Ministry of education and science of the Republic of Kazakhstan

A. Baitursynov Kostanai State university

Department of electricity and physics

M. Dunskiy

PHYSICS

(Magnetism, Optics, Modern physics)

Manual

Kostanay, 2020

UDC 53 (075.8)
LBC 22.3я73
D 83

Author:
Dunskiy Mikhail, master of physics, senior lecturer at the department of Electricity and Physics

Readers:
Jamanbalin Kadyrgali - Doctor of Physics and Mathematics, professor, Rector of Kostanay Social and Technical of Aldamjar University
Poezjalov Vladimir - professor, candidate of physics and mathematics science
Liphenko Valeriy - senior lecturer of department of electricity and physics, candidate of physics and mathematics science

Dunskiy M.
D83 Physics: manual. – Kostanay: A. Baitursynov Kostanai State university, 2020. – 180 p.

ISBN

This manual contains the theoretical material of three parts of general physics: Magnetism, Optics, and Modern physics. It is necessary for specialty 5B071800 – Electrical power engineering.

UDC 53 (075.8)
LBC 22.3я73
D 83

Approved and recommended by the Educational-Methodical Council of A. Baitursynov Kostanay State university, ______ year, Protocol № ____

Content

Introduction

Physics is a science about nature around us. They studies many different things. In this manual we will discuss three parts of general physics:

- Magnetism. It is a theory about magnetic phenomena (permanent magnets, magnetic field around current-carrying conductors).
- Optics which studies light phenomena (as geometrical and wave processes).
- Modern physics. This part includes quantum physics, atomic physics, nucleus physics, physics of elementary particles.

Physics is one of the most fundamental of the sciences. Scientists of all disciplines use the ideas of physics, including chemists who study the structure of molecules, paleontologists who try to reconstruct how dinosaurs lived, and climatologists who study how climate changes. Physics is also the foundation of all engineering and technology. No engineer could design a flat-screen TV, an interplanetary spacecraft, or even a better mousetrap without first understanding the basic laws of physics.

We will:

- discuss the nature of physical theory and the use of idealized models to represent physical systems;

- introduce the systems of units used to describe physical quantities and discuss ways to describe the accuracy of a number

- look at examples of problems for which we can’t (or don’t want to) find a precise answer, but for which rough estimates can be useful and interesting

- study several aspects of vectors and vector algebra

Vectors will be needed throughout our study of physics to describe and analyze physical quantities, such as velocity and force, that have direction as well as magnitude.

Physics is an *experimental* science. Physicists observe the phenomena of nature and try to find patterns that relate these phenomena. These patterns are called physical theories or, when they are very well established and widely used, physical laws or principles.

To develop a physical theory, a physicist has to learn to ask appropriate questions, design experiments to try to answer the questions, and draw appropriate conclusions from the results.

The development of physical theories often takes an indirect path, with blind alleys, wrong guesses, and the discarding of unsuccessful theories in favor of more promising ones. Physics is not simply a collection of facts and principles; it is also the *process* by which we arrive at general principles that describe how the physical universe behaves.

No theory is ever regarded as the final or ultimate truth. The possibility always exists that new observations will require that a theory be revised or discarded. It is in the nature of physical theory that we can disprove a theory by finding behaviour that is inconsistent with it, but we can never prove that a theory is always correct.

In this manual we will use the next basic denotes:

$\vec{F}$	Force	q	Electric charge
$\vec{B}$	Magnetic field	I	Current
$\vec{E}$	Electric field	R	Resistance
$\vec{v}$	Velocity	U	Voltage
m	Mass	C	Capacitance
n	Refractive index	L	Inductance
S	Area	μ	Magnetic moment
P	Electric power	λ	Wavelength
ν	Frequency	c	Speed of the light
f	Focal length	R	Radius
h	Constant of Planck	e	Charge of electron

Topic 1 Magnetism

1.1 Magnetic field and force

1.1.1 Magnetism and magnetic field

Everybody uses magnetic forces. They are at the heart of electric motors, microwave ovens, loudspeakers, computer printers, and disk drives. The most familiar examples of magnetism are permanent magnets, which attract unmagnetized iron objects and can also attract or repel other magnets. A compass needle aligning itself with the earth’s magnetism is an example of this interaction. But the *fundamental* nature of magnetism is the interaction of moving electric charges. Unlike electric forces, which act on electric charges whether they are moving or not, magnetic forces act only on *moving* charges.

We saw that the electric force arises in two stages: (1) a charge produces an electric field in the space around it, and (2) a second charge responds to this field. Magnetic forces also arise in two stages. First, a *moving* charge or a collection of moving charges (that is, an electric current) produces a *magnetic* field. Next, a second current or moving charge responds to this magnetic field, and so experiences a magnetic force.

In this chapter we study the second stage in the magnetic interaction - that is, how moving charges and currents *respond* to magnetic fields. In particular, we will see how to calculate magnetic forces and torques, and we will discover why magnets can pick up iron objects like paper clips. We will complete our picture of the magnetic interaction by examining how moving charges and currents *produce* magnetic fields.

Magnetic phenomena were first observed at least 2500 years ago in fragments of magnetized iron ore found near the ancient city of Magnesia (now Manisa, in western Turkey). These fragments were examples of what are now called **permanent magnets;** you probably have several permanent magnets on your refrigerator door at home. Permanent magnets were found to exert forces on each other as well as on pieces of iron that were not magnetized. It was discovered that when an iron rod is brought in contact with a natural magnet, the rod also becomes magnetized. When such a rod is floated on water or suspended by a string from its center, it tends to line itself up in a north-south direction. The needle of an ordinary compass is just such a piece of magnetized iron.

Before the relationship of magnetic interactions to moving charges was understood, the interactions of permanent magnets and compass needles were described in terms of *magnetic poles.* If a bar-shaped permanent magnet, or *bar magnet,* is free to rotate, one end points north. This end is called a *north pole* or *N pole;* the other end is a *south pole* or *S pole.* Opposite poles attract each other, and like poles repel each other (Fig. 1). An object that contains iron but is not itself magnetized (that is, it shows no tendency to point north or south) is attracted by *either* pole of a permanent magnet (Fig. 2). This is the attraction that acts between a magnet and the unmagnetized steel door of a refrigerator. By analogy to electric interactions, we describe the interactions in Figs. 1 and 2 by saying that a bar magnet sets up a *magnetic field* in the space around it and a second body responds to that field. A compass needle tends to align with the magnetic field at the needle's position.

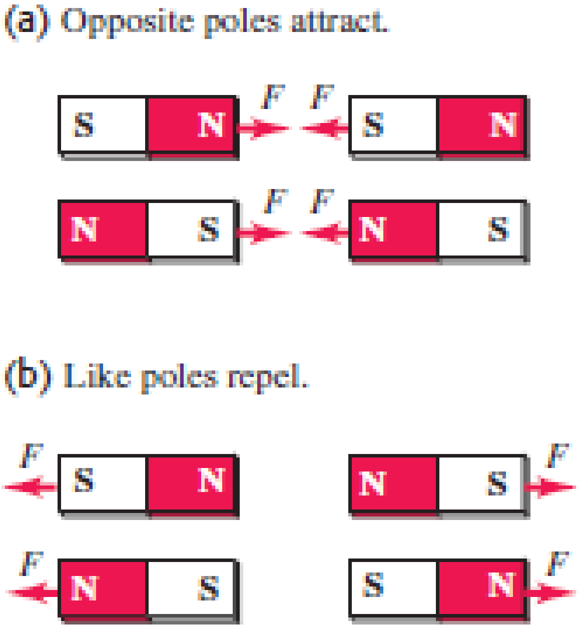


Figure 1 – (a) Two bar magnets attract when opposite poles (N and S, or S or N) are next to each other. (b) The bar magnets repel when like poles (N and N, or S and S) are next to each other

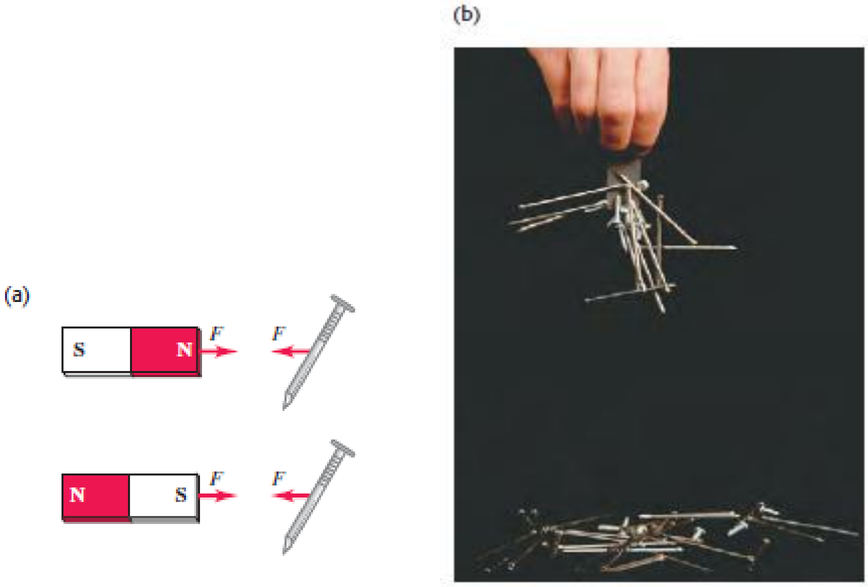


Figure 2 – (a) Either pole a bar magnet attracts an unmagnetized object that contains iron, such as a nail. (b) A real-life example of this effect

The earth itself is a magnet. Its north geographic pole is close to a magnetic *south* pole, which is why the north pole of a compass needle points north. The earth's magnetic axis is not quite parallel to its geographic axis (the axis of rotation), so a compass reading deviates somewhat from geographic north. This deviation, which varies with location, is called *magnetic declination* or *magnetic variation.* Also, the magnetic field is not horizontal at most points on the earth's surface; its angle up or down is called *magnetic inclination.* At the magnetic poles the magnetic field is vertical.

Figure 3 is a sketch of the earth's magnetic field. The lines, called *magnetic field lines,* show the direction that a compass would point at each location.. The direction of the field at any point can be defined as the direction of the force that the field would exert on a magnetic north pole. In follow section we'll describe a more fundamental way to define the direction and magnitude of a magnetic field.

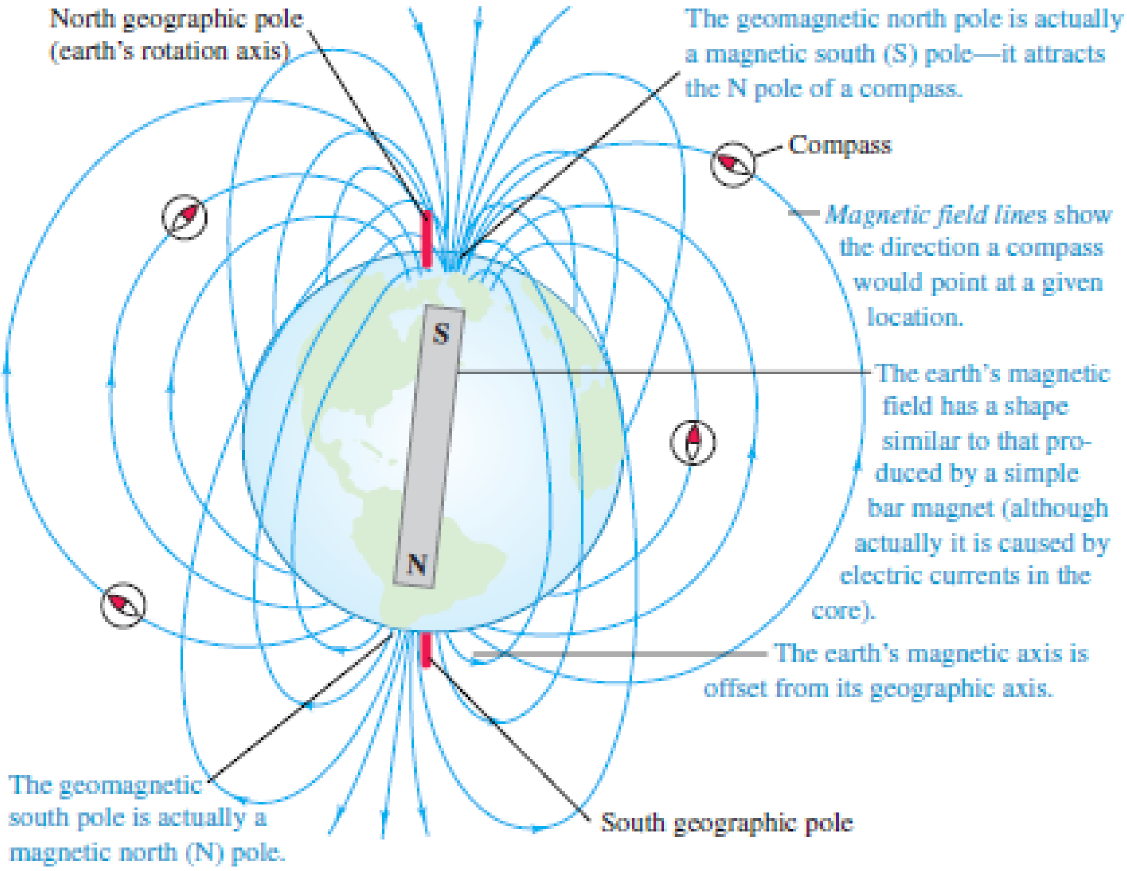

North geographic pole (earth's rotation axis)
The geomagnetic north pole is actually a magnetic south (S) pole—it attracts the N pole of a compass.
Compass
Magnetic field lines show the direction a compass would point at a given location.
S
The earth's magnetic field has a shape similar to that produced by a simple bar magnet (although actually it is caused by electric currents in the core).
N
The earth's magnetic axis is offset from its geographic axis.
The geomagnetic south pole is actually a magnetic north (N) pole.
South geographic pole

Figure 3 – A sketch of the earth's magnetic field. The field, which is caused by currents in the earth's molten core, changes with time; geologic evidence shows that it reverses direction entirely at irregular intervals of 10^4 to 10^6

To introduce the concept of magnetic field properly, let's review our formulation of *electric* interactions, where we introduced the concept of *electric* field. We represented electric interactions in two steps:

1. A distribution of electric charge at rest creates an electric field $\vec{E}$ in the surrounding space.

2. The electric field exerts a force $\vec{F} = q\vec{E}$ on any other charge that is present in the field. We can describe magnetic interactions in a similar way:

1. A moving charge or a current creates a **magnetic field** in the surrounding space (in addition to its *electric* field).

2. The magnetic field exerts a force $\vec{F}$ on any other moving charge or current that is present in the field.

In this chapter we'll concentrate on the *second* aspect of the interaction: Given the presence of a magnetic field, what force does it exert on a moving charge or a current? Next we will come back to the problem of how magnetic fields are *created* by moving charges and currents.

Like electric field, magnetic field is a *vector field* - that is, a vector quantity associated with each point in space. We will use the symbol $\vec{B}$ for magnetic field. At any position the direction of $\vec{B}$ is defined as the direction in which the north pole of a compass needle tends to point. The arrows in Fig. 3 suggest the direction of the earth's magnetic field; for any magnet, $\vec{B}$ points out of its north pole and into its south pole.

There are four key characteristics of the magnetic force on a moving charge. First, its magnitude is proportional to the magnitude of the charge. If a $1 - \mu C$ charge and a $2 - \mu C$ charge move through a given magnetic field with the same velocity, experiments show that the force on the $2-\mu C$ charge is twice as great as the force on the charge. Second, the magnitude of the force is also proportional to the magnitude, or "strength," of the field; if we double the magnitude of the field (for example, by using two identical bar magnets instead of one) without changing the charge or its velocity, the force doubles.

A third characteristic is that the magnetic force depends on the particle's velocity. This is quite different from the electric-field force, which is the same whether the charge is moving or not. A charged particle at rest experiences *no* magnetic force. And fourth, we find by experiment that the magnetic force $\vec{F}$ *does not* have the same direction as the magnetic field but instead is always *perpendicular* to both $\vec{B}$ and the velocity $\vec{v}$. The magnitude F of the force is found to be proportional to the component of $\vec{v}$ perpendicular to the field; when that component is zero (that is, when $\vec{v}$ and $\vec{B}$ are parallel or antiparallel), the force is zero.

Figure 4 shows these relationships. The direction of $\vec{F}$ is always perpendicular to the plane containing $\vec{v}$ and $\vec{B}$. Its magnitude is given by

$$F = |q|v_{\perp}B = |q|vB\sin\alpha \qquad (1)$$

where $|q|$ is the magnitude of the charge and α is the angle measured from the direction of $\vec{v}$ to the direction of $\vec{B}$, as shown in the figure.

This description does not specify the direction of $\vec{F}$ completely; there are always two directions, opposite to each other, that are both perpendicular to the plane of $\vec{v}$ and $\vec{B}$. To complete the description, we use the same right-hand rule . (It would be a good idea to review that section before you go on.) Draw the vectors $\vec{v}$ and $\vec{B}$ with their tails together, as in Fig. 5a. Imagine turning $\vec{v}$ until it points in the direction of $\vec{B}$ (turning through the smaller of the two possible angles). Wrap the fingers of your right hand around the line perpendicular to the plane of $\vec{v}$ and $\vec{B}$ so that they curl around with the sense of rotation from $\vec{v}$ to $\vec{B}$. Your thumb then points in the direction of the force $\vec{F}$ on a *positive* charge. (Alternatively, the direction of the force $\vec{F}$ on a positive charge is the direction in which a right-hand-thread screw would advance if turned the same way.)

This discussion shows that the force on a charge moving with velocity in a magnetic field $\vec{B}$ is given, both in magnitude and in direction, by (magnetic force on a moving charged particle)

$$\vec{F} = q\vec{v} \times \vec{B} \tag{2}$$

This is the first of several vector products we will encounter in our study of magnetic-field relationships. It’s important to note that Eq. (2) was *not* deduced theoretically; it is an observation based on *experiment.*

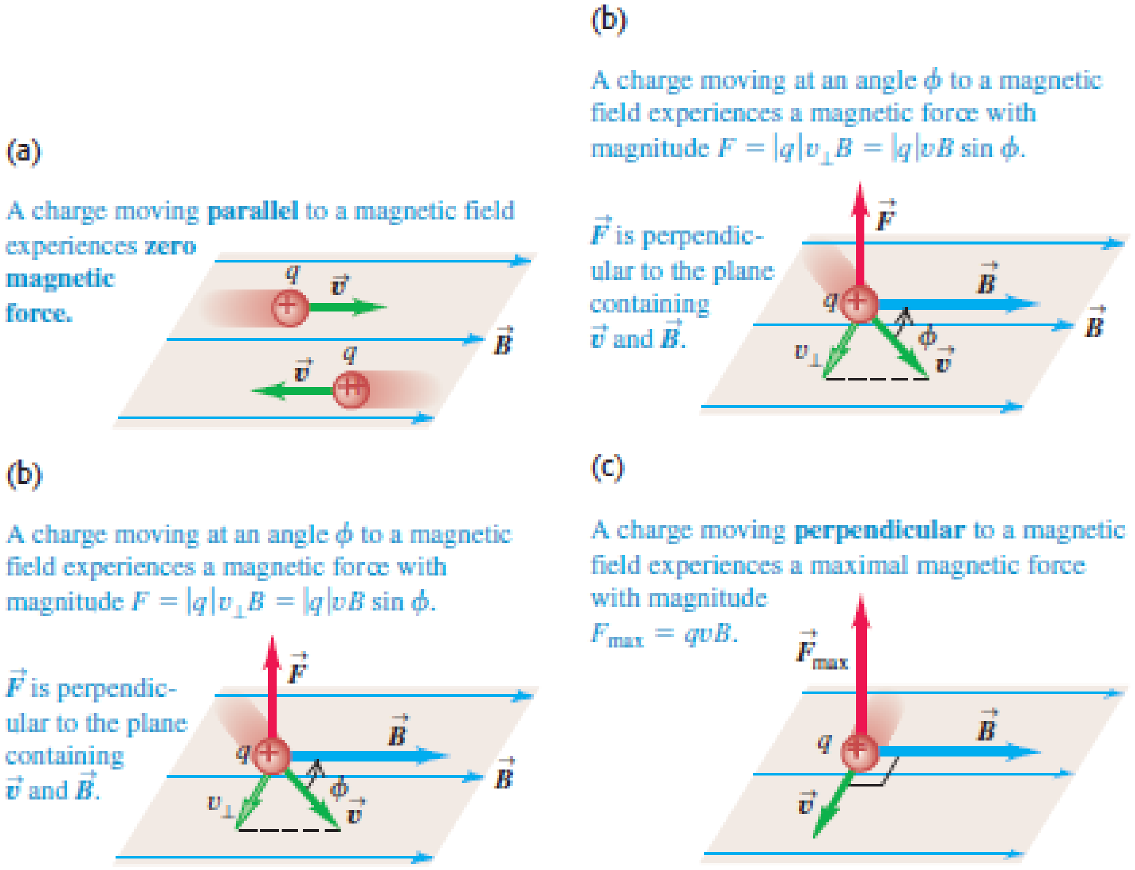

(b)
A charge moving at an angle ϕ to a magnetic field experiences a magnetic force with magnitude $F = |q|v_\perp B = |q|vB \sin\phi$.
$\vec{F}$ is perpendicular to the plane containing $\vec{v}$ and $\vec{B}$.
$\vec{F}$
q
$\vec{B}$
$\vec{B}$
$v_\perp$
ϕ
$\vec{v}$
(a)
A charge moving **parallel** to a magnetic field experiences **zero magnetic force.**
q
$\vec{v}$
$\vec{B}$
q
$\vec{v}$
(b)
A charge moving at an angle ϕ to a magnetic field experiences a magnetic force with magnitude $F = |q|v_\perp B = |q|vB \sin\phi$.
$\vec{F}$ is perpendicular to the plane containing $\vec{v}$ and $\vec{B}$.
$\vec{F}$
q
$\vec{B}$
$\vec{B}$
$v_\perp$
ϕ
$\vec{v}$
(c)
A charge moving **perpendicular** to a magnetic field experiences a maximal magnetic force with magnitude $F_{max} = qvB$.
$\vec{F}_{max}$
q
$\vec{B}$
$\vec{v}$

Figure 4 – The magnetic force $\vec{F}$ acting on a positive charge q moving with velocity $\vec{v}$ and the magnetic field $\vec{B}$. For given values of the speed v and magnetic field strength B, the force is greatest when $\vec{v}$ and $\vec{B}$ are perpendicular

Equation (2) is valid for both positive and negative charges. When q is negative, the direction of the force $\vec{F}$ is opposite to that of $\vec{v} \times \vec{B}$ (Fig. 5b). If two charges with equal magnitude and opposite sign move in the same $\vec{B}$ field with the same velocity (Fig. 6), the forces have equal magnitude and opposite direction. Figures 4, 5, and 6 show several examples of the relationships of the directions of $\vec{F}, \vec{v}$ and $\vec{B}$ for both positive and negative charges. Be sure you understand the relationships shown in these figures.

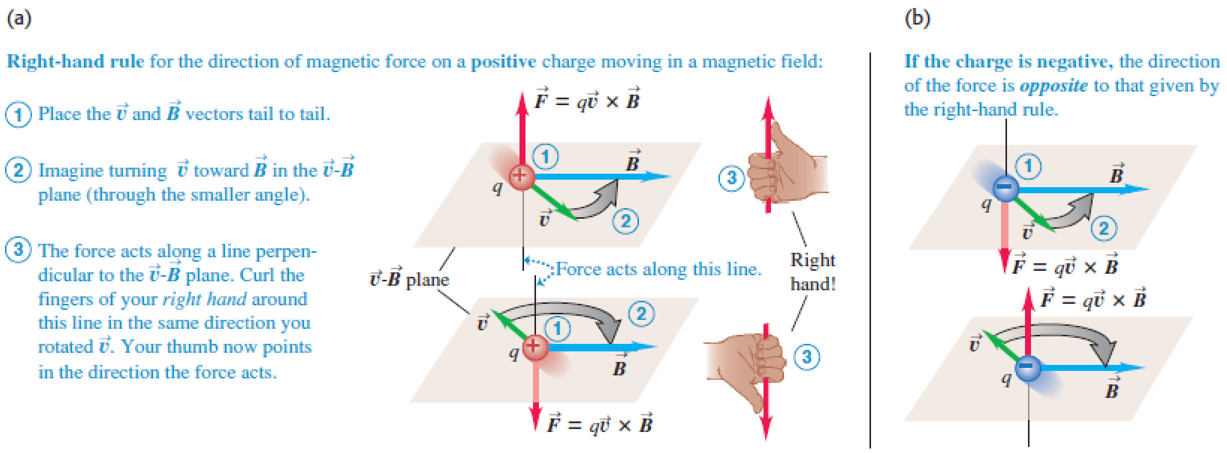

(a)
Right-hand rule for the direction of magnetic force on a positive charge moving in a magnetic field:
1 Place the $\vec{v}$ and $\vec{B}$ vectors tail to tail.
2 Imagine turning $\vec{v}$ toward $\vec{B}$ in the $\vec{v}$-$\vec{B}$ plane (through the smaller angle).
3 The force acts along a line perpendicular to the $\vec{v}$-$\vec{B}$ plane. Curl the fingers of your right hand around this line in the same direction you rotated $\vec{v}$. Your thumb now points in the direction the force acts.
$\vec{F} = q\vec{v} \times \vec{B}$
q
$\vec{B}$
$\vec{v}$
$\vec{v}$-$\vec{B}$ plane
Force acts along this line.
Right hand!
(b)
If the charge is negative, the direction of the force is opposite to that given by the right-hand rule.
$\vec{F} = q\vec{v} \times \vec{B}$

Figure 5 – Finding the direction of the magnetic force on a moving charged particle

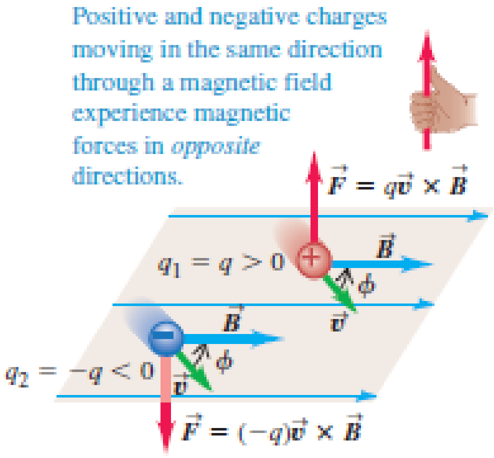

Positive and negative charges moving in the same direction through a magnetic field experience magnetic forces in opposite directions.
$\vec{F} = q\vec{v} \times \vec{B}$
$q_1 = q > 0$
$\vec{B}$
ϕ
$\vec{v}$
$q_2 = -q < 0$
$\vec{F} = (-q)\vec{v} \times \vec{B}$

Figure 6 – Two charges of the same magnitude but opposite sign moving with the same velocity in the same magnetic field. The magnetic force on the charges are equal in magnitude but opposite direction

Equation (1) gives the magnitude of the magnetic force $\vec{F}$ in Eq. (2). We can express this magnitude in a different but equivalent way. Since α is the angle between the directions of vectors $\vec{v}$ and $\vec{B}$, we may interpret $B \sin \alpha$ as the component of $\vec{B}$ perpendicular to $\vec{v}$—that is, $B_\perp$. With this notation the force magnitude is

$$F = |q| v B_\perp \tag{3}$$

This form is sometimes more convenient, especially in problems involving *currents* rather than individual particles. We will discuss forces on currents later in this chapter.

