the first truly direct evidence.

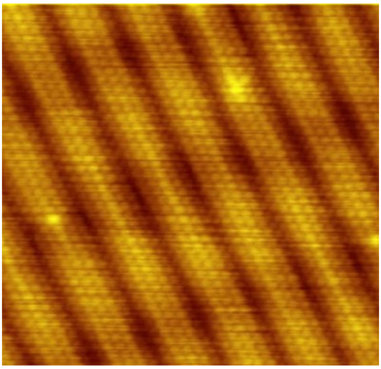


Figure 2. Individual atoms can be detected with devices such as the scanning tunneling electron microscope that produced this image of individual gold atoms on a graphite substrate. (credit: Erwin Rossen, Eindhoven University of Technology, via Wikimedia Commons)

A huge array of direct and indirect evidence for the existence of atoms now exists. For example, it has become possible to accelerate ions (much as electrons are accelerated in cathode-ray tubes) and to detect them individually as well as measure their masses (see More Applications of Magnetism for a discussion of mass spectrometers). Other devices that observe individual atoms, such as the scanning tunneling electron microscope, will be discussed elsewhere. (See

Figure 2.) All of our understanding of the properties of matter is based on and consistent with the atom. The atom's substructures, such as electron shells and the nucleus, are both interesting and important. The nucleus in turn has a substructure, as do the particles of which it is composed. These topics, and the question of whether there is a smallest basic structure to matter, will be explored in later parts of the text.

Section Summary

- Atoms are the smallest unit of elements; atoms combine to form molecules, the smallest unit of compounds.
- The first direct observation of atoms was in Brownian motion.
- Analysis of Brownian motion gave accurate sizes for atoms (10^{-10} m on average) and a precise value for Avogadro's number.

Conceptual Questions

1. Name three different types of evidence for the existence of atoms.
2. Explain why patterns observed in the periodic table of the elements are evidence for the existence of atoms, and why Brownian motion is a more direct type of evidence for their existence.
3. If atoms exist, why can't we see them with visible light?

Problems & Exercises

1. Using the given charge-to-mass ratios for electrons and protons, and knowing the magnitudes of their charges are equal, what is the ratio of the proton's mass to the electron's? (Note that since the charge-to-mass ratios are given to only three-digit accuracy, your answer may differ from the accepted ratio in the fourth digit.)
2. (a) Calculate the mass of a proton using the charge-to-mass ratio given for it in this chapter and its known charge. (b) How does your result compare with the proton mass given in this chapter?
3. If someone wanted to build a scale model of the atom with a nucleus 1.00 m in diameter, how far away would the nearest electron need to be?

Glossary

atom: basic unit of matter, which consists of a central, positively charged nucleus surrounded by negatively charged electrons

Brownian motion: the continuous random movement of particles of matter suspended in a liquid or gas

Selected Solutions to Problems & Exercises

1. 1.84×10^3

3. 50 km

107. Discovery of the Parts of the Atom: Electrons and Nuclei

Learning Objectives

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

- Describe how electrons were discovered.
- Explain the Millikan oil drop experiment.
- Describe Rutherford's gold foil experiment.
- Describe Rutherford's planetary model of the atom.

Just as atoms are a substructure of matter, electrons and nuclei are substructures of the atom. The experiments that were used to discover electrons and nuclei reveal some of the basic properties of atoms and can be readily understood using ideas such as electrostatic and magnetic force, already covered in previous chapters.

Charges and Electromagnetic Forces

In previous discussions, we have noted that positive charge is associated with nuclei and negative charge with electrons. We have also covered many aspects of the electric and magnetic forces that affect charges. We will now explore the discovery of the electron and nucleus as substructures of the atom and examine their contributions to the properties of atoms.

The Electron

Gas discharge tubes, such as that shown in Figure 1, consist of an evacuated glass tube containing two metal electrodes and a rarefied gas. When a high voltage is applied to the electrodes, the gas glows. These tubes were the precursors to today's neon lights. They were first studied seriously by Heinrich Geissler, a German inventor and glassblower, starting in the 1860s. The English scientist William Crookes, among others, continued to study what for some time were called Crookes tubes, wherein electrons are freed from atoms and molecules in the rarefied gas inside the tube and are accelerated from the cathode (negative) to the anode (positive) by the high potential. These "*cathode rays*" collide with the gas atoms and molecules and excite them, resulting in the emission of electromagnetic (EM) radiation that makes the electrons' path visible as a ray that spreads and fades as it moves away from the cathode.

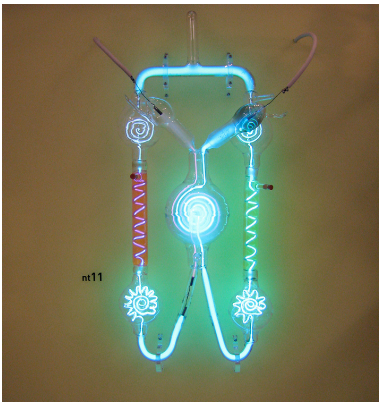


Figure 1. A gas discharge tube glows when a high voltage is applied to it. Electrons emitted from the cathode are accelerated toward the anode; they excite atoms and molecules in the gas, which glow in response. Once called Geissler tubes and later Crookes tubes, they are now known as cathode-ray tubes (CRTs) and are found in older TVs, computer screens, and x-ray machines. When a magnetic field is applied, the beam bends in the direction expected for negative charge. (credit: Paul Downey, Flickr)

Gas discharge tubes today are most commonly called *cathode-ray tubes*, because the rays originate at the cathode. Crookes showed that the electrons carry momentum (they can make a small paddle wheel rotate). He also found that their normally straight path is bent by a magnet in the direction expected for a negative charge moving away from the cathode. These were the first direct indications of electrons and their charge.

Figure 2. J. J. Thomson (credit: www.firstworldwar.com, via Wikimedia Commons)

The English physicist J. J. Thomson (1856–1940) improved and expanded the scope of experiments with gas discharge tubes. (See Figure 2 and Figure 3.) He verified the negative charge of the cathode rays with both magnetic and electric fields. Additionally, he collected the rays in a metal cup and found an excess of negative charge. Thomson was also able to measure the ratio of the charge of the electron to its mass, $\frac{q_e}{m_e}$—an important step to finding the actual values of both q_e and m_e. Figure 4 shows a cathode-ray tube, which produces a narrow beam of electrons that passes through charging plates connected to a high-voltage power supply. An electric field **E** is produced between the charging plates, and the cathode-ray tube is placed between the poles of a magnet so that the electric field E is perpendicular to the magnetic field **B** of the magnet. These fields, being perpendicular to each other, produce opposing forces on the electrons. As discussed for mass spectrometers in More Applications of Magnetism, if the net force due to the fields vanishes, then the velocity of the charged particle is $v = \frac{E}{B}$. In this manner, Thomson determined the velocity of the electrons and then moved the beam up and down by adjusting the electric field.

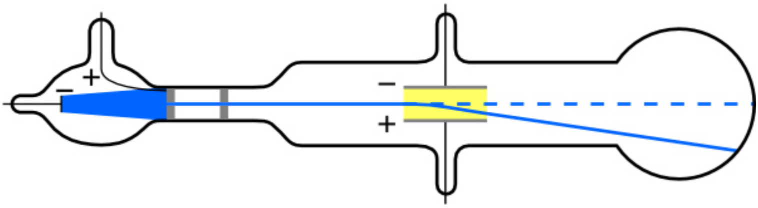


Figure 3. Diagram of Thomson's CRT. (credit: Kurzon, Wikimedia Commons)

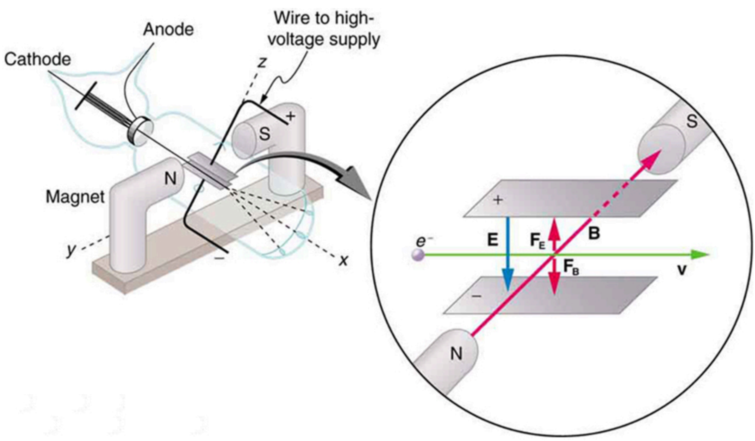


Figure 4. This schematic shows the electron beam in a CRT passing through crossed electric and magnetic fields and causing phosphor to glow when striking the end of the tube.

To see how the amount of deflection is used to calculate $\frac{q_e}{m_e}$, note that the deflection is proportional to the electric force on the electron: $F = q_e E$.

But the vertical deflection is also related to the electron's mass, since the electron's acceleration is $a = \frac{F}{m_e}$.

The value of F is not known, since q_e was not yet known.

Substituting the expression for electric force into the expression for acceleration yields $a = \frac{F}{m_e} = \frac{q_e E}{m_e}$.

Gathering terms, we have $\frac{q_e}{m_e} = \frac{a}{E}$.

The deflection is analyzed to get a, and E is determined from the applied voltage and distance between the plates; thus, $\frac{q_e}{m_e}$ can be determined. With the velocity known, another measurement of $\frac{q_e}{m_e}$ can be obtained by bending the beam of electrons with the magnetic field. Since $F_{mag} = q_e vB = m_e a$, we have $\frac{q_e}{m_e} = \frac{a}{vB}$. Consistent results are obtained using magnetic deflection.

What is so important about $\frac{q_e}{m_e}$, the ratio of the electron's charge to its mass? The value obtained is

$$\frac{q_e}{m_e} = -1.76 \times 10^{11}\ \text{C/kg (electron)}$$

This is a huge number, as Thomson realized, and it implies that the electron has a very small mass. It was known from electroplating that about 10^8 C/kg is needed to plate a material, a factor of about 1000 less than the charge per kilogram of electrons. Thomson went on to do the same experiment for positively charged hydrogen ions (now known to be bare protons) and found a charge per kilogram about 1000 times smaller than that for the electron, implying that the proton is about 1000 times more massive than the electron. Today, we know more precisely that

$$\frac{q_p}{m_p} = 9.58 \times 10^7\ \text{C/kg (proton)},$$

where q_p is the charge of the proton and m_p is its mass. This ratio (to four significant figures) is 1836 times less charge per kilogram than for the electron. Since the charges of electrons and protons are equal in magnitude, this implies $m_p = 1836 m_e$.

Thomson performed a variety of experiments using differing gases in discharge tubes and employing other methods, such as the photoelectric effect, for freeing electrons from atoms. He always found the same properties for the electron, proving it to be an independent particle. For his work, the important pieces of which he began to publish in 1897, Thomson was awarded the 1906 Nobel Prize in Physics. In retrospect, it is difficult to appreciate how astonishing it was to find that the atom has a substructure. Thomson himself said, "It was only when I was convinced that the experiment left no escape from it that I published my belief in the existence of bodies smaller than atoms."

Thomson attempted to measure the charge of individual electrons, but his method could determine its charge only to the order of magnitude expected.

Since Faraday's experiments with electroplating in the 1830s, it had been known that about 100,000 C per mole was needed to plate singly ionized ions. Dividing this by the number of ions per mole (that is, by Avogadro's number), which was approximately known, the charge per ion was calculated to be about 1.6×10^{-19} C, close to the actual value.

An American physicist, Robert Millikan (1868–1953) (see Figure 5), decided to improve upon Thomson's experiment for measuring q_e and was eventually forced to try another approach, which is now a classic experiment performed by students. The Millikan oil drop experiment is shown in Figure 6.

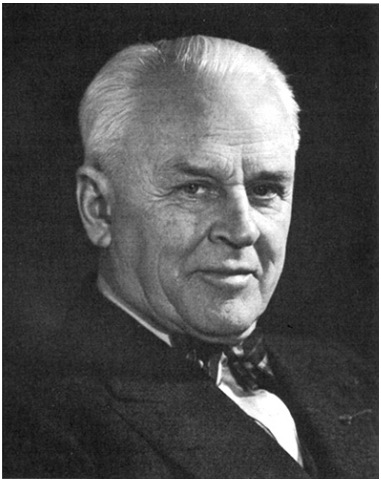


Figure 5. Robert Millikan (credit: Unknown Author, via Wikimedia Commons)

In the Millikan oil drop experiment, fine drops of oil are sprayed from an atomizer. Some of these are charged by the process and can then be suspended between metal plates by a voltage between the plates. In this situation, the weight of the drop is balanced by the electric force:

$$m_{\text{drop}}g = q_e E$$

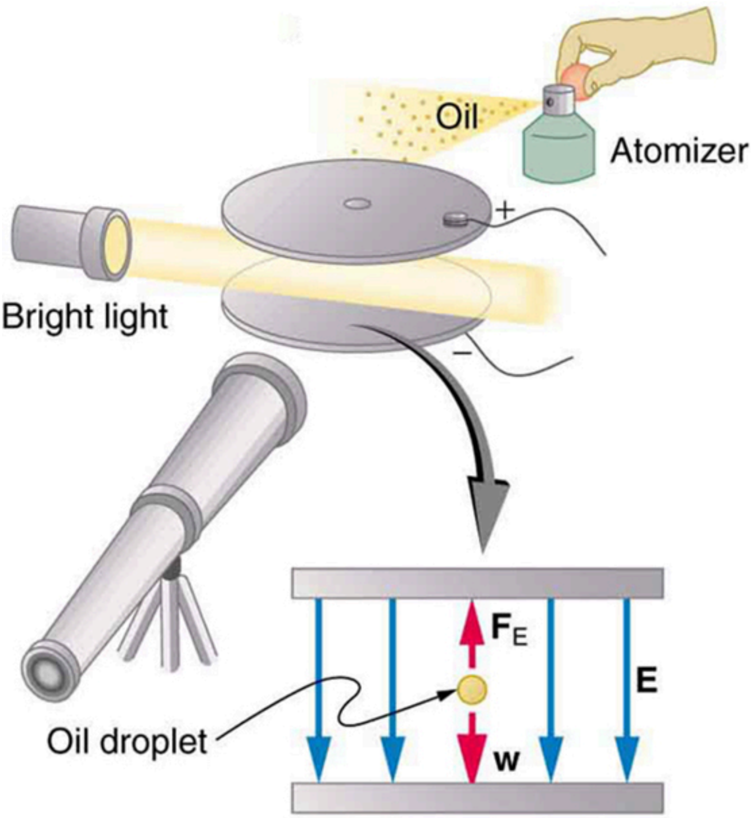


Figure 6. The Millikan oil drop experiment produced the first accurate direct measurement of the charge on electrons, one of the most fundamental constants in nature. Fine drops of oil become charged when sprayed. Their movement is observed between metal plates with a potential applied to oppose the gravitational force. The balance of gravitational and electric forces allows the calculation of the charge on a drop. The charge is found to be quantized in units of -1.6×10^{-19} *C, thus determining directly the charge of the excess and missing electrons on the oil drops.*

The electric field is produced by the applied voltage, hence, $E = \frac{V}{d}$, and V is adjusted to just balance the drop's weight. The drops can be seen as points of reflected light using a microscope, but they are too small to directly measure their size and mass. The mass of the drop is determined by observing how fast it falls when the voltage is turned off. Since air resistance is very significant for these submicroscopic drops, the more massive drops fall faster than the less massive, and sophisticated sedimentation calculations can reveal their mass. Oil is used rather than water, because it does not readily evaporate, and so mass is nearly constant. Once the mass of the drop is known, the charge of the electron is given by rearranging the previous equation:

$$q = \frac{m_{\text{drop}} g}{E} = \frac{m_{\text{drop}} g d}{V},$$

where d is the separation of the plates and V is the voltage that holds the drop motionless. (The same drop can be observed for several hours to see that it really is motionless.) By 1913 Millikan had measured the charge of the electron q_e to an accuracy of 1%, and he improved this by a factor of 10 within a few years to a value of -1.60×10^{-19} C. He also observed that all charges were multiples of the basic electron charge and that sudden changes could occur in which electrons were added or removed from the drops. For this very fundamental direct measurement of q_e and for his studies of the photoelectric effect, Millikan was awarded the 1923 Nobel Prize in Physics.

With the charge of the electron known and the charge-to-mass ratio known, the electron's mass can be calculated. It is

$$m = \frac{q_e}{\left(\frac{q_e}{m_e}\right)}$$

Substituting known values yields

$$m_e = \frac{-1.60 \times 10^{-19}\ \text{C}}{-1.76 \times 10^{11}\ \text{C/kg}}$$

or $m_e = 9.11 \times 10^{-31}$ kg (electron's mass), where the round-off errors have been corrected. The mass of the electron has been verified in many subsequent experiments and is now known to an accuracy of better than one part in one million. It is an incredibly small mass and remains the smallest known mass of any particle that has mass. (Some particles, such as photons, are massless and cannot be brought to rest, but travel at the speed of light.) A similar calculation gives the masses of other particles, including the proton. To three digits, the mass of the proton is now known to be $m_p = 1.67 \times 10^{-27}$ kg (proton's mass), which is nearly identical to the mass of a hydrogen atom. What Thomson and Millikan had done was to prove the existence of one substructure of atoms, the electron, and further to show that it had only a tiny fraction of the mass of an atom. The nucleus of an atom contains most of its mass, and the nature of the nucleus was completely unanticipated.

Another important characteristic of quantum mechanics was also beginning to emerge. All electrons are identical to one another. The charge and mass of electrons are not average values; rather, they are unique values that all electrons have. This is true of other fundamental entities at the submicroscopic level. All protons are identical to one another, and so on.

The Nucleus

Here, we examine the first direct evidence of the size and mass of the nucleus. In later chapters, we will examine many other aspects of nuclear physics, but the basic information on nuclear size and mass is so important to understanding the atom that we consider it here.

Nuclear radioactivity was discovered in 1896, and it was soon the subject of intense study by a number of the best scientists in the world. Among them was New Zealander Lord Ernest Rutherford, who made numerous fundamental discoveries and earned the title of "father of nuclear physics." Born in Nelson, Rutherford did his postgraduate studies at the Cavendish Laboratories in England before taking up a position at McGill University in Canada where he did the work that earned him a Nobel Prize in Chemistry in 1908. In the area of atomic and nuclear physics, there is much overlap between chemistry and physics, with physics providing the fundamental enabling theories. He returned to England in later years and had six future Nobel Prize winners as students. Rutherford used nuclear radiation to directly examine the size and mass of the atomic nucleus. The experiment he devised is shown in Figure 7. A radioactive source that emits alpha radiation was placed in a lead container with a hole in one side to produce a beam of alpha particles, which are a type of ionizing radiation ejected by the nuclei of a radioactive source. A thin gold foil was placed in the beam, and the scattering of the alpha particles was observed by the glow they caused when they struck a phosphor screen.

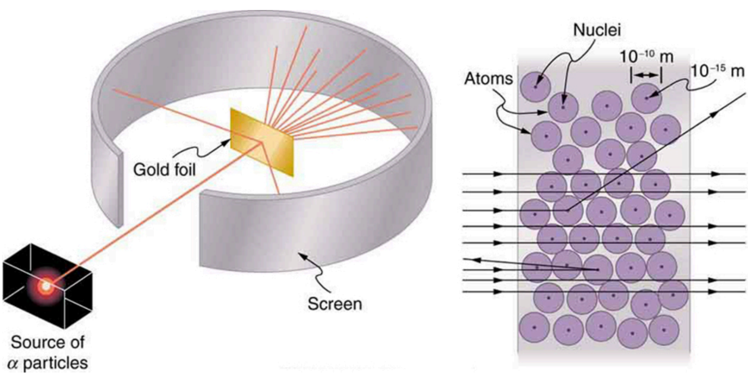


Figure 5. Rutherford's experiment gave direct evidence for the size and mass of the nucleus by scattering alpha particles from a thin gold foil. Alpha particles with energies of about 5MeV are emitted from a radioactive source (which is a small metal container in which a specific amount of a radioactive material is sealed), are collimated into a beam, and fall upon the foil. The number of particles that penetrate the foil or scatter to various angles indicates that gold nuclei are very small and contain nearly all of the gold atom's mass. This is particularly indicated by the alpha particles that scatter to very large angles, much like a soccer ball bouncing off a goalie's head.

Alpha particles were known to be the doubly charged positive nuclei of helium atoms that had kinetic energies on the order of 5 MeV when emitted in nuclear decay, which is the disintegration of the nucleus of an unstable nuclide by the spontaneous emission of charged particles. These particles interact with matter mostly via the Coulomb force, and the manner in which they scatter from nuclei can reveal nuclear size and mass. This is analogous to observing how a bowling ball is scattered by an object you cannot see directly. Because the alpha particle's energy is so large compared with the typical energies associated with atoms (MeV versus eV), you would expect the alpha particles to simply crash through a thin foil much like a supersonic bowling ball would crash through a few dozen rows of bowling pins. Thomson had envisioned the atom to be a small sphere in which equal amounts of positive and negative charge were distributed evenly. The incident massive alpha particles would suffer only small deflections in such a model. Instead, Rutherford and his collaborators found that alpha particles occasionally were scattered to large angles, some even back in the direction from which they came! Detailed analysis using conservation of momentum and energy—particularly of the small number that came straight back—implied that gold nuclei are very small compared with the size of a gold atom, contain almost all of the atom's mass, and are tightly bound. Since the gold nucleus is several times more massive than the alpha particle, a head-on collision would scatter the alpha particle straight back toward the source. In addition, the smaller the nucleus, the fewer alpha particles that would hit one head on.

Although the results of the experiment were published by his colleagues in 1909, it took Rutherford two years to convince himself of their meaning. Like Thomson before him, Rutherford was reluctant to accept such radical results. Nature on a small scale is so unlike our classical world that even those at the forefront of discovery are sometimes surprised. Rutherford later wrote: "It was almost as incredible as if you fired a 15-inch shell at a piece of tissue paper and it came back and hit you. On consideration, I realized that this scattering backwards . . . [meant] . . . the greatest part of the mass of the atom was concentrated in a tiny nucleus."

In 1911, Rutherford published his analysis together with a proposed model of the atom. The size of the nucleus was determined to be about 10^{-15} m, or 100,000 times smaller than the atom. This implies a huge density, on the order of 10^{15} g/cm^3, vastly unlike any macroscopic matter. Also implied is the existence of previously unknown nuclear forces to counteract the huge repulsive Coulomb forces among the positive charges in the nucleus. Huge forces would also be consistent with the large energies emitted in nuclear radiation.

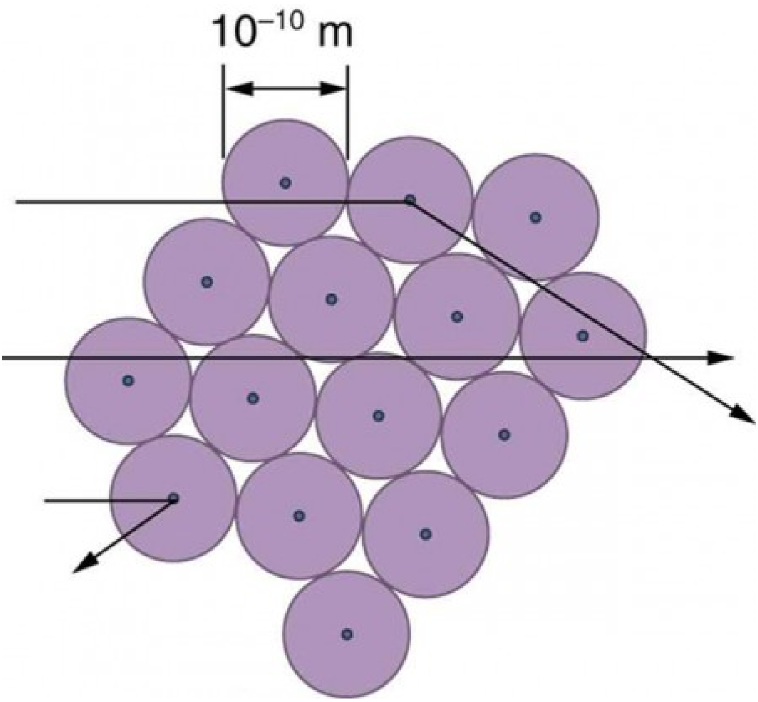


Figure 8. An expanded view of the atoms in the gold foil in Rutherford's experiment. Circles represent the atoms (about 10^{-10} m in diameter), while the dots represent the nuclei (about 10^{-15} m in diameter). To be visible, the dots are much larger than scale. Most alpha particles crash through but are relatively unaffected because of their high energy and the electron's small mass. Some, however, head straight toward a nucleus and are scattered straight back. A detailed analysis gives the size and mass of the nucleus.

The small size of the nucleus also implies that the atom is mostly empty inside. In fact, in Rutherford's experiment, most alphas went straight through the gold foil with very little scattering, since electrons have such small masses and since the atom was mostly empty with nothing for the alpha to hit. There were already hints of this at the time Rutherford performed his experiments, since energetic electrons had been observed to penetrate thin foils more easily than expected. Figure 8 shows a schematic of the atoms in a thin foil with circles representing the size of the atoms (about 10^{-10} m) and dots representing the nuclei. (The dots are not to scale—if they were, you would need a microscope to see them.) Most alpha particles miss the small nuclei and are only slightly scattered by electrons. Occasionally, (about once in 8000 times in Rutherford's experiment), an alpha hits a nucleus head-on and is scattered straight backward.

Based on the size and mass of the nucleus revealed by his experiment, as well as the mass of electrons, Rutherford proposed the *planetary model of the atom*. The planetary model of the atom pictures low-mass electrons orbiting a large-mass nucleus. The sizes of the electron orbits are large compared with the size of the nucleus, with mostly vacuum inside the atom. This picture is analogous to how low-mass planets in our solar system orbit the large-mass Sun at distances large compared with the size of the sun. In the atom, the attractive Coulomb force is analogous to gravitation in the planetary system. (See Figure 9.) Note that a model or mental picture is needed to explain experimental results, since the atom is too small to be directly observed with visible light.

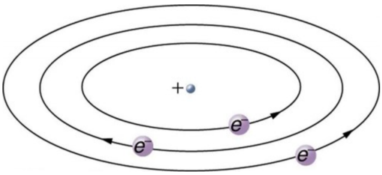
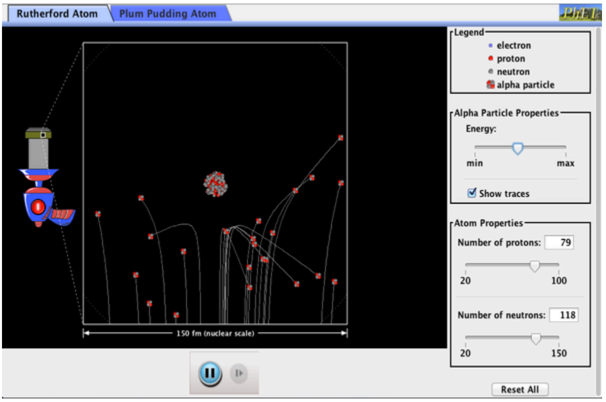


Figure 9. Rutherford's planetary model of the atom incorporates the characteristics of the nucleus, electrons, and the size of the atom. This model was the first to recognize the structure of atoms, in which low-mass electrons orbit a very small, massive nucleus in orbits much larger than the nucleus. The atom is mostly empty and is analogous to our planetary system.

Rutherford's planetary model of the atom was crucial to understanding the characteristics of atoms, and their interactions and energies, as we shall see in the next few sections. Also, it was an indication of how different nature is from the familiar classical world on the small, quantum mechanical scale. The discovery of a substructure to all matter in the form of atoms and molecules was now being taken a step further to reveal a substructure of atoms that was simpler than the 92 elements then known. We have continued to search for deeper substructures, such as those inside the nucleus, with some success. In later chapters, we will follow this quest in the discussion of quarks and other elementary particles, and we will look at the direction the search seems now to be heading.