From Eq. (1) the *units* of B must be the same as the units of . F/qv Therefore the SI unit of B is equivalent to $1\ N \cdot s/C \cdot m$, or, since one ampere is one coulomb per second $(1\ A = 1\ C/s), 1\ N/A \cdot m$. This unit is called the **tesla** (abbreviated T), in honor of Nikola Tesla (1856–1943), the prominent Serbian-American scientist and inventor:

$$1\ tesla = 1\ T = 1\ N/A \cdot m$$

The magnetic field of the earth is of the order of 10^{-4} T.. Magnetic fields of the order of 10 T occur in the interior of atoms and are important in the analysis of atomic spectra. The largest steady magnetic field that can be produced at present in the laboratory is about 45 T. Some pulsed-current electromagnets can produce fields of the order of 120 T for millisecond time intervals.

1.1.2 Motion of Charged Particles in a Magnetic Field

When a charged particle moves in a magnetic field, it is acted on by the magnetic force given by Eq. (2), and the motion is determined by Newton's laws. Figure 7a shows a simple example. A particle with positive charge q is at point O moving with velocity $\vec{v}$ in a uniform magnetic field $\vec{B}$ directed into the plane of the figure. The vectors $\vec{v}$ and $\vec{B}$ are perpendicular, so the magnetic force $\vec{F} = q\vec{v} \times \vec{B}$ has magnitude $F = qvB$ and a direction as shown in the figure. The force is *always* perpendicular to $\vec{v}$ so it cannot change the *magnitude* of the velocity, only its direction. To put it differently, the magnetic force never has a component parallel to the particle's motion, so the magnetic force can never do *work* on the particle. This is true even if the magnetic field is not uniform.

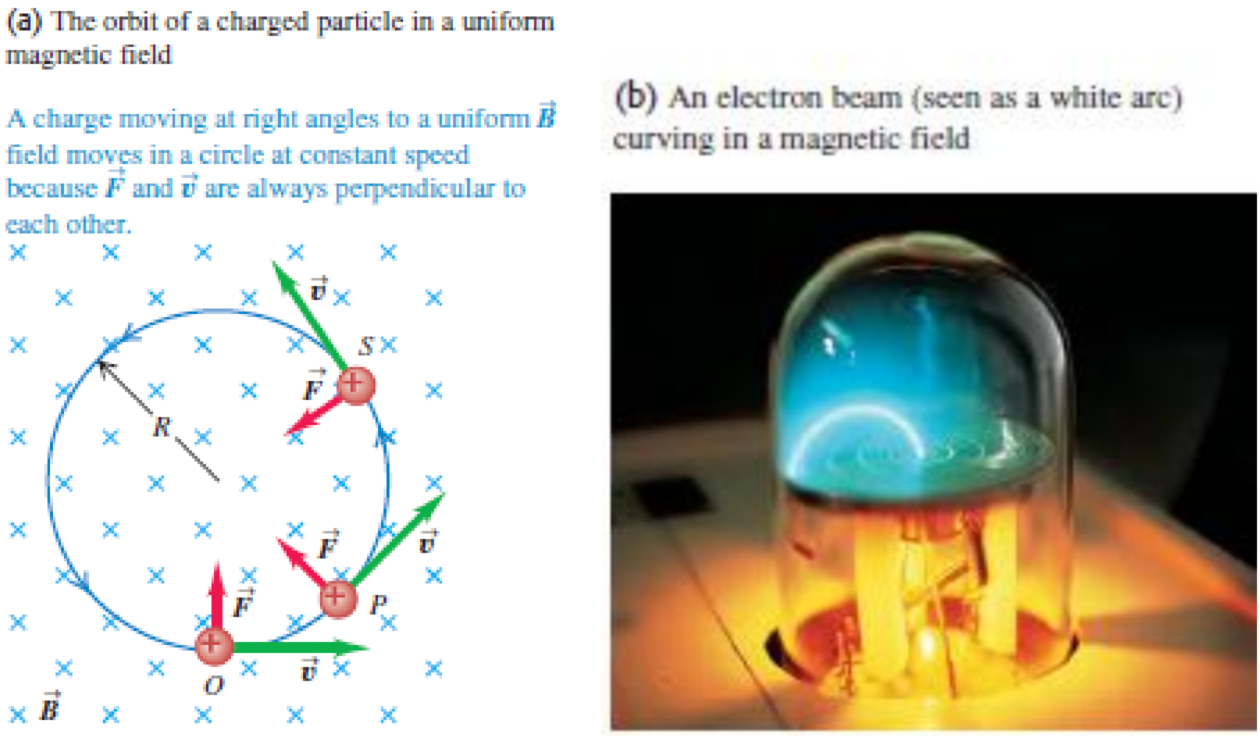
Motion of a charged particle under the action of a magnetic field alone is always motion with constant speed.

Using this principle, we see that in the situation shown in Fig. 7a the magnitudes of both $\vec{F}$ and $\vec{v}$ are constant. At points such as P and S the directions of force and velocity have changed as shown, but their magnitudes are the same. The particle therefore moves under the influence of a constant-magnitude force that is always at right angles to the velocity of the particle. Comparing the discussion of circular motion early, we see that the particle's path is a *circle,* traced out with constant speed v. The centripetal acceleration is v^2/R and only the magnetic force acts, so from Newton's second law,

$$F = |q|vB = m\frac{v^2}{R} \tag{4}$$

where is the mass of the particle. Solving Eq. (4) for the radius R of the circular path, we find

$$R = \frac{mv}{|q|B} \tag{5}$$



(a) The orbit of a charged particle in a uniform magnetic field

A charge moving at right angles to a uniform $\vec{B}$ field moves in a circle at constant speed because $\vec{F}$ and $\vec{v}$ are always perpendicular to each other.

(b) An electron beam (seen as a white arc) curving in a magnetic field

Figure 7 – A charged particle moves in a plane perpendicular to a uniform magnetic field $\vec{B}$

We can also write this as $R = p/|q|B$, where $p = mv$ is the magnitude of the particle’s momentum. If the charge q is negative, the particle moves *clockwise* around the orbit in Fig. 7a.

The angular speed ω of the particle can be found from equation $v = \omega R$. Combining this with Eq. (5), we get

$$\omega = \frac{v}{R} = v\frac{|q|B}{mv} = \frac{|q|B}{m} \tag{6}$$

The number of revolutions per unit time is $\nu = \omega/2\pi$. This frequency ν is independent of the radius R of the path. It is called the **cyclotron frequency;** in a particle accelerator called a *cyclotron*, particles moving in nearly circular paths are given a boost twice each revolution, increasing their energy and their orbital radii but not their angular speed or frequency. Similarly, one type of *magnetron*, a common source of microwave radiation for microwave ovens and radar systems, emits radiation with a frequency equal to the frequency of circular motion of electrons in a vacuum chamber between the poles of a magnet.

If the direction of the initial velocity is *not* perpendicular to the field, the velocity *component* parallel to the field is constant because there is no force parallel to the field. Then the particle moves in a helix (Fig. 8). The radius of the helix is given by Eq. (5), where is now the component of velocity perpendicular to the $\vec{B}$ field.

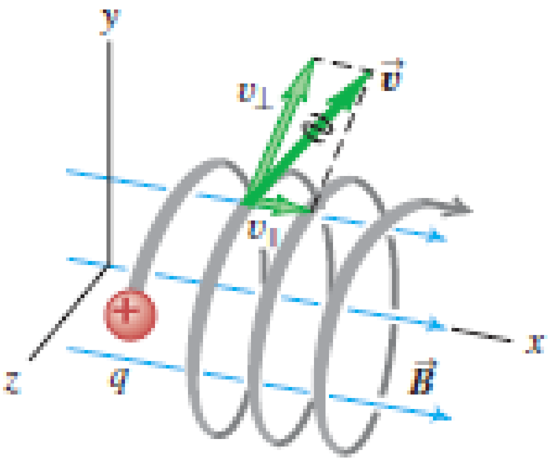

This particle's motion has components both parallel ($v_\parallel$) and perpendicular ($v_\perp$) to the magnetic field, so it moves in a helical path.

Figure 8 – The general case of a charged particle moving in a uniform magnetic field $\vec{B}$. The magnetic field does no work on the particle, so its speed and kinetic energy remain constant

Motion of a charged particle in a nonuniform magnetic field is more complex. Figure 9 shows a field produced by two circular coils separated by some distance. Particles near either coil experience a magnetic force toward the center of the region; particles with appropriate speeds spiral repeatedly from one end of the region to the other and back. Because charged particles can be trapped in such a magnetic field, it is called a *magnetic bottle*. This technique is used to confine very hot plasmas with temperatures of the order of 10^6 K. In a similar way the earth's nonuniform magnetic field traps charged particles coming from the sun in doughnut-shaped regions around the earth, as shown in Fig. 10. These regions, called the *Van Allen radiation belts*, were discovered in 1958 using data obtained by instruments aboard the Explorer I satellite.

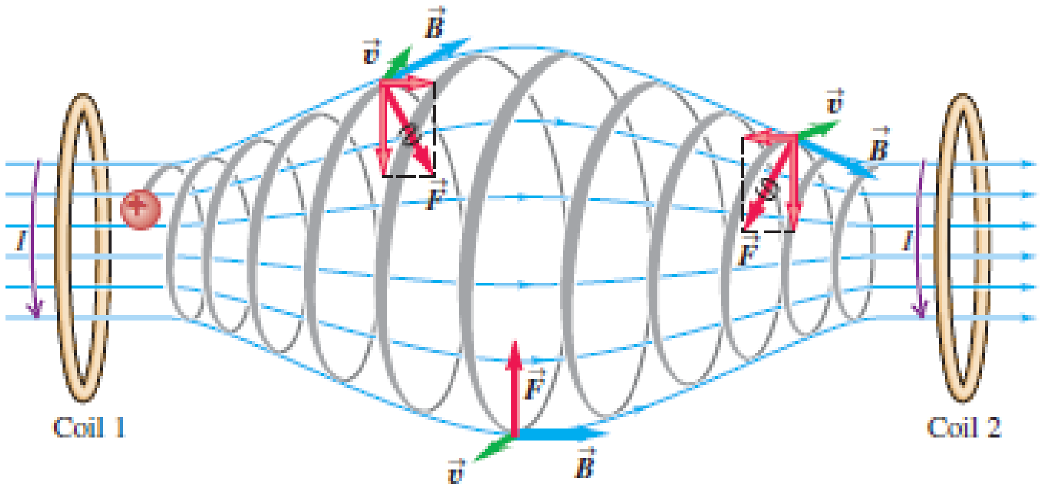


Figure 9 – A magnetic bottle. Particles near either end of the region experience a magnetic force toward the center of the region. This is one way of containing an ionized gas that has a temperature of the order of 10^6 K, which would vaporize any material container

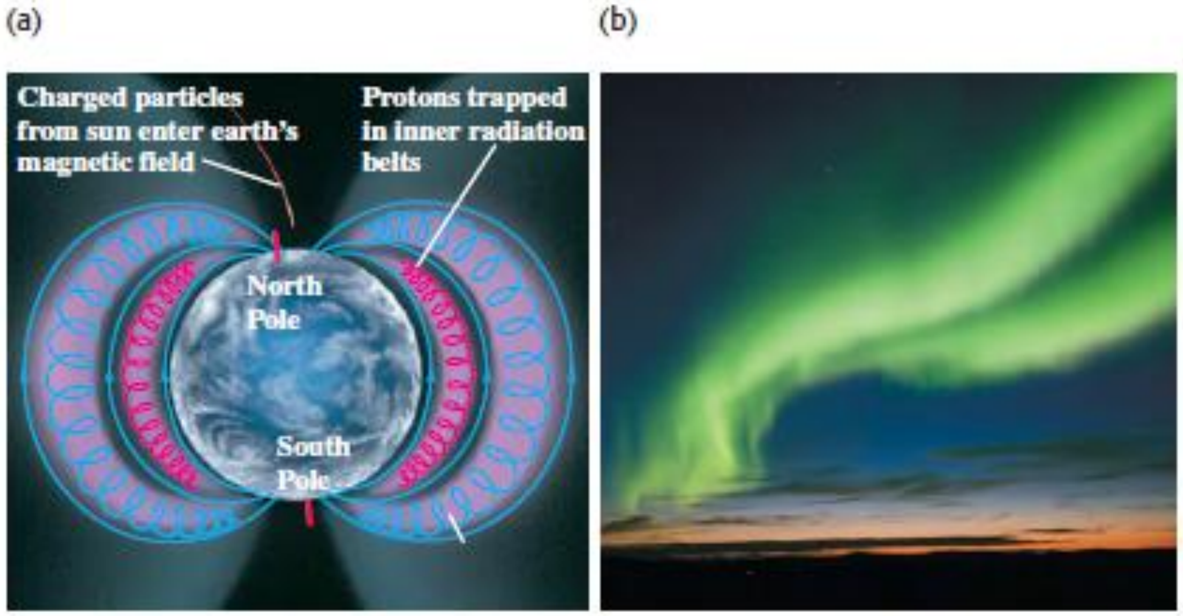

(a)
(b)
Charged particles from sun enter earth's magnetic field
Protons trapped in inner radiation belts
North Pole
South Pole

Figure 10 – (a) The Van Allen radiation belts around the earth. Near the poles, charged particles from these belts can enter the atmosphere, producing the aurora borealis (“northern lights”) and aurora australis (“southern lights”). (b) A photograph of the aurora borealis

Magnetic forces on charged particles play an important role in studies of elementary particles. Figure 11 shows a chamber filled with liquid hydrogen and with a magnetic field directed into the plane of the photograph. A high-energy gamma ray dislodges an electron from a hydrogen atom, sending it off at high speed and creating a visible track in the liquid hydrogen. The track shows the electron curving downward due to the magnetic force. The energy of the collision also produces another electron and a *positron* (a positively charged electron). Because of their opposite charges, the trajectories of the electron and the positron curve in opposite directions. As these particles plow through the liquid hydrogen, they collide with other charged particles, losing energy and speed. As a result, the radius of curvature decreases as suggested by Eq. (5). (The electron's speed is comparable to the speed of light, so Eq. (5) isn't directly applicable here.) Similar experiments allow physicists to determine the mass and charge of newly discovered particles.

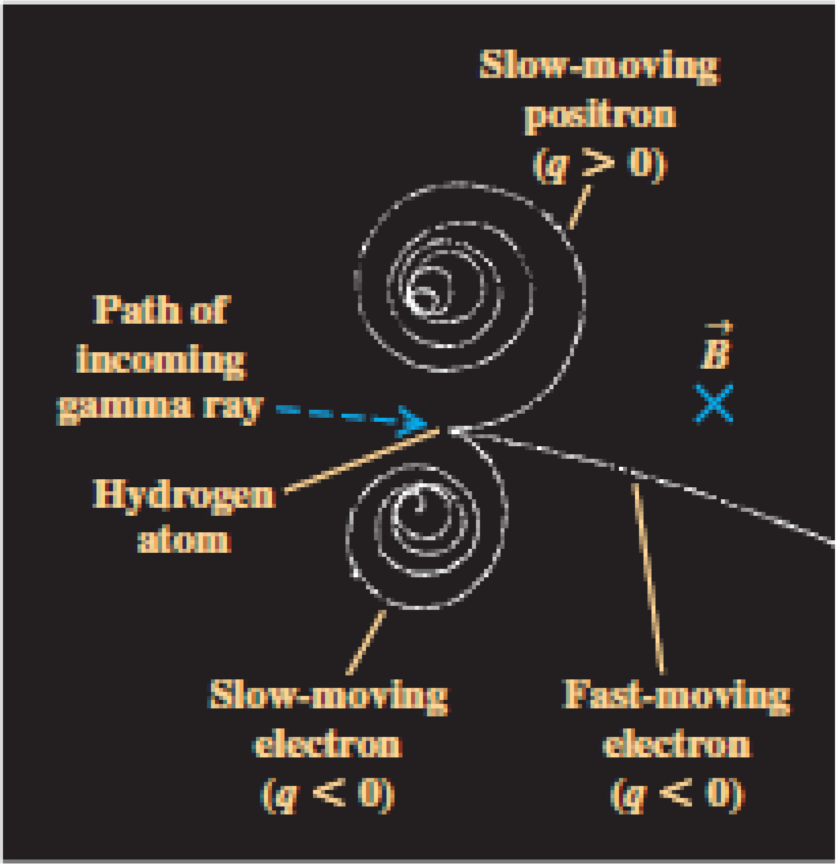

Slow-moving positron ($q > 0$)
Path of incoming gamma ray
$\vec{B}$
Hydrogen atom
Slow-moving electron ($q < 0$)
Fast-moving electron ($q < 0$)

Figure 11 – This bubble chamber image shows the result of a high-energy gamma ray (which does not leave a track) that collides with an electron in a hydrogen atom. This electron flies off to the right at high speed. Some of the energy in the collision is transformed into a second electron and a positron (a positively charged electron). A magnetic field is directed into the plane of the image, which makes the positive and negative particles curve off in different directions

1.1.3 Magnetic Force on a Current-Carrying Conductor

What makes an electric motor work? Within the motor are conductors that carry currents (that is, whose charges are in motion), as well as magnets that exert forces on the moving charges. Hence there is a magnetic force along the length of each current-carrying conductor, and these forces make the motor turn. The moving-coil galvanometer also uses magnetic forces on conductors.

We can compute the force on a current-carrying conductor starting with the magnetic force $\vec{F} = q\vec{v} \times \vec{B}$ on a single moving charge. Figure 12 shows a straight segment of a conducting wire, with length l and cross-sectional area S; the current is from bottom to top. The wire is in a uniform magnetic field $\vec{B}$, perpendicular to the plane of the diagram and directed *into* the plane. Let's assume first that the moving charges are positive. Later we'll see what happens when they are negative.

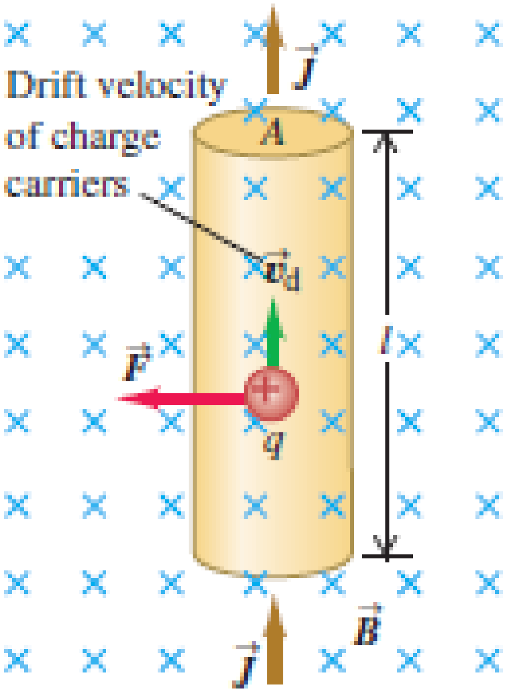


Figure 12 - Forces on a moving positive charge in a current-carrying conductor

The drift velocity $\vec{v}_d$ is upward, perpendicular to $\vec{B}$. The average force on each charge is $\vec{F} = q\vec{v}_d \times \vec{B}$ directed to the left as shown in the figure; since $\vec{v}_d$ and $\vec{B}$ are perpendicular, the magnitude of the force is $F = qv_dB$.

We can derive an expression for the *total* force on all the moving charges in a length l of conductor with cross-sectional area S using the same language we used in Eqs. $I = nqv_dS$ and $j = nqv_d$. The number of charges per unit volume is n; a segment of conductor with length l has volume Sl and contains a number of charges equal to nSl. The total force $\vec{F}$ on *all* the moving charges in this segment has magnitude

$$F = (nAl)(qv_dB) = (nqv_dS)(lB) \qquad (7)$$

The current density is $J = nqv_d$. The product JS is the total current I so we can rewrite Eq. (7) as

$$F = IBl \qquad (8)$$

If the $\vec{B}$ field is not perpendicular to the wire but makes an angle α with it, we handle the situation the same way we did early for a single charge. Only the component of $\vec{B}$ perpendicular to the wire (and to the drift velocities of the charges) exerts a force; this component is $B_{\perp} = B \sin \alpha$. The magnetic force on the wire segment is then

$$F = IB_{\perp}l = IBl \sin \alpha \tag{9}$$

The force is always perpendicular to both the conductor and the field, with the direction determined by the same right-hand rule we used for a moving positive charge (Fig. 13). Hence this force can be expressed as a vector product, just like the force on a single moving charge. We represent the segment of wire with a vector $\vec{l}$ along the wire in the direction of the current; then the force $\vec{F}$ on this segment is

$$\vec{F} = I\vec{B} \times \vec{l} \tag{10}$$

Figure 14 illustrates the directions of $\vec{B}, \vec{l}$ and $\vec{F}$ for several cases.

If the conductor is not straight, we can divide it into infinitesimal segments $d\vec{l}$. The force $d\vec{F}$on each segment is

$$d\vec{F} = Id\vec{l} \times \vec{B} \tag{11}$$

Then we can integrate this expression along the wire to find the total force on a conductor of any shape. The integral is a *line integral*, the same mathematical operation we have used to define work and electric potential.

Force $\vec{F}$ on a straight wire carrying a positive current and oriented at an angle ϕ to a magnetic field $\vec{B}$:

- Magnitude is $F = IlB_{\perp} = IlB \sin \phi$.
- Direction of $\vec{F}$ is given by the right-hand rule.

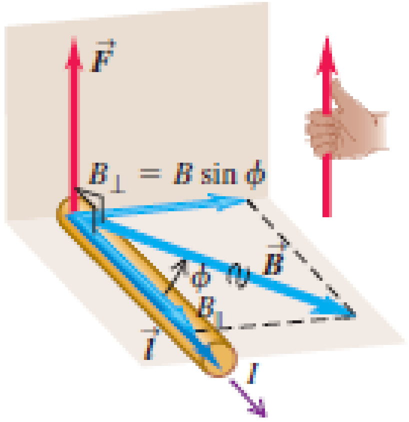


Figure 13 – A straight wire segment of length $\vec{l}$ carries a current I in the direction of $\vec{l}$. The magnetic force on this segment is perpendicular to both $\vec{l}$ ans the magnetic field $\vec{B}$

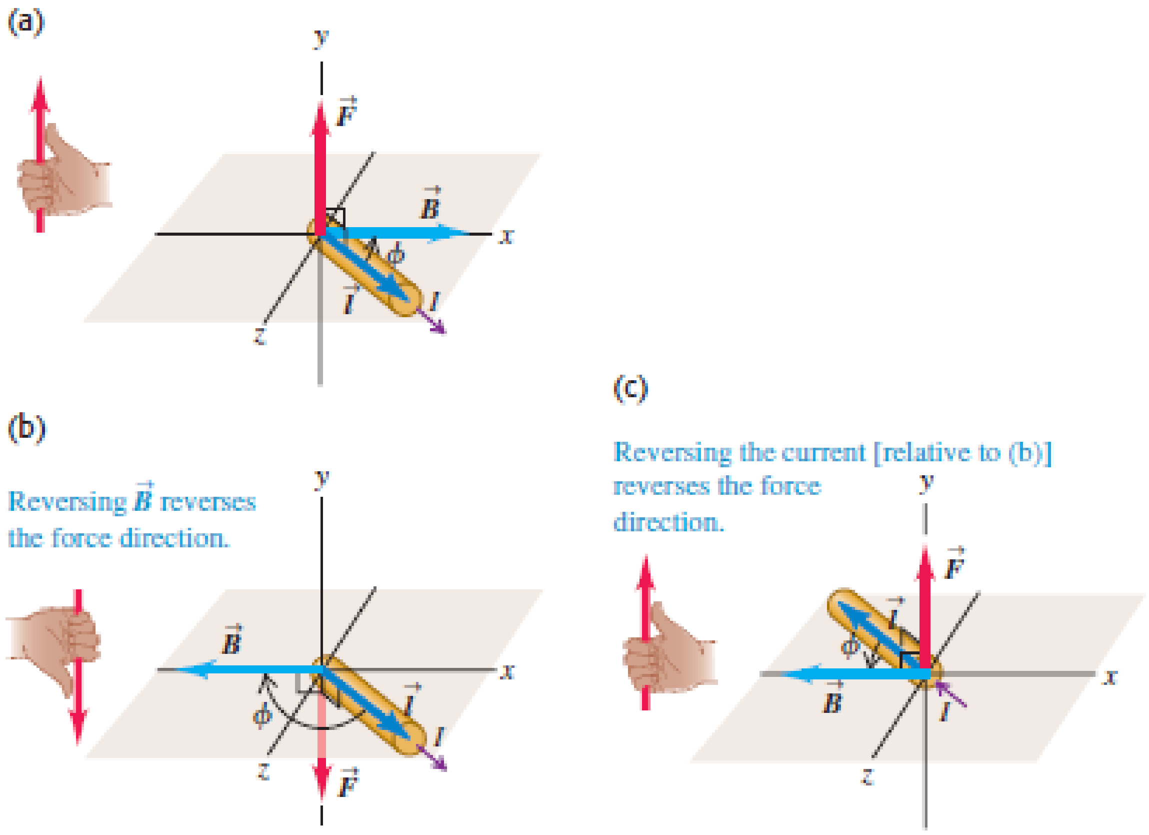


Figure 14 – Magnetic field $\vec{B}$, length $\vec{l}$, and force $\vec{F}$ vectors for a straight wire carrying a current I

Finally, what happens when the moving charges are negative, such as electrons in a metal? Then in Fig. 12 an upward current corresponds to a downward drift velocity. But because is now negative, the direction of the force $\vec{F}$ is the same as before. Thus Eqs. (8) through (11) are valid for *both* positive and negative charges and even when *both* signs of charge are present at once. This happens in some semiconductor materials and in ionic solutions.

A common application of the magnetic forces on a current-carrying wire is found in loudspeakers (Fig. 15). The radial magnetic field created by the permanent magnet exerts a force on the voice coil that is proportional to the current in the coil; the direction of the force is either to the left or to the right, depending on the direction of the current. The signal from the amplifier causes the current to oscillate in direction and magnitude. The coil and the speaker cone to which it is attached respond by oscillating with an amplitude proportional to the amplitude of the current in the coil. Turning up the volume knob on the amplifier increases the current amplitude and hence the amplitudes of the cone's oscillation and of the sound wave produced by the moving cone.

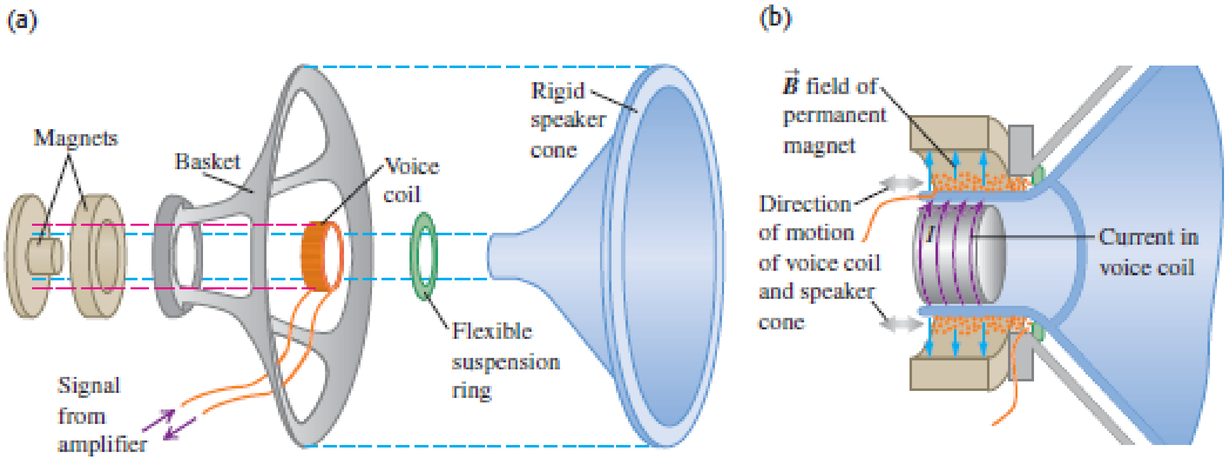


Figure 15 – (a) Components of a loudspeaker. (b) The permanent magnet creates a magnetic field that exerts forces on the current in the voice coil; for a current *I* in the direction shown, the force is to the right. If the electric current in the voice coil oscillates, the speaker cone attached to the voice coil oscillates at the same frequency

Discussion questions

1. Can a charged particle move through a magnetic field without experiencing any force? If so, how? If not, why not?
2. At any point in space, the electric field *E* is defined to be in the direction of the electric force on a positively charged particle at that point. Why don't we similarly define the magnetic field *B* to be in the direction of the magnetic force on a moving, positively charged particle?
3. The magnetic force on a moving charged particle is always perpendicular to the magnetic field *B*. Is the trajectory of a moving charged particle always perpendicular to the magnetic field lines? Explain your reasoning.
4. A charged particle is fired into a cubical region of space where there is a uniform magnetic field. Outside this region, thereis no magnetic field. Is it possible that the particle will remain inside the cubical region? Why or why not?
5. If the magnetic force does no work on a charged particle, how can it have any effect on the particle's motion? Are there other examples of forces that do no work but have a significant effect on a particle's motion?
6. A charged particle moves through a region of space with constant velocity (magnitude and direction). If the external magnetic field is zero in this region, can you conclude that the external electric field in the region is also zero? Explain. (By "external" we mean fields other than those produced by the charged particle.) If the external electric field is zero in the region, can you conclude that the external magnetic field in the region is also zero?
7. How might a loop of wire carrying a current be used as a compass? Could such a compass distinguish between north and south? Why or why not?
8. How could the direction of a magnetic field be determined by making only qualitative observations of the magnetic force on a straight wire carrying a current?

9. A loose, floppy loop of wire is carrying current The loop of wire is placed on a horizontal table in a uniform magnetic field *B* perpendicular to the plane of the table. This causes the loop of wire to expand into a circular shape while still lying on the table. In a diagram, show all possible orientations of the current and magnetic field that could cause this to occur. Explain your reasoning.
10. Several charges enter a uniform magnetic field directed into the page. (a) What path would a positive charge moving with a velocity of magnitude follow through the field? (b) What path would a positive charge q moving with a velocity of magnitude follow through the field? (c) What path would a negative charge –*q* moving with a velocity of magnitude *v* follow through the field? (d) What path would a neutral particle follow through the field?
11. A student claims that if lightning strikes a metal flagpole, the force exerted by the earth's magnetic field on the current in the pole can be large enough to bend it. Typical lightning currents are of the order of 10^4 to 10^5 A. Is the student's opinion justified? Explain your reasoning.
12. Could an accelerator be built in which all the forces on the particles, for steering and for increasing speed, are magnetic forces? Why or why not?
13. The magnetic force acting on a charged particle can never do work because at every instant the force is perpendicular to the velocity. The torque exerted by a magnetic field can do work on a current loop when the loop rotates. Explain how these seemingly contradictory statements can be reconciled.
14. If an emf is produced in a dc motor, would it be possible to use the motor somehow as a generator or source, taking power out of it rather than putting power into it? How might this be done?
15. When the polarity of the voltage applied to a dc motor is reversed, the direction of motion does not reverse. Why not? How could the direction of motion be reversed?
16. In a Hall-effect experiment, is it possible that no transverse potential difference will be observed? Under what circumstances might this happen?
17. Hall-effect voltages are much greater for relatively poor conductors (such as germanium) than for good conductors (such as copper), for comparable currents, fields, and dimensions. Why?

1.2 Sources of magnetic field

1.2.1 Magnetic Field of a Moving Charge

Let's start with the basics, the magnetic field of a single point charge *q* moving with a constant velocity $\vec{v}$. In practical applications, such as the solenoid shown in the photo that opens this chapter, magnetic fields are produced by tremendous numbers of charged particles moving together in a current. But once we understand how to calculate the magnetic field due to a single point charge, it's a small leap to calculate the field due to a current-carrying wire or collection of wires.

As we did for electric fields, we call the location of the moving charge at a given instant the **source point** and the point *P* where we want to find the field the **field point.** Early we found that at a field point a distance *r* from a point charge *q*, the magnitude of the *electric* field $\vec{E}$ caused by the charge is proportional to the charge magnitude $|q|$ and to $1/r^2$, and the direction of $\vec{E}$ (for positive *q*) is along the line from source point to field point. The corresponding relationship for the *magnetic* field $\vec{B}$ of a point charge q moving with constant velocity has some similarities and some interesting differences.

Experiments show that the magnitude of $\vec{B}$ is also proportional to $|q|$ and to $1/r^2$. But the *direction* of $\vec{B}$ is *not* along the line from source point to field point. Instead, $\vec{B}$ is perpendicular to the plane containing this line and the particle's velocity vector $\vec{v}$, as shown in Fig. 16. Furthermore, the field *magnitude B* is also proportional to the particle's speed v and to the sine of the angle α. Thus the magnetic field magnitude at point *P* is given by

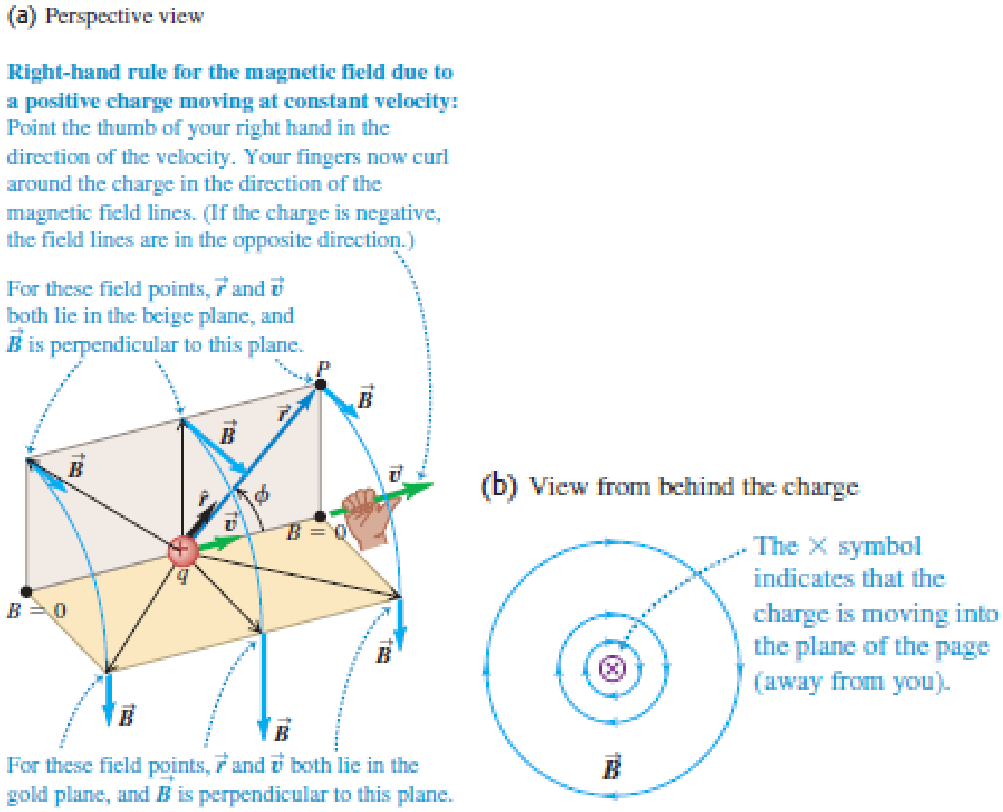
$$B = \frac{\mu_0}{4\pi}\frac{|q|v\sin\alpha}{r^2} \tag{12}$$

where $\frac{\mu_0}{4\pi}$ is a proportionality constant (μ_0 is read as "mu-nought" or "mu-subzero"). The reason for writing the constant in this particular way will emerge shortly. We did something similar with Coulomb's law.

We can incorporate both the magnitude and direction of $\vec{B}$ into a single vector equation using the vector product. To avoid having to say "the direction from the source q to the field point *P*" over and over, we introduce a *unit* vector $\hat{r}$ ("r-hat") that points from the source point to the field point. This unit vector is equal to the vector $\vec{r}$ from the source to the field point divided by its magnitude $\hat{r} = \vec{r}/r$: Then the field of a moving point charge is

$$\vec{B} = \frac{\mu_0}{4\pi}\frac{|q|\vec{v}\times\hat{r}}{r^2} \tag{13}$$

Figure 16 shows the relationship of $\hat{r}$ to *P* and also shows the magnetic field $\vec{B}$ at several points in the vicinity of the charge. At all points along a line through the charge parallel to the velocity $\vec{v}$, the field is zero because $\sin\alpha \neq 0$ at all such points. At any distance r from q, $\vec{B}$ has its greatest magnitude at points lying in the plane perpendicular to $\vec{v}$, because there $\alpha = 90$ and $\sin\alpha = 1$. If q is negative, the directions of $\vec{B}$ are opposite to those shown in Fig. 16.



(a) Perspective view

Right-hand rule for the magnetic field due to a positive charge moving at constant velocity: Point the thumb of your right hand in the direction of the velocity. Your fingers now curl around the charge in the direction of the magnetic field lines. (If the charge is negative, the field lines are in the opposite direction.)

For these field points, $\vec{r}$ and $\vec{v}$ both lie in the beige plane, and $\vec{B}$ is perpendicular to this plane.

For these field points, $\vec{r}$ and $\vec{v}$ both lie in the gold plane, and $\vec{B}$ is perpendicular to this plane.

(b) View from behind the charge

The × symbol indicates that the charge is moving into the plane of the page (away from you).

Figure 16 – (a) Magnetic-field vectors due to moving positive point charge q. At each point, $\vec{B}$ is perpendicular to the plane of $\vec{r}$ and $\vec{v}$, and its magnitude is proportional to the sine of the angle between them. (b) Magnetic field lines in a plane containing moving positive charge

A point charge in motion also produces an *electric* field, with field lines that radiate outward from a positive charge. The *magnetic* field lines are completely different. For a point charge moving with velocity $\vec{v}$, the magnetic field lines are *circles* centered on the line of $\vec{v}$ and lying in planes perpendicular to this line. The field-line directions for a positive charge are given by the following *right-hand rule*, one of several that we will encounter in this chapter: Grasp the velocity vector $\vec{v}$ with your right hand so that your right thumb points in the direction of $\vec{v}$; your fingers then curl around the line of $\vec{v}$ in the same sense as the magnetic field lines, assuming q is positive. Figure 16a shows parts of a few field lines; Fig. 16b shows some field lines in a plane through q, perpendicular to $\vec{v}$. If the point charge is negative, the directions of the field and field lines are the opposite of those shown in Fig. 16.

Equations (12) and (13) describe the $\vec{B}$ field of a point charge moving with *constant* velocity. If the charge *accelerates*, the field can be much more complicated. We won't need these more complicated results for our purposes. (The moving charged particles that make up a current in a wire accelerate at points where the wire bends and the direction of $\vec{v}$ changes. But because the magnitude v_d of the drift velocity in a conductor is typically very small, the centripetal acceleration v_d^2/r is so small that we can ignore its effects.)

As we discussed early, the unit of B is one tesla (1 T):

$$1\,T = 1\,N \cdot \frac{s}{C} \cdot m = 1\,N/A \cdot m$$

Using this with Eq. (12) or (13), we find that the units of the constant μ_0 are

$$1\,N \cdot s^2/C^2 = 1\,N/A^2 = 1\,Wb/A \cdot m = 1\,T \cdot m/A$$

In SI units the numerical value of μ_0 is exactly $4\pi \times 10^{-7}$. Thus

$$\mu_0 = 4\pi \times 10^{-7}\,N \cdot s^2/C^2 = 4\pi \times 10^{-7}\,Wb/A \cdot m = 4\pi \times 10^{-7}\,T \cdot m/A$$

It may seem incredible that μ_0 has *exactly* this numerical value! In fact this is a *defined* value that arises from the definition of the ampere.

The constant in Coulomb's law is related to the speed of light *c*:

$$k = \frac{1}{4\pi\varepsilon_0} = (10^{-7}\,N \cdot s^2/C^2)c^2$$

When we study electromagnetic waves, we will find that their speed of propagation in vacuum, which is equal to the speed of light is given by

$$c^2 = \frac{1}{\varepsilon_0\mu_0} \qquad (14)$$

If we solve the equation $k = \frac{1}{4\pi\varepsilon_0}$ for ε_0, substitute the resulting expression into Eq. (14), and solve for μ_0 we indeed get the value of μ_0 stated above. This discussion is a little premature, but it may give you a hint that electric and magnetic fields are intimately related to the nature of light.

1.2.2 Magnetic Field of a Straight Current-Carrying Conductor

Let's use the law of Biot and Savart to find the magnetic field produced by a straight current-carrying conductor. This result is useful because straight conducting wires are found in essentially all electric and electronic devices. Figure 17 shows such a conductor with length 2*a* carrying a current *I*. We will find $\vec{B}$ at a point a distance *x* from the conductor on its perpendicular bisector.

We first use the law of Biot and Savart, Eq. $B = \frac{\mu_0}{4\pi}\frac{Idl\sin\alpha}{r^2}$, to find the field $d\vec{B}$ caused by the element of conductor of length $dl = dy$ shown in Fig. 17. From the figure, $r = \sqrt{x^2 + y^2}$ and $\sin\alpha = \sin(\pi - \alpha) = \frac{x}{\sqrt{x^2+y^2}}$. The right-hand rule for the vector product $d\vec{l} \times \hat{r}$ shows that the *direction* of $\vec{B}$ is into the plane of the figure, perpendicular to the plane; furthermore, the directions of the $d\vec{B}'s$ from *all* elements of the conductor are the same. Thus in integrating Eq.$\vec{B} = \frac{\mu_0}{4\pi}\int \frac{I d\vec{l} \times \hat{r}}{r^2}$, we can just add the *magnitudes* of the $d\vec{B}'s$ a significant simplification.

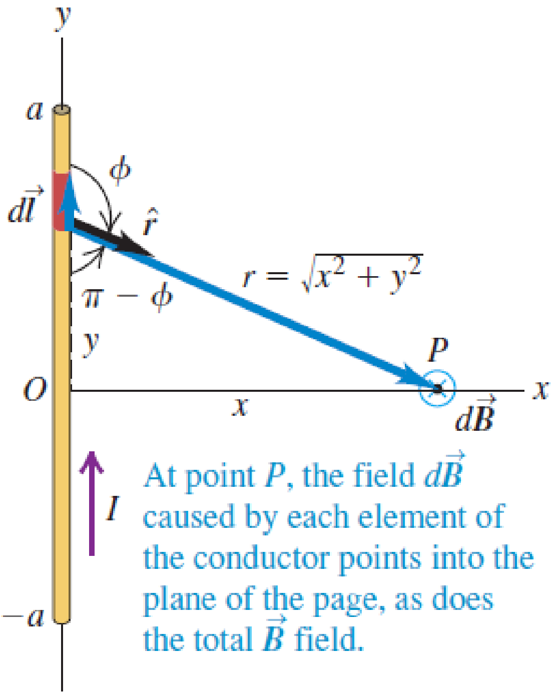

y
a
$d\vec{l}$
ϕ
$\hat{r}$
$r = \sqrt{x^2 + y^2}$
$\pi - \phi$
y
P
O
x
x
$d\vec{B}$
I
−a
At point P, the field $d\vec{B}$ caused by each element of the conductor points into the plane of the page, as does the total $\vec{B}$ field.

Figure 17 – Magnetic produced by a straight current-carrying conductor of 2*a*

Putting the pieces together, we find that the magnitude of the total $\vec{B}$ field is

$$B = \frac{\mu_0 I}{4\pi} \int_{-a}^{a} \frac{x dy}{(x^2 + y^2)^{3/2}} \tag{15}$$

We can integrate this by trigonometric substitution or by using an integral table:

$$B = \frac{\mu_0 I}{4\pi} \frac{2a}{x\sqrt{x^2 + a^2}} \tag{16}$$

When the length 2a of the conductor is very great in comparison to its distance *x* from point *P*, we can consider it to be infinitely long. When a is much larger than *x*, $\sqrt{x^2 + a^2}$ is approximately equal to *a*; hence in the limit $a \to \infty$, Eq. (16) becomes

$$B = \frac{\mu_0 I}{4\pi x} \tag{17}$$

The physical situation has axial symmetry about the *y*-axis. Hence $\vec{B}$ must have the same *magnitude* at all points on a circle centered on the conductor and lying in a plane perpendicular to it, and the *direction* of $\vec{B}$ must be everywhere tangent to such a circle (Fig. 18). Thus, at all points on a circle of radius r around the conductor, the magnitude B is

$$B = \frac{\mu_0 I}{4\pi r} \tag{18}$$

The geometry of this problem is similar to that of Example, in which we solved the problem of the *electric* field caused by an infinite line of charge. The same integral appears in both problems, and the field magnitudes in both problems are proportional to $1/r$. But the lines of $\vec{B}$ in the magnetic problem have completely different shapes than the lines of $\vec{E}$ in the analogous electrical problem. Electric field lines radiate outward from a positive line charge distribution (inward for negative charges). By contrast, magnetic field lines *encircle* the current that acts as their source. Electric field lines due to charges begin and end at those charges, but magnetic field lines always form closed loops and *never* have end points, irrespective of the shape of the current-carrying conductor that sets up the field. As we discussed early, this is a consequence of Gauss's law for magnetism, which states that the total magnetic flux through *any* closed surface is always zero:

$$\oint \vec{B}\, d\vec{S} = 0 \tag{19}$$

Any magnetic field line that enters a closed surface must also emerge from that surface.

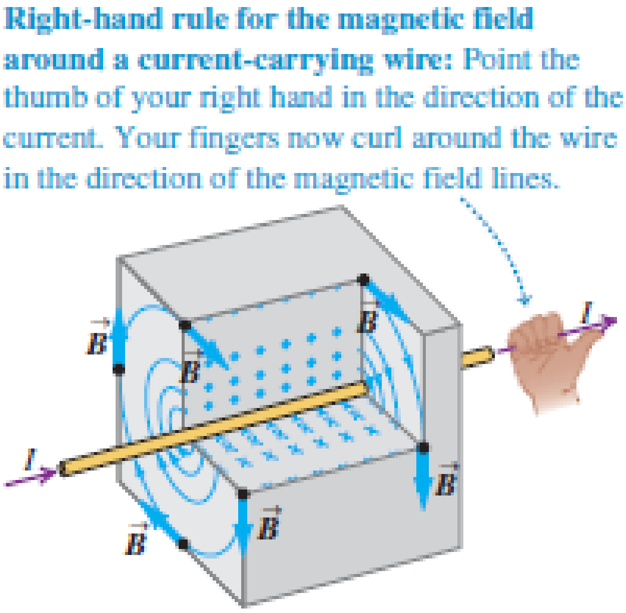

Right-hand rule for the magnetic field around a current-carrying wire: Point the thumb of your right hand in the direction of the current. Your fingers now curl around the wire in the direction of the magnetic field lines.

Figure 18 – Magnetic field around a long, straight, current-carrying conductor. The field lines are circles, with directions determined by the right-hand rule

1.2.3 Magnetic Field of a Circular Current Loop

If you look inside a doorbell, a transformer, an electric motor, or an electromagnet (Fig. 19), you will find coils of wire with a large number of turns, spaced so closely that each turn is very nearly a planar circular loop. A current in such a coil is used to establish a magnetic field. So it is worthwhile to derive an expression for the magnetic field produced by a single circular conducting loop carrying a current or by N closely spaced circular loops forming a coil. Early we considered the force and torque on such a current loop placed in an external magnetic field produced by other currents; we are now about to find the magnetic field produced by the loop itself.

Figure 19 – This electromagnet contains a current-carrying coil with numerous turns of wire. The resulting magnetic field can pick up large quantities of steel bars and other iron-bearing items

Figure 20 shows a circular conductor with radius a. A current I is led into and out of the loop through two long, straight wires side by side; the currents in these straight wires are in opposite directions, and their magnetic fields very nearly cancel each other.

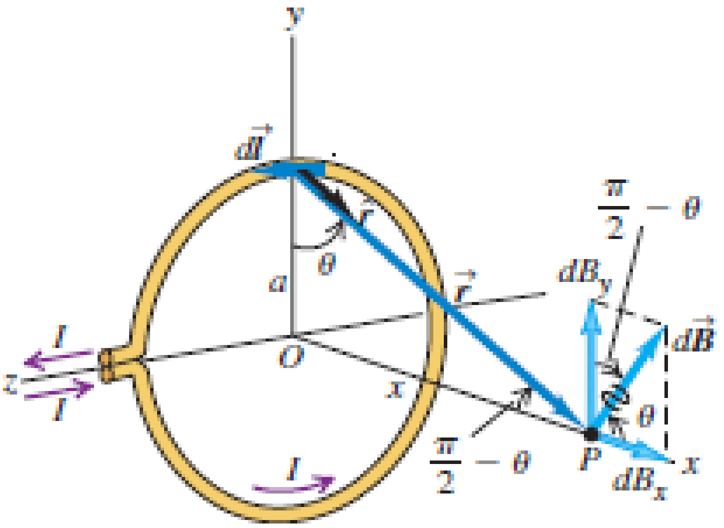


Figure 20 – Magnetic field on the axis of a circular loop. The current in the segment $d\vec{l}$ cause the field $d\vec{B}$, which lies in the xy-plane. The current in other $d\vec{l}$'s cause $d\vec{B}$'s with different components perpendicular to the x-axis; these components add to zero. The x-components of the $d\vec{B}$'s combine to give the total $\vec{B}$ field at point P

We can use the law of Biot and Savart, to find the magnetic field at a point P on the axis of the loop, at a distance x from the center. As the figure shows, $\overrightarrow{dl}$ and $\hat{r}$ are perpendicular, and the direction of the field $d\vec{B}$ caused by this particular element $d\vec{l}$ lies in the *xy*-plane. Since $r^2 = x^2 + a^2$ the magnitude dB of the field due to element $d\vec{l}$ is

$$dB = \frac{\mu_0 I}{4\pi}\frac{dl}{(x^2 + a^2)} \tag{20}$$

The components of the vector are

$$dB_x = Bdl\cos\theta = \frac{\mu_0 I}{4\pi}\frac{dl}{(x^2 + a^2)}\frac{a}{(x^2 + a^2)^{1/2}} \tag{21}$$

$$dB_y = Bdl\sin\theta = \frac{\mu_0 I}{4\pi}\frac{dl}{(x^2 + a^2)}\frac{x}{(x^2 + a^2)^{1/2}} \tag{22}$$

The *total* field $\vec{B}$ at P has only an *x*-component (it is perpendicular to the plane of the loop). Here's why: For every element $d\vec{l}$ there is a corresponding element on the opposite side of the loop, with opposite direction. These two elements give equal contributions to the *x*-component of $\vec{B}$ given by Eq. (21), but *opposite* components perpendicular to the *x*-axis. Thus all the perpendicular components cancel and only the *x*-components survive.

To obtain the x-component of the total field $\vec{B}$, we integrate Eq. (21), including all the $d\vec{l}$'s around the loop. Everything in this expression except dl is constant and can be taken outside the integral, and we have

$$B_x = \int \frac{\mu_0 I}{4\pi} \frac{a dl}{(x^2 + a^2)^{3/2}} = \frac{\mu_0 I a}{4\pi (x^2 + a^2)^{3/2}} \int dl \tag{23}$$

The integral of dl is just the circumference of the circle, $\int dl = 2\pi a$, and we finally get

$$B_x = \frac{\mu_0 I a^2}{4(x^2 + a^2)^{3/2}} \tag{24}$$

The *direction* of the magnetic field on the axis of a current-carrying loop is given by a right-hand rule. If you curl the fingers of your right hand around the loop in the direction of the current, your right thumb points in the direction of the field (Fig. 21).

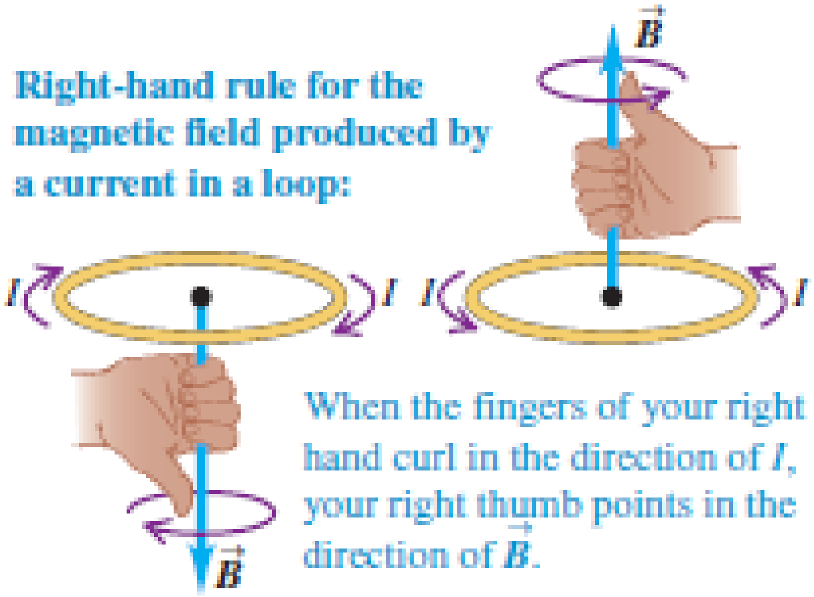

$\vec{B}$
Right-hand rule for the magnetic field produced by a current in a loop:
I
I
I
I
When the fingers of your right hand curl in the direction of I, your right thumb points in the direction of $\vec{B}$.
$\vec{B}$

Figure 21 – the right-hand rule for the direction of the magnetic field produced on the axis of a current-carrying coil

Magnetic Field on the Axis of a Coil. Now suppose that instead of the single loop in Fig. 20 we have a coil consisting of N loops, all with the same radius. The loops are closely spaced so that the plane of each loop is essentially the same distance x from the field point P. Then the total field is N times the field of a single loop:

$$B_x = \frac{\mu_0 N I a^2}{4(x^2 + a^2)^{3/2}} \tag{25}$$

The factor *N* in Eq. (26) is the reason coils of wire, not single loops, are used to produce strong magnetic fields; for a desired field strength, using a single loop might require a current *I* so great as to exceed the rating of the loop's wire.