PhET Explorations: Rutherford Scattering

How did Rutherford figure out the structure of the atom without being able to see it? Simulate the famous experiment in which he disproved the Plum Pudding model of the atom by observing alpha particles bouncing off atoms and determining that they must have a small core.



Click to download the simulation. Run using Java.

Section Summary

- Atoms are composed of negatively charged electrons, first proved to exist in cathode-ray-tube experiments, and a positively charged nucleus.
- All electrons are identical and have a charge-to-mass ratio of $$\frac{q_e}{m_e} = -1.76 \times 10^{11}\ \text{C/kg}.$$
- The positive charge in the nuclei is carried by particles called protons, which have a charge-to-mass ratio of $$\frac{q_p}{m_p} = 9.57 \times 10^{7}\ \text{C/kg}.$$
- Mass of electron, m_e =9.11×10^{-31} kg.
- Mass of proton, $m_p = 1.67 \times 10^{-27}$ kg.

- The planetary model of the atom pictures electrons orbiting the nucleus in the same way that planets orbit the sun.

Conceptual Questions

1. What two pieces of evidence allowed the first calculation of m_e, the mass of the electron? Justify your response.
   1. The ratios $\dfrac{q_e}{m_e}$ and $\dfrac{q_p}{m_p}$.
   2. The values of q_e and E_B.
   3. The ratio $\dfrac{q_e}{m_e}$ and q_e.
2. How do the allowed orbits for electrons in atoms differ from the allowed orbits for planets around the sun? Explain how the correspondence principle applies here.

Problems & Exercises

1. Rutherford found the size of the nucleus to be about 10^{-15} m. This implied a huge density. What would this density be for gold?
2. In Millikan's oil-drop experiment, one looks at a small oil drop held motionless between two plates. Take the voltage between the plates to be 2033 V, and

the plate separation to be 2.00 cm. The oil drop (of density 0.81 g/cm^3) has a diameter of 4.0×10^{-6} m. Find the charge on the drop, in terms of electron units.

3. (a) An aspiring physicist wants to build a scale model of a hydrogen atom for her science fair project. If the atom is 1.00 m in diameter, how big should she try to make the nucleus? (b) How easy will this be to do?

Glossary

cathode-ray tube: a vacuum tube containing a source of electrons and a screen to view images

planetary model of the atom: the most familiar model or illustration of the structure of the atom

Selected Solutions Problems & Exercises

1. 6×10^{20} kg/m^3

3. (a) 10.0 μm; (b) It isn't hard to make one of approximately this size. It would be harder to make it exactly 10.0 μm.

108. Bohr's Theory of the Hydrogen Atom

Learning Objectives

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

- Describe the mysteries of atomic spectra.
- Explain Bohr's theory of the hydrogen atom.
- Explain Bohr's planetary model of the atom.
- Illustrate energy state using the energy-level diagram.
- Describe the triumphs and limits of Bohr's theory.

The great Danish physicist Niels Bohr (1885–1962) made immediate use of Rutherford's planetary model of the atom. (Figure 1). Bohr became convinced of its validity and spent part of 1912 at Rutherford's laboratory. In 1913, after returning to Copenhagen, he began publishing his theory of the simplest atom, hydrogen, based on the planetary model of the atom. For decades, many questions had been asked about atomic characteristics. From their sizes to their spectra, much was known about atoms, but little had been explained in terms of the laws of physics. Bohr's theory explained the atomic spectrum of hydrogen and established new and broadly applicable principles in quantum mechanics.

Figure 1. Niels Bohr, Danish physicist, used the planetary model of the atom to explain the atomic spectrum and size of the hydrogen atom. His many contributions to the development of atomic physics and quantum mechanics, his personal influence on many students and colleagues, and his personal integrity, especially in the face of Nazi oppression, earned him a prominent place in history. (credit: Unknown Author, via Wikimedia Commons)

Mysteries of Atomic Spectra

As noted in Quantization of Energy, the energies of some small systems are quantized. Atomic and molecular emission and absorption spectra have been known for over a century to be discrete (or quantized). (See Figure 2.) Maxwell and others had realized that there must be a connection between the spectrum of an atom and its structure, something like the resonant frequencies of musical instruments. But, in spite of years of efforts by many great minds, no one had a workable theory. (It was a running joke that any theory of atomic and molecular spectra could be destroyed by throwing a book of data at it, so complex were the spectra.) Following Einstein's proposal of photons with quantized energies directly proportional to their wavelengths, it became even more evident that electrons in atoms can exist only in discrete orbits.

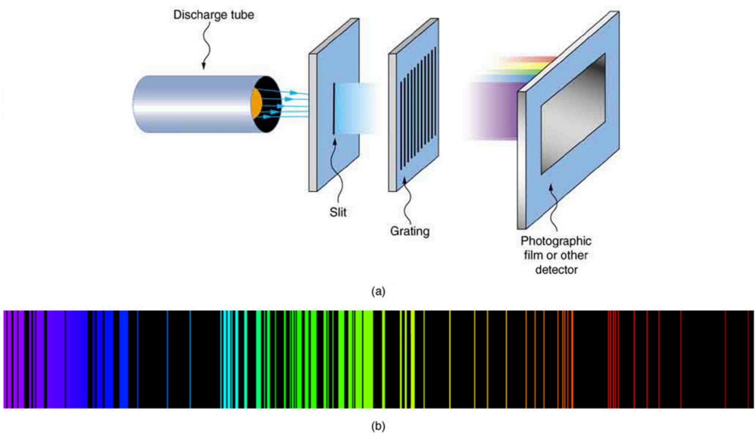


Figure 2. Part (a) shows, from left to right, a discharge tube, slit, and diffraction grating producing a line spectrum. Part (b) shows the emission line spectrum for iron. The discrete lines imply quantized energy states for the atoms that produce them. The line spectrum for each element is unique, providing a powerful and much used analytical tool, and many line spectra were well known for many years before they could be explained with physics. (credit for (b): Yttrium91, Wikimedia Commons)

In some cases, it had been possible to devise formulas that described the emission spectra. As you might expect, the simplest atom—hydrogen, with its single electron—has a relatively simple spectrum. The hydrogen spectrum had been observed in the infrared (IR), visible, and ultraviolet (UV), and several series of spectral lines had been observed. (See Figure 3.) These series are named after early researchers who studied them in particular depth.

The observed *hydrogen-spectrum wavelengths* can be calculated using the following formula:

$$\frac{1}{\lambda} = R\left(\frac{1}{n_{\mathrm{f}}^2} - \frac{1}{n_{\mathrm{i}}^2}\right),$$

where λ is the wavelength of the emitted EM radiation and R is the *Rydberg constant*, determined by the experiment to be $R = 1.097 \times 10^7 \,/\, \mathrm{m}$ (or m^{-1}).

The constant n_f is a positive integer associated with a specific series. For the Lyman series, $n_f = 1$; for the Balmer series, $n_f = 2$; for the Paschen series, $n_f = 3$; and so on. The Lyman series is entirely in the UV, while part of the Balmer series is visible with the remainder UV. The Paschen series and all the rest are entirely IR. There are apparently an unlimited number of series, although they lie progressively farther into the infrared and become difficult to observe as n_f increases. The constant n_i is a positive integer, but it must be greater than n_f. Thus, for the Balmer series, $n_f = 2$ and $n_i = 3, 4, 5, 6, ...$ Note that n_i can approach infinity. While the formula in the wavelengths equation was just a recipe designed to fit data and was not based on physical principles, it did imply a deeper meaning. Balmer first devised the formula for his series alone, and it was later found to describe all the other series by using different values of n_f. Bohr was the first to comprehend the deeper meaning. Again, we see the interplay between experiment and theory in physics. Experimentally, the spectra were well established, an equation was found to fit the experimental data, but the theoretical foundation was missing.

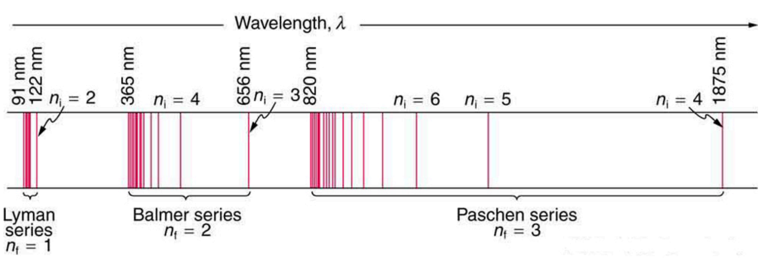


Figure 3. A schematic of the hydrogen spectrum shows several series named for those who contributed most to their determination. Part of the Balmer series is in the visible spectrum, while the Lyman series is entirely in the UV, and the Paschen series and others are in the IR. Values of n_f *and* n_i *are shown for some of the lines.*

Example 1. Calculating Wave Interference of a Hydrogen Line

What is the distance between the slits of a grating that produces a first-order maximum for the second Balmer line at an angle of 15º?

Strategy and Concept

For an Integrated Concept problem, we must first identify the physical principles involved. In this example, we need to know two things:

1. the wavelength of light
2. the conditions for an interference maximum for the pattern from a double slit

Part 1 deals with a topic of the present chapter, while Part 2 considers the wave interference material of Wave Optics.

Solution for Part 1

Hydrogen spectrum wavelength. The Balmer series requires that $n_f = 2$. The first line in the series is taken to be for $n_i = 3$, and so the second would have $n_i = 4$.

The calculation is a straightforward application of the wavelength equation. Entering the determined values for n_f and n_i yields

$$\begin{aligned}\frac{1}{\lambda} &= R\left(\frac{1}{n_{\mathrm{f}}^2} - \frac{1}{n_{\mathrm{i}}^2}\right)\\ &= \left(1.097 \times 10^7 \text{ m}^-1\right)\left(\frac{1}{2^2} - \frac{1}{4^2}\right)\\ &= 2.057 \times 10^6 \text{ m}^{-1}\end{aligned}$$

Inverting to find λ gives

$$\begin{aligned}\lambda &= \frac{1}{2.057 \times 10^6 \text{ m}^-1} = 486 \times 10^{-9} \text{ m}\\ &= 486 \text{ nm}\end{aligned}$$

Discussion for Part 1

This is indeed the experimentally observed wavelength, corresponding to the second (blue-green) line in the Balmer series. More impressive is the fact that the same simple recipe predicts *all* of the hydrogen spectrum lines, including new ones observed in subsequent experiments. What is nature telling us?

Solution for Part 2

Double-slit interference (Wave Optics). To obtain constructive interference for a double slit, the path length difference from two slits must be an integral multiple of the wavelength. This condition was expressed by the equation $d \sin \theta = m\lambda$, where d is the distance between slits and θ is the angle from the original direction of the beam. The number m is the order of the interference; m=1 in this example. Solving for d and entering known values yields

$$d = \frac{(1)\,(486\text{ nm})}{\sin 15^{\circ}} = 1.88 \times 10^{-6}\text{ m}$$

Discussion for Part 2

This number is similar to those used in the interference examples of Introduction to Quantum Physics (and is close to the spacing between slits in commonly used diffraction glasses).

Bohr's Solution for Hydrogen

Bohr was able to derive the formula for the hydrogen spectrum using basic physics, the planetary model of the atom, and some very important new proposals. His first proposal is that only certain orbits are allowed: we say that *the orbits of electrons in atoms are quantized*. Each orbit has a different energy, and electrons can move to a higher orbit by absorbing energy and drop to a lower orbit by emitting energy. If the orbits are quantized, the amount of energy absorbed or emitted is also quantized, producing discrete spectra. Photon absorption and emission are among the primary methods of transferring energy into and out of atoms. The energies of the photons are quantized, and their energy is explained as being equal to the change in energy of the electron when it moves from one orbit to another. In equation form, this is $\Delta E = hf = E_i - E_f$.

Here, ΔE is the change in energy between the initial and final orbits, and *hf* is the energy of the absorbed or emitted photon. It is quite logical (that is, expected from our everyday experience) that energy is involved in changing orbits. A blast of energy is required for the space shuttle, for example, to climb to a higher orbit. What is not expected is that atomic orbits should be quantized. This is not observed for satellites or planets, which can have any orbit given the proper energy. (See Figure 4.)

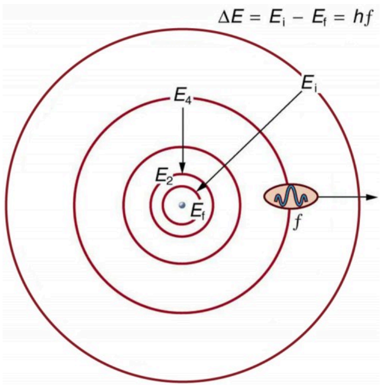


Figure 4. The planetary model of the atom, as modified by Bohr, has the orbits of the electrons quantized. Only certain orbits are allowed, explaining why atomic spectra are discrete (quantized). The energy carried away from an atom by a photon comes from the electron dropping from one allowed orbit to another and is thus quantized. This is likewise true for atomic absorption of photons.

Figure 5 shows an *energy-level diagram*, a convenient way to display energy states. In the present discussion, we take these to be the allowed energy levels of the electron. Energy is plotted vertically with the lowest or ground state at the bottom and with excited states above. Given the energies of the lines in an atomic spectrum, it is possible (although sometimes very difficult) to determine the energy levels of an atom. Energy-level diagrams are used for many systems, including molecules and nuclei. A theory of the atom or any other system must predict its energies based on the physics of the system.

Bohr was clever enough to find a way to calculate the electron orbital energies in hydrogen. This was an important first step that has been improved upon, but it is well worth repeating here, because it does correctly describe many characteristics of hydrogen. Assuming circular orbits, Bohr proposed that the **angular momentum L of an electron in its orbit is quantized**, that is, it has only specific, discrete values. The value for L is given by the formula

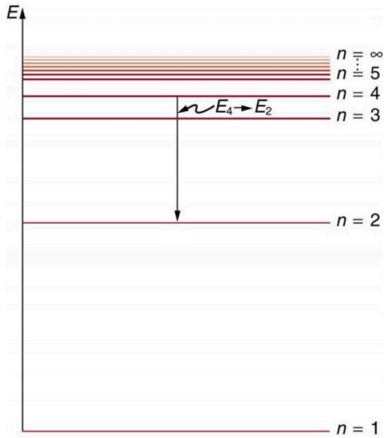


Figure 5. An energy-level diagram plots energy vertically and is useful in visualizing the energy states of a system and the transitions between them. This diagram is for the hydrogen-atom electrons, showing a transition between two orbits having energies E_4 and E_2.

$L = m_e v r_n = n\frac{h}{2\pi}(n = 1, 2, 3, \ldots)$, where L is the angular momentum, m_e is the electron's mass, r_n is the radius of the n th orbit, and h is Planck's constant. Note that angular momentum is $L = I\omega$. For a small object at a radius r, $I = mr^2$ and $\omega = \frac{v}{r}$, so that $L = \left(mr^2\right)\frac{v}{r} = mvr$. Quantization says that this value of mvr can only be equal to $\frac{h}{2}, \frac{2h}{2}, \frac{3h}{2}$, etc. At the time, Bohr himself did not know why angular momentum should be quantized, but using this assumption he was able to calculate the energies in the hydrogen spectrum, something no one else had done at the time.

From Bohr's assumptions, we will now derive a number of important properties of the hydrogen atom from the classical physics we have covered in the text. We start by noting the centripetal force causing the electron to follow a circular path is supplied by the Coulomb force. To be more general, we note that this analysis is valid for any single-electron atom. So, if a nucleus has Z protons (Z = 1 for hydrogen, 2 for helium, etc.) and only one electron, that atom is called a *hydrogen-like atom*. The spectra of hydrogen-like ions are similar to hydrogen, but shifted to higher energy by the greater attractive force between the electron and nucleus. The magnitude of the centripetal force is $\frac{m_e v^2}{r_n}$, while the Coulomb force is $k\frac{(Zq_e)(q_e)}{r_n^2}$. The tacit assumption here is that the nucleus is more massive than the stationary electron, and the electron orbits about it. This is consistent with the planetary model of the atom. Equating these,

$$k\frac{Zq_e^2}{r_n^2} = \frac{m_e v^2}{r_n} \text{ (Coulomb = centripetal)}.$$

Angular momentum quantization is stated in an earlier equation. We solve that equation for v, substitute it into the above, and rearrange the expression to obtain the radius of the orbit. This yields:

$$r_n = \frac{n^2}{Z}a_{\mathrm{B}}, \text{ for allowed orbits } (n = 1, 2, 3\ldots),$$

where a_{B} is defined to be the *Bohr radius*, since for the lowest orbit (n = 1) and for hydrogen (Z = 1), $r_1 = a_{\mathrm{B}}$. It is left for this chapter's Problems and Exercises to show that the Bohr radius is

$$a_{\mathrm{B}} = \frac{h^2}{4\pi^2 m_e k q_e^2} = 0.529 \times 10^{-10} \text{ m}.$$

These last two equations can be used to calculate the **radii of the allowed (quantized) electron orbits in any hydrogen-like atom**. It is impressive that the formula gives the correct size of hydrogen, which is measured experimentally to be very close to the Bohr radius. The earlier equation also tells us that the orbital radius is proportional to n^2, as illustrated in Figure 6.

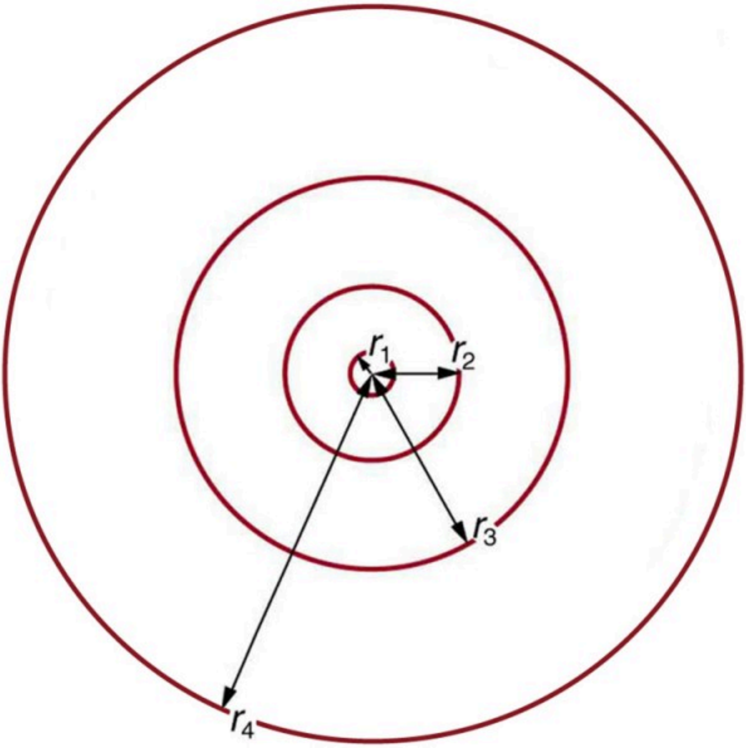


Figure 6. The allowed electron orbits in hydrogen have the radii shown. These radii were first calculated by Bohr and are given by the equation [latex]r_n=\frac{n^2}{Z}a_{\text{B}}\\[/latex]. The lowest orbit has the experimentally verified diameter of a hydrogen atom.

To get the electron orbital energies, we start by noting that the electron energy is the sum of its kinetic and potential energy: E_n = KE + PE.

Kinetic energy is the familiar $KE=\frac{1}{2}m_ev^2$, assuming the electron is not moving at relativistic speeds. Potential energy for the electron is electrical, or PE = q_eV, where V is the potential due to the nucleus, which looks like a point charge. The nucleus has a positive charge Zq_e ; thus, $V = \frac{kZq_e}{r_n}$, recalling an earlier equation for the potential due to a point charge. Since the electron's charge is negative, we see that $PE = -\frac{kZq_e}{r_n}$. Entering the expressions for *KE* and *PE*, we find

$$E_n = \frac{1}{2}m_e v^2 - k\frac{Zq_e^2}{r_n}.$$

Now we substitute r_n and v from earlier equations into the above expression for energy. Algebraic manipulation yields

$$E_n = -\frac{Z^2}{n^2}E_0 \; (n = 1, 2, 3, \ldots)$$

for the orbital *energies of hydrogen-like atoms*. Here, E_0 is the **ground-state energy** (n = 1) for hydrogen (Z = 1) and is given by

$$E_0 = \frac{2\pi q_e^4 m_e k^2}{h^2} = 13.6 \text{ eV}$$

Thus, for hydrogen,

$$E_n = -\frac{13.6 \text{ eV}}{n^2}(n = 1, 2, 3 \ldots)$$

Figure 7 shows an energy-level diagram for hydrogen that also illustrates how the various spectral series for hydrogen are related to transitions between energy levels.

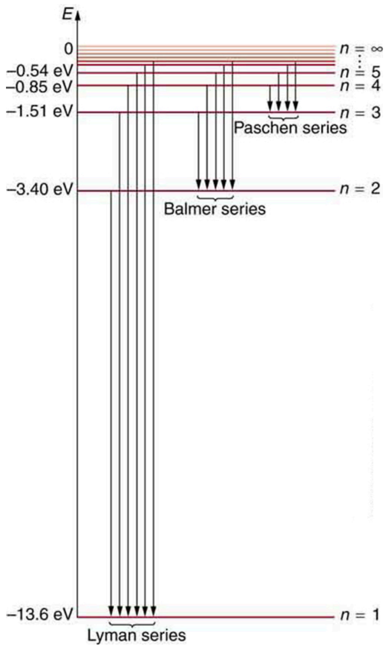


Figure 7. Energy-level diagram for hydrogen showing the Lyman, Balmer, and Paschen series of transitions. The orbital energies are calculated using the above equation, first derived by Bohr.

Electron total energies are negative, since the electron is bound to the nucleus, analogous to being in a hole without enough kinetic energy to escape. As n approaches infinity, the total energy becomes zero. This corresponds to a free electron with no kinetic energy, since r_n gets very large for large n, and the electric potential energy thus becomes zero. Thus, 13.6 eV is needed to ionize hydrogen (to go from –13.6 eV to 0, or unbound), an experimentally verified number. Given more energy, the electron becomes unbound with some kinetic energy. For example, giving 15.0 eV to an electron in the ground state of hydrogen strips it from the atom and leaves it with 1.4 eV of kinetic energy.

Finally, let us consider the energy of a photon emitted in a downward transition, given by the equation to be $\Delta E = hf = E_i - E_f$.

Substituting $E_n = (-13.6\text{ eV}/n^2)$, we see that

$$hf = (13.6\text{ eV})\left(\frac{1}{n_\text{f}^2} - \frac{1}{n_\text{i}^2}\right)$$

Dividing both sides of this equation by hc gives an expression for $\frac{1}{\lambda}$:

$$\frac{hf}{hc} = \frac{f}{c} = \frac{1}{\lambda} = \frac{(13.6\text{ eV})}{hc}\left(\frac{1}{n_{\text{f}}^2} - \frac{1}{n_{\text{i}}^2}\right)$$

It can be shown that

$$\left(\frac{13.6\text{ eV}}{hc}\right) = \frac{(13.6\text{ eV})\left(1.602\times10^{-19}\text{ J/eV}\right)}{\left(6.626\times10^{-34}\text{ J}\cdot\text{s}\right)\left(2.998\times10^{8}\text{ m/s}\right)} = 1.097\times10^{7}\text{ m}^{-1} = R$$

is the *Rydberg constant*. Thus, we have used Bohr's assumptions to derive the formula first proposed by Balmer years earlier as a recipe to fit experimental data.

$$\frac{1}{\lambda} = R\left(\frac{1}{n_{\text{f}}^2} - \frac{1}{n_{\text{i}}^2}\right)$$

We see that Bohr's theory of the hydrogen atom answers the question as to why this previously known formula describes the hydrogen spectrum. It is because the energy levels are proportional to $\frac{1}{n^2}$, where n is a non-negative integer. A downward transition releases energy, and so n_{i} must be greater than n_{f}. The various series are those where the transitions end on a certain level. For the Lyman series, $n_{\text{f}} = 1$—that is, all the transitions end in the ground state (see also Figure 7). For the Balmer series, $n_{\text{f}} = 2$, or all the transitions end in the first excited state; and so on. What was once a recipe is now based in physics, and something new is emerging—angular momentum is quantized.

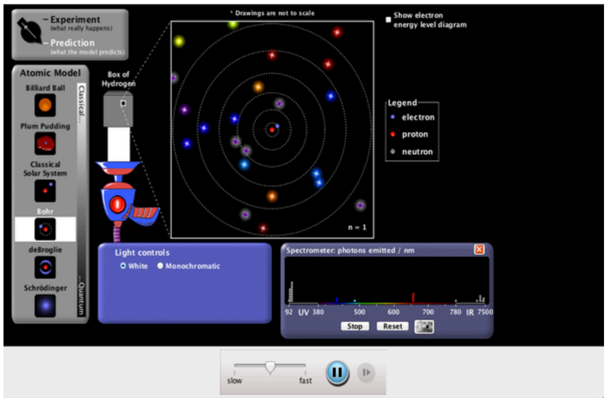
Triumphs and Limits of the Bohr Theory

Bohr did what no one had been able to do before. Not only did he explain the spectrum of hydrogen, he correctly calculated the size of the atom from basic physics. Some of his ideas are broadly applicable. Electron orbital energies are quantized in all atoms and molecules. Angular momentum is quantized. The electrons do not spiral into the nucleus, as expected classically (accelerated charges radiate, so that the electron orbits classically would decay quickly, and the electrons would sit on the nucleus—matter would collapse). These are major triumphs.

But there are limits to Bohr's theory. It cannot be applied to multielectron atoms, even one as simple as a two-electron helium atom. Bohr's model is what we call *semiclassical*. The orbits are quantized (nonclassical) but are assumed to be simple circular paths (classical). As quantum mechanics was developed, it became clear that there are no well-defined orbits; rather, there are clouds of probability. Bohr's theory also did not explain that some spectral lines are doublets (split into two) when examined closely. We shall examine many of these aspects of quantum mechanics in more detail, but it should be kept in mind that Bohr did not fail. Rather, he made very important steps along the path to greater knowledge and laid the foundation for all of atomic physics that has since evolved.

PhET Explorations: Models of the Hydrogen Atom

How did scientists figure out the structure of atoms without looking at them? Try out different models by shooting light at the atom. Check how the prediction of the model matches the experimental results.



Click to download the simulation. Run using Java.

Section Summary

- The planetary model of the atom pictures electrons orbiting the nucleus in the way that planets orbit the sun. Bohr used the planetary model to develop the first reasonable theory of hydrogen, the simplest atom. Atomic and molecular spectra are quantized, with hydrogen spectrum wavelengths given by the formula
$$\frac{1}{\lambda} = R\left(\frac{1}{n_\mathrm{f}^2} - \frac{1}{n_\mathrm{i}^2}\right),$$
where λ is the wavelength of the emitted EM radiation and R is the Rydberg constant, which has the value $R = 1.097 \times 10^7\ \mathrm{m}^{-1}$.
- The constants n_i and n_f are positive integers, and n_i must be

greater than n_f.

- Bohr correctly proposed that the energy and radii of the orbits of electrons in atoms are quantized, with energy for transitions between orbits given by $\Delta E = hf = E_i - E_f$, where ΔE is the change in energy between the initial and final orbits and hf is the energy of an absorbed or emitted photon. It is useful to plot orbital energies on a vertical graph called an energy-level diagram.
- Bohr proposed that the allowed orbits are circular and must have quantized orbital angular momentum given by $L = m_e \mathrm{vr}_n = n\frac{h}{2\pi}(n = 1, 2, 3\ldots)$, where L is the angular momentum, r_n is the radius of the nth orbit, and h is Planck's constant. For all one-electron (hydrogen-like) atoms, the radius of an orbit is given by
$$r_n = \frac{n^2}{Z}a_{\mathrm{B}}\,(\text{allowed orbits } n = 1, 2, 3, \ldots)$$
, Z is the atomic number of an element (the number of electrons is has when neutral) and a_B is defined to be the Bohr radius, which is
$$a_{\mathrm{B}} = \frac{h^2}{4\pi^2 m_e \mathrm{kq}_e^2} = 0.529 \times 10^{-10}\ \mathrm{m}.$$
- Furthermore, the energies of hydrogen-like atoms are given by $E_n = -\frac{Z^2}{n^2}E_0\,(n = 1, 2, 3\ldots)$, where E_0 is the ground-state energy and is given by
$$E_0 = \frac{2\pi^2 q_e^4 m_e k^2}{h^2} = 13.6\ \mathrm{eV}.$$
Thus, for hydrogen,
$$E_n = -\frac{13.6\ \mathrm{eV}}{n^2}(n, =, 1, 2, 3\ldots).$$
- The Bohr Theory gives accurate values for the energy levels in hydrogen-like atoms, but it has been improved upon in several respects.

Conceptual Questions

1. How do the allowed orbits for electrons in atoms differ from the allowed orbits for planets around the sun? Explain how the correspondence principle applies here.
2. Explain how Bohr's rule for the quantization of electron orbital angular momentum differs from the actual rule.
3. What is a hydrogen-like atom, and how are the energies and radii of its electron orbits related to those in hydrogen?

Problems & Exercises

1. By calculating its wavelength, show that the first line in the Lyman series is UV radiation.
2. Find the wavelength of the third line in the Lyman series, and identify the type of EM radiation.
3. Look up the values of the quantities in $a_{\mathrm{B}} = \dfrac{h^2}{4\pi^2 m_e \mathrm{k} \mathrm{q}_e^2}$, and verify that the Bohr radius a_{B} is 0.529×10^{-10} m.
4. Verify that the ground state energy E_0 is 13.6 eV by using $E_0 = \dfrac{2\pi^2 q_e^4 m_e k^2}{h^2}$.

5. If a hydrogen atom has its electron in the $n = 4$ state, how much energy in eV is needed to ionize it?
6. A hydrogen atom in an excited state can be ionized with less energy than when it is in its ground state. What is n for a hydrogen atom if 0.850 eV of energy can ionize it?
7. Find the radius of a hydrogen atom in the $n = 2$ state according to Bohr's theory.
8. Show that

$$\frac{(13.6\text{eV})}{hc} = 1.097 \times 10^7\ \text{m} = R$$

(Rydberg's constant), as discussed in the text.
9. What is the smallest-wavelength line in the Balmer series? Is it in the visible part of the spectrum?
10. Show that the entire Paschen series is in the infrared part of the spectrum. To do this, you only need to calculate the shortest wavelength in the series.
11. Do the Balmer and Lyman series overlap? To answer this, calculate the shortest-wavelength Balmer line and the longest-wavelength Lyman line.
12. (a) Which line in the Balmer series is the first one in the UV part of the spectrum? (b) How many Balmer series lines are in the visible part of the spectrum? (c) How many are in the UV?
13. A wavelength of 4.653 μm is observed in a hydrogen spectrum for a transition that ends in the $n_f = 5$ level. What was n_i for the initial level of the electron?
14. A singly ionized helium ion has only one electron and is denoted He^+. What is the ion's radius in the ground state compared to the Bohr radius of hydrogen atom?

15. A beryllium ion with a single electron (denoted Be^{3+}) is in an excited state with radius the same as that of the ground state of hydrogen. (a) What is n for the Be^{3+} ion? (b) How much energy in eV is needed to ionize the ion from this excited state?

16. Atoms can be ionized by thermal collisions, such as at the high temperatures found in the solar corona. One such ion is C^{+5}, a carbon atom with only a single electron. (a) By what factor are the energies of its hydrogen-like levels greater than those of hydrogen? (b) What is the wavelength of the first line in this ion's Paschen series? (c) What type of EM radiation is this?

17. Verify Equations $r_n = \frac{n^2}{Z} a_{\mathrm{B}}$ and

$$a_B = \frac{h^2}{4\pi^2 m_e k q_e^2} = 0.529 \times 10^{-10}\ \mathrm{m}$$

using the approach stated in the text. That is, equate the Coulomb and centripetal forces and then insert an expression for velocity from the condition for angular momentum quantization.

18. The wavelength of the four Balmer series lines for hydrogen are found to be 410.3, 434.2, 486.3, and 656.5 nm. What average percentage difference is found between these wavelength numbers and those predicted by $\frac{1}{\lambda} = R\left(\frac{1}{n_{\mathrm{f}}^2} - \frac{1}{n_{\mathrm{i}}^2}\right)$? It is amazing how well a simple formula (disconnected originally from theory) could duplicate this phenomenon.

Glossary

hydrogen spectrum wavelengths: the wavelengths of visible light from hydrogen; can be calculated by

$$\frac{1}{\lambda}=R\left(\frac{1}{n_{\mathrm{f}}^2}-\frac{1}{n_{\mathrm{i}}^2}\right)$$

Rydberg constant: a physical constant related to the atomic spectra with an established value of $1.097 \times 10^7\ \mathrm{m}^{-1}$

double-slit interference: an experiment in which waves or particles from a single source impinge upon two slits so that the resulting interference pattern may be observed

energy-level diagram: a diagram used to analyze the energy level of electrons in the orbits of an atom

Bohr radius: the mean radius of the orbit of an electron around the nucleus of a hydrogen atom in its ground state

hydrogen-like atom: any atom with only a single electron

energies of hydrogen-like atoms: Bohr formula for energies of electron states in hydrogen-like atoms:

$$E_n=-\frac{Z^2}{n^2}E_0\,(n=1,2,3,\ldots)$$

Selected Solutions to Problems & Exercises

1.

$$\frac{1}{\lambda}=R\left(\frac{1}{n_{\mathrm{f}}^2}-\frac{1}{n_{\mathrm{i}}^2}\right)\Rightarrow\lambda=\frac{1}{R}\left[\frac{(n_{\mathrm{i}}\cdot n_{\mathrm{f}})^2}{n_{\mathrm{i}}^2-n_{\mathrm{f}}^2}\right];n_{\mathrm{i}}=2,n_{\mathrm{f}}=1$$

, so that

$$\lambda = \left(\frac{m}{1.097 \times 10^7}\right)\left[\frac{(2 \times 1)^2}{2^2 - 1^2}\right] = 1.22 \times 10^{-7}\text{m} = 122 \text{ nm}$$

, which is UV radiation.

3.

$$\begin{aligned} a_{\text{B}} &= \frac{h^2}{4\pi^2 m_e \text{kZq}_e^2} \\ &= \frac{\left(6.626\times10^{-34}\text{J}\cdot\text{s}\right)^2}{4\pi^2\left(9.109\times10^{-31}\text{kg}\right)\left(8.988\times10^9\text{N}\cdot\text{m}^2/C^2\right)(1)\left(1.602\times10^{-19}\text{C}\right)^2} \\ &= 0.529 \times 10^{-10}\text{m} \end{aligned}$$

5. 0.850 eV

7. 2.12×10^{-10} m

9. 365 nm; it is in the ultraviolet.

11. No overlap; 365 nm; 122 nm

13. 7

15. (a) 2; (b) 54.4 eV

17. $\frac{\text{kZq}_e^2}{r_n^2} = \frac{m_e V^2}{r_n}$, so that $r_n = \frac{\text{kZq}_e^2}{m_e V^2} = \frac{\text{kZq}_e^2}{m_e}\frac{1}{V^2}$. From the equation $m_e v r_n = n\frac{h}{2\pi}$, we can substitute for the velocity, giving:

$$r_n = \frac{\text{kZq}_e^2}{m_e} \cdot \frac{4\pi^2 m_e^2 r_n^2}{n^2 h^2}$$

so that

$$r_n = \frac{n^2}{Z}\frac{h^2}{4\pi^2 m_e \text{kq}_e^2} = \frac{n^2}{Z}a_{\text{B}},$$

where

$$a_{\mathrm{B}} = \frac{h^2}{4\pi^2 m_e \mathrm{kq}_e^2}.$$

PART XV

RADIOACTIVITY AND NUCLEAR PHYSICS

109. Introduction to Radioactivity and Nuclear Physics

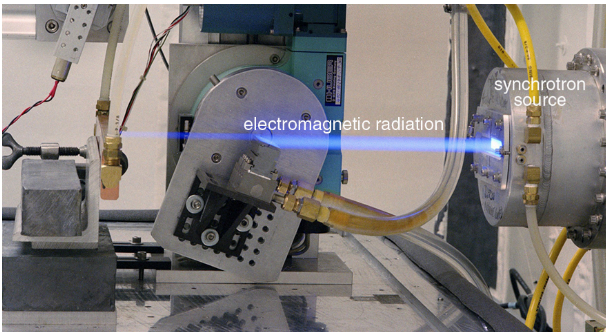


Figure 1. The synchrotron source produces electromagnetic radiation, as evident from the visible glow. (credit: United States Department of Energy, via Wikimedia Commons)

There is an ongoing quest to find substructures of matter. At one time, it was thought that atoms would be the ultimate substructure, but just when the first direct evidence of atoms was obtained, it became clear that they have a substructure and a tiny *nucleus*. The nucleus itself has spectacular characteristics. For example, certain nuclei are unstable, and their decay emits radiations with energies millions of times greater than atomic energies. Some of the mysteries of nature, such as why the core of the earth remains molten and how the sun produces its energy, are explained by nuclear phenomena. The exploration of *radioactivity* and the nucleus revealed fundamental and previously unknown particles, forces, and conservation laws. That exploration has evolved into a search for further underlying structures, such as quarks. In this chapter, the fundamentals of nuclear radioactivity and the nucleus are explored. The following two chapters explore the more important applications of nuclear physics in the field of medicine. We will also explore the basics of what we know about quarks and other substructures smaller than nuclei.

110. Radiation Detection and Detectors

Learning Objectives

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

- Explain the working principle of a Geiger tube.
- Define and discuss radiation detectors.

It is well known that ionizing radiation affects us but does not trigger nerve impulses. Newspapers carry stories about unsuspecting victims of radiation poisoning who fall ill with radiation sickness, such as burns and blood count changes, but who never felt the radiation directly. This makes the detection of radiation by instruments more than an important research tool. This section is a brief overview of radiation detection and some of its applications.

Human Application

The first direct detection of radiation was Becquerel's fogged photographic plate. Photographic film is still the most common detector of ionizing radiation, being used routinely in medical and dental x rays. Nuclear radiation is also captured on film, such as seen in Figure 1. The mechanism for film exposure by ionizing radiation is similar to that by photons. A quantum of energy interacts with the emulsion and alters it chemically, thus exposing the film. The quantum come from an α-particle, β-particle, or photon, provided it has more than the few eV of energy needed to induce the chemical change (as does all ionizing radiation). The process is not 100% efficient, since not all incident radiation interacts and not all interactions produce the chemical change. The amount of film darkening is related to exposure, but the darkening also depends on the type of radiation, so that absorbers and other devices must be used to obtain energy, charge, and particle-identification information.

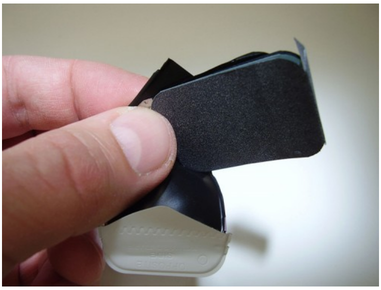


Figure 1. Film badges contain film similar to that used in this dental x-ray film and is sandwiched between various absorbers to determine the penetrating ability of the radiation as well as the amount. (credit: Werneuchen, Wikimedia Commons)

Another very common *radiation detector* is the *Geiger tube*. The clicking and buzzing sound we hear in dramatizations and documentaries, as well as in our own physics labs, is usually an audio output of events detected by a Geiger counter. These relatively inexpensive radiation detectors are based on the simple and sturdy Geiger tube, shown schematically in Figure 2b. A conducting cylinder with a wire along its axis is filled with an insulating gas so that a voltage applied between the cylinder and wire produces almost no current. Ionizing radiation passing through the tube produces free ion pairs that are attracted to the wire and cylinder, forming a current that is detected as a count. The word count implies that there is no information on energy, charge, or type of radiation with a simple Geiger counter. They do not detect every particle, since some radiation can pass through without producing enough ionization to be detected. However, Geiger counters are very useful in producing a prompt output that reveals the existence and relative intensity of ionizing radiation.

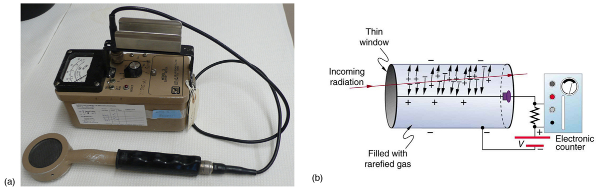


Figure 2. (a) Geiger counters such as this one are used for prompt monitoring of radiation levels, generally giving only relative intensity and not identifying the type or energy of the radiation. (credit: TimVickers, Wikimedia Commons) (b) Voltage applied between the cylinder and wire in a Geiger tube causes ions and electrons produced by radiation passing through the gas-filled cylinder to move towards them. The resulting current is detected and registered as a count.

Another radiation detection method records light produced when radiation interacts with materials. The energy of the radiation is sufficient to excite atoms in a material that may fluoresce, such as the phosphor used by Rutherford's group. Materials called *scintillators* use a more complex collaborative process to convert radiation energy into light. Scintillators may be liquid or solid, and they can be very efficient. Their light output can provide information about the energy, charge, and type of radiation. Scintillator light flashes are very brief in duration, enabling the detection of a huge number of particles in short periods of time. Scintillator detectors are used in a variety of research and diagnostic applications. Among these are the detection by satellite-mounted equipment of the radiation from distant galaxies, the analysis of radiation from a person indicating body burdens, and the detection of exotic particles in accelerator laboratories.

Light from a scintillator is converted into electrical signals by devices such as the *photomultiplier* tube shown schematically in Figure 3. These tubes are based on the photoelectric effect, which is multiplied in stages into a cascade of electrons, hence the name photomultiplier. Light entering the photomultiplier strikes a metal plate, ejecting an electron that is attracted by a positive potential difference to the next plate, giving it enough energy to eject two or more electrons, and so on. The final output current can be made proportional to the energy of the light entering the tube, which is in turn proportional to the energy deposited in the scintillator. Very sophisticated information can be obtained with scintillators, including energy, charge, particle identification, direction of motion, and so on.

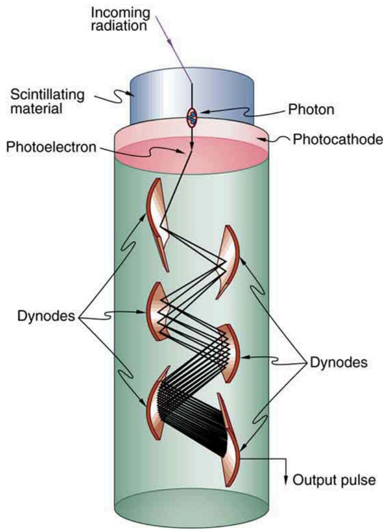
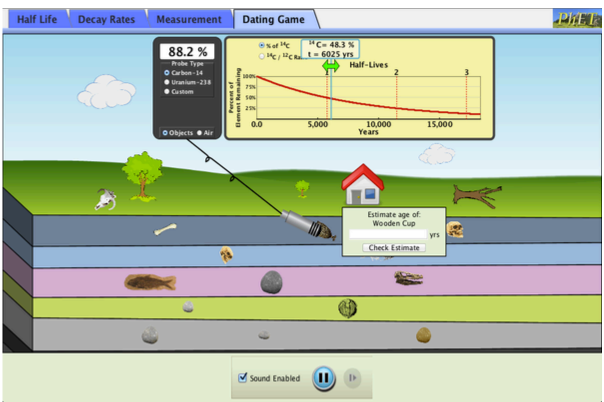


Figure 3. Photomultipliers use the photoelectric effect on the photocathode to convert the light output of a scintillator into an electrical signal. Each successive dynode has a more-positive potential than the last and attracts the ejected electrons, giving them more energy. The number of electrons is thus multiplied at each dynode, resulting in an easily detected output current.

Solid-state radiation detectors convert ionization produced in a semiconductor (like those found in computer chips) directly into an electrical signal. Semiconductors can be constructed that do not conduct current in one particular direction. When a voltage is applied in that direction, current flows only when ionization is produced by radiation, similar to what happens in a Geiger tube. Further, the amount of current in a solid-state detector is closely related to the energy deposited and, since the detector is solid, it can have a high efficiency (since ionizing radiation is stopped in a shorter distance in solids fewer particles escape detection). As with scintillators, very sophisticated information can be obtained from solid-state detectors.

PhET Explorations: Radioactive Dating Game

Learn about different types of radiometric dating, such as carbon dating. Understand how decay and half life work to enable radiometric dating to work. Play a game that tests your ability to match the percentage of the dating element that remains to the age of the object.



Click to download the simulation. Run using Java.

Section Summary

- Radiation detectors are based directly or indirectly upon the ionization created by radiation, as are the effects of radiation on living and inert materials.

Conceptual Questions

1. Is it possible for light emitted by a scintillator to be too low in frequency to be used in a photomultiplier tube? Explain.

Problems & Exercises

1. The energy of 30.0 eV is required to ionize a molecule of the gas inside a Geiger tube, thereby producing an ion pair. Suppose a particle of ionizing radiation deposits 0.500 MeV of energy in this Geiger tube. What maximum number of ion pairs can it create?
2. A particle of ionizing radiation creates 4000 ion pairs in the gas inside a Geiger tube as it passes through. What minimum energy was deposited, if 30.0 eV is required to create each ion pair?
3. (a) Repeat Question 2, and convert the energy to joules or calories. (b) If all of this energy is converted

to thermal energy in the gas, what is its temperature increase, assuming 50.0 cm^3 of ideal gas at 0.250-atm pressure? (The small answer is consistent with the fact that the energy is large on a quantum mechanical scale but small on a macroscopic scale.)

4. Suppose a particle of ionizing radiation deposits 1.0 MeV in the gas of a Geiger tube, all of which goes to creating ion pairs. Each ion pair requires 30.0 eV of energy. (a) The applied voltage sweeps the ions out of the gas in 1.00 μs. What is the current? (b) This current is smaller than the actual current since the applied voltage in the Geiger tube accelerates the separated ions, which then create other ion pairs in subsequent collisions. What is the current if this last effect multiplies the number of ion pairs by 900?

Glossary

Geiger tube: a very common radiation detector that usually gives an audio output

photomultiplier: a device that converts light into electrical signals

radiation detector: a device that is used to detect and track the radiation from a radioactive reaction

scintillators: a radiation detection method that records light produced when radiation interacts with materials

solid-state radiation detectors: semiconductors fabricated to directly convert incident radiation into electrical current

Selected Solutions to Problems & Exercises

1. 1.67×10^{4}

111. Substructure of the Nucleus

Learning Objectives

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

- Define and discuss the nucleus in an atom.
- Define atomic number.
- Define and discuss isotopes.
- Calculate the density of the nucleus.
- Explain nuclear force.

What is inside the nucleus? Why are some nuclei stable while others decay? (See Figure 1.) Why are there different types of decay (α, β and γ)? Why are nuclear decay energies so large? Pursuing natural questions like these has led to far more fundamental discoveries than you might imagine.

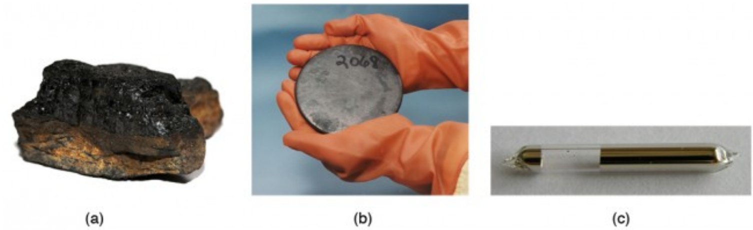


Figure 1. Why is most of the carbon in this coal stable (a), while the uranium in the disk (b) slowly decays over billions of years? Why is cesium in this ampule (c) even less stable than the uranium, decaying in far less than 1/1,000,000 the time? What is the reason uranium and cesium undergo different types of decay (α and β, respectively)? (credits: (a) Bresson Thomas, Wikimedia Commons; (b) U.S. Department of Energy; (c) Tomihahndorf, Wikimedia Commons)

We have already identified *protons* as the particles that carry positive charge in the nuclei. However, there are actually *two* types of particles in the nuclei—the *proton* and the *neutron*, referred to collectively as *nucleons*, the constituents of nuclei. As its name implies, the *neutron* is a neutral particle ($q = 0$) that has nearly the same mass and intrinsic spin as the proton. Table 1 compares the masses of protons, neutrons, and electrons. Note how close the proton and neutron masses are, but the neutron is slightly more massive once you look past the third digit. Both nucleons are much more massive than an electron. In fact, $m_p = 1836m_e$ (as noted in Medical Applications of Nuclear Physics and $m_n = 1839m_e$.

Table 1 also gives masses in terms of mass units that are more convenient than kilograms on the atomic and nuclear scale. The first of these is the *unified atomic mass unit* (u), defined as $1\text{ u} = 1.6605 \times 10^{-27}\text{ kg}$.

This unit is defined so that a neutral carbon ^{12}C atom has a mass of exactly 12 u. Masses are also expressed in units of MeV/c^2. These units are very convenient when considering the conversion of mass into energy (and vice versa), as is so prominent in nuclear processes. Using $E=mc^2$ and units of m in MeV/c^2, we find that c^2 cancels and E comes out conveniently in MeV. For example, if the rest mass of a proton is converted entirely into energy, then $E = mc^2 = (938.27 \text{ MeV}/c^2)c^2 = 938.27 \text{ MeV}$.

It is useful to note that 1 u of mass converted to energy produces 931.5 MeV, or $1 \text{ u} = 931.5 \text{ MeV}/c^2$.

All properties of a nucleus are determined by the number of protons and neutrons it has. A specific combination of protons and neutrons is called a *nuclide* and is a unique nucleus. The following notation is used to represent a particular nuclide: ${}^{A}_{Z}\mathrm{X}_{N}$, where the symbols A, X, Z , and N are defined as follows:

> The *number of protons in a nucleus* is the *atomic number* Z, as defined in Medical Applications of Nuclear Physics. X is the *symbol for the element*, such as Ca for calcium. However, once Z is known, the element is known; hence, Z and X are redundant. For example, Z = 20 is always calcium, and calcium always has Z = 20. N is the *number of neutrons* in a nucleus. In the notation for a nuclide, the subscript N is usually omitted. The symbol A is defined as the number of nucleons or the *total number of protons and neutrons*, A = N + Z, where A is also called the *mass number*.

This name for A is logical; the mass of an atom is nearly equal to the mass of its nucleus, since electrons have so little mass. The mass of the nucleus turns out to be nearly equal to the sum of the masses of the protons and neutrons in it, which is proportional to A. In this context, it is particularly convenient to express masses in units of u. Both protons and neutrons have masses close to 1 u, and so the mass of an atom is close to A u. For example, in an oxygen nucleus with eight protons and eight neutrons, A = 16, and its mass is 16 u. As noticed, the unified atomic mass unit is defined so that a neutral carbon atom (actually a ^{12}C atom) has a mass of *exactly* 12 u. Carbon was chosen as the standard, partly because of its importance in organic chemistry (see Appendix A).

Table 1. Masses of the Proton, Neutron, and Electron				
Particle	**Symbol**	**kg**	**u**	**MeVc^2**
Proton	p	1.67262×10^{-27}	1.007276	938.27
Neutron	n	1.67493×10^{-27}	1.008665	939.57
Electron	e	9.1094×10^{-31}	0.00054858	0.511

Let us look at a few examples of nuclides expressed in the $^{A}_{Z}\mathrm{X}_{N}$ notation. The nucleus of the simplest atom, hydrogen, is a single proton, or $^{1}_{1}\mathrm{H}$ (the zero for no neutrons is often omitted). To check this symbol, refer to the periodic table—you see that the atomic number Z of hydrogen is 1. Since you are given that there are no neutrons, the mass number A is also 1. Suppose you are told that the helium nucleus or α particle has two protons and two neutrons. You can then see that it is written $^{4}_{2}\mathrm{He}_{2}$. There is a scarce form of hydrogen found in nature called deuterium; its nucleus has one proton and one neutron and, hence, twice the mass of common hydrogen. The symbol for deuterium is, thus, $^{2}_{1}\mathrm{H}_{1}$ (sometimes D is used, as for deuterated water D_2 O). An even rarer—and radioactive—form of hydrogen is called tritium, since it has a single proton and two neutrons, and it is written $^{3}_{1}\mathrm{H}_{2}$. These three varieties of hydrogen have nearly identical chemistries, but the nuclei differ greatly in mass, stability, and other characteristics. Nuclei (such as those of hydrogen) having the same Z and different N s are defined to be *isotopes* of the same element.

There is some redundancy in the symbols A, X, Z, and N . If the element X is known, then Z can be found in a periodic table and is always the same for a given element. If both A and X are known, then N can also be determined (first find Z; then, N = A − Z). Thus the simpler notation for nuclides is $^{A}\mathrm{X}$, which is sufficient and is most commonly used. For example, in this simpler notation, the three isotopes of hydrogen are $^{1}\mathrm{H}$, $^{2}\mathrm{H}$, and $^{3}\mathrm{H}$, while the α particle is $^{4}\mathrm{He}$. We read this backward, saying helium-4 for $^{4}\mathrm{He}$, or uranium-238 for $^{238}\mathrm{U}$. So for $^{238}\mathrm{U}$, should we need to know, we can determine that

Z = 92 for uranium from the periodic table, and, thus, N = 238 − 92 = 146.

A variety of experiments indicate that a nucleus behaves something like a tightly packed ball of nucleons, as illustrated in Figure 2. These nucleons have large kinetic energies and, thus, move rapidly in very close contact. Nucleons can be separated by a large force, such as in a collision with another nucleus, but resist strongly being pushed closer together. The most compelling evidence that nucleons are closely packed in a nucleus is that the *radius of a nucleus*, r, is found to be given approximately by $r = r_0 A^{1/3}$, where $r_0 = 1.2$ fm and A is the mass number of the nucleus. Note that $r^3 \propto A$. Since many nuclei are spherical, and the volume of a sphere is $V = (4/3)\pi r^3$, we see that $V \propto A$ —that is, the volume of a nucleus is proportional to the number of nucleons in it. This is what would happen if you pack nucleons so closely that there is no empty space between them.

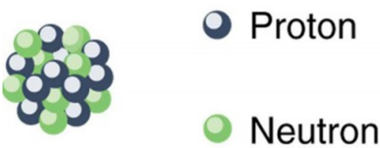


Figure 2. A model of the nucleus.

Nucleons are held together by nuclear forces and resist both being pulled apart and pushed inside one another. The volume of the nucleus is the sum of the volumes of the nucleons in it, here shown in different colors to represent protons and neutrons.

Example 1. How Small and Dense Is a Nucleus?

1. Find the radius of an iron-56 nucleus.
2. Find its approximate density in kg/m^3, approximating the mass of ^{56}Fe to be 56 u.

Strategy and Concept

1. Finding the radius of ^{56}Fe is a straightforward application of $r = r_0 A^{1/3}$, given A = 56.
2. To find the approximate density, we assume the nucleus is spherical (this one actually is), calculate its volume using the radius found in Part 1, and then find its density from $\rho = m/V$. Finally, we will need to convert density from units of u/fm^3 to kg/m^3.