Figure 22 shows a graph of as a function of x. The maximum value of the field is at the center of the loop or coil:

$$B_x = \frac{\mu_0 I}{2a} \tag{26}$$

As we go out along the axis, the field decreases in magnitude.

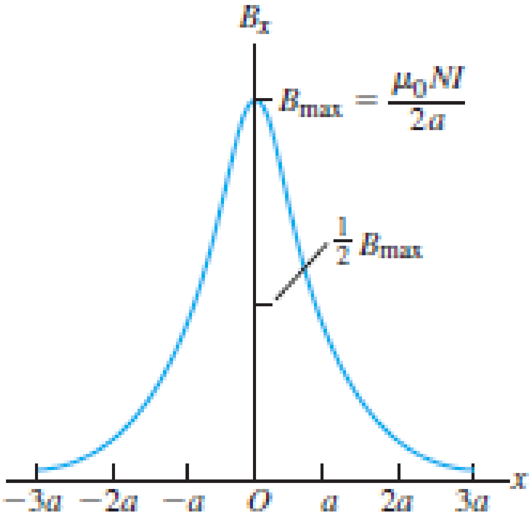


Figure 22 – Graph of the magnetic field along the axis of a circular coil with *N* turns. When x is much larger than a, the field magnitude decreases approximately as $1/x^3$

Soon we defined the *magnetic dipole moment* (or *magnetic moment*) of a current-carrying loop to be equal to *IA*, where *A* is the cross-sectional area of the loop. If there are *N* loops, the total magnetic moment is *NIA*. The circular loop in Fig. 20 has area so the magnetic moment of a single loop is for *N* loops, Substituting these results into Eqs. (24) and (25), we find that both of these expressions can be written as

$$B_x = \frac{\mu_0 \mu}{2\pi(x^2 + a^2)^{3/2}} \tag{27}$$

We described a magnetic dipole other section in terms of its response to a magnetic field produced by currents outside the dipole. But a magnetic dipole is also a *source* of magnetic field; Eq. (27) describes the magnetic field *produced* by a magnetic dipole for points along the dipole axis. This field is directly proportional to the magnetic dipole moment Note that the field along the *x*-axis is in the same direction as the vector magnetic moment this is true on both the positive and negative *x*-axis.

Figure 23 shows some of the magnetic field lines surrounding a circular current loop (magnetic dipole) in planes through the axis. The directions of the field lines are given by the same right-hand rule as for a long, straight conductor. Grab the conductor with your right hand, with your thumb in the direction of the current; your fingers curl around in the same direction as the field lines. The field lines for the circular current loop are closed curves that encircle the conductor; they are *not* circles, however.

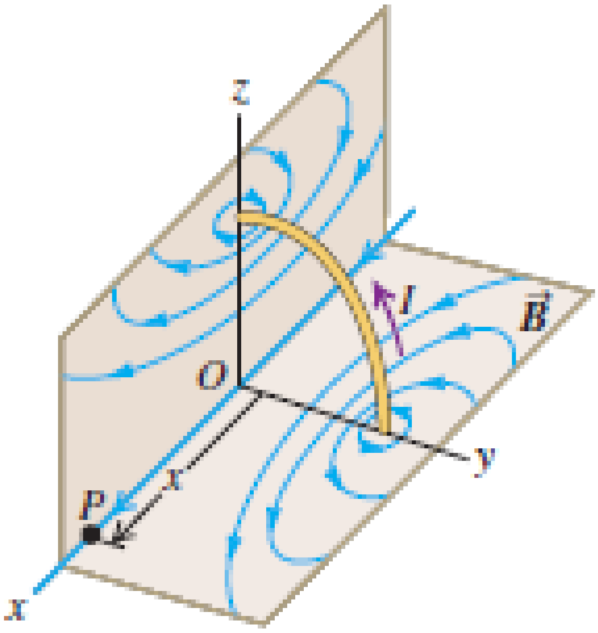


Figure 23 – Magnetic field lines produced by the current in a circular loop. At points on the axis the $\vec{B}$ field has the same direction as the magnetic moment of the loop

1.2.4 Amperes's force

What makes an electric motor work? Within the motor are conductors that carry currents (that is, whose charges are in motion), as well as magnets that exert forces on the moving charges. Hence there is a magnetic force along the length of each current-carrying conductor, and these forces make the motor turn. The moving-coil galvanometer also uses magnetic forces on conductors.

We can compute the force on a current-carrying conductor starting with the magnetic force $\vec{F} = q\vec{v} \times \vec{B}$ on a single moving charge. Figure 24 shows a straight segment of a conducting wire, with length and cross-sectional area S the current is from bottom to top. The wire is in a uniform magnetic field $\vec{B}$, perpendicular to the plane of the diagram and directed into the plane. Let's assume first that the moving charges are positive. Later we'll see what happens when they are negative.

The drift velocity $\vec{v}_d$ is upward, perpendicular to $\vec{B}$. The average force on each charge is $\vec{F} = q\vec{v}_d \times \vec{B}$ directed to the left as shown in the figure; since and are perpendicular, the magnitude of the force is $F = qv_dB$.

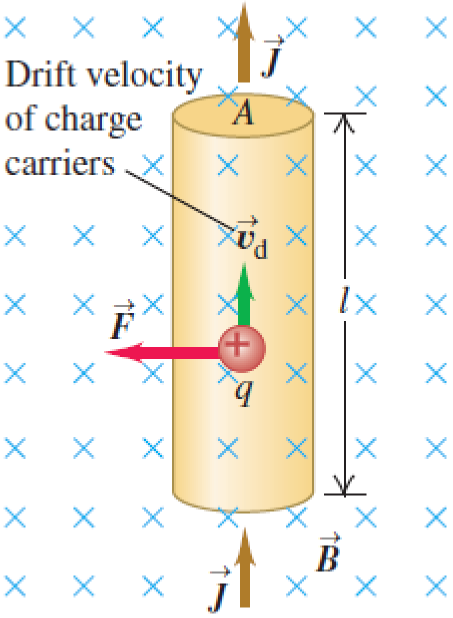


Figure 24 – Forces on a moving positive charge in a current-carrying conductor

We can derive an expression for the total force on all the moving charges in a length l of conductor with cross-sectional area S using the same language we used in Eqs. $I = nqv_dS$ and $j = nqv_d$. The number of charges per unit volume is n; a segment of conductor with length l has volume Sl and contains a number of charges equal to nSl. The total force $\vec{F}$ on all the moving charges in this segment has magnitude

$$F = (nSl)(qv_dB) = (nqv_dS)(lB) \tag{28}$$

The current density is $j = nqv_d$. The product is the total current so we can rewrite Eq. (7) as

$$F = IBl \tag{29}$$

If the field $\vec{B}$ is not perpendicular to the wire but makes an angle φ with it, we handle the situation the same way for a single charge. Only the component of $\vec{B}$ perpendicular to the wire (and to the drift velocities of the charges) exerts a force; this component is $B_\perp = B\sin\varphi$. The magnetic force on the wire segment is then

$$F = IB_\perp l = IBl\sin\varphi \tag{30}$$

The force is always perpendicular to both the conductor and the field, with the direction determined by the same right-hand rule we used for a moving positive charge (Fig. 25). Hence this force can be expressed as a vector product, just like the force on a single moving charge. We represent the segment of wire with a vector along the wire in the direction of the current; then the force on this segment is

$$\vec{F} = I\vec{l} \times \vec{B} \qquad (31)$$

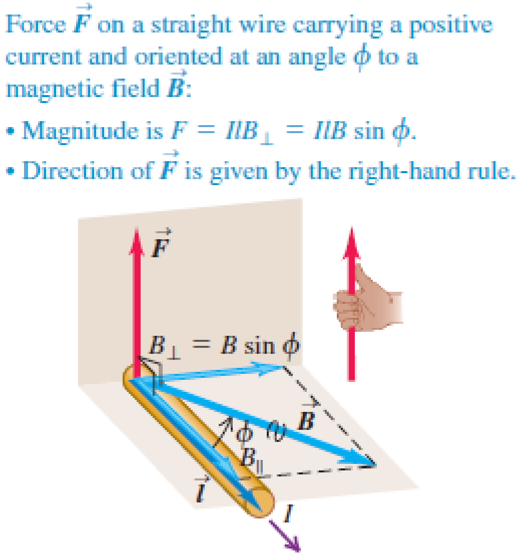

Force $\vec{F}$ on a straight wire carrying a positive current and oriented at an angle ϕ to a magnetic field $\vec{B}$:
• Magnitude is $F = IlB_{\perp} = IlB \sin \phi$.
• Direction of $\vec{F}$ is given by the right-hand rule.
$\vec{F}$
$B_{\perp} = B \sin \phi$
ϕ
$\vec{B}$
$B_{\parallel}$
$\vec{l}$
I

Figure 25 – A straight wire segment of length $\vec{l}$ carries a current I in the direction of $\vec{l}$. The magnetic force on this segment is perpendicular to both $\vec{l}$ and the magnetic field $\vec{B}$

Figure 26 illustrates the directions of and for several cases.

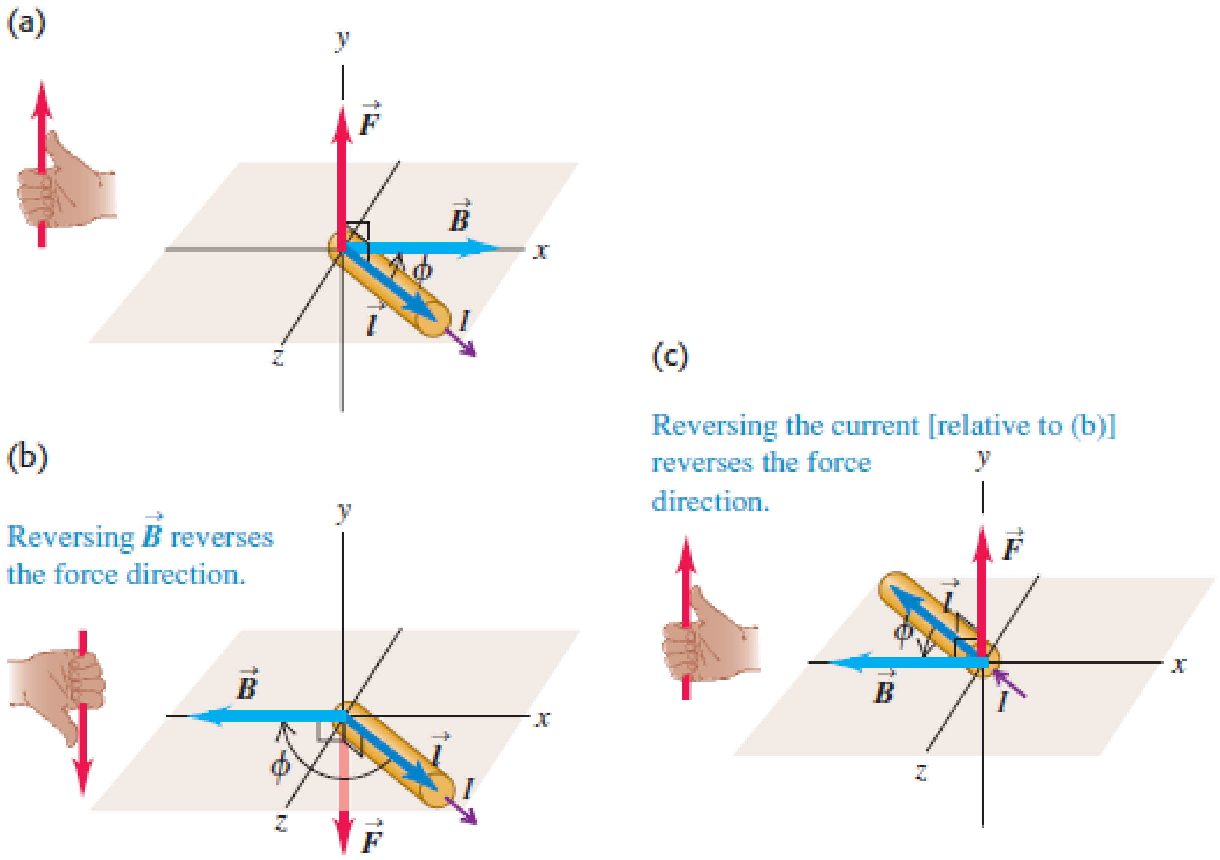


Figure 26 – Magnetic field $\vec{B}$, length $\vec{l}$, and force $\vec{F}$ vectors for a straight wire carrying a current I

If the conductor is not straight, we can divide it into infinitesimal segments $d\vec{l}$. The force $d\vec{F}$ on each segment is

$$d\vec{F} = I\vec{B} \times d\vec{l} \tag{32}$$

Then we can integrate this expression along the wire to find the total force on a conductor of any shape. The integral is a line integral, the same mathematical operation we have used to define work and electric potential.

Finally, what happens when the moving charges are negative, such as electrons in a metal? Then in Fig. 24 an upward current corresponds to a downward drift velocity. But because q is now negative, the direction of the force $\vec{F}$ is the same as before. Thus Eqs. (8) through (11) are valid for both positive and negative charges and even when both signs of charge are present at once. This happens in some semiconductor materials and in ionic solutions.

A common application of the magnetic forces on a current-carrying wire is found in loudspeakers (Fig. 27). The radial magnetic field created by the permanent magnet exerts a force on the voice coil that is proportional to the current in the coil; the direction of the force is either to the left or to the right, depending on the direction of the current. The signal from the amplifier causes the current to oscillate in direction and magnitude. The coil and the speaker cone to which it is attached respond by oscillating with an amplitude proportional to the amplitude of the current in the coil. Turning up the volume knob on the amplifier increases the current amplitude and hence the amplitudes of the cone's oscillation and of the sound wave produced by the moving cone.

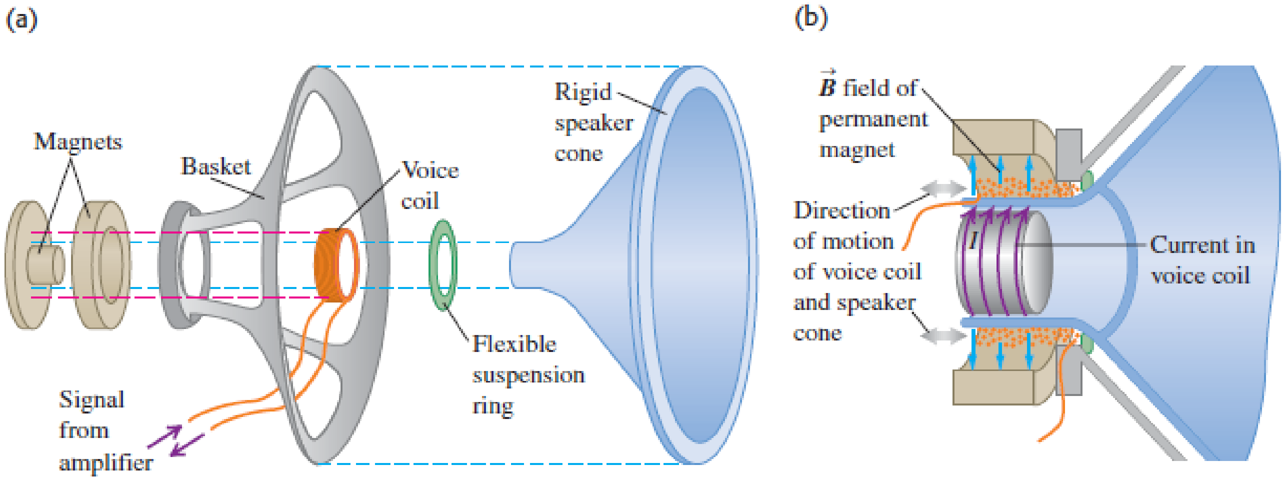


Figure 27 – (a) Components of a loudspeaker. (b) The permanent magnet creates a magnetic field that exerts forces on the current in the voice coil; for a current I in the direction shown, the force is to the right. If the electric current in the voice coil oscillates, the speaker cone attached to the voice coil oscillates at the same frequency.

1.2.5 Magnetic materials

In discussing how currents cause magnetic fields, we have assumed that the conductors are surrounded by vacuum. But the coils in transformers, motors, generators, and electromagnets nearly always have iron cores to increase the magnetic field and confine it to desired regions. Permanent magnets, magnetic recording tapes, and computer disks depend directly on the magnetic properties of materials; when you store information on a computer disk, you are actually setting up an array of microscopic permanent magnets on the disk. So it is worthwhile to examine some aspects of the magnetic properties of materials. After describing the atomic origins of magnetic properties, we will discuss three broad classes of magnetic behavior that occur in materials; these are called *paramagnetism*, *diamagnetism*, and *ferromagnetism.*

The Bohr Magneton. The atoms that make up all matter contain moving electrons, and these electrons form microscopic current loops that produce magnetic fields of their own. In many materials these currents are randomly oriented and cause no net magnetic field. But in some materials an external field (a field produced by currents outside the material) can cause these loops to become oriented preferentially with the field, so their magnetic fields *add* to the external field. We then say that the material is *magnetized.*

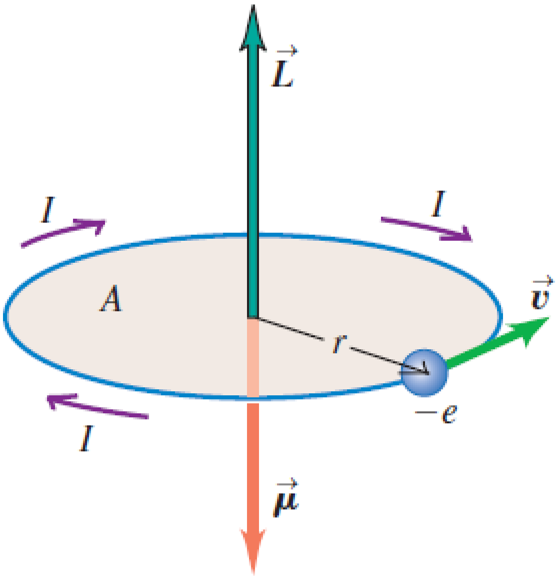


Figure 28 – An electron moving with speed *v* in a circular orbit of radius *r* has an angular momentum $\vec{L}$ and an oppositely directed orbital magnetic dipole moment $\vec{\mu}$. It also has a spin angular momentum and an oppositely directed spin magnetic dipole moment.

Let’s look at how these microscopic currents come about. Figure 28 shows a primitive model of an electron in an atom. We picture the electron (mass *m*, charge -*e*) as moving in a circular orbit with radius *r* and speed This moving charge is equivalent to a current loop. Early we found that a current loop with area *S* and current *I* has a magnetic dipole moment μ given by $\mu = IS$ for the orbiting electron the area of the loop is $S = \pi R^2$. To find the current associated with the electron, we note that the orbital period *T* (the time for the electron to make one complete orbit) is the orbit circumference divided by the electron speed: $T = 2\pi r/v$. The equivalent current *I* is the total charge passing any point on the orbit per unit time, which is just the magnitude *e* of the electron charge divided by the orbital period *T*:

$$I = \frac{e}{T} = \frac{ev}{2\pi r} \tag{33}$$

The magnetic moment $\mu = IS$ is then

$$\mu = \frac{ev}{2\pi r}(\pi r^2) = \frac{evr}{2} \tag{34}$$

It is useful to express μ in terms of the *angular momentum* L of the electron. For a particle moving in a circular path, the magnitude of angular momentum equals the magnitude of momentum multiplied by the radius r - that is, $L = mvr$. Comparing this with Eq. (34), we can write

$$\mu = \frac{e}{2m}L \tag{35}$$

Equation (35) is useful in this discussion because atomic angular momentum is *quantized;* its component in a particular direction is always an integer multiple of $h/2\pi$ where h is a fundamental physical constant called *Planck's constant.* The numerical value of *h* is

$$h = 6.626 \times 10^{-34}\, J \cdot s$$

The quantity $h/2\pi$ thus represents a fundamental unit of angular momentum in atomic systems, just as e is a fundamental unit of charge. Associated with the quantization of $\vec{L}$ is a fundamental uncertainty in the *direction* of $\vec{L}$ and therefore of $\vec{\mu}$. In the following discussion, when we speak of the magnitude of a magnetic moment, a more precise statement would be "maximum component in a given direction." Thus, to say that a magnetic moment $\vec{\mu}$ is aligned with a magnetic field $\vec{B}$ really means that $\vec{\mu}$ has its maximum possible component in the direction of $\vec{B}$, such components are always quantized.

Equation (35) shows that associated with the fundamental unit of angular momentum is a corresponding fundamental unit of magnetic moment. If $L = h/2$ then

$$\mu = \frac{e}{2\pi}\left(\frac{h}{2\pi}\right) = \frac{he}{2\pi m} \tag{36}$$

This quantity is called the **Bohr magneton,** denoted by μ_B. Its numerical value is

$$\mu_B = 9.274 \times 10^{-24} A \cdot m^2 = 9.274 \times 10^{-24} J/T$$

You should verify that these two sets of units are consistent. The second set is useful when we compute the potential energy $U = -\vec{\mu} \cdot \vec{B}$ for a magnetic moment in a magnetic field.

Electrons also have an intrinsic angular momentum, called *spin,* that is not related to orbital motion but that can be pictured in a classical model as spinning on an axis. This angular momentum also has an associated magnetic moment, and its magnitude turns out to be almost exactly one Bohr magneton. (Effectshaving to do with quantization of the electromagnetic field cause the spin magnetic moment to be about 1.001 μ_B

Paramagnetism. In an atom, most of the various orbital and spin magnetic moments of the electrons add up to zero. However, in some cases the atom has a net magnetic moment that is of the order of μ_B.. When such a material is placed in a magnetic field, the field exerts a torque on each magnetic moment, as given by: $\vec{\tau} = \vec{\mu} \times \vec{B}$. These torques tend to align the magnetic moments with the field. In this position, the directions of the current loops are such as to *add* to the externally applied magnetic field.

We saw that the $\vec{B}$ field produced by a current loop is proportional to the loop's magnetic dipole moment. In the same way, the additional $\vec{B}$ field produced by microscopic electron current loops is proportional to the total magnetic moment $\vec{\mu}_{total}$ per unit volume V in the material. We call this vector quantity the **magnetization** of the material, denoted by

$$\vec{M} = \frac{\vec{\mu}_{total}}{V} \tag{37}$$

The additional magnetic field due to magnetization of the material turns out to be equal simply to $\mu_0 \vec{M}$, where μ_0 is the same constant that appears in the law of Biot and Savart and Ampere's law. When such a material completely surrounds a current-carrying conductor, the total magnetic field in the material is

$$\vec{B} = \vec{B}_0 + \mu_0 \vec{M} \tag{38}$$

where is $\vec{B}_0$ the field caused by the current in the conductor.

A material showing the behavior just described is said to be **paramagnetic.** The result is that the magnetic field at any point in such a material is greater by a dimensionless factor K_m, called the **relative permeability** of the material, than it would be if the material were replaced by vacuum. The value of K_m is different for different materials; for common paramagnetic solids and liquids at room temperature, K_m typically ranges from 1.00001 to 1.003.

All of the equations in this chapter that relate magnetic fields to their sources can be adapted to the situation in which the current-carrying conductor is embedded in a paramagnetic material. All that need be done is to replace μ_0 by $K_m\mu_0$ This product is usually denoted as and is called the **permeability** of the material:

$$\mu = K_m\mu_0 \tag{39}$$

The amount by which the relative permeability differs from unity is called the **magnetic susceptibility,** denoted by χ_m:

$$\chi_m = K_m - 1 \tag{40}$$

Both K_m and χ_m are dimensionless quantities. Table 1 lists values of magnetic susceptibility for several materials. For example, for aluminum, $\chi_m = 2.2 \times 10^{-5}$ and $K_m = 1.000022$. The first group of materials in the table are paramagnetic; we'll discuss the second group of materials, which are called *diamagnetic,* very shortly.

Table 1 – Magnetic susceptibilities of paramagnetic and diamagnetic materials al T=20 C

Material	$\chi_m = K_m - 1(\times 10^{-5})$
Paramagnetic	
Iron ammonium alum	66
Uranium	40
Platinum	26
Aluminium	2.2
Sodium	0.72
Oxygen gas	0.19
Diamagnetic	
Bismuth	-16.6
Mercury	-2.9
Silver	-2.6
Carbon (diamond)	-2.1
Lead	-1.8
Sodium chloride	-1.4
Copper	-1.0

The tendency of atomic magnetic moments to align themselves parallel to the magnetic field (where the potential energy is minimum) is opposed by random thermal motion, which tends to randomize their orientations. For this reason, paramagnetic susceptibility always decreases with increasing temperature. In many cases it is inversely proportional to the absolute temperature *T*, and the magnetization *M* can be expressed as

$$M = C\frac{B}{T} \tag{41}$$

This relationship is called *Curie's law,* after its discoverer, Pierre Curie (1859–1906). The quantity C is a constant, different for different materials, called the *Curie constant.*

A a body with atomic magnetic dipoles is attracted to the poles of a magnet. In most paramagnetic substances this attraction is very weak due to thermal randomization of the atomic magnetic moments. But at very low temperatures the thermal effects are reduced, the magnetization in creases in accordance with Curie's law, and the attractive forces are greater.

Diamagnetism. In some materials the total magnetic moment of all the atomic current loops is zero when no magnetic field is present. But even these materials have magnetic effects because an external field alters electron motions within the atoms, causing additional current loops and induced magnetic dipoles comparable to the induced *electric* dipoles we studied early. In this case the additional field caused by these current loops is always *opposite* in direction to that of the external field.

Such materials are said to be **diamagnetic.** They always have negative susceptibility, as shown in Table 1, and relative permeability slightly *less* than unity, typically of the order of 0.99990 to 0.99999 for solids and liquids. Diamagnetic susceptibilities are very nearly temperature independent.

Ferromagnetism. There is a third class of materials, called **ferromagnetic** materials, that includes iron, nickel, cobalt, and many alloys containing these elements. In these materials, strong interactions between atomic magnetic moments cause them to line up parallel to each other in regions called **magnetic domains,** even when no external field is present. Figure 29 shows an example of magnetic domain structure. Within each domain, nearly all of the atomic magnetic moments are parallel.

When there is no externally applied field, the domain magnetizations are randomly oriented. But when a field $\vec{B}_0$ (caused by external currents) is present, the domains tend to orient themselves parallel to the field. The domain boundaries also shift; the domains that are magnetized in the field direction grow, and those that are magnetized in other directions shrink. Because the total magnetic moment of a domain may be many thousands of Bohr magnetons, the torques that tend to align the domains with an external field are much stronger than occur with paramagnetic materials. The relative permeability K_m is *much* larger than unity, typically of the order of 1000 to 100,000. As a result, an object made of a ferromagnetic material such as iron is strongly magnetized by the field from a permanent magnet and is attracted to the magnet. A paramagnetic material such as aluminum is also attracted to a permanent magnet, but K_m for paramagnetic materials is so much smaller for such a material than for ferromagnetic materials that the attraction is very weak. Thus a magnet can pick up iron nails, but not aluminum cans.