Solution for Part 1

The radius of a nucleus is given by $r = r_0 A^{1/3}$. Substituting the values for r_0 and A yields

$$\begin{aligned} r &= (1.2 \text{ fm})(56)^{1/3} = (1.2 \text{ fm})\,(3.83) \\ &= 4.6 \text{ fm} \end{aligned}$$

Solution for Part 2

Density is defined to be $\rho = m/V$, which for a sphere of radius r is

$$\rho = \frac{m}{V} = \frac{m}{(4/3)\pi r^3}$$

Substituting known values gives

$$\rho = \frac{56\ \text{u}}{(1.33)(3.14)(4.6\ \text{fm})^3} = 0.138\ \text{u/fm}^3$$

Converting to units of kg/m^3, we find

$$\begin{aligned}\rho &= \left(0.138\ \text{u/fm}^3\right)\left(1.66 \times 10^{-27}\ \text{kg/u}\right)\left(\frac{1\ \text{fm}}{10^{-15}\ \text{m}}\right) \\ &= 2.3 \times 10^{17}\ \text{kg/m}^3\end{aligned}$$

Discussion for Part 1

The radius of this medium-sized nucleus is found to be approximately 4.6 fm, and so its diameter is about 10 fm, or 10^{-14} m. In our discussion of Rutherford's discovery of the nucleus, we noticed that it is about 10^{-15} m in diameter (which is for lighter nuclei), consistent with this result to an order of magnitude. The nucleus is much smaller in diameter than the typical atom, which has a diameter of the order of 10^{-10} m.

Discussion for Part 2

The density found here is so large as to cause disbelief. It is consistent with earlier discussions we have had about the nucleus being very small and containing nearly all of the mass of the atom. Nuclear densities, such as found here, are about 2×10^{14} times greater than that of water, which has a density of "only" $10^3\ \text{kg/m}^3$. One cubic meter of nuclear matter, such as found in a neutron star, has the same mass as a cube of water 61 km on a side.

Nuclear Forces and Stability

What forces hold a nucleus together? The nucleus is very small and its protons, being positive, exert tremendous repulsive forces on one another. (The Coulomb force increases as charges get closer, since it is proportional to $\frac{1}{r^2}$, even at the tiny distances found in nuclei.) The answer is that two previously unknown forces hold the nucleus together and make it into a tightly packed ball of nucleons. These forces are called the *weak and strong nuclear forces*. Nuclear forces are so short ranged that they fall to zero strength when nucleons are separated by only a few fm. However, like glue, they are strongly attracted when the nucleons get close to one another. The strong nuclear force is about 100 times more attractive than the repulsive EM force, easily holding the nucleons together. Nuclear forces become extremely repulsive if the nucleons get too close, making nucleons strongly resist being pushed inside one another, something like ball bearings.

The fact that nuclear forces are very strong is responsible for the very large energies emitted in nuclear decay. During decay, the forces do work, and since work is force times the distance ($W = Fd \cos \theta$), a large force can result in a large emitted energy. In fact, we know that there are *two* distinct nuclear forces because of the different types of nuclear decay—the strong nuclear force is responsible for α decay, while the weak nuclear force is responsible for β decay.

The many stable and unstable nuclei we have explored, and the hundreds we have not discussed, can be arranged in a table called the *chart of the nuclides*, a simplified version of which is shown in Figure 3. Nuclides are located on a plot of N versus Z. Examination of a detailed chart of the nuclides reveals patterns in the characteristics of nuclei, such as stability, abundance, and types of decay, analogous to but more complex than the systematics in the periodic table of the elements.

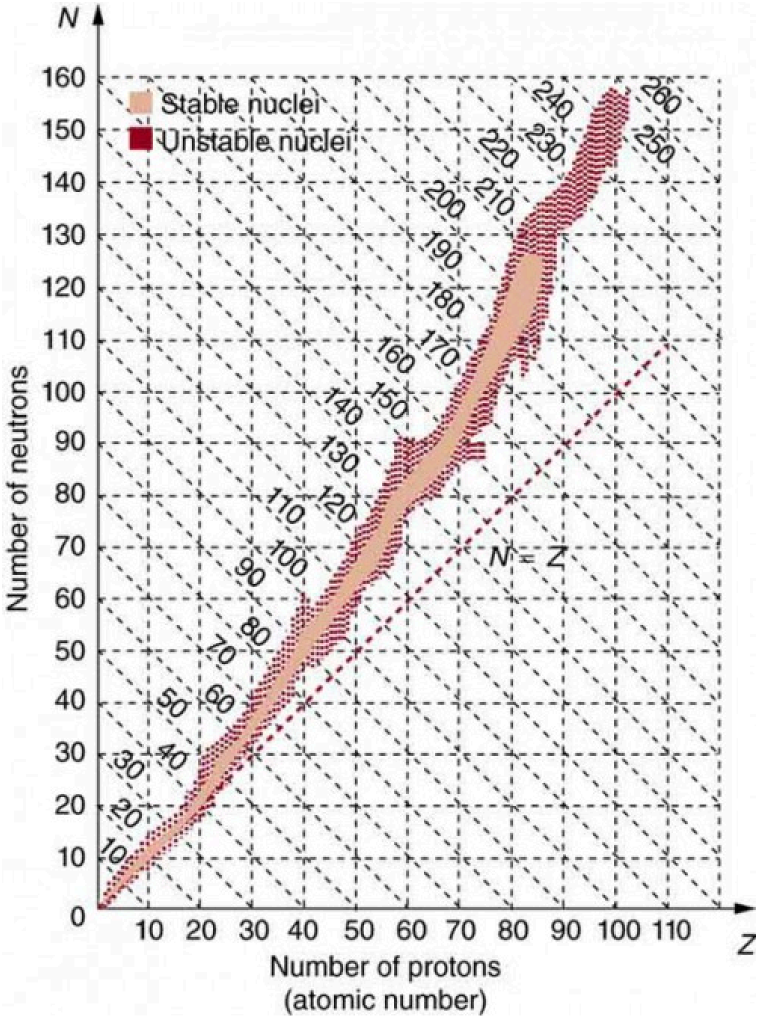


Figure 3. Simplified chart of the nuclides, a graph of N versus Z for known nuclides. The patterns of stable and unstable nuclides reveal characteristics of the nuclear forces. The dashed line is for N = Z. Numbers along diagonals are mass numbers A.

In principle, a nucleus can have any combination of protons and neutrons, but Figure 3 shows a definite pattern for those that are stable. For low-mass nuclei, there is a strong tendency for N and Z to be nearly equal. This means that the nuclear force is more attractive when N = Z. More detailed examination reveals greater stability when N and Z are even numbers—nuclear forces are more attractive when neutrons and protons are in pairs. For increasingly higher masses, there are progressively more neutrons than protons in stable nuclei. This is due to the ever-growing repulsion between protons. Since nuclear forces are short ranged, and the Coulomb force is long ranged, an excess of neutrons keeps the protons a little farther apart, reducing Coulomb repulsion. Decay modes of nuclides out of the region of stability consistently produce nuclides closer to the region of stability. There are more stable nuclei having certain numbers of protons and neutrons, called *magic numbers*. Magic numbers indicate a shell structure for the nucleus in which closed shells are more stable. Nuclear shell theory has been very successful in explaining nuclear energy levels, nuclear decay, and the greater stability of nuclei with closed shells. We have been producing ever-heavier transuranic elements since the early 1940s, and we have now produced the element with Z = 118. There

Figure 4. The German-born American physicist Maria Goeppert Mayer (1906–1972) shared the 1963 Nobel Prize in physics with J. Jensen for the creation of the nuclear shell model. This successful nuclear model has nucleons filling shells analogous to electron shells in atoms. It was inspired by patterns observed in nuclear properties. (credit: Nobel Foundation via Wikimedia Commons)

are theoretical predictions of an island of relative stability for nuclei with such high Zs.

Section Summary

- Two particles, both called nucleons, are found inside nuclei. The two types of nucleons are protons and neutrons; they are very similar, except that the proton is positively charged while the neutron is neutral. Some of their characteristics are given in Table 1 and compared with those of the electron. A mass unit convenient to atomic and nuclear processes is the unified atomic mass unit (u), defined to be
$1\text{ u} = 1.6605 \times 10^{-27}\text{ kg} = 931.46\text{ MeV}/c^2$.
- A nuclide is a specific combination of protons and neutrons, denoted by
$^{A}_{Z}\text{X}_{N}$ or simply ^{A}X, Z is the number of protons or atomic number, X is the symbol for the element, N is the number of neutrons, and A is the mass number or the total number of protons and neutrons, A = N + Z.
- Nuclides having the same Z but different N are isotopes of the same element.
- The radius of a nucleus, r, is approximately $r = r_0 A^{1/3}$, where r_0 = 1.2 fm. Nuclear volumes are proportional to A. There are two nuclear forces, the weak and the strong. Systematics in nuclear stability seen on the chart of the nuclides indicate that there are shell closures in nuclei for values of Z and N equal to the magic numbers, which correspond to highly stable nuclei.

Conceptual Questions

1. The weak and strong nuclear forces are basic to the structure of matter. Why we do not experience them directly?
2. Define and make clear distinctions between the terms neutron, nucleon, nucleus, nuclide, and neutrino.
3. What are isotopes? Why do different isotopes of the same element have similar chemistries?

Problems & Exercises

1. Verify that a 2.3×10^{17} kg mass of water at normal density would make a cube 60 km on a side, as claimed in Example 1. (This mass at nuclear density would make a cube 1.0 m on a side.)
2. Find the length of a side of a cube having a mass of 1.0 kg and the density of nuclear matter, taking this to be $2.3 \times 10^{17}\ \text{kg/m}^3$.
3. What is the radius of an α particle?
4. Find the radius of a ^{238}Pu nucleus. ^{238}Pu is a manufactured nuclide that is used as a power source on some space probes.
5. (a) Calculate the radius of ^{58}Ni, one of the most tightly bound stable nuclei. (b) What is the ratio of the radius of ^{58}Ni to that of ^{258}Ha, one of the largest

nuclei ever made? Note that the radius of the largest nucleus is still much smaller than the size of an atom.

6. The unified atomic mass unit is defined to be $1\ \text{u} = 1.6605 \times 10^{-27}$ kg. Verify that this amount of mass converted to energy yields 931.5 MeV. Note that you must use four-digit or better values for c and $|q_e|$.

7. What is the ratio of the velocity of a β particle to that of an α particle, if they have the same nonrelativistic kinetic energy?

8. If a 1.50-cm-thick piece of lead can absorb 90.0% of the γ rays from a radioactive source, how many centimeters of lead are needed to absorb all but 0.100% of the γ rays?

9. The detail observable using a probe is limited by its wavelength. Calculate the energy of a γ-ray photon that has a wavelength of 1×10^{-16} m, small enough to detect details about one-tenth the size of a nucleon. Note that a photon having this energy is difficult to produce and interacts poorly with the nucleus, limiting the practicability of this probe.

10. (a) Show that if you assume the average nucleus is spherical with a radius $r = r_0 A^{1/3}$, and with a mass of A u, then its density is independent of A. (b) Calculate that density in u/fm^3 and kg/m^3, and compare your results with those found in Example 1 for ^{56}Fe.

11. What is the ratio of the velocity of a 5.00-MeV β ray to that of an α particle with the same kinetic energy? This should confirm that βs travel much faster than αs even when relativity is taken into consideration. (See also Question 7.)

12. (a) What is the kinetic energy in MeV of a β ray that is traveling at 0.998c? This gives some idea of how

energetic a β ray must be to travel at nearly the same speed as a γ ray. (b) What is the velocity of the γ ray relative to the β ray?

Glossary

atomic mass: the total mass of the protons, neutrons, and electrons in a single atom

atomic number: number of protons in a nucleus

chart of the nuclides: a table comprising stable and unstable nuclei

isotopes: nuclei having the same Z and different Ns

magic numbers: a number that indicates a shell structure for the nucleus in which closed shells are more stable

mass number: number of nucleons in a nucleus

neutron: a neutral particle that is found in a nucleus

nucleons: the particles found inside nuclei

nucleus: a region consisting of protons and neutrons at the center of an atom

nuclide: a type of atom whose nucleus has specific numbers of protons and neutrons

protons: the positively charged nucleons found in a nucleus

radius of a nucleus: the radius of a nucleus is $r = r_0 A^{1/3}$

Selected Solutions to Problems & Exercises

1.

$$\begin{aligned} m = \rho V = \rho d^3 \quad \Rightarrow \quad a &= \left(\frac{m}{\rho}\right)^{1/3} = \left(\frac{2.3\times10^{17}\ \text{kg}}{1000\ \text{kg/m}^3}\right)^{1/3} \\ &= \quad 61 \times 10^3\ \text{m} = 61\ \text{km} \end{aligned}$$

3. 1.9 fm

5. (a) 4.6 fm; (b) 0.61 to 1

7. 85.4 to 1

9. 12.4 GeV

11. 19.3 to 1

112. Nuclear Decay and Conservation Laws

Learning Objectives

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

- Define and discuss nuclear decay.
- State the conservation laws.
- Explain parent and daughter nucleus.
- Calculate the energy emitted during nuclear decay.

Nuclear *decay* has provided an amazing window into the realm of the very small. Nuclear decay gave the first indication of the connection between mass and energy, and it revealed the existence of two of the four basic forces in nature. In this section, we explore the major modes of nuclear decay; and, like those who first explored them, we will discover evidence of previously unknown particles and conservation laws.

Some nuclides are stable, apparently living forever. Unstable nuclides decay (that is, they are radioactive), eventually producing a stable nuclide after many decays. We call the original nuclide the *parent* and its decay products the *daughters*. Some radioactive nuclides decay in a single step to a stable nucleus. For example, ^{60}Co is unstable and decays directly to ^{60}Ni, which is stable. Others, such as ^{238}U, decay to another unstable nuclide, resulting in a *decay series* in which each subsequent nuclide decays until a stable nuclide is finally produced.

The decay series that starts from ^{238}U is of particular interest, since it produces the radioactive isotopes ^{226}Ra and ^{210}Po, which the Curies first discovered (see Figure 1). Radon gas is also produced (^{222}Rn in the series), an increasingly recognized naturally occurring hazard. Since radon is a noble gas, it emanates from materials, such as soil, containing even trace amounts of ^{238}U and can be inhaled. The decay of radon and its daughters produces internal damage. The ^{238}U decay series ends with ^{206}Pb, a stable isotope of lead.

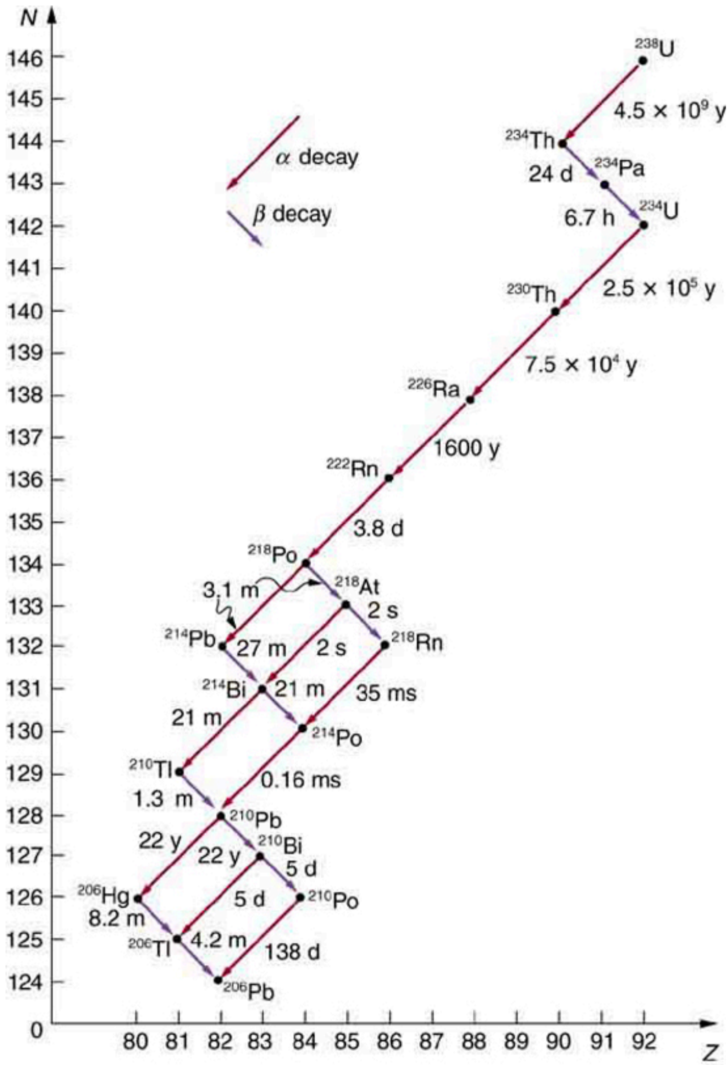


Figure 1. The decay series produced by 238U, the most common uranium isotope. Nuclides are graphed in the same manner as in the chart of nuclides. The type of decay for each member of the series is shown, as well as the half-lives. Note that some nuclides decay by more than one mode. You can see why radium and polonium are found in uranium ore. A stable isotope of lead is the end product of the series.

Note that the daughters of α decay shown in Figure 1 always have two fewer protons and two fewer neutrons than the parent. This seems reasonable, since we know that α decay is the emission of a ^{4}He nucleus, which has two protons and two neutrons. The daughters of β decay have one less neutron and one more proton than their parent. Beta decay is a little more subtle, as we shall see. No γ decays are shown in the figure, because they do not produce a daughter that differs from the parent.

Alpha Decay

In *alpha decay*, a ^{4}He nucleus simply breaks away from the parent nucleus, leaving a daughter with two fewer protons and two fewer neutrons than the parent (see Figure 2). One example of α decay is shown in Figure 1 for ^{238}U. Another nuclide that undergoes α decay is ^{239}Pu. The decay equations for these two nuclides are

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

and

$$^{238}\mathbf{Pu} \rightarrow {}^{235}\mathrm{U} + {}^{4}\mathrm{He}$$

Figure 2. Alpha decay is the separation of a 4He nucleus from the parent. The daughter nucleus has two fewer protons and two fewer neutrons than the parent. Alpha decay occurs spontaneously only if the daughter and 4He nucleus have less total mass than the parent.

If you examine the periodic table of the elements, you will find that Th has Z = 90, two fewer than U, which has Z=92. Similarly, in the second *decay equation*, we see that U has two fewer protons than Pu, which has Z = 94. The general rule for α decay is best written in the format ${}^{A}_{Z}\mathrm{X}_{N}$. If a certain nuclide is known to α decay (generally this information must be looked up in a table of isotopes, such as in Appendix B), its *α decay equation* is

$${}^{A}_{Z}\mathrm{X}_{N} \rightarrow {}^{A-4}_{Z-2}\mathrm{Y}_{N-2} + {}^{4}_{2}\mathrm{He}_2 \; (\alpha \text{ decay})$$

where Y is the nuclide that has two fewer protons than X, such as Th having two fewer than U. So if you were told that ^{239}Pu α decays and were asked to write the complete decay equation, you would first look up which element has two fewer protons (an atomic number two lower) and find that this is uranium. Then since four nucleons have broken away from the original 239, its atomic mass would be 235.

It is instructive to examine conservation laws related to α decay. You can see from the equation ${}_{Z}^{A}\mathrm{X}_{N} \rightarrow {}_{Z-2}^{A-4}\mathrm{Y}_{N-2} + {}_{2}^{4}\mathrm{He}_{2}$ that total charge is conserved. Linear and angular momentum are conserved, too. Although conserved angular momentum is not of great consequence in this type of decay, conservation of linear momentum has interesting consequences. If the nucleus is at rest when it decays, its momentum is zero. In that case, the fragments must fly in opposite directions with equal-magnitude momenta so that total momentum remains zero. This results in the α particle carrying away most of the energy, as a bullet from a heavy rifle carries away most of the energy of the powder burned to shoot it. Total mass–energy is also conserved: the energy produced in the decay comes from conversion of a fraction of the original mass. As discussed in Atomic Physics, the general relationship is $E = (\Delta m)c^2$.

Here, E is the *nuclear reaction energy* (the reaction can be nuclear decay or any other reaction), and Δm is the difference in mass between initial and final products. When the final products have less total mass, Δm is positive, and the reaction releases energy (is exothermic). When the products have greater total mass, the reaction is endothermic (Δm is negative) and must be induced with an energy input. For α decay to be spontaneous, the decay products must have smaller mass than the parent.

Example 1. Alpha Decay Energy Found from Nuclear Masses

Find the energy emitted in the α decay of ^{239}Pu.

Strategy

Nuclear reaction energy, such as released in α decay, can be found using the equation $E = (\Delta m)c^2$. We must first find Δm, the difference in mass between the parent nucleus and the products of the decay. This is easily done using masses given in Appendix A.

Solution

The decay equation was given earlier for ^{239}Pu; it is

$$^{239}\text{Pu} \rightarrow {}^{235}\text{U} + {}^{4}\text{He}.$$

Thus the pertinent masses are those of ^{239}Pu, ^{235}U, and the α particle or ^{4}He, all of which are listed in Appendix A. The initial mass was $m(^{239}\text{Pu})$=239.052157 u. The final mass is the sum $m(^{235}\text{U}) + m(^{4}\text{He})$ = 235.043924 u + 4.002602 u = 239.046526 u. Thus,

$$\begin{aligned} \Delta m &= m\left(^{239}\text{Pu}\right) - \left[m\left(^{235}\text{U}\right) + m\left(^{4}\text{He}\right)\right] \\ &= 239.052157\ \text{u} - 239.046526\ \text{u} \\ &= 0.0005631\ \text{u} \end{aligned}$$

Now we can find E by entering Δm into the equation: $E = (\Delta m)c^2 = (0.005631 \text{ u})c^2$.

We know $1 \text{ u} = 931.5 \text{ MeV}/c^2$, and so $E = (0.005631)(931.5 \text{ MeV}/c^2)(c^2) = 5.25 \text{ MeV}$.

Discussion

The energy released in this α decay is in the MeV range, about 10^6 times as great as typical chemical reaction energies, consistent with many previous discussions. Most of this energy becomes kinetic energy of the α particle (or ^{4}He nucleus), which moves away at high speed. The energy carried away by the recoil of the ^{235}U nucleus is much smaller in order to conserve momentum. The ^{235}U nucleus can be left in an excited state to later emit photons (γ rays). This decay is spontaneous and releases energy, because the products have less mass than the parent nucleus. The question of why the products have less mass will be discussed in Binding Energy. Note that the masses given in Appendix A are atomic masses of neutral atoms, including their electrons. The mass of the electrons is the same before and after α decay, and so their masses subtract out when finding Δm. In this case, there are 94 electrons before and after the decay.

Beta Decay

There are actually *three* types of *beta decay*. The first discovered was "ordinary" beta decay and is called β^- decay or electron emission.

The symbol β^- represents *an electron emitted in nuclear beta decay*. Cobalt-60 is a nuclide that β^- decays in the following manner: $^{60}Co \rightarrow {}^{60}Ni + \beta^- +$ neutrino.

The *neutrino* is a particle emitted in beta decay that was unanticipated and is of fundamental importance. The neutrino was not even proposed in theory until more than 20 years after beta decay was known to involve electron emissions. Neutrinos are so difficult to detect that the first direct evidence of them was not obtained until 1953. Neutrinos are nearly massless, have no charge, and do not interact with nucleons via the strong nuclear force. Traveling approximately at the speed of light, they have little time to affect any nucleus they encounter. This is, owing to the fact that they have no charge (and they are not EM waves), they do not interact through the EM force. They do interact via the relatively weak and very short range weak nuclear force. Consequently, neutrinos escape almost any detector and penetrate almost any shielding. However, neutrinos do carry energy, angular momentum (they are fermions with half-integral spin), and linear momentum away from a beta decay. When accurate measurements of beta decay were made, it became apparent that energy, angular momentum, and linear momentum were not accounted for by the daughter nucleus and electron alone. Either a previously unsuspected particle was carrying them away, or three conservation laws were being violated. Wolfgang Pauli made a formal proposal for the existence of neutrinos in 1930. The Italian-born American physicist Enrico Fermi (1901–1954) gave neutrinos their name, meaning little neutral ones, when he developed a sophisticated theory of beta decay (see Figure 3). Part of Fermi's theory was the identification of the weak nuclear force as being distinct from the strong nuclear force and in fact responsible for beta decay.

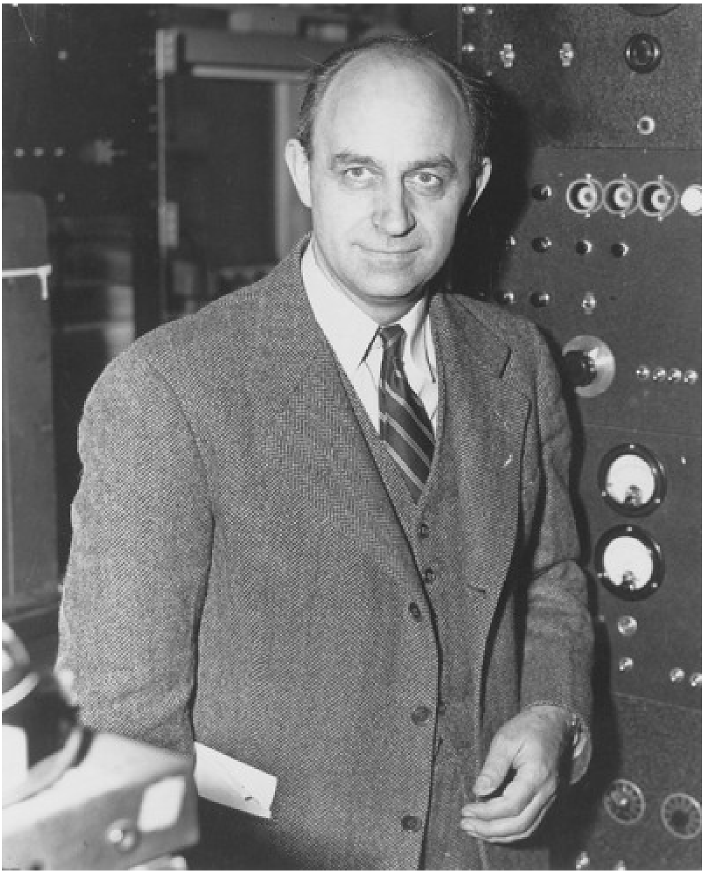


Figure 3. Enrico Fermi was nearly unique among 20th-century physicists—he made significant contributions both as an experimentalist and a theorist. His many contributions to theoretical physics included the identification of the weak nuclear force. The fermi (fm) is named after him, as are an entire class of subatomic particles (fermions), an element (Fermium), and a major research laboratory (Fermilab). His experimental work included studies of radioactivity, for which he won the 1938 Nobel Prize in physics, and creation of the first nuclear chain reaction. (credit: United States Department of Energy, Office of Public Affairs)

The neutrino also reveals a new conservation law. There are various families of particles, one of which is the electron family. We propose that the number of members of the electron family is constant in any process or any closed system. In our example of beta decay, there are no members of the electron family present before the decay, but after, there is an electron and a neutrino. So electrons are given an electron family number of +1. The neutrino in β^- decay is an *electron's antineutrino*, given the symbol $\bar{\nu}_e$, where ν is the Greek letter nu, and the subscript e means this neutrino is related to the electron. The bar indicates this is a particle of *antimatter*. (All particles have antimatter counterparts that are nearly identical except that they have the opposite charge. Antimatter is almost entirely absent on Earth, but it is found in nuclear decay and other nuclear and particle reactions as well as in outer space.) The electron's antineutrino $\bar{\nu}_e$, being antimatter, has an electron family number of −1. The total is zero, before and after the decay. The new conservation law, obeyed in all circumstances, states that the *total electron family number is constant*. An electron cannot be created without also creating an antimatter family member. This law is analogous to the conservation of charge in a situation where total charge is originally zero, and equal amounts of positive and negative charge must be created in a reaction to keep the total zero.