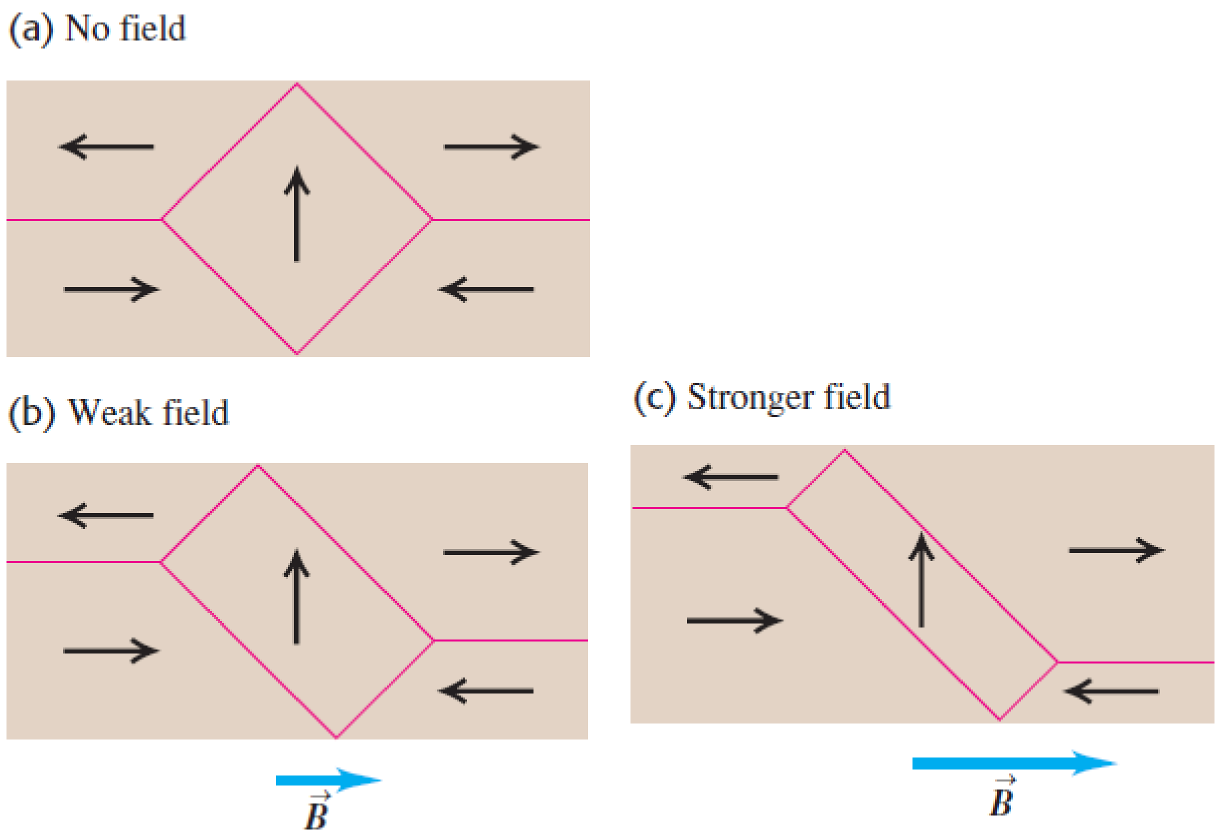


Figure 29 - In this drawing adapted from a magnified photo, the arrows show the directions of magnetization in the domains of a single crystal of nickel. Domains that are magnetized in the direction of an applied magnetic field grow larger.

As the external field is increased, a point is eventually reached at which nearly *all* the magnetic moments in the ferromagnetic material are aligned parallel to the external field. This condition is called *saturation magnetization;* after it is reached, further increase in the external field causes no increase in magnetization or in the additional field caused by the magnetization.

Figure 30 shows a “magnetization curve,” a graph of magnetization M as a function of external magnetic field $\vec{B}_0$ for soft iron. An alternative description of this behavior is that K_m is not constant but decreases as $\vec{B}_0$ increases. (Paramagnetic materials also show saturation at sufficiently strong fields. But the magnetic fields required are so large that departures from a linear relationship between M and $\vec{B}_0$ in these materials can be observed only at very low temperatures, 1 K or so.)

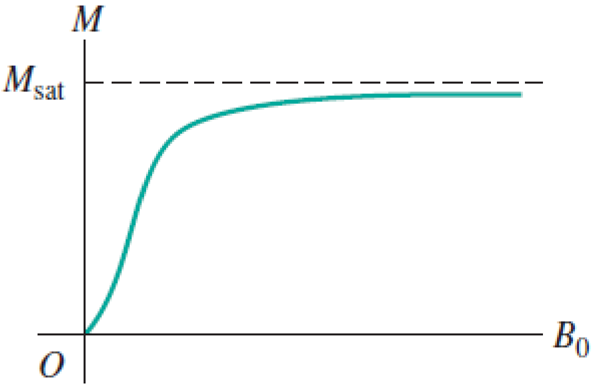


Figure 30 - A magnetization curve for a ferromagnetic material. The magnetization M approaches its saturation value M_{sat} as the magnetic field B_0 (caused by external currents) becomes large.

For many ferromagnetic materials the relationship of magnetization to external magnetic field is different when the external field is increasing from when it is decreasing. Figure 31a shows this relationship for such a material. When the material is magnetized to saturation and then the external field is reduced to zero, some magnetization remains. This behavior is characteristic of permanent magnets, which retain most of their saturation magnetization when the magnetizing field is removed. To reduce the magnetization to zero requires a magnetic field in the reverse direction.

This behavior is called **hysteresis,** and the curves in Fig. 31 are called *hysteresis loops.* Magnetizing and demagnetizing a material that has hysteresis involve the dissipation of energy, and the temperature of the material increases during such a process.

Ferromagnetic materials are widely used in electromagnets, transformer cores, and motors and generators, in which it is desirable to have as large a magnetic field as possible for a given current. Because hysteresis dissipates energy, materials that are used in these applications should usually have as narrow a hysteresis loop as possible. Soft iron is often used; it has high permeability without appreciable hysteresis. For permanent magnets a broad hysteresis loop is usually desirable, with large zero-field magnetization and large reverse field needed to demagnetize. Many kinds of steel and many alloys, such as Alnico, are commonly used for permanent magnets. The remaining magnetic field in such a material, after it has been magnetized $M = B/\mu_0$ to near saturation, is typically of the order of 1 T, corresponding to a remaining magnetization of about 800,000 A/m.

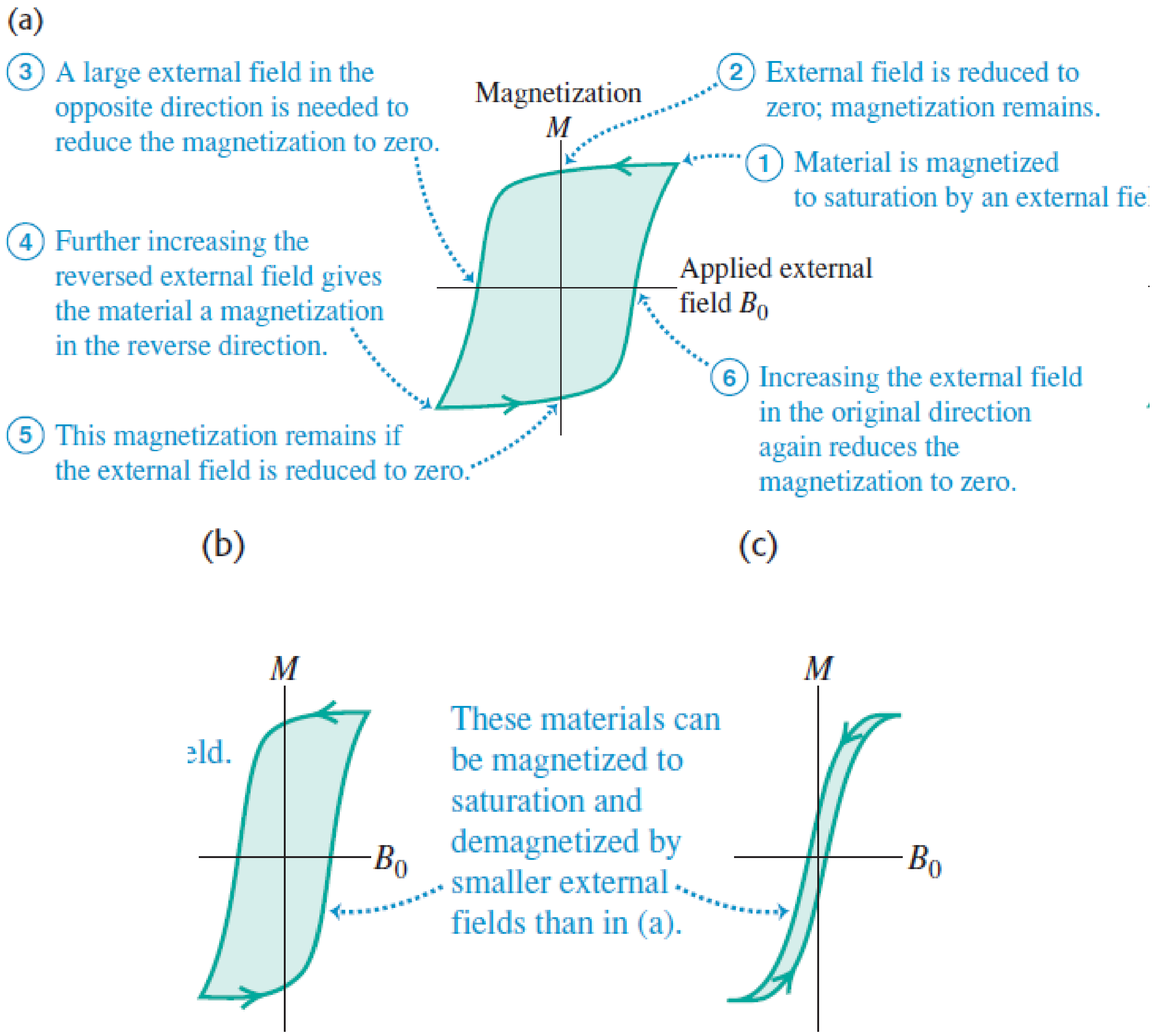


Figure 31 - Hysteresis loops. The materials of both (a) and (b) remain strongly magnetized when is reduced to zero. Since (a) is also hard to demagnetize, it would be good for permanent magnets. Since (b) magnetizes and demagnetizes more easily, it could be used as a computer memory material. The material of (c) would be useful for transformers and other alternating-current devices where zero hysteresis would be optimal.

Discussion questions

1. A topic of current interest in physics research is the search (thus far unsuccessful) for an isolated magnetic pole, or magnetic *monopole*. If such an entity were found, how could it be recognized? What would its properties be?
2. Streams of charged particles emitted from the sun during periods of solar activity create a disturbance in the earth's magnetic field. How does this happen?
3. The text discussed the magnetic field of an infinitely long, straight conductor carrying a current. Of course, there is no such thing as an infinitely long

anything. How do you decide whether a particular wire is long enough to be considered infinite?

4. Two parallel conductors carrying current in the same direction attract each other. If they are permitted to move toward each other, the forces of attraction do work. From where does the energy come?
5. Pairs of conductors carrying current into or out of the powersupply components of electronic equipment are sometimes twisted together to reduce magnetic-field effects. Why does this help?
6. Suppose you have three long, parallel wires arranged so that in cross section they are at the corners of an equilateral triangle. Is there any way to arrange the currents so that all three wires attracteach other? So that all three wires repel each other? Explain.
7. Two concentric, coplanar, circular loops of wire of different diameter carry currents in the same direction. Describe the nature of the force exerted on the inner loop by the outer loop and on the outer loop by the inner loop.
8. A current was sent through a helical coil spring. The spring contracted, as though it had been compressed. Why?
9. What are the relative advantages and disadvantages of Ampere’s law and the law of Biot and Savart for practical calculations of magnetic fields?
10. Magnetic field lines never have a beginning or an end. Use this to explain why it is reasonable for the field of a toroidal solenoid to be confined entirely to its interior, while a straight solenoid *must* have some field outside.
11. If the magnitude of the magnetic field a distance *R* from a very long, straight, current-carrying wire is *B*, at what distance from the wire will the field have magnitude *3B*?
12. Two very long, parallel wires carry equal currents in opposite directions. (a) Is there any place that their magnetic fields completely cancel? If so, where? If not, why not? (b) How would the answer to part (a) change if the currents were in the same direction?
13. A metal ring carries a current that causes a magnetic field at the center of the ring and a field *B* at point *P* a distance *x* from the center along the axis of the ring. If the radius of the ring is doubled, find the magnetic field at the center. Will the field at point *P* change by the same factor? Why?
14. Why should the permeability of a paramagnetic material be expected to decrease with increasing temperature?
15. If a magnet is suspended over a container of liquid air, it attracts droplets to its poles. The droplets contain only liquid oxygen; even though nitrogen is the primary constituent of air, it is not attracted to the magnet. Explain what this tells you about the magnetic susceptibilities of oxygen and nitrogen, and explain why a magnet in ordinary, room-temperature air doesn’t attract molecules of oxygen *gas* to its poles.
16. What features of atomic structure determine whether an element is diamagnetic or paramagnetic? Explain.

17. The magnetic susceptibility of paramagnetic materials is quite strongly temperature dependent, but that of diamagnetic materials is nearly independent of temperature. Why the difference?
18. A cylinder of iron is placed so that it is free to rotate around its axis. Initially the cylinder is at rest, and a magnetic field is applied to the cylinder so that it is magnetized in a direction parallel to its axis. If the direction of the *external* field is suddenly reversed, the direction of magnetization will also reverse and the cylinder will begin rotating around its axis. (This is called the *Einstein–de Haas effect*.) Explain why the cylinder begins to rotate.

1.3 Electromagnetic induction and inductance

1.3.1 Induction experiments

During the 1830s, several pioneering experiments with magnetically induced emf were carried out in England by Michael Faraday and in the United States by Joseph Henry (1797–1878), later the first director of the Smithsonian Institution. Figure 32 shows several examples. In Fig. 32a, a coil of wire is connected to a galvanometer. When the nearby magnet is stationary, the meter shows no current. This isn't surprising; there is no source of emf in the circuit. But when we *move* the magnet either toward or away from the coil, the meter shows current in the circuit, but *only* while the magnet is moving (Fig. 32b). If we keep the magnet stationary and move the coil, we again detect a current during the motion. We call this an **induced current,** and the corresponding emf required to cause this current is called an **induced emf.**

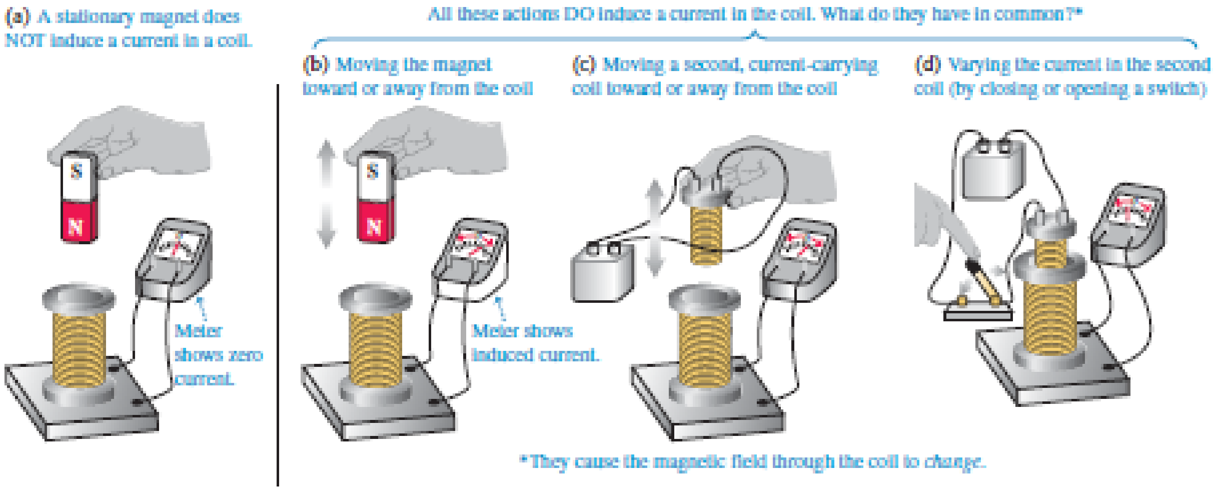

(a) A stationary magnet does NOT induce a current in a coil.
All these actions DO induce a current in the coil. What do they have in common?*
(b) Moving the magnet toward or away from the coil
(c) Moving a second, current-carrying coil toward or away from the coil
(d) Varying the current in the second coil (by closing or opening a switch)
S
N
Meter shows zero current.
S
N
Meter shows induced current.
*They cause the magnetic field through the coil to *change*.

Figure 32 – Demonstrating the phenomenon of induced current

In Fig. 32c we replace the magnet with a second coil connected to a battery. When the second coil is stationary, there is no current in the first coil. However, when we move the second coil toward or away from the first or move the first toward or away from the second, there is current in the first coil, but again *only* while one coil is moving relative to the other.

Finally, using the two-coil setup in Fig. 32d, we keep both coils stationary and vary the current in the second coil, either by opening and closing the switch or by changing the resistance of the second coil with the switch closed (perhaps by changing the second coil's temperature). We find that as we open or close the switch, there is a momentary current pulse in the first circuit. When we vary the resistance (and thus the current) in the second coil, there is an induced current in the first circuit, but only while the current in the second circuit is changing.

To explore further the common elements in these observations, let's consider a more detailed series of experiments (Fig. 33). We connect a coil of wire to a galvanometer and then place the coil between the poles of an electromagnet whose magnetic field we can vary. Here's what we observe:

1. When there is no current in the electromagnet, so that $\vec{B} = 0$ the galvanometer shows no current.
2. When the electromagnet is turned on, there is a momentary current through the meter as $\vec{B}$ increases.
3. When $\vec{B}$ levels off at a steady value, the current drops to zero, no matter how large $\vec{B}$ is.
4. With the coil in a horizontal plane, we squeeze it so as to decrease the cross-sectional area of the coil. The meter detects current only *during* the deformation, not before or after. When we increase the area to return the coil to its original shape, there is current in the opposite direction, but only while the area of the coil is changing.
5. If we rotate the coil a few degrees about a horizontal axis, the meter detects current during the rotation, in the same direction as when we decreased the area. When we rotate the coil back, there is a current in the opposite direction during this rotation.
6. If we jerk the coil out of the magnetic field, there is a current during the motion, in the same direction as when we decreased the area.
7. If we decrease the number of turns in the coil by unwinding one or more turns, there is a current during the unwinding, in the same direction as when we decreased the area. If we wind more turns onto the coil, there is a current in the opposite direction during the winding.
8. When the magnet is turned off, there is a momentary current in the direction opposite to the current when it was turned on.
9. The faster we carry out any of these changes, the greater the current.
10. If all these experiments are repeated with a coil that has the same shape but different material and different resistance, the current in each case is inversely proportional to the total circuit resistance. This shows that the induced emfs that are causing the current do not depend on the material of the coil but only on its shape and the magnetic field.

The common element in all these experiments is changing *magnetic flux* Φ_B through the coil connected to the galvanometer. In each case the flux changes either because the magnetic field changes with time or because the coil is moving through a nonuniform magnetic field. Faraday's law of induction, the subject of the next section, states that in all of these situations the induced emf is proportional to the *rate of change* of magnetic flux Φ_B through the coil. The *direction* of the induced emf depends on whether the flux is increasing or decreasing. If the flux is constant, there is no induced emf.

Induced emfs are not mere laboratory curiosities but have a tremendous number of practical applications. If you are reading these words indoors, you are making use of induced emfs right now! At the power plant that supplies your neighborhood, an electric generator produces an emf by varying the magnetic flux through coils of wire. (In the next section we'll see in detail how this is done.) This emf supplies the voltage between the terminals of the wall sockets in your home, and this voltage supplies the power to your reading lamp. Indeed, any appliance that you plug into a wall socket makes use of induced emfs.

Magnetically induced emfs are the result of *nonelectrostatic* forces. We have to distinguish carefully between the electrostatic electric fields produced by charges (according to Coulomb's law) and the nonelectrostatic electric fields produced by changing magnetic fields. We'll return to this distinction later in this chapter and the next.

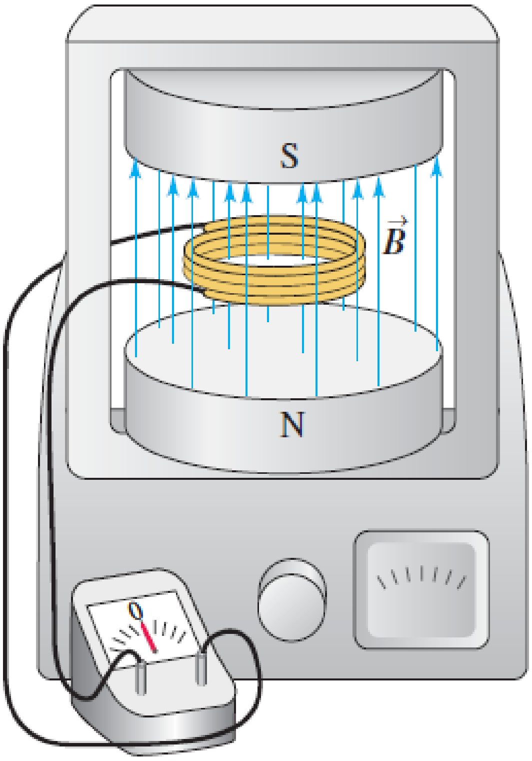


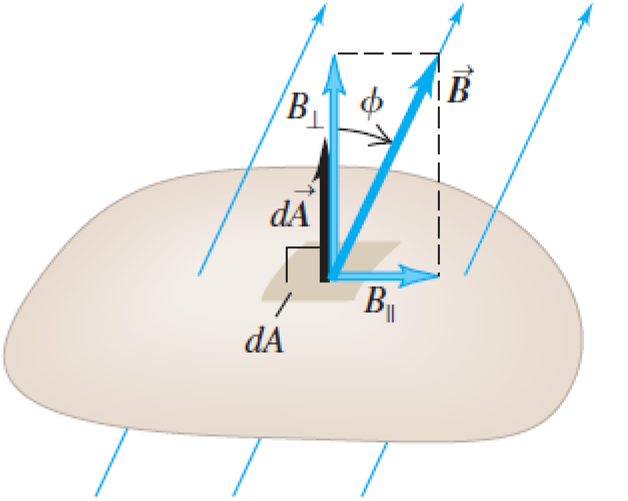
Figure 33 – A coil in a magnetic field. When the $\vec{B}$ field is constant and the shape, location, and orientation of the coil do not change, no current is induced in the coil. Acurrent is induced when any of these factors change

1.3.2 Faraday's law and Lenz's law

The common element in all induction effects is changing magnetic flux through a circuit. Before stating the simple physical law that summarizes all of the kinds of experiments described early, let’s first review the concept of magnetic flux Φ_B. For an infinitesimalarea element $d\vec{S}$ in a magnetic field $\vec{B}$ (Fig. 34), the magnetic flux $d\Phi_B$ through the area is

$$d\Phi_B = \vec{B} \cdot d\vec{S} = B_\perp dA = BdA\cos\varphi \qquad (42)$$

where $B_\perp$ is the component of $\vec{B}$ perpendicular to the surface of the area element and φ is the angle between $\vec{B}$ and $d\vec{S}$.



Magnetic flux through element of area $d\vec{A}$:
$d\Phi_B = \vec{B} \cdot d\vec{A} = B_\perp dA = B\,dA\cos\phi$

Figure 34 - Calculating the magnetic flux through an area element

$$\Phi_B = \int \vec{B} \cdot d\vec{S} = \int BdA\cos\varphi \qquad (43)$$

If $\vec{B}$ is uniform over a flat area $\vec{S}$, then

$$\Phi_B = \vec{B} \cdot d\vec{S} = BdA\cos\varphi \qquad (44)$$

Figure 35 reviews the rules for using Eq. (44).

Faraday’s law of induction states:

The induced emf in a closed loop equals the negative of the time rate of change of magnetic flux through the loop.

In symbols, Faraday’s law is

$$\mathcal{E} = -\frac{d\Phi_B}{dt} \qquad (45)$$

To understand the negative sign, we have to introduce a sign convention for the induced emf $\mathcal{E}$. But first let's look at a simple example of this law in action.

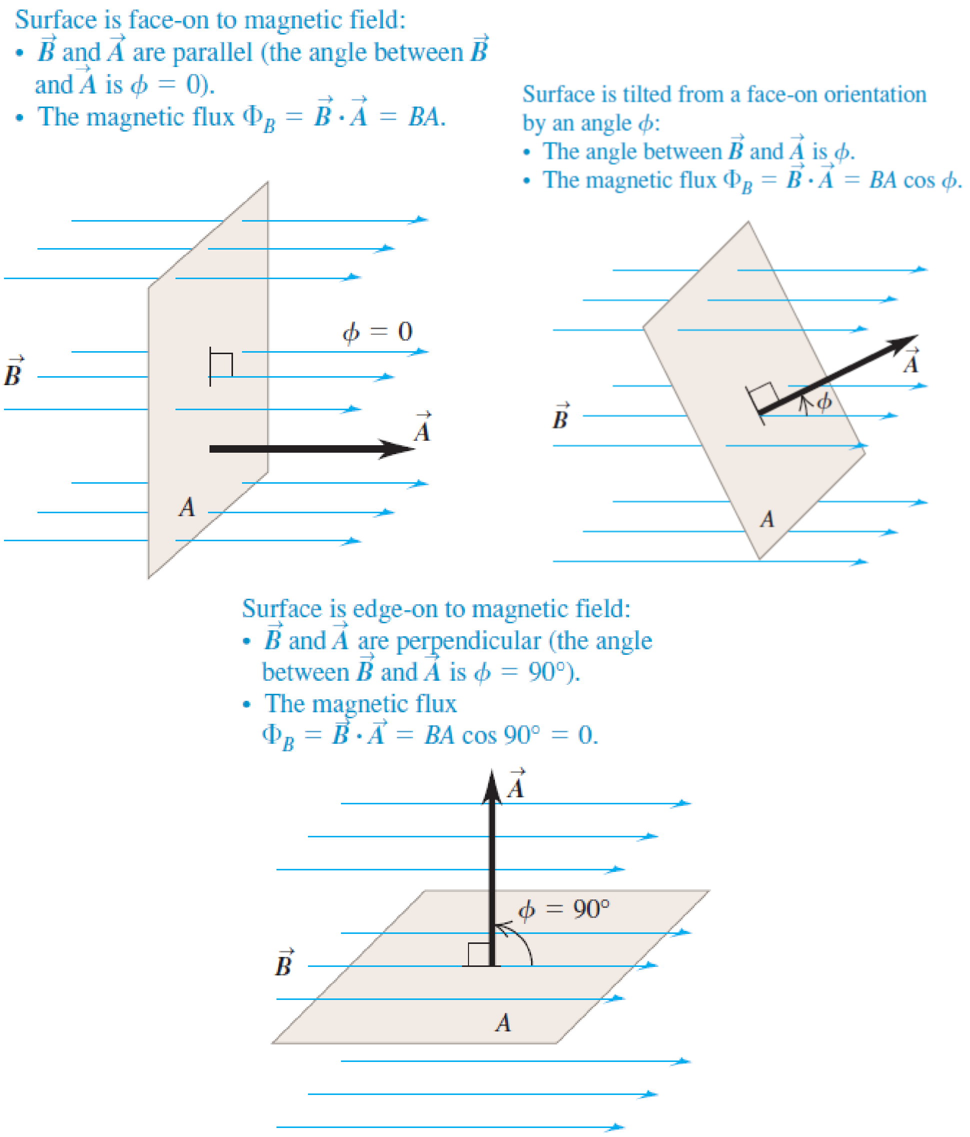

Surface is face-on to magnetic field:
• $\vec{B}$ and $\vec{A}$ are parallel (the angle between $\vec{B}$ and $\vec{A}$ is $\phi = 0$).
• The magnetic flux $\Phi_B = \vec{B}\cdot\vec{A} = BA$.
$\phi = 0$
$\vec{B}$
$\vec{A}$
A
Surface is tilted from a face-on orientation by an angle ϕ:
• The angle between $\vec{B}$ and $\vec{A}$ is ϕ.
• The magnetic flux $\Phi_B = \vec{B}\cdot\vec{A} = BA\cos\phi$.
$\vec{A}$
ϕ
$\vec{B}$
A
Surface is edge-on to magnetic field:
• $\vec{B}$ and $\vec{A}$ are perpendicular (the angle between $\vec{B}$ and $\vec{A}$ is $\phi = 90°$).
• The magnetic flux $\Phi_B = \vec{B}\cdot\vec{A} = BA\cos 90° = 0$.
$\vec{A}$
$\phi = 90°$
$\vec{B}$
A

Figure 35 - Calculating the flux of a uniform magnetic field through a flat area.