If a nuclide ${}^{A}_{Z}\mathrm{X}_N$ is known to β^- decay, then its β^- decay equation is

$$\mathrm{X}_N \rightarrow \mathrm{Y}_{N-1} + \beta^- + \bar{\nu}_e \text{ } (\beta^- \text{ decay}),$$

where Y is the nuclide having one more proton than X (see Figure 4). So if you know that a certain nuclide β^- decays, you can find the daughter nucleus by first looking up Z for the parent and then determining which element has atomic number Z + 1. In the example of the β^- decay of $^{60}\mathrm{Co}$ given earlier, we see that Z = 27 for Co and Z = 28 is Ni. It is as if one of the neutrons in the parent nucleus decays into a proton, electron, and neutrino. In fact, neutrons outside of nuclei do just that—they live only an average of a few minutes and β^- decay in the following manner:

$$\mathrm{n} \rightarrow \mathrm{p} + \beta^- + \bar{\nu}_e$$

We see that charge is conserved in β^- decay, since the total charge is Z before and after the decay. For example, in ^{60}Co decay, total charge is 27 before decay, since cobalt has Z = 27. After decay, the daughter nucleus is Ni, which has Z = 28, and there is an electron, so that the total charge is also 28 + (−1) or 27. Angular momentum is conserved, but not obviously (you have to examine the spins and angular momenta of the final products in detail to verify this). Linear momentum is also conserved, again imparting most of the decay energy to the electron and the antineutrino, since they are of low and zero mass, respectively. Another new conservation law is obeyed here and elsewhere in nature. *The total number of nucleons* A *is conserved.* In ^{60}Co decay, for example, there are 60 nucleons before and after the decay. Note that total A is also conserved in α decay. Also note that the total number of protons changes, as does the total number of neutrons, so that total Z and total N are *not* conserved in β^- decay, as they are in α decay. Energy released in β^- decay can be calculated given the masses of the parent and products.

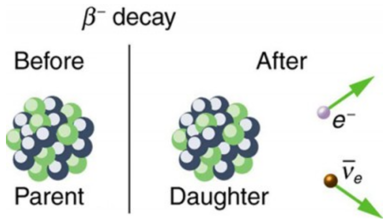


Figure 4. In $\beta-$ decay, the parent nucleus emits an electron and an antineutrino. The daughter nucleus has one more proton and one less neutron than its parent. Neutrinos interact so weakly that they are almost never directly observed, but they play a fundamental role in particle physics.

Example 2. β^- Decay Energy from Masses

Find the energy emitted in the β^- decay of ^{60}Co.

Strategy and Concept

As in the preceding example, we must first find Δm, the difference in mass between the parent nucleus and the products of the decay, using masses given in Appendix A. Then the emitted energy is calculated as before, using $E = (\Delta m)c^2$. The initial mass is just that of the parent nucleus, and the final mass is that of the daughter nucleus and the electron created in the decay. The neutrino is massless, or nearly so. However, since the masses given in Appendix A are for neutral atoms, the daughter nucleus has one more electron than the parent, and so the extra electron mass that corresponds to the β^- is included in the atomic mass of Ni. Thus, $\Delta m = m(^{60}\text{Co}) - m(^{60}\text{Ni})$.

Solution

The β^- decay equation for ^{60}Co is

$$^{60}_{27}\text{Co}_{33} \rightarrow {}^{60}_{28}\text{Ni}_{32} + \beta^- + \bar{\nu}_e$$

As noticed, $\Delta m = m(^{60}\text{Co}) - m(^{60}\text{Ni})$.

Entering the masses found in Appendix A gives $\Delta m = 59.933820\ \text{u} - 59.930789\ \text{u} = 0.003031\ \text{u}$.

Thus, $E = (\Delta m)c^2 = (0.003031\ \text{u})c^2$.

Using $1\ \text{u} = 931.5\ \text{MeV}/c^2$, we obtain $E = (0.003031)(931.5\ \text{MeV}/c^2)(c^2) = 2.82\ \text{MeV}$.

Discussion and Implications

Perhaps the most difficult thing about this example is convincing yourself that the β^- mass is included in the atomic mass of ^{60}Ni. Beyond that are other implications. Again the decay energy is in the MeV range. This energy is shared by all of the products of the decay. In many ^{60}Co decays, the daughter nucleus ^{60}Ni is left in an excited state and emits photons (γ rays). Most of the remaining energy goes to the electron and neutrino, since the recoil kinetic energy of the daughter nucleus is small. One final note: the electron emitted in β^- decay is created in the nucleus at the time of decay.

The second type of beta decay is less common than the first. It is β^+decay. Certain nuclides decay by the emission of a *positive* electron. This is *antielectron* or *positron decay* (see Figure 5).

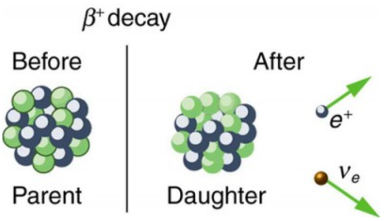


Figure 5. β+ decay is the emission of a positron that eventually finds an electron to annihilate, characteristically producing gammas in opposite directions.

The antielectron is often represented by the symbol e^+, but in beta decay it is written as β^+ to indicate the antielectron was emitted in a nuclear decay. Antielectrons are the antimatter counterpart to electrons, being nearly identical, having the same mass, spin, and so on, but having a positive charge and an electron family number of −1. When a *positron* encounters an electron, there is a mutual annihilation in which all the mass of the antielectron-electron pair is converted into pure photon energy. (The reaction, $e^+ + e^- \rightarrow \gamma + \gamma$, conserves electron family number as well as all other conserved quantities.) If a nuclide ${}^{A}_{Z}\mathrm{X}_N$ is known to β^+ decay, then its β^+*decay equation* is

$${}^{A}_{Z}\mathrm{X}_N \rightarrow \mathrm{Y}_{N+1} + \beta^+ + v_e \ (\beta^+ \text{ decay}),$$

where Y is the nuclide having one less proton than X (to conserve charge) and v_e is the symbol for the *electron's neutrino*, which has an electron family number of +1. Since an antimatter member of the electron family (the β^+) is created in the decay, a matter member of the family (here the v_e) must also be created. Given, for example, that ^{22}Na β^+ decays, you can write its full decay equation by first finding that Z = 11 for ^{22}Na, so that the daughter nuclide will have Z = 10, the atomic number for neon. Thus the β^+ decay equation for ^{22}Na is

$${}^{11}_{22}\mathrm{Na}_{11} \rightarrow {}^{22}_{10}\mathrm{Ne}_{12} + \beta^+ + v_e$$

In β^+ decay, it is as if one of the protons in the parent nucleus decays into a neutron, a positron, and a neutrino. Protons do not do this outside of the nucleus, and so the decay is due to the complexities of the nuclear force. Note again that the total number of nucleons is constant in this and any other reaction. To find the energy emitted in β^+ decay, you must again count the number of electrons in the neutral atoms, since atomic masses are used. The daughter has one less electron than the parent, and one electron mass is created in the decay. Thus, in β^+ decay,

$$\Delta m = m(\text{parent}) - [m(\text{daughter}) + 2m_e],$$

since we use the masses of neutral atoms.

Electron capture is the third type of beta decay. Here, a nucleus captures an inner-shell electron and undergoes a nuclear reaction that has the same effect as β^+ decay. Electron capture is sometimes denoted by the letters EC. We know that electrons cannot reside in the nucleus, but this is a nuclear reaction that consumes the electron and occurs spontaneously only when the products have less mass than the parent plus the electron. If a nuclide ${}^{A}_{Z}\mathrm{X}_N$

is known to undergo electron capture, then its *electron capture equation* is

$$^{A}_{Z}\mathrm{X}_{N} + e^{-} \rightarrow \mathrm{Y}_{N+1} + v_{e} \text{ (electron capture, or EC)}$$

Any nuclide that can β^{+} decay can also undergo electron capture (and often does both). The same conservation laws are obeyed for EC as for β^{+} decay. It is good practice to confirm these for yourself.

All forms of beta decay occur because the parent nuclide is unstable and lies outside the region of stability in the chart of nuclides. Those nuclides that have relatively more neutrons than those in the region of stability will β^{-} decay to produce a daughter with fewer neutrons, producing a daughter nearer the region of stability. Similarly, those nuclides having relatively more protons than those in the region of stability will β^{-} decay or undergo electron capture to produce a daughter with fewer protons, nearer the region of stability.

Gamma Decay

Gamma decay is the simplest form of nuclear decay—it is the emission of energetic photons by nuclei left in an excited state by some earlier process. Protons and neutrons in an excited nucleus are in higher orbitals, and they fall to lower levels by photon emission (analogous to electrons in excited atoms). Nuclear excited states have lifetimes typically of only about 10^{-14} s, an indication of the great strength of the forces pulling the nucleons to lower states. The γ decay equation is simply

$$^{A}_{Z}\mathrm{X}^{*}_{N} \rightarrow \mathrm{X}_{N} + \gamma_{1} + \gamma_{2} \ldots (\gamma \text{ decay})$$

where the asterisk indicates the nucleus is in an excited state. There may be one or more γ s emitted, depending on how the nuclide de-excites. In radioactive decay, γ emission is common and is preceded by γ or β decay. For example, when ^{60}Co β^{-} decays, it most often leaves the daughter nucleus in an excited state, written

$^{60}Ni^*$. Then the nickel nucleus quickly γ decays by the emission of two penetrating γs: $^{60}Ni^* \rightarrow {^{60}Ni} + \gamma_1 + \gamma_2$.

These are called cobalt γ rays, although they come from nickel—they are used for cancer therapy, for example. It is again constructive to verify the conservation laws for gamma decay. Finally, since γ decay does not change the nuclide to another species, it is not prominently featured in charts of decay series, such as that in Figure 1.

There are other types of nuclear decay, but they occur less commonly than α, β, and γ decay. Spontaneous fission is the most important of the other forms of nuclear decay because of its applications in nuclear power and weapons. It is covered in the next chapter.

Section Summary

- When a parent nucleus decays, it produces a daughter nucleus following rules and conservation laws. There are three major types of nuclear decay, called alpha (α), beta (β), and gamma (γ). The α decay equation is ${}^{A}_{Z}\mathrm{X}_{N} \rightarrow {}^{A-4}_{Z-2}\mathrm{Y}_{N-2} + {}^{4}_{2}\mathrm{He}_{2}$.
- Nuclear decay releases an amount of energy E related to the mass destroyed Δm by $E = (\Delta m)c^2$.
- There are three forms of beta decay. The β^- decay equation is ${}^{A}_{Z}\mathrm{X}_{N} \rightarrow {}^{A}_{Z+1}\mathrm{Y}_{N-1} + \beta^- + \bar{\nu}_e$.
- The β^+ decay equation is ${}^{A}_{Z}\mathrm{X}_{N} \rightarrow {}^{A}_{Z-1}\mathrm{Y}_{N+1} + \beta^+ + \nu_e$.
- The electron capture equation is ${}^{A}_{Z}\mathrm{X}_{N} + e^- \rightarrow {}^{A}_{Z-1}\mathrm{Y}_{N+1} + \nu_e$.
- β^- is an electron, β^+ is an antielectron or positron, ν_e represents an electron's neutrino, and $\overline{\nu}_e$ is an electron's antineutrino. In addition to all previously known conservation laws, two new ones arise—conservation of electron family

number and conservation of the total number of nucleons. The γ decay equation is

$$Z^{A}\text{X}_{N}^{*} \rightarrow {}_{Z}^{A}\text{X}_{N} + \gamma_1 + \gamma_2 + \cdots$$

γ is a high-energy photon originating in a nucleus.

Conceptual Questions

1. Star Trek fans have often heard the term "antimatter drive." Describe how you could use a magnetic field to trap antimatter, such as produced by nuclear decay, and later combine it with matter to produce energy. Be specific about the type of antimatter, the need for vacuum storage, and the fraction of matter converted into energy.
2. What conservation law requires an electron's neutrino to be produced in electron capture? Note that the electron no longer exists after it is captured by the nucleus.
3. Neutrinos are experimentally determined to have an extremely small mass. Huge numbers of neutrinos are created in a supernova at the same time as massive amounts of light are first produced. When the 1987A supernova occurred in the Large Magellanic Cloud, visible primarily in the Southern Hemisphere and some 100,000 light-years away from Earth, neutrinos from the explosion were observed at about the same time as the light from the blast. How could the relative arrival times of neutrinos and light be used to place limits on the mass of neutrinos?
4. What do the three types of beta decay have in common that is distinctly different from alpha decay?

Problems & Exercises

In the following eight problems, write the complete decay equation for the given nuclide in the complete ${}^{A}_{Z}\text{X}_{N}$ notation. Refer to the periodic table for values of Z.

1. β^{-} decay of ^{3}H (tritium), a manufactured isotope of hydrogen used in some digital watch displays, and manufactured primarily for use in hydrogen bombs.
2. β^{-} decay of ^{40}K, a naturally occurring rare isotope of potassium responsible for some of our exposure to background radiation.
3. β^{+} decay of ^{50}Mn.
4. β^{+} decay of ^{52}Fe.
5. Electron capture by ^{7}Be.
6. Electron capture by ^{106}In.
7. α decay of ^{210}Po, the isotope of polonium in the decay series of ^{238}U that was discovered by the Curies. A favorite isotope in physics labs, since it has a short half-life and decays to a stable nuclide.
8. α decay of ^{226}Ra, another isotope in the decay series of ^{238}U, first recognized as a new element by the Curies. Poses special problems because its daughter is a radioactive noble gas.

In the following four problems, identify the parent nuclide and write the complete decay equation in the ${}^{A}_{Z}\text{X}_{N}$ notation. Refer to the periodic table for values of Z.

1. β^{-} decay producing ^{137}Ba . The parent nuclide is a major waste product of reactors and has chemistry similar to potassium and sodium, resulting in its

concentration in your cells if ingested.

2. β^{-} decay producing ^{90}Y. The parent nuclide is a major waste product of reactors and has chemistry similar to calcium, so that it is concentrated in bones if ingested (^{90}Y is also radioactive.)
3. α decay producing ^{228}Ra. The parent nuclide is nearly 100% of the natural element and is found in gas lantern mantles and in metal alloys used in jets (^{228}Ra is also radioactive).
4. α decay producing ^{208}Pb. The parent nuclide is in the decay series produced by ^{232}Th, the only naturally occurring isotope of thorium.

Answer the remaining questions.

1. When an electron and positron annihilate, both their masses are destroyed, creating two equal energy photons to preserve momentum. (a) Confirm that the annihilation equation $e^{+} + e^{-} \rightarrow \gamma + \gamma$ conserves charge, electron family number, and total number of nucleons. To do this, identify the values of each before and after the annihilation. (b) Find the energy of each γ ray, assuming the electron and positron are initially nearly at rest. (c) Explain why the two γ rays travel in exactly opposite directions if the center of mass of the electron-positron system is initially at rest.
2. Confirm that charge, electron family number, and the total number of nucleons are all conserved by the rule for α decay given in the equation $^{A}_{Z}\mathrm{X}_{N} \rightarrow {}^{A-4}_{Z-2}\mathrm{Y}_{N-2} + {}^{4}_{2}\mathrm{He}_{2}$. To do this, identify the values of each before and after the decay.

3. Confirm that charge, electron family number, and the total number of nucleons are all conserved by the rule for β^- decay given in the equation $^{A}_{Z}\mathrm{X}_N \rightarrow {}^{A}_{Z+1}\mathrm{Y}_{N-1} + \beta^- + \bar{\nu}_e$. To do this, identify the values of each before and after the decay.
4. Confirm that charge, electron family number, and the total number of nucleons are all conserved by the rule for β^- decay given in the equation $^{A}_{Z}\mathrm{X}_N \rightarrow {}^{A}_{Z-1}\mathrm{Y}_{N-1} + \beta^- + \nu_e$. To do this, identify the values of each before and after the decay.
5. Confirm that charge, electron family number, and the total number of nucleons are all conserved by the rule for electron capture given in the equation $^{A}_{Z}\mathrm{X}_N + e^- \rightarrow {}^{A}_{Z-1}\mathrm{Y}_{N+1} + \nu_e$. To do this, identify the values of each before and after the capture.
6. A rare decay mode has been observed in which ^{222}Ra emits a ^{14}C nucleus. (a) The decay equation is $^{222}\mathrm{Ra} \rightarrow {}^{A}\mathrm{X} + {}^{14}\mathrm{C}$. Identify the nuclide ^{A}X. (b) Find the energy emitted in the decay. The mass of ^{222}Ra is 222.015353 u.
7. (a) Write the complete α decay equation for ^{226}Ra. (b) Find the energy released in the decay.
8. (a) Write the complete α decay equation for ^{249}Cf. (b) Find the energy released in the decay.
9. (a) Write the complete β^- decay equation for the neutron. (b) Find the energy released in the decay.
10. (a) Write the complete β^- decay equation for ^{90}Sr, a major waste product of nuclear reactors. (b) Find the energy released in the decay.
11. Calculate the energy released in the β^+ decay of

^{22}Na, the equation for which is given in the text. The masses of ^{22}Na and ^{22}Na are 21.994434 and 21.991383 u, respectively.

12. (a) Write the complete β^{+} decay equation for ^{11}C. (b) Calculate the energy released in the decay. The masses of ^{11}C and ^{11}B are 11.011433 and 11.009305 u, respectively.
13. (a) Calculate the energy released in the α decay of ^{238}U. (b) What fraction of the mass of a single ^{238}U is destroyed in the decay? The mass of ^{234}Th is 234.043593 u. (c) Although the fractional mass loss is large for a single nucleus, it is difficult to observe for an entire macroscopic sample of uranium. Why is this?
14. (a) Write the complete reaction equation for electron capture by ^{7}Be. (b) Calculate the energy released.
15. (a) Write the complete reaction equation for electron capture by ^{15}O. (b) Calculate the energy released.

Glossary

parent: the original state of nucleus before decay

daughter: the nucleus obtained when parent nucleus decays and produces another nucleus following the rules and the conservation laws

positron: the particle that results from positive beta decay; also known as an antielectron

decay: the process by which an atomic nucleus of an unstable atom loses mass and energy by emitting ionizing particles

alpha decay: type of radioactive decay in which an atomic nucleus emits an alpha particle

beta decay: type of radioactive decay in which an atomic nucleus emits a beta particle

gamma decay: type of radioactive decay in which an atomic nucleus emits a gamma particle

decay equation: the equation to find out how much of a radioactive material is left after a given period of time

nuclear reaction energy: the energy created in a nuclear reaction

neutrino: an electrically neutral, weakly interacting elementary subatomic particle

electron's antineutrino: antiparticle of electron's neutrino

positron decay: type of beta decay in which a proton is converted to a neutron, releasing a positron and a neutrino

antielectron: another term for positron

decay series: process whereby subsequent nuclides decay until a stable nuclide is produced

electron's neutrino: a subatomic elementary particle which has no net electric charge

antimatter: composed of antiparticles

electron capture: the process in which a proton-rich nuclide absorbs an inner atomic electron and simultaneously emits a neutrino

electron capture equation: equation representing the electron capture

Selected Solutions to Problems & Exercises

Write the complete decay equation for the given nuclide in *the complete $^A_Z\mathrm{X}_N$ notation. Refer to the periodic table for values of Z.*

1. $^3_1\mathrm{H}_2 \rightarrow {}^3_2\mathrm{He}_1 + \beta^- + \bar{\nu}_e$
3. $^{50}_{25}\mathrm{M}_{25} \rightarrow {}^{50}_{24}\mathrm{Cr}_{26} + \beta^+ + \nu_e$
5. $^7_4\mathrm{Be}_3 + e^- \rightarrow {}^7_3\mathrm{Li}_4 + \nu_e$
7. $^{210}_{84}\mathrm{Po}_{126} \rightarrow {}^{206}_{82}\mathrm{Pb}_{124} + {}^4_2\mathrm{He}_2$

Identify the parent nuclide and write the complete decay equation in the $^A_Z\mathrm{X}_N$ notation. Refer to the periodic table for values of Z.

1. $^{137}_{55}\mathrm{Cs}_{82} \rightarrow {}^{137}_{56}\mathrm{Ba}_{81} + \beta^- + \bar{\nu}_e$
3. $^{232}_{90}\mathrm{Th}_{142} \rightarrow {}^{228}_{88}\mathrm{Ra}_{140} + {}^4_2\mathrm{He}_2$

Answer the remaining questions.

1. (a) charge: (+1) + (−1) = 0; electron family number: (+1) + (−1) = 0; A: 0 + 0 = 0; (b) 0.511 MeV; (c) The two γ rays must travel in exactly opposite directions in order to conserve momentum, since initially there is zero momentum if the center of mass is initially at rest.

3. Z = (Z + 1) − 1; A = A; efn: 0 = (+1) + (−1)

5. Z − 1 = Z − 1; A = A; efn : (+1) = (+1)

7. (a) $^{226}_{88}\mathrm{Ra}_{138} \rightarrow {}^{222}_{86}\mathrm{Rn}_{136} + {}^4_2\mathrm{He}_2$; (b) 4.87 MeV

9. (a) $\mathrm{n} \rightarrow \mathrm{p} + \beta^- + \bar{\nu}_e$; (b) 0.783 MeV

11. 1.82 MeV

13. (a) 4.274 MeV; (b) 1.927×10^{-5}; (c) Since U-238 is a slowly decaying substance, only a very small number of nuclei decay on human timescales; therefore, although those nuclei that decay lose a noticeable fraction of their mass,

the change in the total mass of the sample is not detectable for a macroscopic sample.

15. (a) ${}^{15}_{8}\text{O}_7 + e^- \rightarrow {}^{15}_{7}\text{N}_8 + \nu_e$; (b) 2.754 MeV

113. Half-Life and Activity

Learning Objectives

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

- Define half-life.
- Define dating.
- Calculate age of old objects by radioactive dating.

Unstable nuclei decay. However, some nuclides decay faster than others. For example, radium and polonium, discovered by the Curies, decay faster than uranium. This means they have shorter lifetimes, producing a greater rate of decay. In this section we explore half-life and activity, the quantitative terms for lifetime and rate of decay.

Half-Life

Why use a term like half-life rather than lifetime? The answer can be found by examining Figure 1, which shows how the number of radioactive nuclei in a sample decreases with time. The *time in which half of the original number of nuclei decay* is defined as the *half-life*, $t_{1/2}$. Half of the remaining nuclei decay in the next half-life. Further, half of that amount decays in the following half-life. Therefore, the number of radioactive nuclei decreases from N to $\frac{N}{2}$ in one half-life, then to $\frac{N}{4}$ in the next, and to $\frac{N}{8}$ in the next, and so on. If N is a large number, then *many* half-lives (not just two) pass before all of the nuclei decay. Nuclear decay is an example of a purely statistical process. A more precise definition of half-life is that *each nucleus has a 50% chance of living for a time equal to one half-life* $t_{1/2}$. Thus, if N is reasonably large, half of the original nuclei decay in a time of one half-life. If an individual nucleus makes it through that time, it still has a 50% chance of surviving through another half-life. Even if it happens to make it through hundreds of half-lives, it still has a 50% chance of surviving through one more. The probability of decay is the same no matter when you start counting. This is like random coin flipping. The chance of heads is 50%, no matter what has happened before.

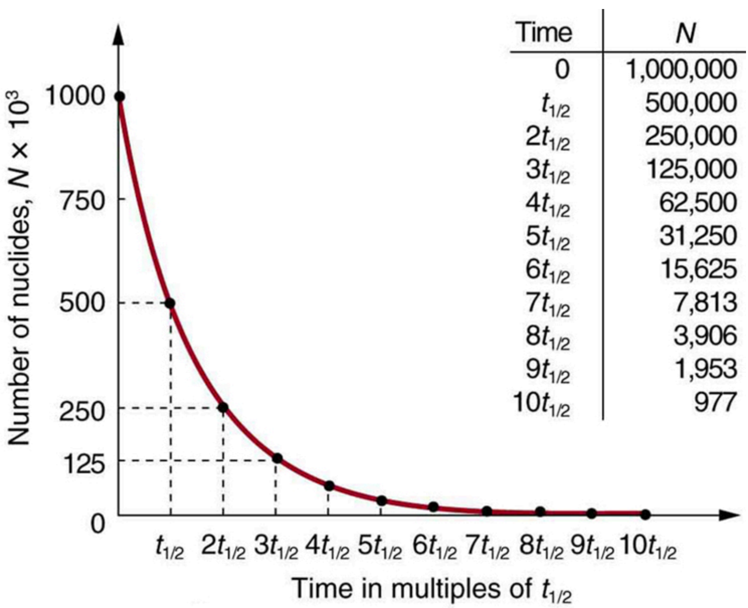


Figure 1. Radioactive decay reduces the number of radioactive nuclei over time. In one half-life $t_{1/2}$, the number decreases to half of its original value. Half of what remains decay in the next half-life, and half of those in the next, and so on. This is an exponential decay, as seen in the graph of the number of nuclei present as a function of time.

There is a tremendous range in the half-lives of various nuclides, from as short as 10^{-23} s for the most unstable, to more than 10^{16} y for the least unstable, or about 46 orders of magnitude. Nuclides with the shortest half-lives are those for which the nuclear forces are least attractive, an indication of the extent to which the nuclear force can depend on the particular combination of neutrons and protons. The concept of half-life is applicable to other subatomic particles, as will be discussed in Particle Physics. It is also applicable to the decay of excited states in atoms and nuclei. The following equation gives the quantitative relationship between the original number of nuclei present at time zero (N_0) and the number (N) at a later time t: $N = N_0 e^{-\lambda t}$, where e=2.71828... is the base of the natural logarithm, and λ is the *decay constant* for the nuclide. The shorter the half-life, the larger is the value of λ, and the faster the exponential $e^{-\lambda t}$ decreases with time. The relationship between the decay constant λ and the half-life $t_{1/2}$ is

$$\lambda = \frac{\ln(2)}{t_{1/2}} \approx \frac{0.693}{t_{1/2}}.$$

To see how the number of nuclei declines to half its original value in one half-life, let $t = t_{1/2}$ in the exponential in the equation $N = N_0 e^{-\lambda t}$. This gives $N = N_0 e^{-\lambda t} = N_0 e^{-0.693} = 0.500 N_0$. For integral numbers of half-lives, you can just divide the original number by 2 over and over again, rather than using the exponential relationship. For example, if ten half-lives have passed, we divide N by 2 ten times. This reduces it to $\dfrac{N}{1024}$. For an arbitrary time, not just a multiple of the half-life, the exponential relationship must be used.

Radioactive dating is a clever use of naturally occurring radioactivity. Its most famous application is *carbon-14 dating*. Carbon-14 has a half-life of 5730 years and is produced in a nuclear reaction induced when solar neutrinos strike ^{14}N in the atmosphere. Radioactive carbon has the same chemistry as stable carbon, and so it mixes into the ecosphere, where it is consumed and becomes part of every living organism. Carbon-14 has an abundance of 1.3 parts per trillion of normal carbon. Thus, if you know the number of carbon nuclei in an object (perhaps determined by mass and Avogadro's number), you multiply that number by 1.3×10^{-12} to find the number of ^{14}C nuclei in the object. When an organism dies, carbon exchange with the environment ceases, and ^{14}C is not replenished as it decays. By comparing the abundance of ^{14}C in an artifact, such as mummy wrappings, with the normal abundance in living tissue, it is possible to determine the artifact's age (or time since death). Carbon-14 dating can be used for biological tissues as old as 50 or 60 thousand years, but is most accurate for younger samples, since the abundance of ^{14}C nuclei in them is greater. Very old biological materials contain no ^{14}C at all. There are instances in which the date of an artifact can be determined by other means, such as historical knowledge or tree-ring counting. These cross-references have confirmed the validity of carbon-14 dating and permitted us to calibrate the technique as well. Carbon-14 dating revolutionized parts of archaeology and is of such importance that it earned the 1960 Nobel Prize in chemistry for its developer, the American chemist Willard Libby (1908–1980).

One of the most famous cases of carbon-14 dating involves the Shroud of Turin, a long piece of fabric purported to be the burial shroud of Jesus (see Figure 2). This relic was first displayed in Turin in 1354 and was denounced as a fraud at that time by a French bishop. Its remarkable negative imprint of an apparently crucified body resembles the then-accepted image of Jesus, and so the shroud was never disregarded completely and remained controversial over the centuries. Carbon-14 dating was not performed on the shroud until 1988, when the process had been refined to the point where only a small amount of material needed to be destroyed. Samples were tested at three independent laboratories, each being given four pieces of cloth, with only one unidentified piece from the shroud, to avoid prejudice. All three laboratories found samples of the shroud contain 92% of the ^{14}C found in living tissues, allowing the shroud to be dated (see Example 1).

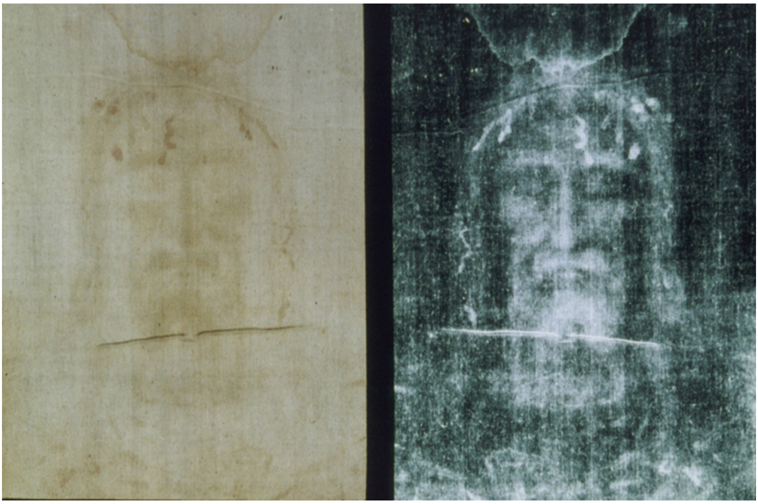


Figure 2. Part of the Shroud of Turin, which shows a remarkable negative imprint likeness of Jesus complete with evidence of crucifixion wounds. The shroud first surfaced in the 14th century and was only recently carbon-14 dated. It has not been determined how the image was placed on the material. (credit: Butko, Wikimedia Commons)

Example 1. How Old Is the Shroud of Turin?

Calculate the age of the Shroud of Turin given that the amount of ^{14}C found in it is 92% of that in living tissue.

Strategy

Knowing that 92% of the ^{14}C remains means that

$\frac{N}{N_0} = 0.92$. Therefore, the equation $N = N_0 e^{-\lambda t}$ can be used to find λt. We also know that the half-life of ^{14}C is 5730 y, and so once λt is known, we can use the equation $\lambda = \frac{0.693}{t_{1/2}}$ to find λ and then find t as requested. Here, we postulate that the decrease in ^{14}C is solely due to nuclear decay.

Solution

Solving the equation $N = N_0 e^{-\lambda t}$ for $\frac{N}{N_0}$ gives

$$\frac{N}{N_0} = e^{-\lambda t}.$$

Thus, $0.92 = e^{-\lambda t}$.

Taking the natural logarithm of both sides of the equation yields

$$\ln 0.92 = -\lambda t$$

so that

$$-0.0834 = -\lambda t.$$

Rearranging to isolate t gives

$$t = \frac{0.0834}{\lambda}$$

Now, the equation $\lambda = \frac{0.693}{t_{1/2}}$ can be used to find λ

for ^{14}C. Solving for λ and substituting the known half-life gives

$$\lambda = \frac{0.693}{t_{1/2}} = \frac{0.693}{5730\text{ y}}.$$

We enter this value into the previous equation to find t:

$$t = \frac{0.0834}{\frac{0.693}{5730\text{ y}}} = 690\text{ y}$$

Discussion

This dates the material in the shroud to 1988 − 690 = AD 1300. Our calculation is only accurate to two digits, so that the year is rounded to 1300. The values obtained at the three independent laboratories gave a weighted average date of AD 1320 ± 60. The uncertainty is typical of carbon-14 dating and is due to the small amount of ^{14}C in living tissues, the amount of material available, and experimental uncertainties (reduced by having three independent measurements). It is meaningful that the date of the shroud is consistent with the first record of its existence and inconsistent with the period in which Jesus lived.

There are other forms of radioactive dating. Rocks, for example, can sometimes be dated based on the decay of ^{238}U. The decay series for ^{238}U ends with ^{206}Pb, so that the ratio of these nuclides in a rock is an indication of how long it has been since the rock solidified. The original composition of the rock, such as the absence of lead, must be known with some confidence. However, as with carbon-14 dating, the technique can be verified by a consistent body of knowledge. Since ^{238}U has a half-life of 4.5×10^9 y, it is useful for dating only very old materials, showing, for example, that the oldest rocks on Earth solidified about 3.5×10^9 years ago.

Activity, the Rate of Decay

What do we mean when we say a source is highly radioactive? Generally, this means the number of decays per unit time is very high. We define *activity* R to be the *rate of decay* expressed in decays per unit time. In equation form, this is

$$R = \frac{\Delta N}{\Delta t}$$

where ΔN is the number of decays that occur in time Δt. The SI unit for activity is one decay per second and is given the name *becquerel* (Bq) in honor of the discoverer of radioactivity. That is, 1 Bq = 1 decay/s.

Activity R is often expressed in other units, such as decays per minute or decays per year. One of the most common units for activity is the *curie* (Ci), defined to be the activity of 1 g of ^{226}Ra, in honor of Marie Curie's work with radium. The definition of curie is 1 Ci = 3.70×10^{10} Bq, or 3.70×10^{10} decays per second. A curie is a large unit of activity, while a becquerel is a relatively small unit. 1 MBq = 100 microcuries (μCi). In countries like Australia and New Zealand that adhere more to SI units, most radioactive sources, such as those used in medical diagnostics or in physics laboratories, are labeled in Bq or megabecquerel (MBq).

Intuitively, you would expect the activity of a source to depend on two things: the amount of the radioactive substance present, and its half-life. The greater the number of radioactive nuclei present in the sample, the more will decay per unit of time. The shorter the half-life, the more decays per unit time, for a given number of nuclei. So activity R should be proportional to the number of radioactive nuclei, N, and inversely proportional to their half-life, $t_{1/2}$. In fact, your intuition is correct. It can be shown that the activity of a source is

$$R = \frac{0.693N}{t_{1/2}}$$

where N is the number of radioactive nuclei present, having half-life $t_{1/2}$. This relationship is useful in a variety of calculations, as the Examples 2 and 3 illustrate.

Example 2. How Great Is the ^{14}C Activity in Living Tissue?

Calculate the activity due to ^{14}C in 1.00 kg of carbon found in a living organism. Express the activity in units of Bq and Ci.

Strategy

To find the activity R using the equation $R = \frac{0.693N}{t_{1/2}}$, we must know N and $t_{1/2}$. The half-life of ^{14}C can be found in Appendix B, and was stated above as 5730 y. To find N, we first find the number of ^{12}C nuclei in 1.00 kg of carbon using the concept of a mole. As indicated, we then multiply by 1.3×10^{-12} (the abundance of ^{14}C in a carbon sample from a living organism) to get the number of ^{14}C nuclei in a living organism.

Solution

One mole of carbon has a mass of 12.0 g, since it is nearly pure ^{12}C. (A mole has a mass in grams equal in magnitude to A found in the periodic table.) Thus the number of carbon nuclei in a kilogram is

$$N\left(^{12}\mathrm{C}\right)=\frac{6.02\times10^{23}\ \mathrm{mol}^{-1}}{12.0\ \mathrm{g/mol}}\times(1000\ \mathrm{g})=5.02\times10^{25}$$

So the number of ^{14}C nuclei in 1 kg of carbon is

$$N(^{14}\mathrm{C}) = (5.02 \times 10^{25})(1.3 \times 10^{-12}) = 6.52 \times 10^{13}.$$

Now the activity R is found using the equation

$$R=\frac{0.693N}{t_{1/2}}.$$

Entering known values gives

$$R=\frac{0.693\left(6.52\times10^{13}\right)}{5730\ \mathrm{y}}=7.89\times10^{9}\ \mathrm{y}^{-1}$$

or 7.89×10^9 decays per year. To convert this to the unit Bq, we simply convert years to seconds. Thus,

$$R=\left(7.89\times10^{9}\ \mathrm{y}^{-1}\right)\frac{1.00\ \mathrm{y}}{3.16\times10^{7}\ \mathrm{s}}=250\ \mathrm{Bq}$$

or 250 decays per second. To express R in curies, we use the definition of a curie,

$$R=\frac{250\ \mathrm{Bq}}{3.7\times10^{10}\ \mathrm{Bq/Ci}}=6.76\times10^{-9}\ \mathrm{Ci}$$

Thus, $R = 6.76$ nCi.

Discussion

Our own bodies contain kilograms of carbon, and it is intriguing to think there are hundreds of ^{14}C decays per second taking place in us. Carbon-14 and other naturally occurring radioactive substances in our bodies contribute to the background radiation we receive. The small number of decays per second found for a kilogram of carbon in this example gives you some idea of how difficult it is to detect ^{14}C in a small sample of material. If there are 250 decays per second in a kilogram, then there are 0.25 decays per second in a gram of carbon in living tissue. To observe this, you must be able to distinguish decays from other forms of radiation, in order to reduce background noise. This becomes more difficult with an old tissue sample, since it contains less ^{14}C, and for samples more than 50 thousand years old, it is impossible.

Human and Medical Applications

Figure 3. The Chernobyl reactor. More than 100 people died soon after its meltdown, and there will be thousands of deaths from radiation-induced cancer in the future. While the accident was due to a series of human errors, the cleanup efforts were heroic. Most of the immediate fatalities were firefighters and reactor personnel. (credit: Elena Filatova)

Human-made (or artificial) radioactivity has been produced for decades and has many uses. Some of these include medical therapy for cancer, medical imaging and diagnostics, and food preservation by irradiation. Many applications as well as the biological effects of radiation are explored in the chapter Medical Applications of Nuclear Physics, but it is clear that radiation is hazardous. A number of tragic examples of this exist, one of the most disastrous being the meltdown and fire at the Chernobyl reactor complex in the Ukraine (see Figure 3).

Several radioactive isotopes were released in huge quantities, contaminating many thousands of square kilometers and directly affecting hundreds of thousands of people. The most significant releases were of ^{131}I, ^{90}Sr, ^{137}Cs, ^{239}Pu, ^{238}U, and ^{235}U. Estimates are that the total amount of radiation released was about 100 million curies.

Example 3. What Mass of ^{137}Cs Escaped Chernobyl?

It is estimated that the Chernobyl disaster released 6.0

MCi of ^{137}Cs into the environment. Calculate the mass of ^{137}Cs released.

Strategy

We can calculate the mass released using Avogadro's number and the concept of a mole if we can first find the number of nuclei N released. Since the activity R is given, and the half-life of ^{137}Cs is found in Appendix B to be 30.2 y, we can use the equation $R = \frac{0.693N}{t_{1/2}}$ to find N.

Solution

Solving the equation $R = \frac{0.693N}{t_{1/2}}$ for N gives

$$N = \frac{Rt_{1/2}}{0.693}$$

Entering the given values yields

$$N = \frac{(6.0\text{ MCi})\,(30.2\text{ y})}{0.693}$$

Converting curies to becquerels and years to seconds, we get

$$\begin{aligned} N &= \frac{\left(6.0\times10^{6}\text{ Ci}\right)\left(3.7\times10^{10}\text{ Bq/Ci}\right)(30.2\text{ y})\left(3.16\times10^{7}\text{ s/y}\right)}{0.693} \\ &= 3.1\times10^{26} \end{aligned}$$

One mole of a nuclide ^{A}X has a mass of A grams, so that one mole of ^{137}Cs has a mass of 137 g. A mole has 6.02 ×10^{23} nuclei. Thus the mass of ^{137}Cs released was

$$\begin{aligned} m &= \left(\frac{137\text{ g}}{6.02\times10^{23}}\right)\left(3.1\times10^{26}\right) = 70\times10^{3}\text{ g} \\ &= 70\text{ kg} \end{aligned}$$

Discussion

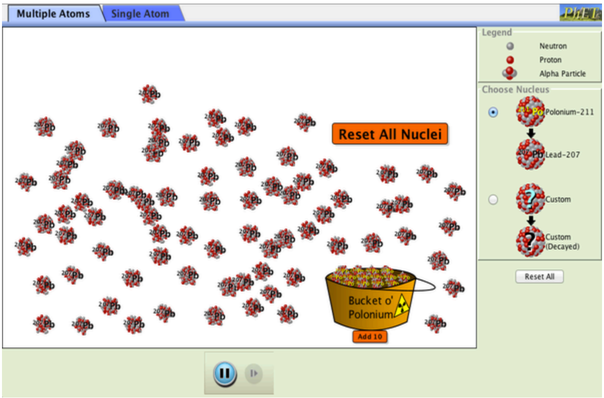
While 70 kg of material may not be a very large mass compared to the amount of fuel in a power plant, it is extremely radioactive, since it only has a 30-year half-life. Six megacuries (6.0 MCi) is an extraordinary amount of activity but is only a fraction of what is produced in nuclear reactors. Similar amounts of the other isotopes were also released at Chernobyl. Although the chances of such a disaster may have seemed small, the consequences were extremely severe, requiring greater caution than was used. More will be said about safe reactor design in the next chapter, but it should be noted that Western reactors have a fundamentally safer design.

Activity R decreases in time, going to half its original value in one half-life, then to one-fourth its original value in the next half-life, and so on. Since $R=\dfrac{0.693N}{t_{1/2}}$, the activity decreases as the number of radioactive nuclei decreases. The equation for R as a function of time is found by combining the equations $N=N_0e^{-\lambda t}$ and $R=\dfrac{0.693N}{t_{1/2}}$, yielding $R=R_0e^{-\lambda t}$, where R_0 is the activity at t=0. This equation shows exponential decay of radioactive nuclei.

For example, if a source originally has a 1.00-mCi activity, it declines to 0.500 mCi in one half-life, to 0.250 mCi in two half-lives, to 0.125 mCi in three half-lives, and so on. For times other than whole half-lives, the equation $R=R_0e^{-\lambda t}$ must be used to find R.

PhET Explorations: Alpha Decay

Watch alpha particles escape from a polonium nucleus, causing radioactive alpha decay. See how random decay times relate to the half life.



Click to download the simulation. Run using Java.

Section Summary

- Half-life $t_{1/2}$ is the time in which there is a 50% chance that a nucleus will decay. The number of nuclei N as a function of time is $N = N_0 e^{-\lambda t}$, where N_0 is the number present at $t = 0$, and λ is the decay constant, related to the half-life by $\lambda = \frac{0.693}{t_{1/2}}$.
- One of the applications of radioactive decay is radioactive dating, in which the age of a material is determined by the amount of radioactive decay that occurs. The rate of decay is called the activity R: $R = \frac{\Delta N}{\Delta t}$.
- The SI unit for R is the becquerel (Bq), defined by 1 Bq = 1 decay/s.
- R is also expressed in terms of curies (Ci), where 1 Ci = 3.70×10^{10} Bq.
- The activity R of a source is related to N and $t_{1/2}$ by $R = \frac{0.693N}{t_{1/2}}$.
- Since N has an exponential behavior as in the equation $N = N_0 e^{-\lambda t}$, the activity also has an exponential behavior, given by $R = R_0 e^{-\lambda t}$, where R_0 is the activity at $t = 0$.

Conceptual Questions

1. In a 3×10^9-year-old rock that originally contained some ^{238}U, which has a half-life of 4.5×10^9 years, we expect to find some ^{238}U remaining in it. Why are

^{226}Ra, ^{222}Rn, and ^{210}Po also found in such a rock, even though they have much shorter half-lives (1600 years, 3.8 days, and 138 days, respectively)?

2. Does the number of radioactive nuclei in a sample decrease to exactly half its original value in one half-life? Explain in terms of the statistical nature of radioactive decay.
3. Radioactivity depends on the nucleus and not the atom or its chemical state. Why, then, is one kilogram of uranium more radioactive than one kilogram of uranium hexafluoride?
4. Explain how a bound system can have less mass than its components. Why is this not observed classically, say for a building made of bricks?
5. Spontaneous radioactive decay occurs only when the decay products have less mass than the parent, and it tends to produce a daughter that is more stable than the parent. Explain how this is related to the fact that more tightly bound nuclei are more stable. (Consider the binding energy per nucleon.)
6. To obtain the most precise value of BE from the equation $\text{BE} = [ZM(^1\text{H}) + Nm_\text{n}]c^2 - m(^A\text{X})c^2$, we should take into account the binding energy of the electrons in the neutral atoms. Will doing this produce a larger or smaller value for BE? Why is this effect usually negligible?
7. How does the finite range of the nuclear force relate to the fact that $\frac{\text{BE}}{A}$ is greatest for A near 60?

Problems & Exercises

Data from the appendices and the periodic table may be needed for these problems.

1. An old campfire is uncovered during an archaeological dig. Its charcoal is found to contain less than 1/1000 the normal amount of ^{14}C. Estimate the minimum age of the charcoal, noting that $2^{10} = 1024$.
2. A ^{60}Co source is labeled 4.00 mCi, but its present activity is found to be 1.85×10^7 Bq. (a) What is the present activity in mCi? (b) How long ago did it actually have a 4.00-mCi activity?
3. (a) Calculate the activity R in curies of 1.00 g of ^{226}Ra. (b) Discuss why your answer is not exactly 1.00 Ci, given that the curie was originally supposed to be exactly the activity of a gram of radium.
4. Show that the activity of the ^{14}C in 1.00 g of ^{12}C found in living tissue is 0.250 Bq.
5. Mantles for gas lanterns contain thorium, because it forms an oxide that can survive being heated to incandescence for long periods of time. Natural thorium is almost 100% ^{232}Th, with a half-life of 1.405×10^{10} y. If an average lantern mantle contains 300 mg of thorium, what is its activity?
6. Cow's milk produced near nuclear reactors can be tested for as little as 1.00 pCi of ^{131}I per liter, to check for possible reactor leakage. What mass of ^{131}I has this activity?
7. (a) Natural potassium contains ^{40}K, which has a

half-life of 1.277×10^9 y. What mass of ^{40}K in a person would have a decay rate of 4140 Bq? (b) What is the fraction of ^{40}K in natural potassium, given that the person has 140 g in his body? (These numbers are typical for a 70-kg adult.)

8. There is more than one isotope of natural uranium. If a researcher isolates 1.00 mg of the relatively scarce ^{235}U and finds this mass to have an activity of 80.0 Bq, what is its half-life in years?
9. ^{50}V has one of the longest known radioactive half-lives. In a difficult experiment, a researcher found that the activity of 1.00 kg of ^{50}V is 1.75 Bq. What is the half-life in years?
10. You can sometimes find deep red crystal vases in antique stores, called uranium glass because their color was produced by doping the glass with uranium. Look up the natural isotopes of uranium and their half-lives, and calculate the activity of such a vase assuming it has 2.00 g of uranium in it. Neglect the activity of any daughter nuclides.
11. A tree falls in a forest. How many years must pass before the ^{14}C activity in 1.00 g of the tree's carbon drops to 1.00 decay per hour?
12. What fraction of the ^{40}K that was on Earth when it formed 4.5×10^9 years ago is left today?
13. A 5000-Ci ^{60}Co source used for cancer therapy is considered too weak to be useful when its activity falls to 3500 Ci. How long after its manufacture does this happen?
14. Natural uranium is 0.7200% ^{235}U and 99.27% ^{238}U. What were the percentages of ^{235}U and ^{238}U in natural uranium when Earth formed 4.5×10^9 years

ago?

15. β^- particles emitted in the decay of ^{3}H (tritium) interact with matter to create light in a glow-in-the-dark exit sign. At the time of manufacture, such a sign contains 15.0 Ci of ^{3}H. (a) What is the mass of the tritium? (b) What is its activity 5.00 y after manufacture?
16. World War II aircraft had instruments with glowing radium-painted dials (see [link]). The activity of one such instrument was 1.0×10^5 Bq when new. (a) What mass of ^{226}Ra was present? (b) After some years, the phosphors on the dials deteriorated chemically, but the radium did not escape. What is the activity of this instrument 57.0 years after it was made?
17. (a) The ^{210}Po source used in a physics laboratory is labeled as having an activity of 1.0 μCi on the date it was prepared. A student measures the radioactivity of this source with a Geiger counter and observes 1500 counts per minute. She notices that the source was prepared 120 days before her lab. What fraction of the decays is she observing with her apparatus? (b) Identify some of the reasons that only a fraction of the αs emitted are observed by the detector.
18. Armor-piercing shells with depleted uranium cores are fired by aircraft at tanks. (The high density of the uranium makes them effective.) The uranium is called depleted because it has had its ^{235}U removed for reactor use and is nearly pure ^{238}U. Depleted uranium has been erroneously called non-radioactive. To demonstrate that this is wrong: (a) Calculate the activity of 60.0 g of pure ^{238}U. (b) Calculate the activity of 60.0 g of natural uranium, neglecting the

^{234}U and all daughter nuclides.

19. The ceramic glaze on a red-orange Fiestaware plate is U_2O_3 and contains 50.0 grams of ^{238}U, but very little ^{235}U. (a) What is the activity of the plate? (b) Calculate the total energy that will be released by the ^{238}U decay. (c) If energy is worth 12.0 cents per $kW \cdot h$, what is the monetary value of the energy emitted? (These plates went out of production some 30 years ago, but are still available as collectibles.)
20. Large amounts of depleted uranium (^{238}U) are available as a by-product of uranium processing for reactor fuel and weapons. Uranium is very dense and makes good counter weights for aircraft. Suppose you have a 4000-kg block of ^{238}U. (a) Find its activity. (b) How many calories per day are generated by thermalization of the decay energy? (c) Do you think you could detect this as heat? Explain.
21. The *Galileo* space probe was launched on its long journey past several planets in 1989, with an ultimate goal of Jupiter. Its power source is 11.0 kg of ^{238}Pu, a by-product of nuclear weapons plutonium production. Electrical energy is generated thermoelectrically from the heat produced when the 5.59-MeVα particles emitted in each decay crash to a halt inside the plutonium and its shielding. The half-life of ^{238}Pu is 87.7 years. (a) What was the original activity of the ^{238}Pu in becquerel? (b) What power was emitted in kilowatts? (c) What power was emitted 12.0 y after launch? You may neglect any extra energy from daughter nuclides and any losses from escaping γ rays.
22. **Construct Your Own Problem.** Consider the

generation of electricity by a radioactive isotope in a space probe, such as described in Question 21. Construct a problem in which you calculate the mass of a radioactive isotope you need in order to supply power for a long space flight. Among the things to consider are the isotope chosen, its half-life and decay energy, the power needs of the probe and the length of the flight.

23. **Unreasonable Results.** A nuclear physicist finds 1.0 μg of ^{236}U in a piece of uranium ore and assumes it is primordial since its half-life is 2.3×10^7 y. (a) Calculate the amount of ^{236}U that would had to have been on Earth when it formed 4.5×10^9 y ago for 1.0 μg to be left today. (b) What is unreasonable about this result? (c) What assumption is responsible?
24. **Unreasonable Results.** (a) Repeat Question 14 but include the 0.0055% natural abundance of ^{234}U with its 2.45×10^5 y half-life. (b) What is unreasonable about this result? (c) What assumption is responsible? (d) Where does the ^{234}U come from if it is not primordial?
25. **Unreasonable Results.** The manufacturer of a smoke alarm decides that the smallest current of α radiation he can detect is 1.00 μA. (a) Find the activity in curies of an α emitter that produces a 1.00 μA current of α particles. (b) What is unreasonable about this result? (c) What assumption is responsible?

Glossary

becquerel: SI unit for rate of decay of a radioactive material

half-life: the time in which there is a 50% chance that a nucleus will decay

radioactive dating: an application of radioactive decay in which the age of a material is determined by the amount of radioactivity of a particular type that occurs

decay constant: quantity that is inversely proportional to the half-life and that is used in equation for number of nuclei as a function of time

carbon-14 dating: a radioactive dating technique based on the radioactivity of carbon-14

activity: the rate of decay for radioactive nuclides

rate of decay: the number of radioactive events per unit time

curie: the activity of 1g of ^{226}Ra, equal to 3.70×10^{10} Bq

Selected Solutions to Problems & Exercises

1. 57,300 y

3. (a) 0.988 Ci; (b) The half-life of ^{226}Ra is now better known.

5. 1.22×10^{3} Bq

7. (a) 16.0 mg; (b) 0.0114%

9. 1.48×10^{17} y

11. 5.6×10^{4} y

13. 2.71 y

15. (a) 1.56 mg; (b) 11.3 Ci

17. (a) 1.23×10^{-3}; (b) Only part of the emitted radiation goes in the direction of the detector. Only a fraction of that causes a response in the detector. Some of the emitted radiation (mostly α particles) is observed within the source. Some is absorbed within the source, some is absorbed by the detector, and some does not penetrate the detector.

19. (a) 1.68×10^{-5} Ci; (b) 8.65×10^{10} J; (c) $\$2.9 \times 10^{3}$

21. (a) 6.97×10^{15} Bq; (b) 6.24 kW; (c) 5.67 kW

25. (a) 84.5 Ci; (b) An extremely large activity, many orders of magnitude greater than permitted for home use; (c) The assumption of 1.00 μA is unreasonably large. Other methods can detect much smaller decay rates.

114. Binding Energy

Learning Objectives

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

- Define and discuss binding energy.
- Calculate the binding energy per nucleon of a particle.

The more tightly bound a system is, the stronger the forces that hold it together and the greater the energy required to pull it apart. We can therefore learn about nuclear forces by examining how tightly bound the nuclei are. We define the *binding energy* (BE) of a nucleus to be *the energy required to completely disassemble it into separate protons and neutrons*. We can determine the BE of a nucleus from its rest mass. The two are connected through Einstein's famous relationship $E = (\Delta m)c^2$. A bound system has a *smaller* mass than its separate constituents; the more tightly the nucleons are bound together, the smaller the mass of the nucleus.

Imagine pulling a nuclide apart as illustrated in Figure 1. Work done to overcome the nuclear forces holding the nucleus together puts energy into the system. By definition, the energy input equals the binding energy BE. The pieces are at rest when separated, and so the energy put into them increases their total rest mass compared with what it was when they were glued together as a nucleus. That mass increase is thus $\Delta m = \text{BE}/c^2$. This difference in mass is known as *mass defect*. It implies that the mass of the nucleus is less than the sum of the masses of its constituent protons and neutrons. A nuclide ^{A}X has Z protons and N neutrons, so that the difference in mass is $\Delta m = (Zm_p + Nm_n) - m_{tot}$.Thus, $BE = (\Delta m)c^2 = [(Zm_p + Nm_n) - m_{tot}]c^2$, where m_{tot} is the mass of the nuclide ${}^{A}X$, m_p is the mass of a proton, and m_n is the mass of a neutron. Traditionally, we deal with the masses of neutral atoms. To get atomic masses into the last equation, we first add Z electrons to m_{tot}, which gives $m({}^{A}X)$, the atomic mass of the nuclide. We then add Z electrons to the Z protons, which gives $Zm({}^{1}H)$, or Z times the mass of a hydrogen atom. Thus the binding energy of a nuclide ${}^{A}X$ is $BE = \{[Zm({}^{1}H) + Nm_n] - m({}^{A}X)\}c^2$.