The total magnetic flux Φ_B through a finite area is the integral of this expression over the area:

Lenz's law is a convenient alternative method for determining the direction of an induced current or emf. Lenz's law is not an independent principle; it can be derived from Faraday's law. It always gives the same results as the sign rules we introduced in connection with Faraday's law, but it is often easier to use. Lenz's law also helps us gain intuitive understanding of various induction effects and of the role of energy conservation. H. F. E. Lenz (1804–1865) was a Russian scientist who duplicated independently many of the discoveries of Faraday and Henry. **Lenz's law** states:

The direction of any magnetic induction effect is such as to oppose the cause of the effect.

The “cause” may be changing flux through a stationary circuit due to a varying magnetic field, changing flux due to motion of the conductors that make up the circuit, or any combination. If the flux in a stationary circuit changes, the induced current sets up a magnetic field of its own. Within the area bounded by the circuit, this field is *opposite* to the original field if the original field is *increasing* but is in the *same* direction as the original field if the latter is *decreasing*. That is, the induced current opposes the *change in flux* through the circuit (*not* the flux itself).

If the flux change is due to motion of the conductors the direction of the induced current in the moving conductor is such that the direction of the magnetic-field force on the conductor is opposite in direction to its motion. Thus the motion of the conductor, which caused the induced current, is opposed.

Lenz’s law is also directly related to energy conservation.

1.3.3 Induced electric field. Eddy currents

When a conductor moves in a magnetic field, we can understand the induced emf on the basis of magnetic forces on charges in the conductor. But an induced emf also occurs when there is a changing flux through a stationary conductor. What is it that pushes the charges around the circuit in this type of situation?

As an example, let’s consider the situation shown in Fig. 36. A long, thin solenoid with cross-sectional area A and turns per unit length is encircled at its center by a circular conducting loop. The galvanometer G measures the current in the loop. A current I in the winding of the solenoid sets up a magnetic field $\vec{B}$ along the solenoid axis, as shown, with magnitude B as calculated early: $B = \mu_0 n I$, where n is the number of turns per unit length.

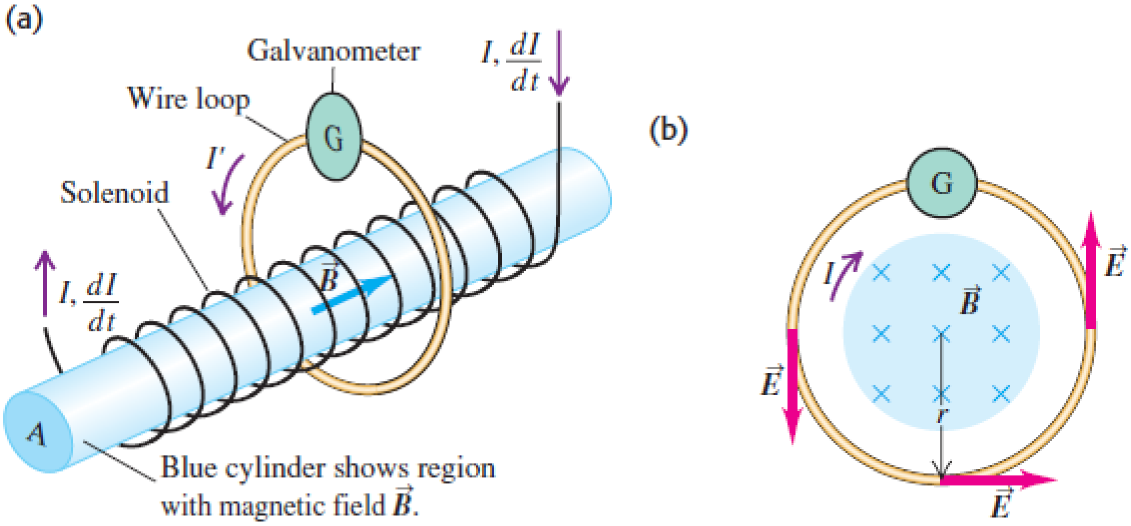


Figure 36 - (a) The windings of a long solenoid carry a current I that is increasing at a rate dI/dt. The magnetic flux in the solenoid is increasing at a rate $\frac{d\Phi_B}{dt}$ and this changing flux passes through a wire loop. An emf $\mathcal{E} = -\frac{d\Phi_B}{dt}$ is induced in the loop,

inducing a current I' that is measured by the galvanometer G. (b) Cross-sectional view.

If we neglect the small field outside the solenoid and take the area vector $\vec{S}$ to point in the same direction as $\vec{B}$, then the magnetic flux Φ_B through the loop is

$$\Phi_B = BS = \mu_0 nIS \tag{46}$$

When the solenoid current I changes with time, the magnetic flux Φ_B also changes, and according to Faraday's law the induced emf in the loop is given by

$$\mathcal{E} = -\frac{d\Phi_B}{dt} = -\mu_0 nS\frac{dI}{dt} \tag{47}$$

If the total resistance of the loop is R, the induced current in the loop, which we may call I', is $I' = \mathcal{E}/R$

But what *force* makes the charges move around the wire loop? It can't be a magnetic force because the loop isn't even *in* a magnetic field. We are forced to conclude that there has to be an **induced electric field** in the conductor *caused by the changing magnetic flux*. This may be a little jarring; we are accustomed to thinking about electric field as being caused by electric charges, and now we are saying that a changing magnetic field somehow acts as a source of electric field. Furthermore, it's a strange sort of electric field. When a charge q goes once around the loop, the total work done on it by the electric field must be equal to q times the emf $\mathcal{E}$. That is, the electric field in the loop *is not conservative*, because the line integral of $\vec{E}$ around a closed path is not zero. Indeed, this line integral, representing the work done by the induced $\vec{E}$ field per unit charge, is equal to the induced emf $\mathcal{E}$:

$$\oint \vec{E}\cdot d\vec{l} = \mathcal{E} \tag{48}$$

From Faraday's law the emf $\mathcal{E}$ is also the negative of the rate of change of magnetic flux through the loop. Thus for this case we can restate Faraday's law as

$$\oint \vec{E}\cdot d\vec{l} = -\frac{d\Phi_B}{dt} \tag{49}$$

Note that Faraday's law is *always* true in the form the form $\mathcal{E} = -\frac{d\Phi_B}{dt}$ given in Eq. (49) is valid *only* if the path around which we integrate is *stationary*.

As an example of a situation to which Eq. (49) can be applied, consider the stationary circular loop in Fig. 36b, which we take to have radius r. Because of cylindrical symmetry, the electric field $\vec{E}$ has the same magnitude at every point on the circle and is tangent to it at each point. (Symmetry would also permit the field to be *radial,* but then Gauss's law would require the presence of a net charge inside the circle, and there is none.) The line integral in Eq. (49) becomes simply the magnitude E times the circumference $2\pi r$ of the loop, $\oint \vec{E} \cdot d\vec{l} = 2\pi r$ and Eq. (49) gives

$$\mathcal{E} = \frac{1}{2\pi r}\left|\frac{d\Phi_B}{dt}\right| \tag{50}$$

The directions of $\vec{E}$ at points on the loop are shown in Fig. 36b. We know that $\vec{E}$ has to have the direction shown when $\vec{B}$ in the solenoid is increasing, because $\oint \vec{E} \cdot d\vec{l}$ has to be negative when $\frac{d\Phi_B}{dt}$ is positive. The same approach can be used to find the induced electric field *inside* the solenoid when the solenoid $\vec{B}$ field is changing; we leave the details to you.

Now let's summarize what we've learned. Faraday's law, Eq. (45), is valid for two rather different situations. In one, an emf is induced by magnetic forces on charges when a conductor moves through a magnetic field. In the other, a time-varying magnetic field induces an electric field in a stationary conductor and hence induces an emf; in fact, the $\vec{E}$ field is induced even when no conductor is present. This $\vec{E}$ field differs from an electro*static* field in an important way. It is *nonconservative;* the line integral $\oint \vec{E} \cdot d\vec{l}$ around a closed path is not zero, and when a charge moves around a closed path, the field does a nonzero amount of work on it. It follows that for such a field the concept of *potential* has no meaning. We call such a field a **nonelectrostatic field.** In contrast, an *electrostatic* fieldis *always* conservative and always has an associated potential function. Despite this difference, the fundamental effect of *any* electric field is to exert a force $\vec{F} = q\vec{E}$ on a charge q. This relationship is valid whether $\vec{E}$ is a conservative field produced by a charge distribution or a nonconservative field caused by changing magnetic flux.

So a changing magnetic field acts as a source of electric field of a sort that we *cannot* produce with any static charge distribution. This may seem strange, but it's the way nature behaves. What's more that a changing *electric* field acts as a source of *magnetic* field. We'll explore this symmetry between the two fields in greater detail in our study of electromagnetic waves.

If any doubt remains in your mind about the reality of magnetically induced electric fields, consider a few of the many practical applications. Pickups in electric guitars use currents induced in stationary pickup coils by the vibration of nearby ferromagnetic strings. Alternators in most cars use rotating magnets to induce currents in stationary coils. Whether we realize it or not, magnetically induced electric fields play an important role in everyday life.

In the examples of induction effects that we have studied, the induced currents have been confined to well-defined paths in conductors and other components forming a circuit. However, many pieces of electrical equipment contain masses of metal moving in magnetic fields or located in changing magnetic fields. In situations like these we can have induced currents that circulate throughout the volume of a material. Because their flow patterns resemble swirling eddies in a river, we call these **eddy currents.**

As an example, consider a metallic disk rotating in a magnetic field perpendicular to the plane of the disk but confined to a limited portion of the disk's area, as shown in Fig. 37a. Sector *Ob* is moving across the field and has an emf induced in it. Sectors *Oa* and *Oc* are not in the field, but they provide return conducting paths for charges displaced along *Ob* to return from *b* to *O*. The result is a circulation of eddy currents in the disk, somewhat as sketched in Fig. 37b.

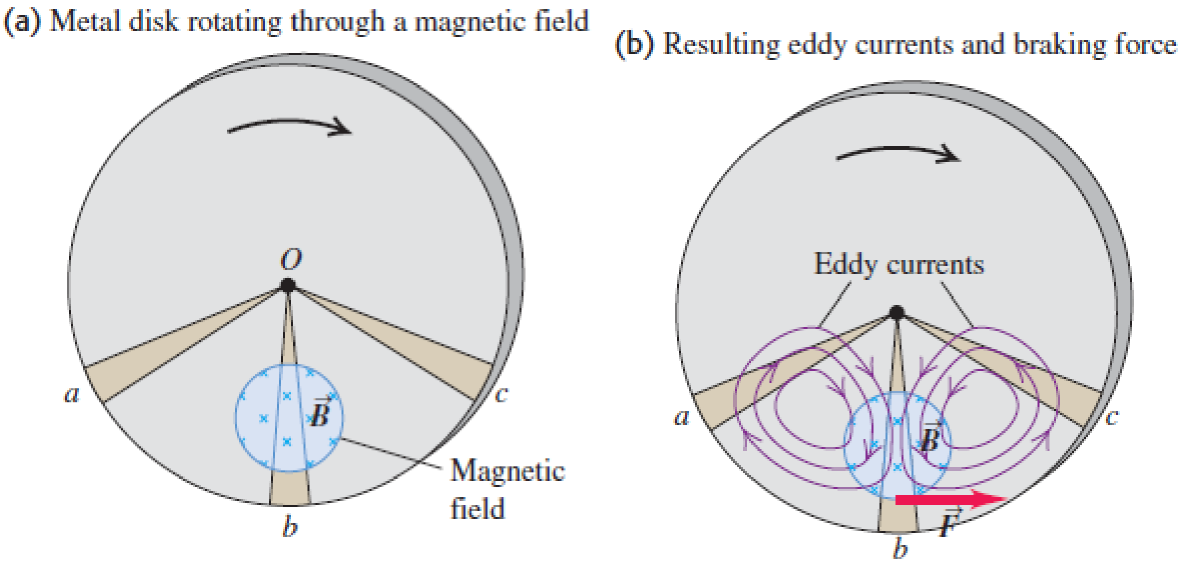

(a) Metal disk rotating through a magnetic field
O
a
c
b
$\vec{B}$
Magnetic field
(b) Resulting eddy currents and braking force
Eddy currents
a
c
b
$\vec{B}$
$\vec{F}$

Figure 37 - Eddy currents induced in a rotating metal disk

We can use Lenz's law to decide on the direction of the induced current in the neighborhood of sector *Ob*. This current must experience a magnetic force $\vec{F} = I\vec{L} \times \vec{B}$ that *opposes* the rotation of the disk, and so this force must be to the right in Fig. 37b. Since $\vec{B}$ is directed into the plane of the disk, the current and hence $\vec{L}$ have downward components. The return currents lie outside the field, so they do not experience magnetic forces. The interaction between the eddy currents and the field causes a braking action on the disk. Such effects can be used to stop the rotation of a circular saw quickly when the power is turned off. Some sensitive balances use this effect to damp out vibrations. Eddy current braking is used on some electrically powered rapid-transit vehicles. Electromagnets mounted in the cars induce eddy currents in the rails; the resulting magnetic fields cause braking forces on the electromagnets and thus on the cars.

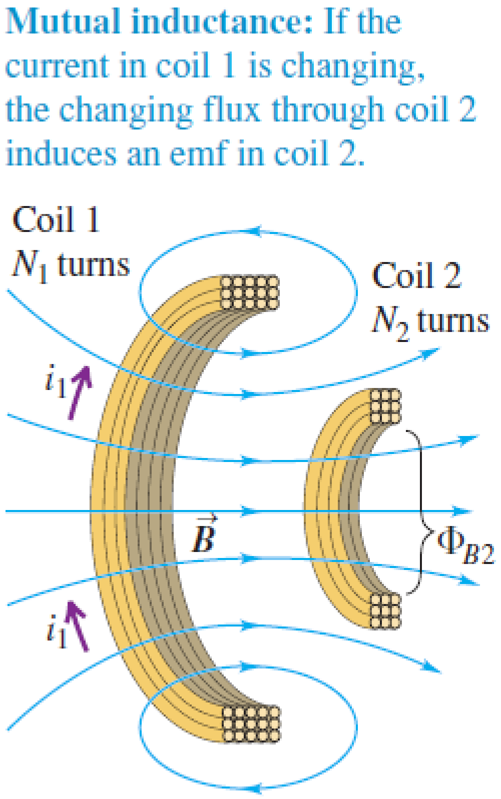
Eddy currents have many other practical uses. The shiny metal disk in the electric power company's meter outside your house rotates as a result of eddy currents. These currents are induced in the disk by magnetic fields caused by sinusoidally varying currents in a coil. In induction furnaces, eddy currents are used to heat materials in completely sealed containers for processes in which it is essential to avoid the slightest contamination of the materials. The metal detectors used at airport security checkpoints operate by detecting eddy currents induced in metallic objects. Similar devices are used to find buried treasure such as bottlecaps and lost pennies.

Eddy currents also have undesirable effects. In an alternating-current transformer, coils wrapped around an iron core carry a sinusoidally varying current. The resulting eddy currents in the core waste energy through heating and themselves set up an unwanted opposing emf in the coils. To minimize these effects, the core is designed so that the paths for eddy currents are as narrow as possible. We'll describe how this is done when we discuss transformers.

1.3.4 Magnetic-field energy

Take a length of copper wire and wrap it around a pencil to form a coil. If you put this coil in a circuit, does it behave any differently than a straight piece of wire? Remarkably, the answer is yes. In an ordinary gasoline-powered car, a coil of this kind makes it possible for the 12-volt car battery to provide thousands of volts to the spark plugs, which in turn makes it possible for the plugs to fire and make the engine run. Other coils of this type are used to keep fluorescent light fixtures shining. Larger coils placed under city streets are used to control the operation of traffic signals. All of these applications, and many others, involve the *induction.*

A changing current in a coil induces an emf in an adjacent coil. The coupling between the coils is described by their *mutual inductance.* A changing current in a coil also induces an emf in that same coil. Such a coil is called an *inductor,* and the relationship of current to emf is described by the *inductance* (also called *selfinductance*) of the coil. If a coil is initially carrying a current, energy is released when the current decreases; this principle is used in automotive ignition systems. We'll find that this released energy was stored in the magnetic field caused by the current that was initially in the coil, and we'll look at some of the practical applications of magnetic-field energy.



Mutual inductance: If the current in coil 1 is changing, the changing flux through coil 2 induces an emf in coil 2.

Figure 38 - A current i_1 in coil 1 gives rise to a magnetic flux through coil 2

Early we considered the magnetic interaction between two wires carrying *steady* currents; the current in one wire causes a magnetic field, which exerts a force on the current in the second wire. But an additional interaction arises between two circuits when there is a *changing* current in one of the circuits. Consider two neighboring coils of wire, as in Fig. 38. A current flowing in coil 1 produces a magnetic field $\vec{B}$ and hence a magnetic flux through coil 2. If the current in coil 1 changes, the flux through coil 2 changes as well; according to Faraday's law, this induces an emf in coil 2. In this way, a change in the current in one circuit can induce a current in a second circuit.

Let's analyze the situation shown in Fig. 38 in more detail. We will use lowercase letters to represent quantities that vary with time; for example, a timevarying current is often with a subscript to identify the circuit. In Fig. 38 a current in coil 1 sets up a magnetic field (as indicated by the blue lines), and some of these field lines pass through coil 2. We denote the magnetic flux through *each* turn of coil 2, caused by the current i_1 in coil 1, as Φ_{B2} (If the flux is different through different turns of the coil, then denotes the *average* flux.) The magnetic field is proportional to i_1 so Φ_{B2} is also proportional to i_1. When i_1 changes, Φ_{B2} changes; this changing flux induces an emf $\mathcal{E}_2$ in coil 2, given by

$$\mathcal{E}_2 = -N_2 \frac{d\Phi_{B2}}{dt} \tag{51}$$

We could represent the proportionality of Φ_{B2} and i_1 in the form $\Phi_{B2} = (\text{constant})i_1$, but instead it is more convenient to include the number of turns N_2 in the relationship. Introducing a proportionality constant M_{21} called the **mutual inductance** of the two coils, we write

$$N_2\Phi_{B2} = M_{21}i_1 \tag{52}$$

Where Φ_{B2} is the flux through a *single* turn of coil 2. From this,

$$N_2\frac{d\Phi_{B2}}{dt} = M_{21}\frac{di_1}{dt} \tag{53}$$

and we can rewrite Eq. (51) as

$$\mathcal{E}_2 = -M_{21}\frac{di_1}{dt} \tag{54}$$

That is, a change in the current in coil 1 induces an emf in coil 2 that is directly proportional to the rate of change of i_1.

We may also write the definition of mutual inductance, Eq. (52), as

$$M_{21} = \frac{N_2\Phi_{B2}}{i_1} \tag{55}$$

If the coils are in vacuum, the flux Φ_{B2} through each turn of coil 2 is directly proportional to the current i_1. Then the mutual inductance M_{21} is a constant that depends only on the geometry of the two coils (the size, shape, number of turns, and orientation of each coil and the separation between the coils). If a magnetic material is present, M_{21} also depends on the magnetic properties of the material. If the material has nonlinear magnetic properties - that is, if the relative permeability K_m is not constant and magnetization is not proportional to magnetic field – then Φ_{B2} is no longer directly proportional to i_1. In that case the mutual inductance also depends on the value of In this discussion we will assume that any magnetic material present has constant K_m so that flux *is* directly proportional to current and M_{21} depends on geometry only.

We can repeat our discussion for the opposite case in which a changing current in coil 2 causes a changing flux Φ_{B1} and an emf $\mathcal{E}_1$ in coil 1. We might expect that the corresponding constant M_{12} would be different from M_{21} because in general the two coils are not identical and the flux through them is not the same. It turns out, however, that M_{12} is *always* equal to M_{21} even when the two coils are not symmetric. We call this common value simply the mutual inductance, denoted by the symbol *M* without subscripts; it characterizes completely the induced-emf interaction of two coils. Then we can write

$$\mathcal{E}_2 = -M\frac{di_1}{dt} \text{ and } \mathcal{E}_1 = -M\frac{di_2}{dt} \tag{56}$$

where the mutual inductance M is

$$M = \frac{N_2\Phi_{B2}}{i_1} = \frac{N_1\Phi_{B1}}{i_2} \tag{57}$$

The negative signs in Eq. (56) are a reflection of Lenz's law. The first equation says that a change in current in coil 1 causes a change in flux through coil 2, inducing an emf in coil 2 that opposes the flux change; in the second equation the roles of the two coils are interchanged.

The SI unit of mutual inductance is called the **henry** (1 H), in honor of the American physicist Joseph Henry (1797–1878), one of the discoverers of electromagnetic induction. From Eq. (57), one henry is equal to *one weber per ampere*. Other equivalent units, obtained by using Eq. (56), are *one volt-second per ampere, one ohm-second*, and *one joule per ampere squared:*

$$1\,H = 1\frac{Wb}{A} = 1\,V\cdot\frac{s}{A} = 1\,\Omega\cdot s = 1\,J/A^2$$

Establishing a current in an inductor requires an input of energy, and an inductor carrying a current has energy stored in it. Let's see how this comes about. In Fig. 39, an increasing current i in the inductor causes an emf between its terminals and a corresponding potential difference V_{ab} between the terminals of the source, with point a at higher potential than point b. Thus the source must be adding energy to the inductor, and the instantaneous power (rate of transfer of energy into the inductor) is $P = V_{ab}i$.

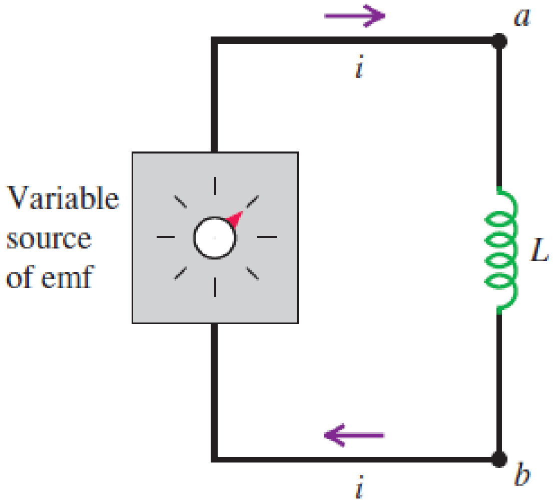


Figure 39 – A circuit containing a source of emf and an inductor. The source is variable, so the current i and its rate of change

We can calculate the total energy input U needed to establish a final current I in an inductor with inductance L if the initial current is zero. We assume that the inductor has zero resistance, so no energy is dissipated within the inductor. Let the current at some instant be i and let its rate of change be di/dt the current is increasing, so $di/dt > 0$. The voltage between the terminals a and b of the inductor at this instant is $V_{ab} = L\, di/dt$ and the rate P at which energy is being delivered to the inductor (equal to the instantaneous power supplied by the external source) is

$$P = V_{ab}i = Li\, di/dt \tag{58}$$

The energy dU supplied to the inductor during an infinitesimal time interval dt Is $dU = Pdt$ so

$$dU = Li\, di \tag{59}$$

The total energy U supplied while the current increases from zero to a final value I is

$$U = L\int_0^I idi = \frac{1}{2}LI^2 \tag{60}$$

After the current has reached its final steady value $I, di/dt = 0$ and no more energy is input to the inductor. When there is no current, the stored energy U is zero; when the current is I, the energy is $\frac{1}{2}LI^2$.

When the current decreases from to zero, the inductor acts as a source that supplies a total amount of energy $\frac{1}{2}LI^2$ to the external circuit. If we interrupt the circuit suddenly by opening a switch or yanking a plug from a wall socket, the current decreases very rapidly, the induced emf is very large, and the energy may be dissipated in an arc across the switch contacts. This large emf is the electrical analog of the large force exerted by a car running into a brick wall and stopping very suddenly.

Discussion questions

1. In an electric trolley or bus system, the vehicle’s motor draws current from an overhead wire by means of a long arm with an attachment at the end that slides along the overhead wire. Abrilliant electric spark is often seen when the attachment crosses a junction in the wires where contact is momentarily lost. Explain this phenomenon.
2. The tightly wound toroidal solenoid is one of the few configurations for which it is easy to calculate self-inductance. What features of the toroidal solenoid give it this simplicity?
3. Two identical, closely wound, circular coils, each having self-inductance are placed next to each other, so that they are coaxial and almost touching. If they

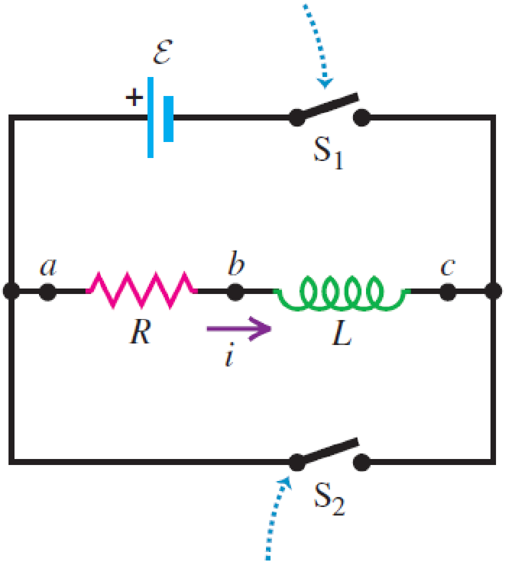
are connected in series, what is the self-inductance of the combination? What if they are connected in parallel? Can they be connected so that the total inductance is zero? Explain.