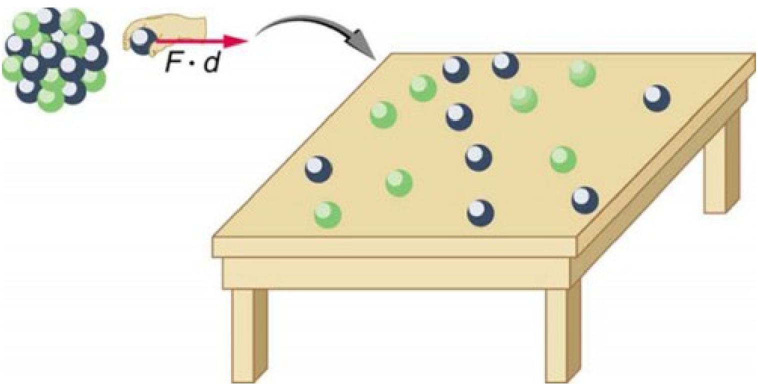


Figure 1. Work done to pull a nucleus apart into its constituent protons and neutrons increases the mass of the system. The work to disassemble the nucleus equals its binding energy BE. A bound system has less mass than the sum of its parts, especially noticeable in the nuclei, where forces and energies are very large.

The atomic masses can be found in [Appendix A](), most conveniently expressed in unified atomic mass units u ($1\ u = 931.5\ MeV/c^2$). BE is thus calculated from known atomic masses.

Things Great and Small

Nuclear Decay Helps Explain Earth's Hot Interior

A puzzle created by radioactive dating of rocks is resolved by radioactive heating of Earth's interior. This intriguing story is another example of how small-scale physics can explain large-scale phenomena.

Radioactive dating plays a role in determining the approximate age of the Earth. The oldest rocks on Earth solidified about 3.5×10^9 years ago—a number determined by uranium-238 dating. These rocks could only have solidified once the surface of the Earth had cooled sufficiently. The temperature of the Earth at formation can be estimated based on gravitational potential energy of the assemblage of pieces being converted to thermal energy. Using heat transfer concepts discussed in Thermodynamics it is then possible to calculate how long it would take for the surface to cool to rock-formation temperatures. The result is about 10^9 years. The first rocks formed have been solid for 3.5×10^9 years, so that the age of the Earth is approximately 4.5×10^9 years. There is a large body of other types of evidence (both Earth-bound and solar system characteristics are used) that supports this age. The puzzle is that, given its age and initial temperature, the center of the Earth should be much cooler than it is today (see Figure 2).

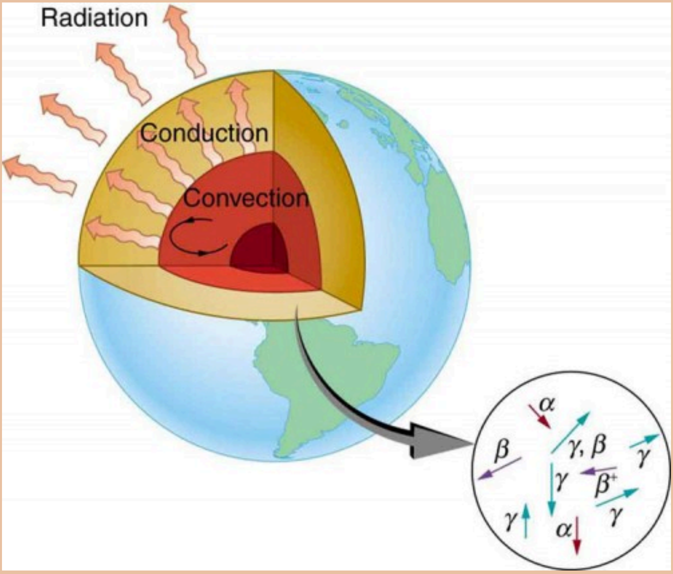


Figure 2. The center of the Earth cools by well-known heat transfer methods. Convection in the liquid regions and conduction move thermal energy to the surface, where it radiates into cold, dark space. Given the age of the Earth and its initial temperature, it should have cooled to a lower temperature by now. The blowup shows that nuclear decay releases energy in the Earth's interior. This energy has slowed the cooling process and is responsible for the interior still being molten.

We know from seismic waves produced by earthquakes that parts of the interior of the Earth are liquid. Shear or transverse waves cannot travel through a liquid and are not transmitted through the Earth's core. Yet compression or longitudinal waves can pass through a liquid and do go through the core. From this information, the temperature of the interior can be estimated. As noticed, the interior should have cooled more from its initial temperature in the

4.5×10^9 years since its formation. In fact, it should have taken no more than about 10^9 years to cool to its present temperature. What is keeping it hot? The answer seems to be radioactive decay of primordial elements that were part of the material that formed the Earth (see the blowup in Figure 2).

Nuclides such as ^{238}U and ^{40}K have half-lives similar to or longer than the age of the Earth, and their decay still contributes energy to the interior. Some of the primordial radioactive nuclides have unstable decay products that also release energy—^{238}U has a long decay chain of these. Further, there were more of these primordial radioactive nuclides early in the life of the Earth, and thus the activity and energy contributed were greater then (perhaps by an order of magnitude). The amount of power created by these decays per cubic meter is very small. However, since a huge volume of material lies deep below the surface, this relatively small amount of energy cannot escape quickly. The power produced near the surface has much less distance to go to escape and has a negligible effect on surface temperatures.

A final effect of this trapped radiation merits mention. Alpha decay produces helium nuclei, which form helium atoms when they are stopped and capture electrons. Most of the helium on Earth is obtained from wells and is produced in this manner. Any helium in the atmosphere will escape in geologically short times because of its high thermal velocity.

What patterns and insights are gained from an examination of the binding energy of various nuclides? First, we find that BE is approximately proportional to the number of nucleons A in any nucleus. About twice as much energy is needed to pull apart a nucleus like ^{24}Mg compared with pulling apart ^{12}C, for example. To help us look at other effects, we divide BE by A and consider the *binding energy per nucleon*, BE/A. The graph of BE/A in Figure 3 reveals some very interesting aspects of nuclei. We see that the binding energy per nucleon averages about 8 MeV, but is lower for both the lightest and heaviest nuclei. This overall trend, in which nuclei with A equal to about 60 have the greatest BE/A and are thus the most tightly bound, is due to the combined characteristics of the attractive nuclear forces and the repulsive Coulomb force. It is especially important to note two things—the strong nuclear force is about 100 times stronger than the Coulomb force, *and* the nuclear forces are shorter in range compared to the Coulomb force. So, for low-mass nuclei, the nuclear attraction dominates and each added nucleon forms bonds with all others, causing progressively heavier nuclei to have progressively greater values of BE/A. This continues up to A ≈ 60, roughly corresponding to the mass number of iron. Beyond that, new nucleons added to a nucleus will be too far from some others to feel their nuclear attraction. Added protons, however, feel the repulsion of all other protons, since the Coulomb force is longer in range. Coulomb repulsion grows for progressively heavier nuclei, but nuclear attraction remains about the same, and so BE/A becomes smaller. This is why stable nuclei heavier than A ≈ 40 have more neutrons than protons. Coulomb repulsion is reduced by having more neutrons to keep the protons farther apart (see Figure 4).

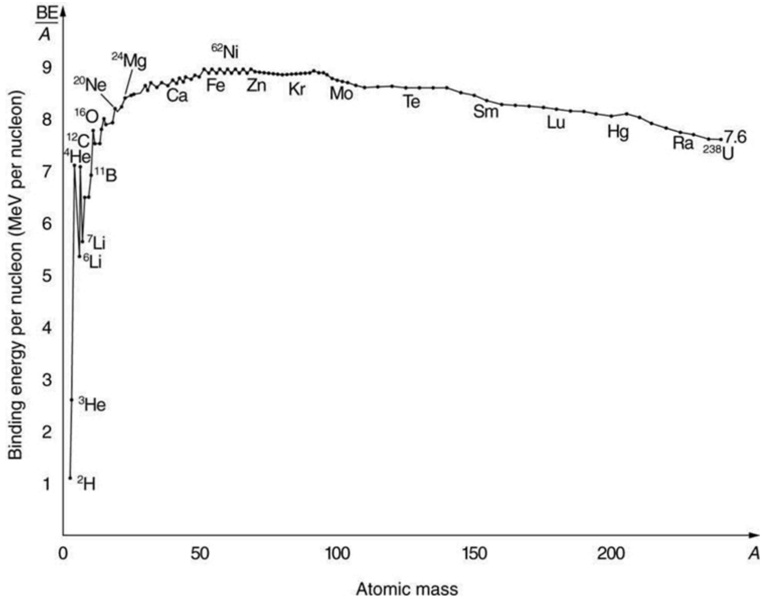


Figure 3. A graph of average binding energy per nucleon, BE/A, *for stable nuclei. The most tightly bound nuclei are those with* A *near* 60, *where the attractive nuclear force has its greatest effect. At higher* As, *the Coulomb repulsion progressively reduces the binding energy per nucleon, because the nuclear force is short ranged. The spikes on the curve are very tightly bound nuclides and indicate shell closures.*

There are some noticeable spikes on the BE/A graph, which represent particularly tightly bound nuclei. These spikes reveal further details of nuclear forces, such as confirming that closed-shell nuclei (those with magic numbers of protons or neutrons or both) are more tightly bound. The spikes also indicate that some nuclei with even numbers for Z and N, and with Z=N, are exceptionally tightly bound. This finding can be correlated with some of the cosmic abundances of the elements.

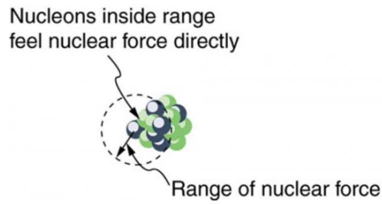


Figure 4. The nuclear force is attractive and stronger than the Coulomb force, but it is short ranged. In low-mass nuclei, each nucleon feels the nuclear attraction of all others. In larger nuclei, the range of the nuclear force, shown for a single nucleon, is smaller than the size of the nucleus, but the Coulomb repulsion from all protons reaches all others. If the nucleus is large enough, the Coulomb repulsion can add to overcome the nuclear attraction.

The most common elements in the universe, as determined by observations of atomic spectra from outer space, are hydrogen, followed by ^{4}He, with much smaller amounts of ^{12}C and other elements. It should be noted that the heavier elements are created in supernova explosions, while the lighter ones are produced by nuclear fusion during the normal life cycles of stars, as will be discussed in subsequent chapters. The most common elements have the most tightly bound nuclei. It is also no accident that one of the most tightly bound light nuclei is ^{4}He, emitted in α decay.

Example 1. What Is BE/A for an Alpha Particle?

Calculate the binding energy per nucleon of ^{4}He, the α particle.

Strategy

To find BE/A, we first find BE using the equation

$$\text{BE} = \{[Zm(^1\text{H}) + Nm_\text{n}] - m(^A\text{X})\}c^2$$

and then divide by A. This is straightforward once we have looked up the appropriate atomic masses in Appendix A.

Solution

The binding energy for a nucleus is given by the equation

$$\text{BE} = \{[Zm(^1\text{H}) + Nm_\text{n}] - m(^A\text{X})\}c^2.$$

For ^{4}He, we have

$$Z = N = 2;\ \text{thus, BE} = \{[2m(^1\text{H}) + 2m_n] - m(^4\text{He})\}c^2.$$

Appendix A gives these masses as $m\,(^4\text{He}) = 4.002602$ u, $m(^1\text{H}) = 1.007825$ u, and $m_n = 1.008665$ u. Thus,

$$\text{BE} = (0.030378\ \text{u})c^2.$$

Noting that 1 u=931.5 MeV/c^2, we find

$$\text{BE} = (0.030378)(931.5\ \text{MeV}/c^2)c^2 = 28.3\ \text{MeV}.$$

Since A = 4, we see that BE/A is this number divided by 4, or

$$\text{BE/A} = 7.07\ \text{MeV/nucleon}.$$

Discussion

This is a large binding energy per nucleon compared with those for other low-mass nuclei, which have BE/A ≈ 3 MeV/nucleon. This indicates that ^{4}He is tightly bound compared with its neighbors on the chart of the nuclides. You can see the spike representing this value of BE/A for ^{4}He on the graph in Figure 3. This is why ^{4}He is stable. Since ^{4}He is tightly bound, it has less mass than other A = 4 nuclei and, therefore, cannot spontaneously decay into them. The large binding energy also helps to explain why some nuclei undergo α decay. Smaller mass in the decay products can mean energy release, and such decays can be spontaneous. Further, it can happen that two protons and two neutrons in a nucleus can randomly find themselves together, experience the exceptionally large nuclear force that binds this combination, and act as a ^{4}He unit within the nucleus, at least for a while. In some cases, the ^{4}He escapes, and α decay has then taken place.

There is more to be learned from nuclear binding energies. The general trend in BE/A is fundamental to energy production in stars, and to fusion and fission energy sources on Earth, for example. This is one of the applications of nuclear physics covered in Medical Applications of Nuclear Physics. The abundance of elements on Earth, in stars, and in the universe as a whole is related to the binding energy of nuclei and has implications for the continued expansion of the universe.

Problem-Solving Strategies

For Reaction And Binding Energies and Activity Calculations in Nuclear Physics

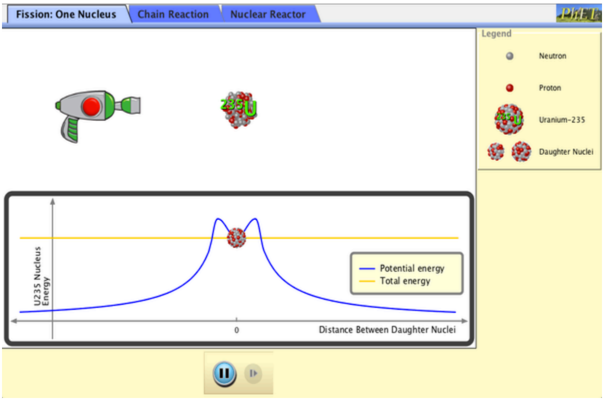
1. *Identify exactly what needs to be determined in the problem (identify the unknowns).* This will allow you to decide whether the energy of a decay or nuclear reaction is involved, for example, or whether the problem is primarily concerned with activity (rate of decay).
2. *Make a list of what is given or can be inferred from the problem as stated (identify the knowns).*
3. *For reaction and binding-energy problems, we use atomic rather than nuclear masses.* Since the masses of neutral atoms are used, you must count the number of electrons involved. If these do not balance (such as in β^+ decay), then an energy adjustment of 0.511 MeV per electron must be made. Also note that atomic masses may not be given in a problem; they can be found in tables.
4. *For problems involving activity, the relationship of activity to half-life, and the number of nuclei given in the equation* $R = \frac{0.693N}{t_{1/2}}$ *can be very useful.* Owing to the fact that number of nuclei is involved, you will also need to be familiar with moles and Avogadro's number.
5. *Perform the desired calculation; keep careful track of*

plus and minus signs as well as powers of 10.

6. *Check the answer to see if it is reasonable: Does it make sense?* Compare your results with worked examples and other information in the text. (Heeding the advice in Step 5 will also help you to be certain of your result.) You must understand the problem conceptually to be able to determine whether the numerical result is reasonable.

PhET Explorations: Nuclear Fission

Start a chain reaction, or introduce non-radioactive isotopes to prevent one. Control energy production in a nuclear reactor! Click on the image to download the simulation.



Click to download the simulation. Run using Java.

Section Summary

- The binding energy (BE) of a nucleus is the energy needed to separate it into individual protons and neutrons. In terms of atomic masses, $BE = \{[Zm(^1H) + Nm_n] - m(^AX)\}c^2$, where $m(^1H)$ is the mass of a hydrogen atom, $m(^AX)$ is the atomic mass of the nuclide, and m_n is the mass of a neutron. Patterns in the binding energy per nucleon, BE/A, reveal details of the nuclear force. The larger the BE/A, the more stable the nucleus.

Conceptual Questions

1. Why is the number of neutrons greater than the number of protons in stable nuclei having A greater than about 40, and why is this effect more pronounced for the heaviest nuclei?

Problems & Exercises

1. ^{2}H is a loosely bound isotope of hydrogen. Called deuterium or heavy hydrogen, it is stable but relatively rare—it is 0.015% of natural hydrogen. Note that deuterium has Z = N, which should tend to make it more tightly bound, but both are odd numbers. Calculate BE/A, the binding energy per nucleon, for ^{2}H and compare it with the approximate value obtained from the graph in Figure 3.
2. ^{56}Fe is among the most tightly bound of all nuclides. It is more than 90% of natural iron. Note that ^{56}Fe has even numbers of both protons and neutrons. Calculate BE/A, the binding energy per nucleon, for ^{56}Fe and compare it with the approximate value obtained from the graph in Figure 3.
3. ^{209}Bi is the heaviest stable nuclide, and its BE/A is low compared with medium-mass nuclides. Calculate BE/A, the binding energy per nucleon, for ^{209}Bi and

compare it with the approximate value obtained from the graph in [link].

4. (a) Calculate BE/A for ^{235}U, the rarer of the two most common uranium isotopes. (b) Calculate BE/A for ^{238}U. (Most of uranium is ^{238}U.) Note that ^{238}U has even numbers of both protons and neutrons. Is the BE/A of ^{238}U significantly different from that of ^{235}U?
5. (a) Calculate BE/A for ^{12}C. Stable and relatively tightly bound, this nuclide is most of natural carbon. (b) Calculate BE/A for ^{14}C. Is the difference in BE/A between ^{12}C and ^{14}C significant? One is stable and common, and the other is unstable and rare.
6. The fact that BE/A is greatest for A near 60 implies that the range of the nuclear force is about the diameter of such nuclides. (a) Calculate the diameter of an A = 60 nucleus. (b) Compare BE/A for ^{58}Ni and ^{90}Sr. The first is one of the most tightly bound nuclides, while the second is larger and less tightly bound.
7. The purpose of this problem is to show in three ways that the binding energy of the electron in a hydrogen atom is negligible compared with the masses of the proton and electron. (a) Calculate the mass equivalent in u of the 13.6-eV binding energy of an electron in a hydrogen atom, and compare this with the mass of the hydrogen atom obtained from Appendix A. (b) Subtract the mass of the proton given in Table 1 in Substructure of the Nucleus from the mass of the hydrogen atom given in Appendix A. You will find the difference is equal to the electron's mass to three digits, implying the binding energy is small in comparison. (c) Take the ratio of the binding energy

of the electron (13.6 eV) to the energy equivalent of the electron's mass (0.511 MeV). (d) Discuss how your answers confirm the stated purpose of this problem.

8. **Unreasonable Results.** A particle physicist discovers a neutral particle with a mass of 2.02733 u that he assumes is two neutrons bound together. (a) Find the binding energy. (b) What is unreasonable about this result? (c) What assumptions are unreasonable or inconsistent?

Glossary

binding energy: the energy needed to separate nucleus into individual protons and neutrons

binding energy per nucleon: the binding energy calculated per nucleon; it reveals the details of the nuclear force—larger the BE/A, the more stable the nucleus

Selected Solutions to Problems & Exercises

1. 1.112 MeV, consistent with graph

3. 7.848 MeV, consistent with graph

5. (a) 7.680 MeV, consistent with graph; (b) 7.520 MeV, consistent with graph. Not significantly different from value

for ^{12}C, but sufficiently lower to allow decay into another nuclide that is more tightly bound

7. (a) 1.46×10^{-8} u vs. 1.007825 u for ^{1}H; (b) 0.000549 u; (c) 2.66×10^{-5}

8. (a) −9.315 MeV; (b) The negative binding energy implies an unbound system; (c) This assumption that it is two bound neutrons is incorrect.

PART XVI

APPENDICES

115. Appendix A. Atomic Masses

Table 1. Atomic Masses

Atomic Number, Z	Name	Atomic Mass Number, A	Symbol	Atomic Mass (u)	Percent Abundance or Decay Mode	Half-life, $t_{1/2}$
0	neutron	1	n	1.008 665	β^-	10.37 min
1	Hydrogen	1	^1H	1.007 825	99.985%	
	Deuterium	2	^2H or D	2.014 102	0.015%	
	Tritium	3	^3H or T	3.016 050	β^-	12.33 y
2	Helium	3	^3He	3.016 030	1.38\times 10^{-4}%	
		4	^4He	4.002 603	\approx 100%	
3	Lithium	6	^6Li	6.015 121	7.5%	
		7	^7Li	7.016 003	92.5%	
4	Beryllium	7	^7Be	7.016 928	EC	53.29 d
		9	^9Be	9.012 182	100%	
5	Boron	10	^{10}B	10.012 937	19.9%	
		11	^{11}B	11.009 305	80.1%	
6	Carbon	11	^{11}C	11.011 432	EC, β^+	
		12	^{12}C	12.000 000	98.90%	
		13	^{13}C	13.003 355	1.10%	
		14	^{14}C	14.003 241	β^-	5730 y
7	Nitrogen	13	^{13}N	13.005 738	β^+	9.96 min

Atomic Number, Z	Name	Atomic Mass Number, A	Symbol	Atomic Mass (u)	Percent Abundance or Decay Mode	Half-life, $t_{1/2}$
		14	^{14}N	14.003 074	99.63%	
		15	^{15}N	15.000 108	0.37%	
8	Oxygen	15	^{15}O	15.003 065	EC, β^+	122 s
		16	^{16}O	15.994 915	99.76%	
		18	^{18}O	17.999 160	0.200%	
9	Fluorine	18	^{18}F	18.000 937	EC, β^+	1.83 h
		19	^{19}F	18.998 403	100%	
10	Neon	20	^{20}Ne	19.992 435	90.51%	
		22	^{22}Ne	21.991 383	9.22%	
11	Sodium	22	^{22}Na	21.994 434	β^+	2.602 y
		23	^{23}Na	22.989 767	100%	
		24	^{24}Na	23.990 961	β^-	14.96 h
12	Magnesium	24	^{24}Mg	23.985 042	78.99%	
13	Aluminum	27	^{27}Al	26.981 539	100%	
14	Silicon	28	^{28}Si	27.976 927	92.23%	2.62h
		31	^{31}Si	30.975 362	β^-	
15	Phosphorus	31	^{31}P	30.973 762	100%	
		32	^{32}P	31.973 907	β^-	14.28 d

Atomic Number, Z	Name	Atomic Mass Number, A	Symbol	Atomic Mass (u)	Percent Abundance or Decay Mode	Half-life, $t_{1/2}$
16	Sulfur	32	^{32}S	31.972 070	95.02%	
		35	^{35}S	34.969 031	β^{-}	87.4 d
17	Chlorine	35	^{35}Cl	34.968 852	75.77%	
		37	^{37}Cl	36.965 903	24.23%	
18	Argon	40	^{40}Ar	39.962 384	99.60%	
19	Potassium	39	^{39}K	38.963 707	93.26%	
		40	^{40}K	39.963 999	0.0117%, EC, β^{-}	1.28×10
20	Calcium	40	^{40}Ca	39.962 591	96.94%	
21	Scandium	45	^{45}Sc	44.955 910	100%	
22	Titanium	48	^{48}Ti	47.947 947	73.8%	
23	Vanadium	51	^{51}V	50.943 962	99.75%	
24	Chromium	52	^{52}Cr	51.940 509	83.79%	
25	Manganese	55	^{55}Mn	54.938 047	100%	
26	Iron	56	^{56}Fe	55.934 939	91.72%	
27	Cobalt	59	^{59}Co	58.933 198	100%	
		60	^{60}Co	59.933 819	β^{-}	5.271 y
28	Nickel	58	^{58}Ni	57.935 346	68.27%	

Atomic Number, Z	Name	Atomic Mass Number, A	Symbol	Atomic Mass (u)	Percent Abundance or Decay Mode	Half-life, $t_{1/2}$
		60	^{60}Ni	59.930 788	26.10%	
29	Copper	63	^{63}Cu	62.939 598	69.17%	
		65	^{65}Cu	64.927 793	30.83%	
30	Zinc	64	^{64}Zn	63.929 145	48.6%	
		66	^{66}Zn	65.926 034	27.9%	
31	Gallium	69	^{69}Ga	68.925 580	60.1%	
32	Germanium	72	^{72}Ge	71.922 079	27.4%	
		74	^{74}Ge	73.921 177	36.5%	
33	Arsenic	75	^{75}As	74.921 594	100%	
34	Selenium	80	^{80}Se	79.916 520	49.7%	
35	Bromine	79	^{79}Br	78.918 336	50.69%	
36	Krypton	84	^{84}Kr	83.911 507	57.0%	
37	Rubidium	85	^{85}Rb	84.911 794	72.17%	
38	Strontium	86	^{86}Sr	85.909 267	9.86%	
		88	^{88}Sr	87.905 619	82.58%	
		90	^{90}Sr	89.907 738	β^-	28.8 y
39	Yttrium	89	^{89}Y	88.905 849	100%	
		90	^{90}Y	89.907 152	β^-	64.1 h

Atomic Number, Z	Name	Atomic Mass Number, A	Symbol	Atomic Mass (u)	Percent Abundance or Decay Mode	Half-life, $t_{1/2}$
40	Zirconium	90	^{90}Zr	89.904 703	51.45%	
41	Niobium	93	^{93}Nb	92.906 377	100%	
42	Molybdenum	98	^{98}Mo	97.905 406	24.13%	
43	Technetium	98	^{98}Tc	97.907 215	β^-	4.2×10^6
44	Ruthenium	102	^{102}Ru	101.904 348	31.6%	
45	Rhodium	103	^{103}Rh	102.905 500	100%	
46	Palladium	106	^{106}Pd	105.903 478	27.33%	
47	Silver	107	^{107}Ag	106.905 092	51.84%	
		109	^{109}Ag	108.904 757	48.16%	
48	Cadmium	114	^{114}Cd	113.903 357	28.73%	
49	Indium	115	^{115}In	114.903 880	95.7%, β^-	4.4×10^1
50	Tin	120	^{120}Sn	119.902 200	32.59%	
51	Antimony	121	^{121}Sb	120.903 821	57.3%	
52	Tellurium	130	^{130}Te	129.906 229	33.8%, β^-	2.5×10^2
53	Iodine	127	^{127}I	126.904 473	100%	
		131	^{131}I	130.906 114	β^-	8.040 d
54	Xenon	132	^{132}Xe	131.904 144	26.9%	

Atomic Number, Z	Name	Atomic Mass Number, A	Symbol	Atomic Mass (u)	Percent Abundance or Decay Mode	Half-life, $t_{1/2}$
		136	^{136}Xe	135.907 214	8.9%	
55	Cesium	133	^{133}Cs	132.905 429	100%	
		134	^{134}Cs	133.906 696	EC, β^{-}	2.06 y
56	Barium	137	^{137}Ba	136.905 812	11.23%	
		138	^{138}Ba	137.905 232	71.70%	
57	Lanthanum	139	^{139}La	138.906 346	99.91%	
58	Cerium	140	^{140}Ce	139.905 433	88.48%	
59	Praseodymium	141	^{141}Pr	140.907 647	100%	
60	Neodymium	142	^{142}Nd	141.907 719	27.13%	
61	Promethium	145	^{145}Pm	144.912 743	EC, α	17.7 y
62	Samarium	152	^{152}Sm	151.919 729	26.7%	
63	Europium	153	^{153}Eu	152.921 225	52.2%	
64	Gadolinium	158	^{158}Gd	157.924 099	24.84%	
65	Terbium	159	^{159}Tb	158.925 342	100%	
66	Dysprosium	164	^{164}Dy	163.929 171	28.2%	
67	Holmium	165	^{165}Ho	164.930 319	100%	
68	Erbium	166	^{166}Er	165.930 290	33.6%	
69	Thulium	169	^{169}Tm	168.934 212	100%	

Atomic Number, Z	Name	Atomic Mass Number, A	Symbol	Atomic Mass (u)	Percent Abundance or Decay Mode	Half-life, $t_{1/2}$
70	Ytterbium	174	^{174}Yb	173.938 859	31.8%	
71	Lutecium	175	^{175}Lu	174.940 770	97.41%	
72	Hafnium	180	^{180}Hf	179.946 545	35.10%	
73	Tantalum	181	^{181}Ta	180.947 992	99.98%	
74	Tungsten	184	^{184}W	183.950 928	30.67%	
75	Rhenium	187	^{187}Re	186.955 744	62.6%, β^-	4.6×10^1
76	Osmium	191	^{191}Os	190.960 920	β^-	15.4 d
		192	^{192}Os	191.961 467	41.0%	
77	Iridium	191	^{191}Ir	190.960 584	37.3%	
		193	^{193}Ir	192.962 917	62.7%	
78	Platinum	195	^{195}Pt	194.964 766	33.8%	
79	Gold	197	^{197}Au	196.966 543	100%	
		198	^{198}Au	197.968 217	β^-	2.696 d
80	Mercury	199	^{199}Hg	198.968 253	16.87%	
		202	^{202}Hg	201.970 617	29.86%	
81	Thallium	205	^{205}Tl	204.974 401	70.48%	
82	Lead	206	^{206}Pb	205.974 440	24.1%	