4. Two closely wound circular coils have the same number of turns, but one has twice the radius of the other. How are the selfinductances of the two coils related? Explain your reasoning.
5. You are to make a resistor by winding a wire around a cylindrical form. To make the inductance as small as possible, it is proposed that you wind half the wire in one direction and the other half in the opposite direction. Would this achieve the desired result? Why or why not?
6. In the *R*-*L* circuit, when switch is closed, the potential changes suddenly and discontinuously,but the current does not. Explain why the voltage can change suddenly but the current can't.
7. In the *R*-*L* circuit is the current in the resistor always the same as the current in the inductor? How do you know?
8. Suppose there is a steady current in an inductor. If you attempt to reduce the current to zero instantaneously by quickly opening a switch, an arc can appear at the switch contacts. Why? Is it physically possible to stop the current instantaneously? Explain.
9. In an *L*-*R*-*C* series circuit, what criteria could be used to decide whether the system is overdamped or underdamped? For example, could we compare the maximum energy stored during one cycle to the energy dissipated during one cycle? Explain.

1.4 Electromagnetic oscillations and waves

1.4.1 The R-L circuit

Let's look at some examples of the circuit behavior of an inductor. One thing is clear already; an inductor in a circuit makes it difficult for rapid changes in current to occur, thanks to the effects of self-induced emf. Equation $\mathcal{E} = -L\frac{di}{dt}$ shows that the greater the rate of change of current di/dt, the greater the self-induced emf and the greater the potential difference between the inductor terminals. This equation, together with Kirchhoff's rules, gives us the principles we need to analyze circuits containing inductors.



Closing switch S_1 connects the R-L combination in series with a source of emf $\mathcal{E}$.

Closing switch S_2 while opening switch S_1 disconnnects the combination from the source.

Figure 40 – An R-L circuit

We can learn several basic things about inductor behavior by analyzing the circuit of Fig. 40. A circuit that includes both a resistor and an inductor, and possibly a source of emf, is called an ***R-L* circuit.** The inductor helps to prevent rapid changes in current, which can be useful if a steady current is required but the external source has a fluctuating emf. The resistor *R* may be a separate circuit element, or it may be the resistance of the inductor windings; every real-life inductor has some resistance unless it is made of superconducting wire. By closing switch we can connect the *R*-*L* combination to a source with constant emf $\mathcal{E}$. (We assume that the source has zero internal resistance, so the terminal voltage equals the emf.)

Suppose both switches are open to begin with, and then at some initial time $t = 0$ we close switch S_1. The current cannot change suddenly from zero to some final value, since di/dt and the induced emf in the inductor would both be infinite. Instead, the current begins to grow at a rate that depends only on the value of *L* in the circuit.

Let *i* be the current at some time *t* after switch S_1 is closed, and let di/dt be its rate of change at that time. The potential difference v_{ab} across the resistor at that time is

$$v_{ab} = iR \tag{61}$$

and the potential difference v_{bc} across the inductor is

$$v_{ab} = L\frac{di}{dt} \tag{62}$$

Note that if the current is in the direction shown in Fig. 40 and is increasing, then both and are positive; a is at a higher potential than b, which in turn is at a higher potential than c. We apply Kirchhoff's loop rule, starting at the negative terminal and proceeding counterclockwise around the loop:

$$\mathcal{E} - iR - L\frac{di}{dt} = 0 \tag{63}$$

Solving this for di/dt we find that the rate of increase of current is

$$\frac{di}{dt} = \frac{\mathcal{E} - iR}{L} = \frac{\mathcal{E}}{R} - \frac{R}{L}i \tag{64}$$

At the instant that switch S_1 is first closed, $i = 0$ and the potential drop across R is zero. The initial rate of change of current is

$$\left(\frac{di}{dt}\right)_{initial} = \frac{\mathcal{E}}{L} \tag{65}$$

As we would expect, the greater the inductance L, the more slowly the current increases.

As the current increases, the term $(R/L)i$ in Eq. (64) also increases, and the *rate* of increase of current given by Eq. (64) becomes smaller and smaller. This means that the current is approaching a final, steady-state value I. When the current reaches this value, its rate of increase is zero. Then Eq. (64) becomes

$$\left(\frac{di}{dt}\right)_{final} = 0 = \frac{\mathcal{E}}{L} - \frac{R}{L}I \tag{66}$$

and

$$I = \frac{\mathcal{E}}{R} \tag{67}$$

The *final* current I does not depend on the inductance L it is the same as it would be if the resistance R alone were connected to the source with emf $\mathcal{E}$.

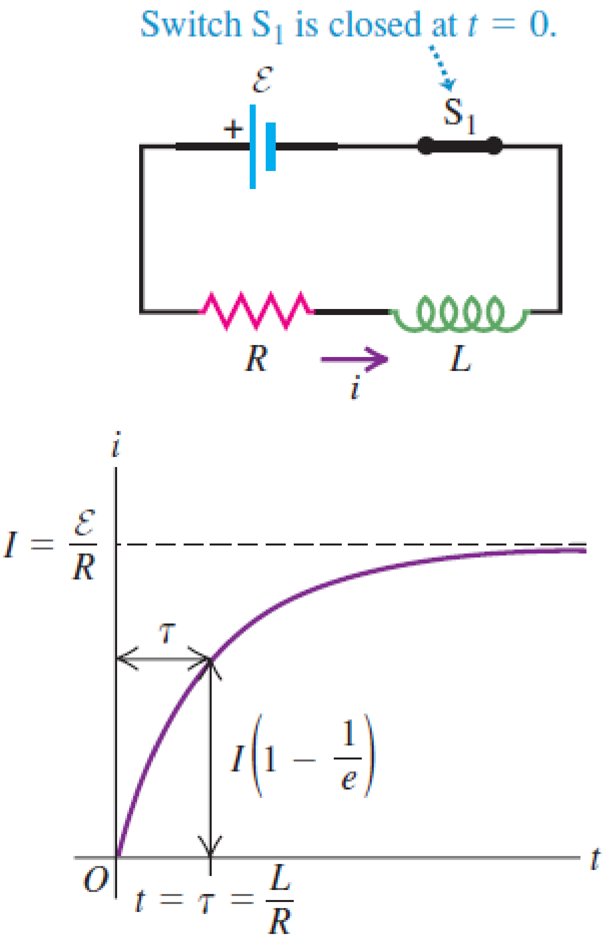

Switch S_1 is closed at $t = 0$.
$\mathcal{E}$
+
S_1
R
i
L
i
$I = \frac{\mathcal{E}}{R}$
τ
$I\left(1 - \frac{1}{e}\right)$
O
t
$t = \tau = \frac{L}{R}$

Figure 41 - Graph of i versus t for growth of current in an R-L circuit with an emf in series. The final current is $I = \mathcal{E}/R$; after one time constant τ, the current is $1 - 1/e$ of this value

Figure 41 shows the behavior of the current as a function of time. To derive the equation for this curve (that is, an expression for current as a function of time), we proceed just as we did for the charging capacitor. First we rearrange Eq. (64) to the form

$$\frac{di}{i - (\mathcal{E}/R)} = -\frac{R}{L}dt \tag{68}$$

This separates the variables, with i on the left side and t on the right. Then we integrate both sides, renaming the integration variables i' andt t' so that we can use i and t as the upper limits. (The lower limit for each integral is zero, corresponding to zero current at the initial time $t = 0$). We get

$$\int_0^{i'} \frac{di'}{i' - \left(\frac{\mathcal{E}}{R}\right)} = -\int_0^{t} \frac{R}{L}dt' \tag{69}$$

(70)

$$\ln\left(\frac{i-\left(\frac{\mathcal{E}}{R}\right)}{-\frac{\mathcal{E}}{R}}\right)=-\frac{R}{L}t$$

Now we take exponentials of both sides and solve for i. We leave the details for you to work out; the final result is

$$i=\frac{\mathcal{E}}{R}\left(1-e^{-\left(\frac{R}{L}\right)t}\right) \tag{71}$$

This is the equation of the curve in Fig. 41. Taking the derivative of Eq. (71), we find

$$\frac{di}{dt}=\frac{R}{L}e^{-\left(\frac{R}{L}\right)t} \tag{72}$$

At time $t=0, i=0$ and $\frac{di}{dt}=\mathcal{E}/L$. As $t\rightarrow\infty, i\rightarrow\mathcal{E}/R$ and $di/dt\rightarrow 0$, as we predicted.

As Fig. 41 shows, the instantaneous current i first rises rapidly, then increases more slowly and approaches the final value $I=\mathcal{E}/r$ asymptotically. At a time equal to L/R, the current has risen to $(1-1/e)$, or about 63%, of its final value. The quantity L/R is therefore a measure of how quickly the current builds toward its final value; this quantity is called the **time constant** for the circuit, denoted by τ:

$$\tau=\frac{L}{R} \tag{73}$$

In a time equal to 2τ the current reaches 86% of its final value; i n 5τ, 99.3%; and in 10τ, 99.995%.

The graphs of *i* versus *t* have the same general shape for all values of L. For a given value of R, the time constant τ is greater for greater values of L. When L is small, the current rises rapidly to its final value; when L is large, it rises more slowly. For example, if $R=100\ \Omega$ and $L=10\ H$

$$\tau=\frac{L}{R}=\frac{10\ H}{100\ \Omega}=0.01\ s \tag{74}$$

and the current increases to about 63% of its final value in 0.10 s. But if $=0.01\ H$, $\tau=10^{-4}s=0.1\ ms$, and the rise is much more rapid.

Energy considerations offer us additional insight into the behavior of an *R*-*L* circuit. The instantaneous rate at which the source delivers energy to the circuit is $P=\mathcal{E}i$. The instantaneous rate at which energy is dissipated in the resistor is i^2R, and the rate at which energy is stored in the inductor is $iv_{bc} = Li\,di/dt$ [or, equivalently, $(d/dt)\left(\frac{1}{2}Li^2\right) = Li\,di/dt$]. When we multiply Eq. (63) by i and rearrange, we find

$$\mathcal{E}i = i^2R + Li\frac{di}{dt} \tag{75}$$

Of the power $\mathcal{E}i$ supplied by the source, part (i^2R) is dissipated in the resistor and part $\left(Li\frac{di}{dt}\right)$ goes to store energy in the inductor. This discussion is completely analogous to our power analysis for a charging capacitor.

1.4.2 The L-C circuit

A circuit containing an inductor and a capacitor shows an entirely new mode of behavior, characterized by *oscillating* current and charge. This is in sharp contrast to the *exponential* approach to a steady-state situation that we have seen with both *R-C* and *R-L* circuits. In the ***L-C* circuit** in Fig. 42a we charge the capacitor to a potential difference and initial charge $Q = CV_m$ on its left-hand plate and then close the switch. What happens?

The capacitor begins to discharge through the inductor. Because of the induced emf in the inductor, the current cannot change instantaneously; it starts at zero and eventually builds up to a maximum value I_m. During this buildup the capacitor is discharging. At each instant the capacitor potential equals the induced emf, so as the capacitor discharges, the *rate of change* of current decreases. When the capacitor potential becomes zero, the induced emf is also zero, and the current has leveled off at its maximum value I_m. Figure 42b shows this situation; the capacitor has completely discharged. The potential difference between its terminals (and those of the inductor) has decreased to zero, and the current has reached its maximum value I_m.

During the discharge of the capacitor, the increasing current in the inductor has established a magnetic field in the space around it, and the energy that was initially stored in the capacitor's electric field is now stored in the inductor's magnetic field.

Although the capacitor is completely discharged in Fig. 42b, the current persists (it cannot change instantaneously), and the capacitor begins to charge with polarity opposite to that in the initial state. As the current decreases, the magnetic field also decreases, inducing an emf in the inductor in the *same* direction as the current; this slows down the decrease of the current. Eventually, the current and the magnetic field reach zero, and the capacitor has been charged in the sense *opposite* to its initial polarity (Fig. 42c), with potential difference $-V_m$ and charge $-Q$ on its left-hand plate.

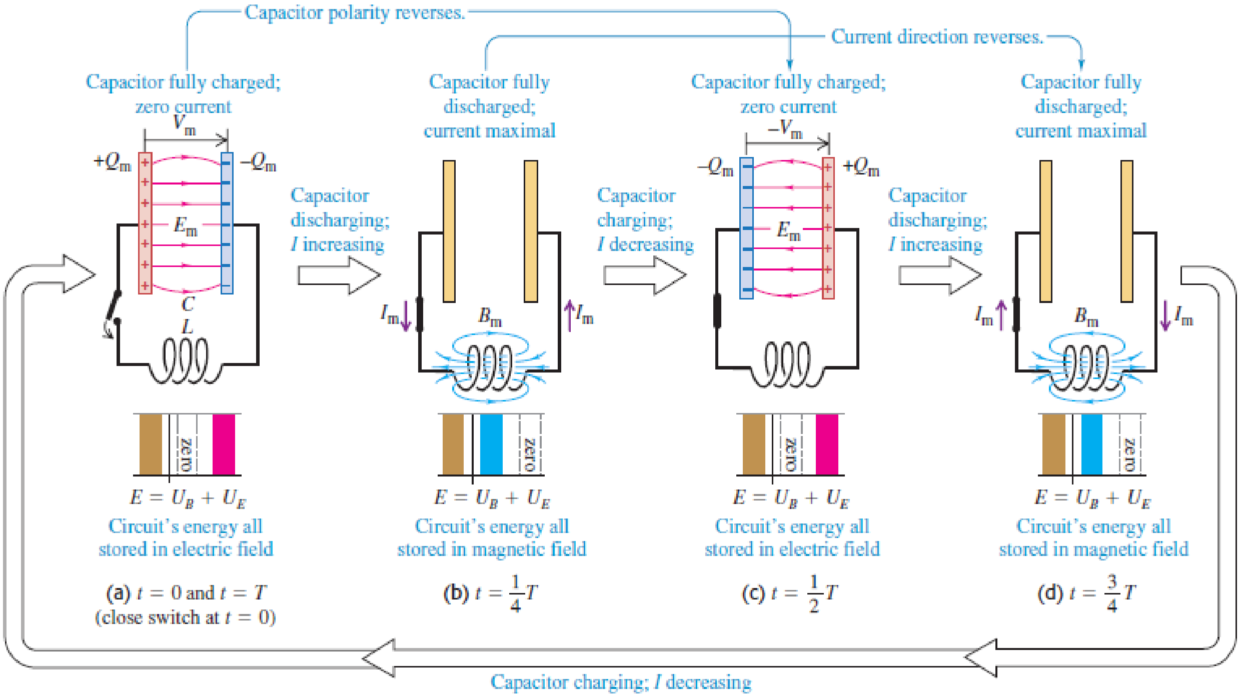

Capacitor polarity reverses.
Current direction reverses.
Capacitor fully charged; zero current
V_m
$+Q_m$
$-Q_m$
E_m
C
L
Capacitor discharging; I increasing
Capacitor fully discharged; current maximal
I_m
B_m
I_m
Capacitor charging; I decreasing
Capacitor fully charged; zero current
$-V_m$
$-Q_m$
$+Q_m$
E_m
Capacitor discharging; I increasing
Capacitor fully discharged; current maximal
I_m
B_m
I_m
zero
zero
zero
zero
$E = U_B + U_E$
Circuit's energy all stored in electric field
(a) $t = 0$ and $t = T$
(close switch at $t = 0$)
$E = U_B + U_E$
Circuit's energy all stored in magnetic field
(b) $t = \frac{1}{4}T$
$E = U_B + U_E$
Circuit's energy all stored in electric field
(c) $t = \frac{1}{2}T$
$E = U_B + U_E$
Circuit's energy all stored in magnetic field
(d) $t = \frac{3}{4}T$
Capacitor charging; I decreasing

Figure 42 - In an oscillating *L-C* circuit, the charge on the capacitor and the current through the inductor both vary sinusoidally with time. Energy is transferred between magnetic energy in the inductor U_B and electric energy in the capacitor U_E. As in simple harmonic motion, the total energy E remains constant

The process now repeats in the reverse direction; a little later, the capacitor has again discharged, and there is a current in the inductor in the opposite direction (Fig. 42d). Still later, the capacitor charge returns to its original value (Fig. 42a), and the whole process repeats. If there are no energy losses, the charges on the capacitor continue to oscillate back and forth indefinitely. This process is called an **electrical oscillation.**

From an energy standpoint the oscillations of an electrical circuit transfer energy from the capacitor's electric field to the inductor's magnetic field and back. The *total* energy associated with the circuit is constant. This is analogous to the transfer of energy in an oscillating mechanical system from potential energy to kinetic energy and back, with constant total energy. As we will see, this analogy goes much further.

To study the flow of charge in detail, we proceed just as we did for the *R-L* circuit. Figure 43 shows our definitions of *q* and *i*.

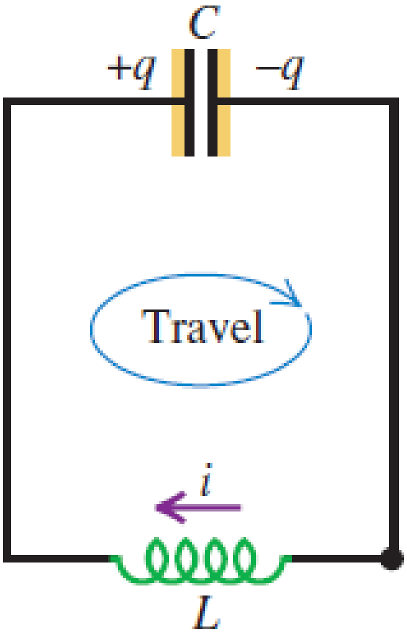


Figure 43 - Applying Kirchhoff's loop rule to the *L*-*C* circuit. The direction of travel around the loop in the loop equation is shown. Just after the circuit is completed and the capacitor first begins to discharge, as in Fig. 42a, the current is negative (opposite to the direction shown)

We apply Kirchhoff's loop rule to the circuit in Fig. 43. Starting at the lower-right corner of the circuit and adding voltages as we go clockwise around the loop, we obtain

$$-L\frac{di}{dt}-\frac{q}{C}=0 \tag{76}$$

Since $i=dq/dt$, it follows that $\frac{di}{dt}=\frac{d^2q}{dt^2}$. We substitute this expression into the above equation and divide by $-L$ to obtain

$$\frac{d^2q}{dt^2}+\frac{1}{LC}=0 \tag{77}$$

Equation (52) has exactly the same form as the equation we derived for simple harmonic motion. That equation is $\frac{d^2x}{dt^2}=-\left(\frac{k}{m}\right)x$, or

$$\frac{d^2x}{dt^2}+\frac{k}{m}x=0x \tag{78}$$

In the *L*-*C* circuit the capacitor charge q plays the role of the displacement x, and the current $i=dq/dt$ is analogous to the particle's velocity $v_x=dx/dt$. The inductance

L is analogous to the mass m, and the reciprocal of the capacitance, $1/C$, is analogous to the force constant k.

Pursuing this analogy, we recall that the angular frequency $\omega = 2\pi f$ of the harmonic oscillator is equal to $\sqrt{k/m}$ and the position is given as a function of time by

$$x = A\cos(\omega t + \varphi) \tag{79}$$

where the amplitude A and the phase angle φ depend on the initial conditions. In the analogous electrical situation the capacitor charge q is given by

$$q = Q\cos(\omega t + \varphi) \tag{80}$$

and the angular frequency ω of oscillation is given by

$$\omega = \sqrt{\frac{1}{LC}} \tag{81}$$

You should verify that Eq. (80) satisfies the loop equation, Eq. (77), when ω has the value given by Eq. (81). In doing this, you will find that the instantaneous current $i = dq/dt$ is given by

$$i = -\omega Q\sin(\omega t + \varphi) \tag{82}$$

Thus the charge and current in an *L-C* circuit oscillate sinusoidally with time, with an angular frequency determined by the values of L and C. The ordinary frequency f the number of cycles per second, is equal to $\omega/2\pi$ as always. The constants Q and φ in Eqs. (63) and (82) are determined by the initial conditions. If at time $t = 0$ the left-hand capacitor plate in Fig. 43 has its maximum charge Q and the current i is zero, then $\varphi = 0$. If $q = 0$ at time $t = 0$, then $\varphi = \pm\frac{\pi}{2}$ rad.

1.4.3 The L-R-C circuit

In our discussion of the *L-C* circuit we assumed that there was no *resistance* in the circuit. This is an idealization, of course; every real inductor has resistance in its windings, and there may also be resistance in the connecting wires. Because of resistance, the electromagnetic energy in the circuit is dissipated and converted to other forms, such as internal energy of the circuit materials. Resistance in an electric circuit is analogous to friction in a mechanical system.

Suppose an inductor with inductance and a resistor of resistance are connected in series across the terminals of a charged capacitor, forming an ***L-R-C*** **series circuit.** As before, the capacitor starts to discharge as soon as the circuit is completed. But because of i^2R losses in the resistor, the magnetic-field energy acquired by the inductor when the capacitor is completely discharged is *less* than the original electric-field energy of the capacitor. In the same way, the energy of the capacitor when the magnetic field has decreased to zero is still smaller, and so on.

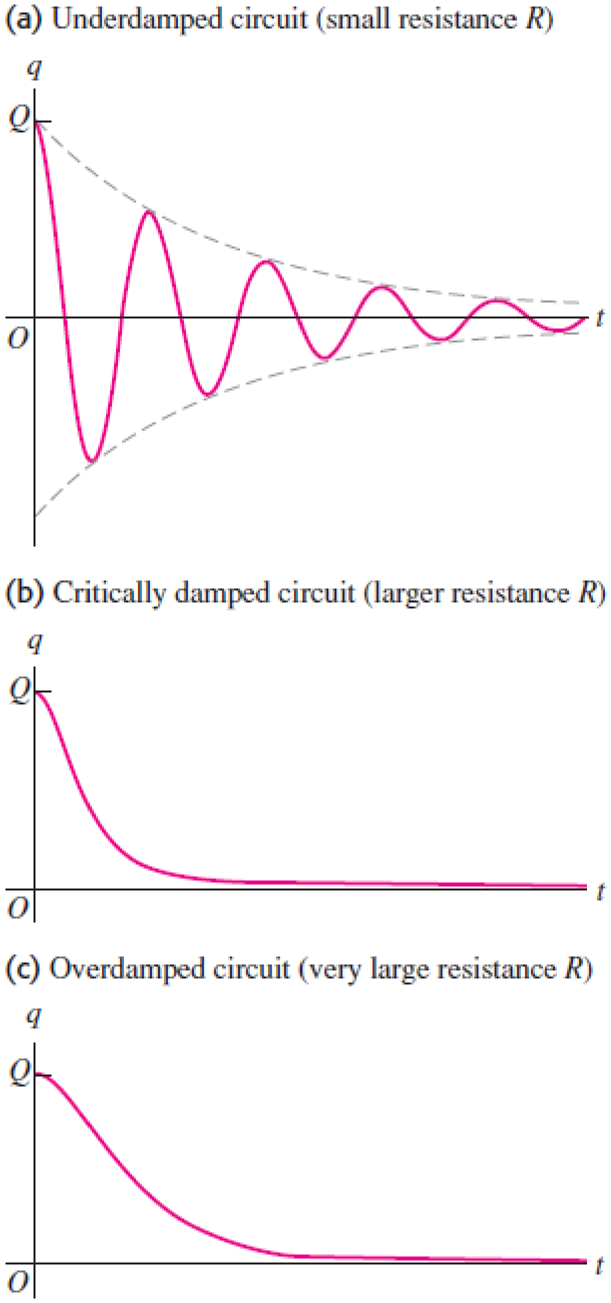


Figure 44 - Graphs of capacitor charge as a function of time in an *L*-*R*-*C* series circuit with initial charge *Q*

If the resistance is relatively small, the circuit still oscillates, but with **damped harmonic motion** (Fig. 44a), and we say that the circuit is **underdamped.** If we increase *R* the oscillations die out more rapidly. When *R* reaches a certain value, the circuit no longer oscillates; it is **critically damped** (Fig. 44b). For still larger values of *R* the circuit is **overdamped** (Fig. 44c), and the capacitor charge approaches zero even more slowly. We used these same terms to describe the behavior of the analogous mechanical system, the damped harmonic oscillator.

To analyze *L*-*R*-*C* series circuit behavior in detail, we consider the circuit shown in Fig. 45. It is like the *L*-*C* circuit of Fig. 43 except for the added resistor we also show the source that charges the capacitor initially. The labeling of the positive senses of *q* and *i* are the same as for the *L*-*C* circuit.

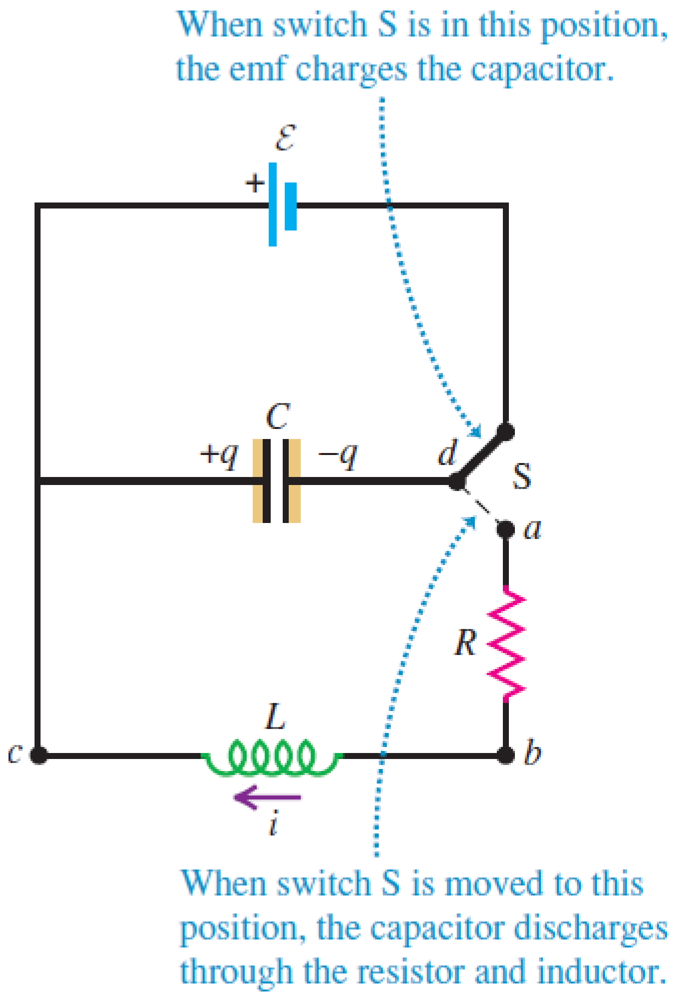

When switch S is in this position, the emf charges the capacitor.
$\mathcal{E}$
+
C
$+q$
$-q$
d
S
a
R
L
c
b
i
When switch S is moved to this position, the capacitor discharges through the resistor and inductor.

Figure 45 – L-R-C series circuit

First we close the switch in the upward position, connecting the capacitor to a source of emf $\mathcal{E}$ for a long enough time to ensure that the capacitor acquires its final charge $Q = C\mathcal{E}$ and any initial oscillations have died out. Then at time $t = 0$ we flip the switch to the downward position, removing the source from the circuit and placing the capacitor in series with the resistor and inductor. Note that the initial current is negative, opposite to the direction of *i* shown in Fig. 45.