Atomic Number, Z	Name	Atomic Mass Number, A	Symbol	Atomic Mass (u)	Percent Abundance or Decay Mode	Half-life, $t_{1/2}$
		207	^{207}Pb	206.975 872	22.1%	
		208	^{208}Pb	207.976 627	52.4%	
		210	^{210}Pb	209.984 163	α, β^-	22.3 y
		211	^{211}Pb	210.988 735	β^-	36.1 min
		212	^{212}Pb	211.991 871	β^-	10.64 h
83	Bismuth	209	^{209}Bi	208.980 374	100%	
		211	^{211}Bi	210.987 255	α, β^-	2.14 min
84	Polonium	210	^{210}Po	209.982 848	α	138.38 d
85	Astatine	218	^{218}At	218.008 684	α, β^-	1.6 s
86	Radon	222	^{222}Rn	222.017 570	α	3.82 d
87	Francium	223	^{223}Fr	223.019 733	α, β^-	21.8 min
88	Radium	226	^{226}Ra	226.025 402	α	1.60×10
89	Actinium	227	^{227}Ac	227.027 750	α, β^-	21.8 y
90	Thorium	228	^{228}Th	228.028 715	α	1.91 y
		232	^{232}Th	232.038 054	100%, α	1.41×10
91	Protactinium	231	^{231}Pa	231.035 880	α	3.28×10
92	Uranium	233	^{233}U	233.039 628	α	1.59×10
		235	^{235}U	235.043 924	0.720%, α	7.04×10

Atomic Number, Z	Name	Atomic Mass Number, A	Symbol	Atomic Mass (u)	Percent Abundance or Decay Mode	Half-life, $t_{1/2}$
		236	^{236}U	236.045 562	α	2.34×10
		238	^{238}U	238.050 784	99.2745%, α	4.47×10
		239	^{239}U	239.054 289	β^-	23.5 min
93	Neptunium	239	^{239}Np	239.052 933	β^-	2.355 d
94	Plutonium	239	^{239}Pu	239.052 157	α	2.41×10
95	Americium	243	^{243}Am	243.061 375	α, fission	7.37×10
96	Curium	245	^{245}Cm	245.065 483	α	8.50×10
97	Berkelium	247	^{247}Bk	247.070 300	α	1.38×10
98	Californium	249	^{249}Cf	249.074 844	α	351 y
99	Einsteinium	254	^{254}Es	254.088 019	α, β^-	276 d
100	Fermium	253	^{253}Fm	253.085 173	EC, α	3.00 d
101	Mendelevium	255	^{255}Md	255.091 081	EC, α	27 min
102	Nobelium	255	^{255}No	255.093 260	EC, α	3.1 min
103	Lawrencium	257	^{257}Lr	257.099 480	EC, α	0.646 s
104	Rutherfordium	261	^{261}Rf	261.108 690	α	1.08 min
105	Dubnium	262	^{262}Db	262.113 760	α, fission	34 s
106	Seaborgium	263	^{263}Sg	263.11 86	α, fission	0.8 s
107	Bohrium	262	^{262}Bh	262.123 1	α	0.102 s

Atomic Number, Z	Name	Atomic Mass Number, A	Symbol	Atomic Mass (u)	Percent Abundance or Decay Mode	Half-life, $t_{1/2}$
108	Hassium	264	^{264}Hs	264.1285	α	0.08 ms
109	Meitnerium	266	^{266}Mt	266.1378	α	3.4 ms

116. Appendix B. Selected Radioactive Isotopes

Decay modes are α, β^-, β^+, electron capture (EC) and isomeric transition (IT). EC results in the same daughter nucleus as would β^+ decay. IT is a transition from a metastable excited state. Energies for $\beta^{\pm}$ decays are the maxima; average energies are roughly one-half the maxima.

Table 1. Selected Radioactive Isotopes

Isotope	$t_{1/2}$	DecayMode(s)	Energy(MeV)	Percent		γ-Ray Energy(MeV)	Percent
^{3}H	12.33 y	β^-	0.0186	100%			
^{14}C	5730 y	β^-	0.156	100%			
^{13}N	9.96 min	β^+	1.20	100%			
^{22}Na	2.602 y	β^+	0.55	90%	γ	1.27	100%
^{32}P	14.28 d	β^-	1.71	100%			
^{35}S	87.4 d	β^-	0.167	100%			
^{36}Cl	3.00×10^5 y	β^-	0.710	100%			
^{40}K	1.28×10^9 y	β^-	1.31	89%			
^{43}K	22.3 h	β^-	0.827	87%	γs	0.373	87%
						0.618	87%
^{45}Ca	165 d	β^-	0.257	100%			
^{51}Cr	27.70 d	EC			γ	0.320	10%
^{52}Mn	5.59d	β^+	3.69	28%	γs	1.33	28%
						1.43	28%
^{52}Fe	8.27 h	β^+	1.80	43%		0.169	43%
						0.378	43%
^{59}Fe	44.6 d	β^-s	0.273	45%	γs	1.10	57%
			0.466	55%		1.29	43%
^{60}Co	5.271 y	β^-	0.318	100%	γs	1.17	100%
						1.33	100%
^{65}Zn	244.1 d	EC			γ	1.12	51%
^{67}Ga	78.3 h	EC			γs	0.0933	70%

Table 1. Selected Radioactive Isotopes

						0.185	35%
						0.300	19%
						others	
^{75}Se	118.5 d	EC			γs	0.121	20%
						0.136	65%
						0.265	68%
						0.280	20%
						others	
^{86}Rb	18.8 d	β^-s	0.69	9%	γ	1.08	9%
			1.77	91%			
^{85}Sr	64.8 d	EC			γ	0.514	100%
^{90}Sr	28.8 y	β^-	0.546	100%			
^{90}Y	64.1 h	β^-	2.28	100%			
^{99m}Tc	6.02 h	IT			γ	0.142	100%
^{113m}In	99.5 min	IT			γ	0.159	100%
^{123}I	13.0 h	EC			γ	0.159	≈100%
^{131}I	8.040 d	β^-s	0.248	7%	γs	0.364	85%
			0.607	93%		others	
			others				
^{129}Cs	32.3 h	EC			γs	0.0400	35%
						0.372	32%
						0.411	25%
						others	
^{137}Cs	30.17 y	β^-s	0.511	95%	γ	0.662	95%
			1.17	5%			
^{140}Ba	12.79 d	β^-	1.035	≈100%	γs	0.030	25%

Table 1. Selected Radioactive Isotopes

						0.044	65%
						0.537	24%
						others	
^{198}Au	2.696 d	β^-	1.161	≈100%	γ	0.412	≈100%
^{197}Hg	64.1 h	EC			γ	0.0733	100%
^{210}Po	138.38 d	α	5.41	100%			
^{226}Ra	1.60×10^3 y	αs	4.68	5%	γ	0.186	100%
			4.87	95%			
^{235}U	7.038×10^8 y	α	4.68	≈100%	γs	numerous	<0.400%
^{238}U	4.468×10^9 y	αs	4.22	23%	γ	0.050	23%
			4.27	77%			
^{237}Np	2.14×10^6 y	αs	numerous		γs	numerous	<0.250%
			4.96 (max.)				
^{239}Pu	2.41×10^4 y	αs	5.19	11%	γs	7.5×10^{-5}	73%
			5.23	15%		0.0133	15%
			5.24	73%		0.052	10%
						others	
^{243}Am	7.37×10^3 y	αs	Max. 5.44		γs	0.075	
			5.37	88%		others	
			5.32	11%			
			others				

117. Appendix C. Useful Information

- Table 1, Important Constants
- Table 2, Submicroscopic Masses
- Table 3, Solar System Data
- Table 4, Metric Prefixes for Powers of Ten and Their Symbols
- Table 5, The Greek Alphabet
- Table 6, SI units
- Table 7, Selected British Units
- Table 8, Other Units
- Table 9, Useful Formulae

Symbol	Meaning	Best Value	Approximate Value
c	Speed of light in vacuum	2.99792458×10^{8} m/s	3.00×10^{8} m/s
G	Gravitational constant	$6.67384(80) \times 10^{-11}$ $N \cdot m^2/kg^2$	6.67×10^{-11} $N \cdot m^2/kg^2$
N_A	Avogadro's number	$6.02214129(27) \times 10^{23}$	6.02×10^{23}
k	Boltzmann's constant	$1.3806488(13) \times 10^{-23}$ J/K	1.38×10^{-23} J/K
R	Gas constant	$8.3144621(75)$ $J/mol \cdot K$	$8.31\ J/mol \cdot K = 1.99\ cal/mol \cdot K = 0.0821\ atm \cdot L/mol \cdot K$
σ	Stefan-Boltzmann constant	$5.670373(21) \times 10^{-8}$ $W/m^2 \cdot K$	5.67×10^{-8} $W/m^2 \cdot K$
k	Coulomb force constant	$8.987551788... \times 10^{9}$ $N \cdot m^2/C^2$	8.99×10^{9} $N \cdot m^2/C^2$
q_e	Charge on electron	$-1.602176565(35) \times 10^{-19}$ C	-1.60×10^{-19} C
ε_0	Permittivity of free space	$8.854187817... \times 10^{-12}$ $C^2/N \cdot m^2$	8.85×10^{-12} $C^2/N \cdot m^2$
μ_0	Permeability of free space	$4\pi \times 10^{-7}$ $T \cdot m/A$	1.26×10^{-6} $T \cdot m/A$
h	Planck's constant	$6.62606957(29) \times 10^{-34}$ $J \cdot s$	6.63×10^{-34} $J \cdot s$

Table 1. Important Constants[1]

Return to Top

1. Stated values are according to the National Institute of Standards and Technology Reference on Constants, Units, and Uncertainty, www.physics.nist.gov/cuu (accessed May 18, 2012). Values in parentheses are the uncertainties in the last digits. Numbers without uncertainties are exact as defined.

Table 2. Submicroscopic Masses[2]			
Symbol	**Meaning**	**Best Value**	**Approximate Value**
m_e	Electron mass	$9.10938291(40) \times 10^{-31}$ kg	9.11×10^{-31} kg
m_p	Proton mass	$1.672621777(74) \times 10^{-27}$ kg	1.6726×10^{-27} kg
m_n	Neutron mass	$1.674927351(74) \times 10^{-27}$ kg	1.6749×10^{-27} kg
u	Atomic mass unit	$1.660538921(73) \times 10^{-27}$ kg	1.6605×10^{-27} kg

Return to Top

Table 3. Solar System Data		
Sun	mass	1.99×10^{30} kg
	average radius	6.96×10^{8} m
	Earth-sun distance (average)	1.496×10^{11} m
Earth	mass	5.9736×10^{24} kg
	average radius	6.376×10^{6} m
	orbital period	3.16×10^{7} s
Moon	mass	7.35×10^{22} kg
	average radius	1.74×10^{6} m
	orbital period (average)	2.36×10^{6} s
	Earth-moon distance (average)	3.84×10^{8} m

2. Stated values are according to the National Institute of Standards and Technology Reference on Constants, Units, and Uncertainty, www.physics.nist.gov/cuu (accessed May 18, 2012). Values in parentheses are the uncertainties in the last digits. Numbers without uncertainties are exact as defined.

Return to Top

Table 4. Metric Prefixes for Powers of Ten and Their Symbols					
Prefix	**Symbol**	**Value**	**Prefix**	**Symbol**	**Value**
tera	T	10^{12}	deci	d	10^{-1}
giga	G	10^{9}	centi	c	10^{-2}
mega	M	10^{6}	milli	m	10^{-3}
kilo	k	10^{3}	micro	μ	10^{-6}
hecto	h	10^{2}	nano	n	10^{-9}
deka	da	10^{1}	pico	p	10^{-12}
—	—	10^{0} (= 1)	femto	f	10^{-15}

Return to Top

Table 5. The Greek Alphabet								
Alpha	Α	α	Eta	Η	η	Nu	Ν	ν
Beta	Β	β	Tau	Τ	τ	Theta	Θ	θ
Xi	Ξ	ξ	Upsilon	Υ	υ	Gamma	Γ	γ
Iota	Ι	ι	Omicron	Ο	ο	Phi	φ	φ
Delta	Δ	δ	Kappa	Κ	κ	Pi	Π	π
Chi	Χ	χ	Epsilon	Ε	ε	Lambda	Λ	λ
Rho	Ρ	ρ	Psi	Ψ	ψ	Zeta	Ζ	ζ
Mu	Μ	μ	Sigma	Σ	σ	Omega	Ω	ω

Return to Top

Table 6. SI Units

	Entity	Abbreviation	Name
Fundamental units	Length	m	meter
	Mass	kg	kilogram
	Time	s	second
	Current	A	ampere
Supplementary unit	Angle	rad	radian
Derived units	Force	$N = kg \cdot m/s^2$	newton
	Energy	$J = kg \cdot m^2/s^2$	joule
	Power	$W = J/s$	watt
	Pressure	$Pa = N/m^2$	pascal
	Frequency	$Hz = 1/s$	hertz
	Electronic potential	$V = J/C$	volt
	Capacitance	$F = C/V$	farad
	Charge	$C = s \cdot A$	coulomb
	Resistance	$\Omega = V/A$	ohm
	Magnetic field	$T = N/(A \cdot m)$	tesla
	Nuclear decay rate	$Bq = 1/s$	becquerel

Return to Top

Table 7. Selected British Units

Length	1 inch (in.) = 2.54 cm (exactly)
	1 foot (ft) = 0.3048 m
	1 mile (mi) = 1.609 km
Force	1 pound (lb) = 4.448 N
Energy	1 British thermal unit (Btu) = 1.055×10^3 J
Power	1 horsepower (hp) = 746 W
Pressure	$1 \text{ lb/in}^2 = 6.895 \times 10^3$ Pa

Return to Top

Table 8. Other Units

Length	1 light year (ly) = 9.46×10^{15} m
	1 astronomical unit (au) = 1.50×10^{11} m
	1 nautical mile = 1.852 km
	1 angstrom (Å) = 10^{-10} m
Area	1 acre (ac) = 4.05×10^{3} m^2
	1 square foot (ft^2) = 9.29×10^{-2} m^2
	1 barn (b) = 10^{-28} m^2
Volume	1 liter (L) = 10^{-3} m^3
	1 U.S. gallon (gal) = 3.785×10^{-3} m^3
Mass	1 solar mass = 1.99×10^{30} kg
	1metric ton = 10^{3} kg
	1 atomic mass unit (u) = 1.6605×10^{-27} kg
Time	1 year (y) = 3.16×10^{7} s
	1 day (d) = 86,400 s
Speed	1 mile per hour (mph) = 1.609 km/h
	1 nautical mile per hour (naut) = 1.852 km/h
Angle	1 degree (°) = 1.745×10^{-2} rad
	1 minute of arc (′) = 1/60 degree
	1 second of arc (″) = 1/60 minute of arc
	1 grad = 1.571×10^{-2} rad
Energy	1 kiloton TNT (kT) = 4.2×10^{12} J
	1 kilowatt hour (kW · h) = 3.60×10^{6} J
	1 food calorie (kcal) = 4186 J
	1 calorie (cal) = 4.186 J
	1 electron volt (eV) = 1.60×10^{-19} J
Pressure	1 atmosphere (atm) = 1.013×10^{5} Pa
	1 millimeter of mercury (mm Hg) = 133.3 Pa
	1 torricelli (torr) = 1 mm Hg = 133.3 Pa

Table 8. Other Units

Nuclear decay rate	1 curie (Ci) = 3.70×10^{10} Bq

Return to Top

Table 9. Useful Formulae

Circumference of a circle with radius r or diameter d	$C = 2\pi r = \pi d$
Area of a circle with radius r or diameter d	$A = \pi r^2 = \dfrac{\pi d^2}{4}$
Area of a sphere with radius r	$A = 4\pi r^2$
Volume of a sphere with radius r	$V = \dfrac{4}{3}\left(\pi r^3\right)$

Return to Top

118. Appendix D. Glossary of Key Symbols and Notation

In this glossary, key symbols and notation are briefly defined.

Symbol	Definition
$\overline{\text{any symbol}}$	average (indicated by a bar over a symbol—e.g., $\overline{v}$ is average velocity)
°C	Celsius degree
°F	Fahrenheit degree
//	parallel
$\perp$	perpendicular
$\propto$	proportional to
±	plus or minus
0	zero as a subscript denotes an initial value
α	alpha rays
α	angular acceleration
α	temperature coefficient(s) of resistivity
β	beta rays
β	sound level
β	volume coefficient of expansion
β^-	electron emitted in nuclear beta decay
β^+	positron decay
γ	gamma rays
γ	surface tension
$\gamma = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}}$	a constant used in relativity
Δ	change in whatever quantity follows
δ	uncertainty in whatever quantity follows
ΔE	change in energy between the initial and final orbits of an electron in an atom
ΔE	uncertainty in energy
Δm	difference in mass between initial and final products
ΔN	number of decays that occur

Symbol	Definition
Δp	change in momentum
Δp	uncertainty in momentum
ΔPE_g	change in gravitational potential energy
$\Delta \theta$	rotation angle
Δs	distance traveled along a circular path
Δt	uncertainty in time
Δt_0	proper time as measured by an observer at rest relative to the process
ΔV	potential difference
Δx	uncertainty in position
ε_0	permittivity of free space
η	viscosity
θ	angle between the force vector and the displacement vector
θ	angle between two lines
θ	contact angle
θ	direction of the resultant
θ_b	Brewster's angle
θ_c	critical angle
κ	dielectric constant
λ	decay constant of a nuclide
λ	wavelength
λ_n	wavelength in a medium
μ_0	permeability of free space
μ_k	coefficient of kinetic friction
μ_s	coefficient of static friction
v_e	electron neutrino
π^+	positive pion
π^-	negative pion
π^0	neutral pion

Symbol	Definition
ρ	density
ρ_c	critical density, the density needed to just halt universal expansion
ρ_{fl}	fluid density
$\overline{\rho}_{obj}$	average density of an object
$\dfrac{\rho}{\rho_w}$	specific gravity
τ	characteristic time constant for a resistance and inductance (*RL*) or resistance and capacitance (*RC*) circuit
τ	characteristic time for a resistor and capacitor (*RC*) circuit
τ	torque
Y	upsilon meson
Φ	magnetic flux
φ	phase angle
Ω	ohm (unit)
ω	angular velocity
A	ampere (current unit)
A	area
A	cross-sectional area
A	total number of nucleons
a	acceleration
a_B	Bohr radius
a_c	centripetal acceleration
a_t	tangential acceleration
AC	alternating current
AM	amplitude modulation
atm	atmosphere
B	baryon number

Symbol	Definition
B	blue quark color
$\overline{B}$	antiblue (yellow) antiquark color
b	quark flavor bottom or beauty
B	bulk modulus
B	magnetic field strength
B_{int}	electron’s intrinsic magnetic field
B_{orb}	orbital magnetic field
BE	binding energy of a nucleus—it is the energy required to completely disassemble it into separate protons and neutrons
$\frac{\text{BE}}{A}$	binding energy per nucleon
Bq	becquerel—one decay per second
C	capacitance (amount of charge stored per volt)
C	coulomb (a fundamental SI unit of charge)
C_p	total capacitance in parallel
C_s	total capacitance in series
CG	center of gravity
CM	center of mass
c	quark flavor charm
c	specific heat
c	speed of light
Cal	kilocalorie
cal	calorie
COP_{hp}	heat pump’s coefficient of performance
COP_{ref}	coefficient of performance for refrigerators and air conditioners
$\cos\theta$	cosine
$\cot\theta$	cotangent
$\csc\theta$	cosecant

Symbol	Definition
D	diffusion constant
d	displacement
d	quark flavor down
dB	decibel
d_i	distance of an image from the center of a lens
d_o	distance of an object from the center of a lens
DC	direct current
E	electric field strength
ε	emf (voltage) or Hall electromotive force
emf	electromotive force
E	energy of a single photon
E	nuclear reaction energy
E	relativistic total energy
E	total energy
E_0	ground state energy for hydrogen
E_0	rest energy
EC	electron capture
E_{cap}	energy stored in a capacitor
Eff	efficiency—the useful work output divided by the energy input
Eff_C	Carnot efficiency
E_{in}	energy consumed (food digested in humans)
E_{ind}	energy stored in an inductor
E_{out}	energy output
e	emissivity of an object
e^+	antielectron or positron
eV	electron volt
F	farad (unit of capacitance, a coulomb per volt)
F	focal point of a lens

Symbol	Definition
$\mathbf{F}$	force
F	magnitude of a force
F	restoring force
F_B	buoyant force
F_c	centripetal force
F_i	force input
$\mathbf{F}_{net}$	net force
F_o	force output
FM	frequency modulation
f	focal length
f	frequency
f_0	resonant frequency of a resistance, inductance, and capacitance (*RLC*) series circuit
f_0	threshold frequency for a particular material (photoelectric effect)
f_1	fundamental
f_2	first overtone
f_3	second overtone
f_B	beat frequency
f_k	magnitude of kinetic friction
f_s	magnitude of static friction
G	gravitational constant
G	green quark color
$\overline{G}$	antigreen (magenta) antiquark color
g	acceleration due to gravity
g	gluons (carrier particles for strong nuclear force)
h	change in vertical position
h	height above some reference point
h	maximum height of a projectile
h	Planck's constant

Symbol	Definition
hf	photon energy
h_i	height of the image
h_o	height of the object
I	electric current
I	intensity
I	intensity of a transmitted wave
I	moment of inertia (also called rotational inertia)
I_0	intensity of a polarized wave before passing through a filter
I_{ave}	average intensity for a continuous sinusoidal electromagnetic wave
I_{rms}	average current
J	joule
$\frac{J}{\Psi}$	Joules/psi meson
K	kelvin
k	Boltzmann constant
k	force constant of a spring
K_α	x rays created when an electron falls into an $n = 1$ shell vacancy from the $n = 3$ shell
K_β	x rays created when an electron falls into an $n = 2$ shell vacancy from the $n = 3$ shell
kcal	kilocalorie
KE	translational kinetic energy
KE + PE	mechanical energy
KE_e	kinetic energy of an ejected electron
KE_{rel}	relativistic kinetic energy
KE_{rot}	rotational kinetic energy
$\overline{KE}$	thermal energy
kg	kilogram (a fundamental SI unit of mass)

Symbol	Definition
L	angular momentum
L	liter
L	magnitude of angular momentum
L	self-inductance
l	angular momentum quantum number
L_α	x rays created when an electron falls into an n = 2 shell from the n = 3 shell
L_e	electron total family number
L_μ	muon family total number
L_τ	tau family total number
L_f	heat of fusion
L_f and L_v	latent heat coefficients
L_{orb}	orbital angular momentum
L_s	heat of sublimation
L_v	heat of vaporization
L_z	z-component of the angular momentum
M	angular magnification
M	mutual inductance
m	indicates metastable state
m	magnification
m	mass
m	mass of an object as measured by a person at rest relative to the object
m	meter (a fundamental SI unit of length)
m	order of interference
m	overall magnification (product of the individual magnifications)
$m\left({}^{A}\mathrm{X}\right)$	atomic mass of a nuclide
MA	mechanical advantage

Symbol	Definition
m_e	magnification of the eyepiece
m_e	mass of the electron
m_ℓ	angular momentum projection quantum number
m_n	mass of a neutron
m_o	magnification of the objective lens
mol	mole
m_p	mass of a proton
m_s	spin projection quantum number
N	magnitude of the normal force
N	newton
N	normal force
N	number of neutrons
n	index of refraction
n	number of free charges per unit volume
N_A	Avogadro's number
N_r	Reynolds number
$N \cdot m$	newton-meter (work-energy unit)
$N \cdot m$	newtons times meters (SI unit of torque)
OE	other energy
P	power
P	power of a lens
P	pressure
p	momentum
p	momentum magnitude
p	relativistic momentum
$\mathbf{p}_{tot}$	total momentum
$\mathbf{p}_{tot}$	total momentum some time later
P_{abs}	absolute pressure
P_{atm}	atmospheric pressure

Symbol	Definition
P_{atm}	standard atmospheric pressure
PE	potential energy
PE_{el}	elastic potential energy
PE_{elec}	electric potential energy
PE_s	potential energy of a spring
P_g	gauge pressure
P_{in}	power consumption or input
P_{out}	useful power output going into useful work or a desired, form of energy
Q	latent heat
Q	net heat transferred into a system
Q	flow rate—volume per unit time flowing past a point
+Q	positive charge
−Q	negative charge
q	electron charge
q_p	charge of a proton
q	test charge
QF	quality factor
R	activity, the rate of decay
R	radius of curvature of a spherical mirror
R	red quark color
$\overline{R}$	antired (cyan) quark color
R	resistance
R	resultant or total displacement
R	Rydberg constant
R	universal gas constant
r	distance from pivot point to the point where a force is applied
r	internal resistance

Symbol	Definition
$r\perp$	perpendicular lever arm
r	radius of a nucleus
r	radius of curvature
r	resistivity
r or rad	radiation dose unit
rem	roentgen equivalent man
rad	radian
RBE	relative biological effectiveness
RC	resistor and capacitor circuit
rms	root mean square
r_{n}	radius of the nth H-atom orbit
R_{p}	total resistance of a parallel connection
R_{s}	total resistance of a series connection
R_{s}	Schwarzschild radius
S	entropy
S	intrinsic spin (intrinsic angular momentum)
S	magnitude of the intrinsic (internal) spin angular momentum
S	shear modulus
S	strangeness quantum number
s	quark flavor strange
s	second (fundamental SI unit of time)
s	spin quantum number
s	total displacement
$\sec\theta$	secant
$\sin\theta$	sine
s_{z}	z-component of spin angular momentum
T	period—time to complete one oscillation
T	temperature

Symbol	Definition
T_c	critical temperature—temperature below which a material becomes a superconductor
T	tension
T	tesla (magnetic field strength B)
t	quark flavor top or truth
t	time
$t_{1/2}$	half-life—the time in which half of the original nuclei decay
$\tan\theta$	tangent
U	internal energy
u	quark flavor up
u	unified atomic mass unit
$\mathbf{u}$	velocity of an object relative to an observer
$\mathbf{u}'$	velocity relative to another observer
V	electric potential
V	terminal voltage
V	volt (unit)
V	volume
$\mathbf{v}$	relative velocity between two observers
v	speed of light in a material
$\mathbf{v}$	velocity
$\overline{\mathbf{v}}$	average fluid velocity
$V_B - V_A$	change in potential
$\mathbf{v}_d$	drift velocity
V_p	transformer input voltage
V_{rms}	rms voltage
V_s	transformer output voltage
$\mathbf{v}_{tot}$	total velocity
v_w	propagation speed of sound or other wave
$\mathbf{v}_w$	wave velocity

Symbol	Definition
W	work
W	net work done by a system
W	watt
w	weight
w_{fl}	weight of the fluid displaced by an object
W_c	total work done by all conservative forces
W_{nc}	total work done by all nonconservative forces
W_{out}	useful work output
X	amplitude
X	symbol for an element
${}^{A}_{Z}\mathrm{X}_{N}$	notation for a particular nuclide
x	deformation or displacement from equilibrium
x	displacement of a spring from its undeformed position
x	horizontal axis
X_C	capacitive reactance
X_L	inductive reactance
x_{rms}	root mean square diffusion distance
y	vertical axis
Y	elastic modulus or Young's modulus
Z	atomic number (number of protons in a nucleus)
Z	impedance