To find how *q* and *i* vary with time, we apply Kirchhoff's loop rule. Starting at point *a* and going around the loop in the direction *abcda*, we obtain the equation

$$-iR - L\frac{di}{dt} - \frac{q}{C} = 0 \tag{83}$$

Replacing i with dq/dt and rearranging, we get

$$\frac{d^2q}{dt^2}+\frac{R}{L}\frac{dq}{dt}+\frac{1}{LC}q=0 \tag{84}$$

Note that when $R=0$ this reduces to Eq. (77) for an L-C circuit.

There are general methods for obtaining solutions of Eq. (84). The form of the solution is different for the underdamped (small R) and overdamped (large R) cases. When R^2 is less than $\frac{4L}{C}$, the solution has the form

$$q=Ae^{-\left(\frac{R}{2L}\right)t}\cos\left(\sqrt{\frac{1}{LC}-\frac{R^2}{4L^2}}t+\varphi\right) \tag{85}$$

where A and φ are constants. We invite you to take the first and second derivatives of this function and show by direct substitution that it does satisfy Eq. (84).

This solution corresponds to the *underdamped* behavior shown in Fig. 44a; the function represents a sinusoidal oscillation with an exponentially decaying amplitude. (Note that the exponential factor $e^{-\left(\frac{R}{2L}\right)t}$ is *not* the same as the factor $e^{-\left(\frac{R}{L}\right)t}$). When $R=0$ Eq. (85) reduces to Eq. (80) for the oscillations in an L-C circuit. If R is not zero, the angular frequency ω' of the oscillation is *less* than because of the term containing The angular frequency of the damped oscillations is given by

$$\omega'=\sqrt{\frac{1}{LC}-\frac{R^2}{4L^2}} \tag{86}$$

When $R=0$, this reduces to Eq. (81), $\omega=\sqrt{\frac{1}{LC}}$. As R increases, ω' becomes smaller and smaller. When $R^2=\frac{4L}{C}$, the quantity under the radical becomes zero; the system no longer oscillates, and the case of *critical damping* (Fig. 44b) has been reached. For still larger values of R the system behaves as in Fig. 44c. In this case the circuit is *overdamped*, and q is given as a function of time by the sum of two decreasing exponential functions.

In the *underdamped* case the phase constant φ in the cosine function of Eq. (85) provides for the possibility of both an initial charge and an initial current at time $t=0$ analogous to an underdamped harmonic oscillator given both an initial displacement and an initial velocity.

We emphasize once more that the behavior of the *L*-*R*-*C* series circuit is completely analogous to that of the damped harmonic oscillator studied early. We invite you to verify, for example, that if you start with Eq. $-kx-bv_x=m\frac{d^2x}{dt^2}$ and substitute q for x, L for m, 1/C for k, and R for the damping constant b, the result is Eq. (84). Similarly, the cross-over point between underdamping and overdamping occurs at $b^2 = 4km$ for the mechanical system and at $R^2 = 4L/C$ for the electrical one.

The practical applications of the L-R-C series circuit emerge when we include a sinusoidally varying source of emf in the circuit. This is analogous to the *forced oscillation,* and there are analogous *resonance* effects. Such a circuit is called an *alternating-current (ac) circuit;* the analysis of ac circuits is the principal topic of the next chapter.

1.4.4 Maxwell’s equations and electromagnetic waves

In the last several chapters we studied various aspects of electric and magnetic fields. We learned that when the fields don’t vary with time, such as an electric field produced by charges at rest or the magnetic field of a steady current, we can analyze the electric and magnetic fields independently without considering interactions between them. But when the fields vary with time, they are no longer independent. Faraday’s law tells us that a time-varying magnetic field acts as a source of electric field, as shown by induced emfs in inductors and transformers. Ampere’s law, including the displacement current discovered by Maxwell, shows that a time-varying electric field acts as a source of magnetic field. This mutual interaction between the two fields is summarized in Maxwell’s equations.

Thus, when *either* an electric or a magnetic field is changing with time, a field of the other kind is induced in adjacent regions of space. We are led (as Maxwell was) to consider the possibility of an electromagnetic disturbance, consisting of time-varying electric and magnetic fields, that can propagate through space from one region to another, even when there is no matter in the intervening region. Such a disturbance, if it exists, will have the properties of a *wave,* and an appropriate term is **electromagnetic wave.**

Such waves do exist; radio and television transmission, light, x rays, and many other kinds of radiation are examples of electromagnetic waves. Our goal in this chapter is to see how such waves are explained by the principles of electromagnetism that we have studied thus far and to examine the properties of these waves.

As often happens in the development of science, the theoretical understanding of electromagnetic waves evolved along a considerably more devious path than the one just outlined. In the early days of electromagnetic theory (the early 19th century), two different units of electric charge were used: one for electrostatics and the other for magnetic phenomena involving currents. In the system of units used at that time, these two units of charge had different physical dimensions. Their *ratio* had units of velocity, and measurements showed that the ratio had a numerical value that was precisely equal to the speed of light, $3 \cdot 10^8\ m/s$. At the time, physicists regarded this as an extraordinary coincidence and had no idea how to explain it.

In searching to understand this result, Maxwell proved in 1865 that an electromagnetic disturbance should propagate in free space with a speed equal to that of light and hence that light waves were likely to be electromagnetic in nature. At the same time, he discovered that the basic principles of electromagnetism can be expressed in terms of the four equations that we now call **Maxwell's equations.** These four equations are (1) Gauss's law for electric fields; (2) Gauss's law for magnetic fields, showing the absence of magnetic monopoles; (3) Ampere's law, including displacement current; and (4) Faraday's law:

$$\oint \vec{E}\cdot d\vec{S} = \frac{Q_{encl}}{\varepsilon_0} \tag{87}$$

$$\oint \vec{B}\cdot d\vec{S} = 0 \tag{88}$$

$$\oint \vec{B}\cdot d\vec{l} = \mu_0\left(i_C + \varepsilon_0 \frac{d\Phi_E}{dt}\right)_{encl} \tag{89}$$

$$\oint \vec{E}\cdot d\vec{l} = -\frac{d\Phi_B}{dt} \tag{90}$$

These equations apply to electric and magnetic fields *in vacuum.* If a material is present, the permittivity ε_0 and permeability μ_0 of free space are replaced by the permittivity and permeability of the material. If the values of and are different at different points in the regions of integration, then and have to be transferred to the left sides of Eqs. (87) and (89), respectively, and placed inside the integrals. The in Eq. (89) also has to be included in the integral that gives $\frac{d\Phi_E}{dt}$.

According to Maxwell's equations, a point charge at rest produces a static $\vec{E}$ field but no $\vec{B}$ field; a point charge moving with a constant velocity produces both $\vec{E}$ and $\vec{B}$ fields. Maxwell's equations can also be used to show that in order for a point charge to produce electromagnetic waves, the charge must *accelerate.* In fact, it's a general result of Maxwell's equations that *every* accelerated charge radiates electromagnetic energy (Fig. 46).

One way in which a point charge can be made to emit electromagnetic waves is by making it oscillate in simple harmonic motion, so that it has an acceleration at almost every instant (the exception is when the charge is passing through its equilibrium position). Figure 47 shows some of the electric field lines produced by such an oscillating point charge. Field lines are *not* material objects, but you may nonetheless find it helpful to think of them as behaving somewhat like strings that extend from the point charge off to infinity. Oscillating the charge up and down makes waves that propagate outward from the charge along these "strings." Note that the charge does not emit waves equally in all directions; the waves are strongest at 90° to the axis of motion of the charge, while there are *no* waves along this axis. This is just what the "string" picture would lead you to conclude. There is also a *magnetic* disturbance that spreads outward from the charge; this is not shown in Fig. 47. Because the electric and magnetic disturbances spread or radiate away from the source, the name **electromagnetic radiation** is used interchangeably with the phrase "electromagnetic waves."

Figure 46 - Power lines carry a strong alternating current, which means that a substantial amount of charge is accelerating back and forth and generating electromagnetic waves. These waves can produce a buzzing sound from your car radio when you drive near the lines

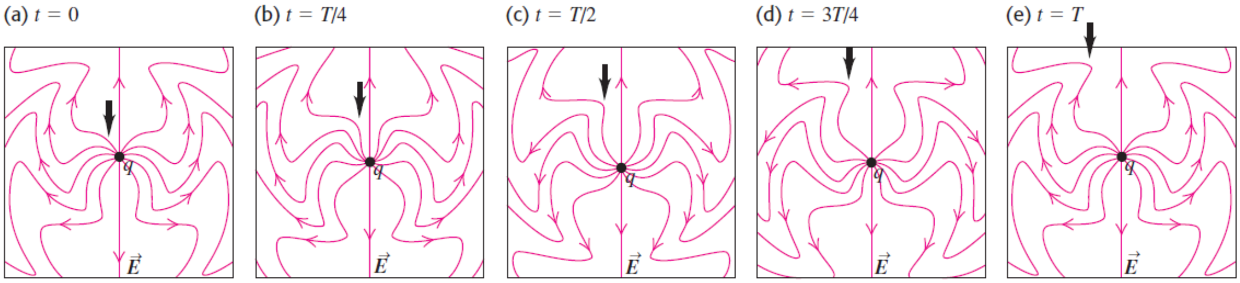


Figure 47 - Electric field lines of a point charge oscillating in simple harmonic motion, seen at five instants during an oscillation period T. The charge's trajectory is in the plane of the drawings. At $t = 0$ the point charge is at its maximum upward displacement. The arrow shows one "kink" in the lines of $\vec{E}$ as it propagates outward from the point charge. The magnetic field (not shown) comprises circles that lie in planes perpendicular to these figures and concentric with the axis of oscillation

Electromagnetic waves with macroscopic wavelengths were first produced in the laboratory in 1887 by the German physicist Heinrich Hertz. As a source of waves, he used charges oscillating in L-C circuits; he detected the resulting electromagnetic waves with other circuits tuned to the same frequency. Hertz also produced electromagnetic *standing waves* and measured the distance between adjacent nodes (one half-wavelength) to determine the wavelength. Knowing the resonant frequency of his circuits, he then found the speed of the waves from the wavelength–frequency relationship $v = \lambda\nu$. He established that their speed was the same as that of light; this verified Maxwell’s theoretical prediction directly. The SI unit of frequency is named in honor of Hertz: One hertz (1 Hz) equals one cycle per second.

The modern value of the speed of light, which we denote by the symbol c, is $299792458\,m/s$. (This value is the basis of our standard of length: One meter is defined to be the distance that light travels in $1/299792458$ second.) For our purposes, is sufficiently accurate. The possible use of electromagnetic waves for long-distance communication does not seem to have occurred to Hertz. It was left to Marconi and others to make radio communication a familiar household experience. In a radio *transmitter*, electric charges are made to oscillate along the length of the conducting antenna, producing oscillating field disturbances like those shown in Fig. 47. Since many charges oscillate together in the antenna, the disturbances are much stronger than those of a single oscillating charge and can be detected at a much greater distance. In a radio *receiver* the antenna is also a conductor; the fields of the wave emanating from a distant transmitter exert forces on free charges within the receiver antenna, producing an oscillating current that is detected and amplified by the receiver circuitry.

For the remainder of this chapter our concern will be with electromagnetic waves themselves, not with the rather complex problem of how they are produced.

The **electromagnetic spectrum** encompasses electromagnetic waves of all frequencies and wavelengths. Figure 48 shows approximate wavelength and frequency ranges for the most commonly encountered portion of the spectrum. Despite vast differences in their uses and means of production, these are all electromagnetic waves with the same propagation speed (in vacuum) $c = 299792458\,m/s$. Electromagnetic waves may differ in frequency ν and wavelength λ, but the relationship $c = \lambda\nu$ in vacuum holds for each.

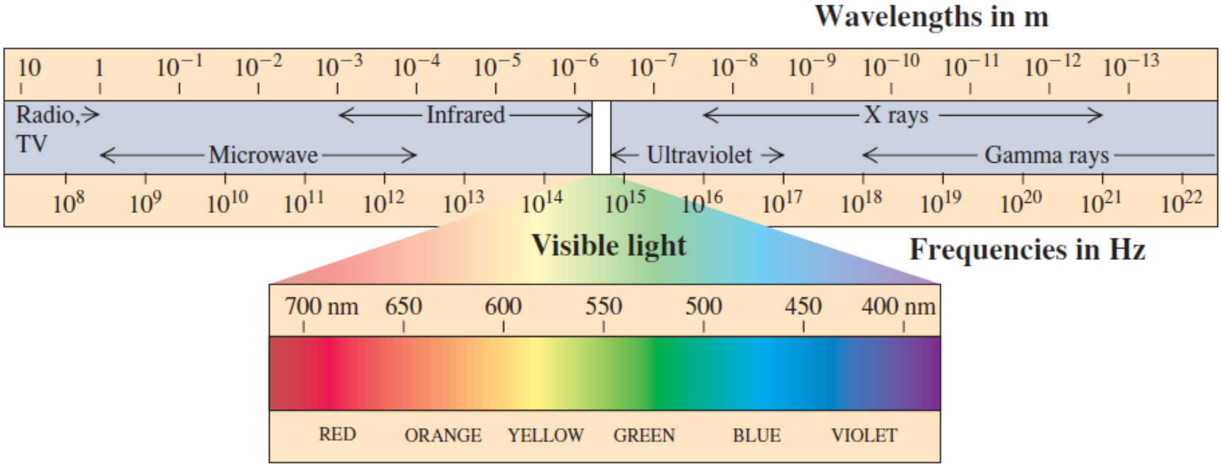


Figure 48 - The electromagnetic spectrum. The frequencies and wavelengths found in nature extend over such a wide range that we have to use a logarithmic scale to show all important bands. The boundaries between bands are somewhat arbitrary

We can detect only a very small segment of this spectrum directly through our sense of sight. We call this range **visible light.** Its wavelengths range from about 380 to 750 nm ($380\ to\ 750 \times 10^{-9}\ m$, with corresponding frequencies from about 790 to

400 THz ($7.9\ to\ 4 \times 10^{14}\ Hz$). Different parts of the visible spectrum evoke in humans the sensations of different colors. Table 2 gives the approximate wavelengths for colors in the visible spectrum.

Table 2 – Wavelengths of visible light

Wavelength, nm	Colour
380 - 450	Violet
450 - 495	Blue
495 - 570	Green
570 - 590	Yellow
590 - 620	Orange
620 - 750	Red

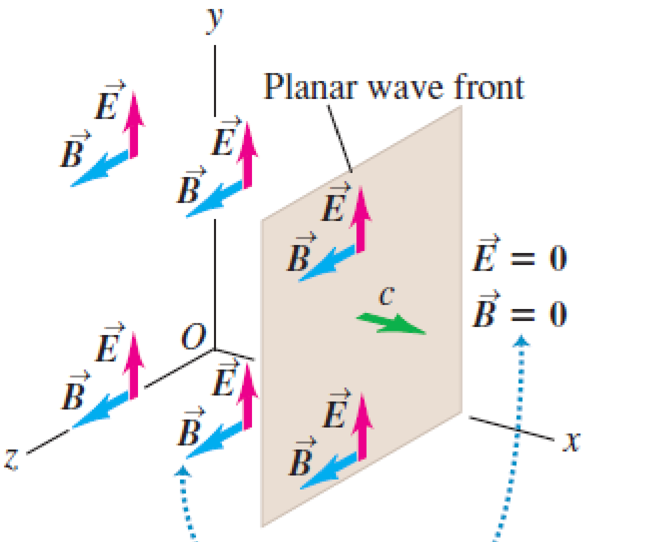
Ordinary white light includes all visible wavelengths. However, by using special sources or filters, we can select a narrow band of wavelengths within a range of a few nm. Such light is approximately *monochromatic* (single-color) light. Absolutely monochromatic light with only a single wavelength is an unattainable idealization. When we use the expression “monochromatic light with $\lambda = 550\ nm$” with reference to a laboratory experiment, we really mean a small band of wavelengths *around* 550 nm. Light from a *laser* is much more nearly monochromatic than is light obtainable in any other way.

Invisible forms of electromagnetic radiation are no less important than visible light. Our system of global communication, for example, depends on radio waves: AM radio uses waves with frequencies from $5.4 \times 10^5\ Hz$ to $1.6 \times 10^6\ Hz$, while FM radio broadcasts are at frequencies from $8.8 \times 10^7\ Hz$ to $1.08 \times 10^8\ Hz$. (Television broadcasts use frequencies that bracket the FM band.) Microwaves are also used for communication (for example, by cellular phones and wireless networks) and for weather radar (at frequencies near $3 \times 10^9\ Hz$).Many cameras have a device that emits a beam of infrared radiation; by analyzing the properties of the infrared radiation reflected from the subject, the camera determines the distance to the subject and automatically adjusts the focus. X rays are able to penetrate through flesh, which makes them invaluable in dentistry and medicine. Gamma rays, the shortest-wavelength type of electromagnetic radiation, are used in medicine to destroy cancer cells.

1.4.5 Plane electromagnetic waves

We are now ready to develop the basic ideas of electromagnetic waves and their relationship to the principles of electromagnetism. Our procedure will be to postulate a simple field configuration that has wavelike behavior. We’ll assume an electric field $\vec{E}$ that has only a and a magnetic field $\vec{B}$ with only a *z*-component, and we’ll assume that both fields move together in the +*x*-direction with a speed that is initially unknown. (As we go along, it will become clear why we choose $\vec{E}$ and $\vec{B}$ to be perpendicular to the direction of propagation as well as to each other.) Then we will test whether these fields are physically possible by asking whether they are consistent with Maxwell's equations, particularly Ampere's law and Faraday's law. We'll find that the answer is yes, provided that c has a particular value. We'll also show that the *wave equation*, which we encountered during our study of mechanical waves, can be derived from Maxwell's equations.

Using an xyz-coordinate system (Fig. 49), we imagine that all space is divided into two regions by a plane perpendicular to the x-axis (parallel to the yz-plane). At every point to the left of this plane there are a uniform electric field $\vec{E}$ in the $+y$-direction and a uniform magnetic field $\vec{B}$ in the $+z$-direction as shown. Furthermore, we suppose that the boundary plane, which we call the *wave front*, moves to the right in the $+x$-direction with a constant speed c the value of which we'll leave undetermined for now. Thus the $\vec{E}$ and $\vec{B}$ fields travel to the right into previously field-free regions with a definite speed. This is a rudimentary electromagnetic wave. A wave such as this, in which at any instant the fields are uniform over any plane perpendicular to the direction of propagation, is called a **plane wave.** In the case shown in Fig. 49, the fields are zero for planes to the right of the wave front and have the same values on all planes to the left of the wave front; later we will consider more complex plane waves.

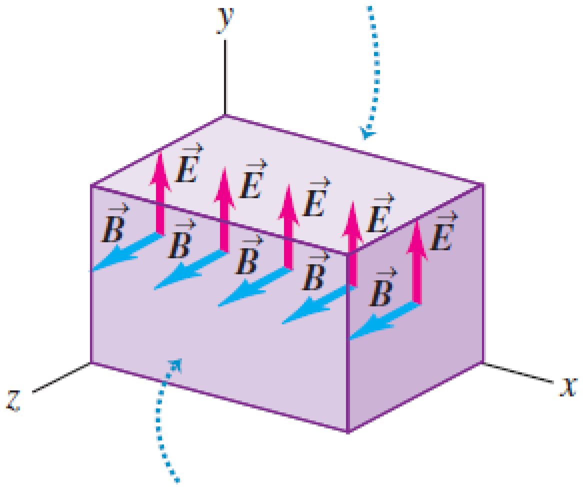


The electric and magnetic fields are uniform behind the advancing wave front and zero in front of it.

Figure 49 - An electromagnetic wave front. The plane representing the wave front moves to the right (in the positive x-direction) with speed c

We won't concern ourselves with the problem of actually *producing* such a field configuration. Instead, we simply ask whether it is consistent with the laws of electromagnetism - that is, with Maxwell's equations. We'll consider each of these four equations in turn.

The electric field is the same on the top and bottom sides of the Gaussian surface, so the total electric flux through the surface is zero.



The magnetic field is the same on the left and right sides of the Gaussian surface, so the total magnetic flux through the surface is zero.

Figure 50 - Gaussian surface for a transverse plane electromagnetic wave

Let us first verify that our wave satisfies Maxwell's first and second equations - that is, Gauss's laws for electric and magnetic fields. To do this, we take as our Gaussian surface a rectangular box with sides parallel to the *xy*, *xz*, and *yz* coordinate planes (Fig. 50). The box encloses no electric charge. The total electric flux and magnetic flux through the box are both zero, even if part of the box is in the region where $E = B = 0$. This would *not* be the case if $\vec{E}$ or $\vec{B}$ had an *x*-component, parallel to the direction of propagation; if the wave front were inside the box, there would be flux through the left-hand side of the box (at *x*=0) but not the right-hand side (at *x*>0). Thus to satisfy Maxwell's first and second equations, the electric and magnetic fields must be perpendicular to the direction of propagation; that is, the wave must be **transverse.**

The next of Maxwell's equations to be considered is Faraday's law:

$$\oint \vec{E} \cdot d\vec{l} = -\frac{d\Phi_B}{dt} \tag{91}$$

To test whether our wave satisfies Faraday's law, we apply this law to a rectangle *efgh* that is parallel to the *xy*-plane (Fig. 51a). As shown in Fig. 51b, a cross section in the *xy*-plane, this rectangle has height *a* and width Δx. At the time shown, the wave front has progressed partway through the rectangle, and $\vec{E}$ is zero along the side *ef*. In applying Faraday's law we take the vector area $d\vec{S}$ of rectangle *efgh* to be in the +*z*-direction. With this choice the right-hand rule requires that we integrate $\vec{E} \cdot d\vec{l}$ *counterclockwise* around the rectangle. At every point on side *ef*, $\vec{E}$ is zero. At every point on sides and *he*, $\vec{E}$ is either zero or perpendicular to $d\vec{l}$. Only side *gh* contributes to the integral. On this side, $\vec{E}$ and $d\vec{l}$ are opposite, and we obtain

$$\oint \vec{E} \cdot d\vec{l} = -Ea \tag{92}$$

Hence, the left-hand side of Eq. $v = \frac{c}{\sqrt{\varepsilon\varepsilon_0}}$ is nonzero.

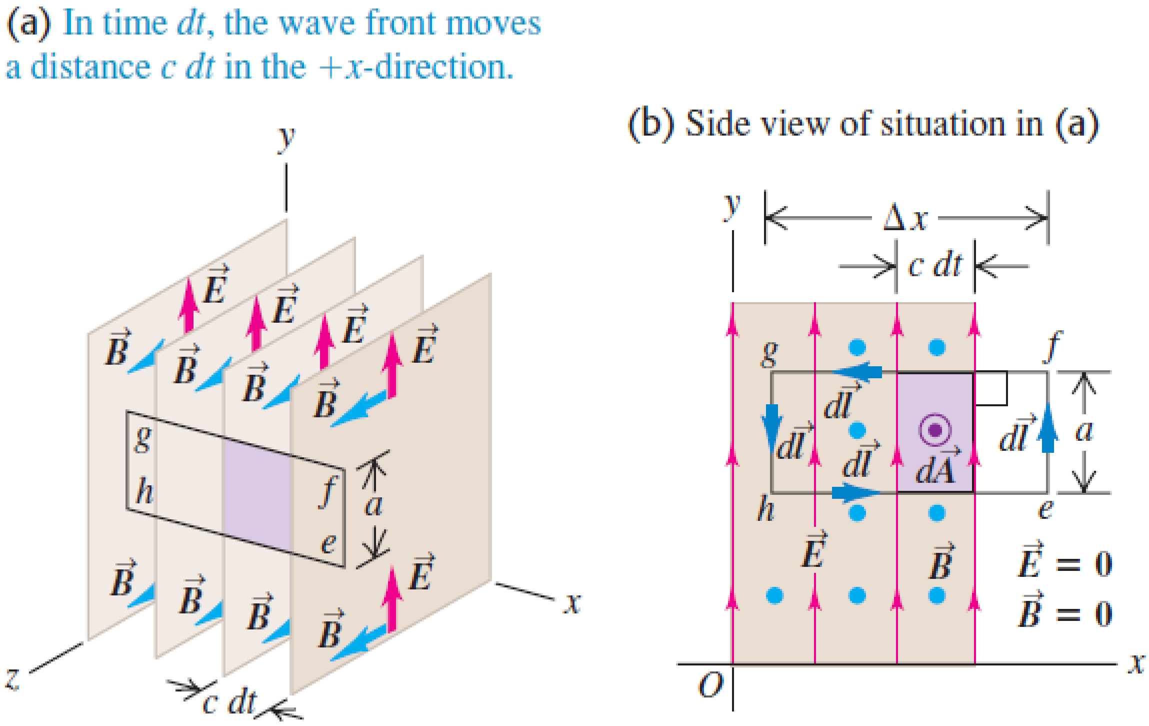
To satisfy Faraday's law, Eq. $v = \frac{c}{\sqrt{\varepsilon\varepsilon_0}}$, there must be a component of $\vec{B}$ in the *z*-component (perpendicular to $\vec{E}$) so that there can be a nonzero magnetic flux $d\Phi_B$ through the rectangle *efgh* and a nonzero derivative $\frac{d\Phi_B}{dt}$. Indeed, in our wave, $\vec{B}$ has *only* a z-component. We have assumed that this component is in the *positive z*-direction. let's see whether this assumption is consistent with Faraday's law. During a time interval dt the wave front moves a distance $c\,dt$ to the right in Fig. 51b, sweeping out an area $ac\,dt$ of the rectangle *efgh*. During this interval the magnetic flux $d\Phi_B$ through the rectangle *efgh* increases by $d\Phi_B = B(ac\,dt)$, so the rate of change of magnetic flux is

$$\frac{d\Phi_B}{dt} = Bac \tag{93}$$

Now we substitute Eqs. (92) and (93) into Faraday's law, Eq. (91); we get

$$\begin{aligned} -Ea &= -Bac \\ E &= cB \end{aligned} \tag{94}$$

This shows that our wave is consistent with Faraday's law only if the wave speed *c* and the magnitudes of the perpendicular vectors $\vec{E}$ and $\vec{B}$ are related as in Eq. (94). Note that if we had assumed that $\vec{B}$ was in the *negative* z-direction, there would have been an additional minus sign in Eq. (94); since $E, c,$ and B are all positive magnitudes, no solution would then have been possible. Furthermore, any component of $\vec{B}$ in the *y*-direction (parallel to $\vec{E}$) would not contribute to the changing magnetic flux Φ_B through the rectangle *efgh* (which is parallel to the *xy*-plane) and so would not be part of the wave.



(a) In time dt, the wave front moves a distance $c\,dt$ in the $+x$-direction.

(b) Side view of situation in (a)

Figure 51 - (a) Applying Faraday's law to a plane wave. (b) In a time *dt*, the magnetic flux through the rectangle in the *xy*-plane increases by an amount $d\Phi_B$. This increase equals the flux through the shaded rectangle with area *ac dt*; that is, $d\Phi_B = Bac\ dt$. Thus $d\Phi_B/dt = Bac$

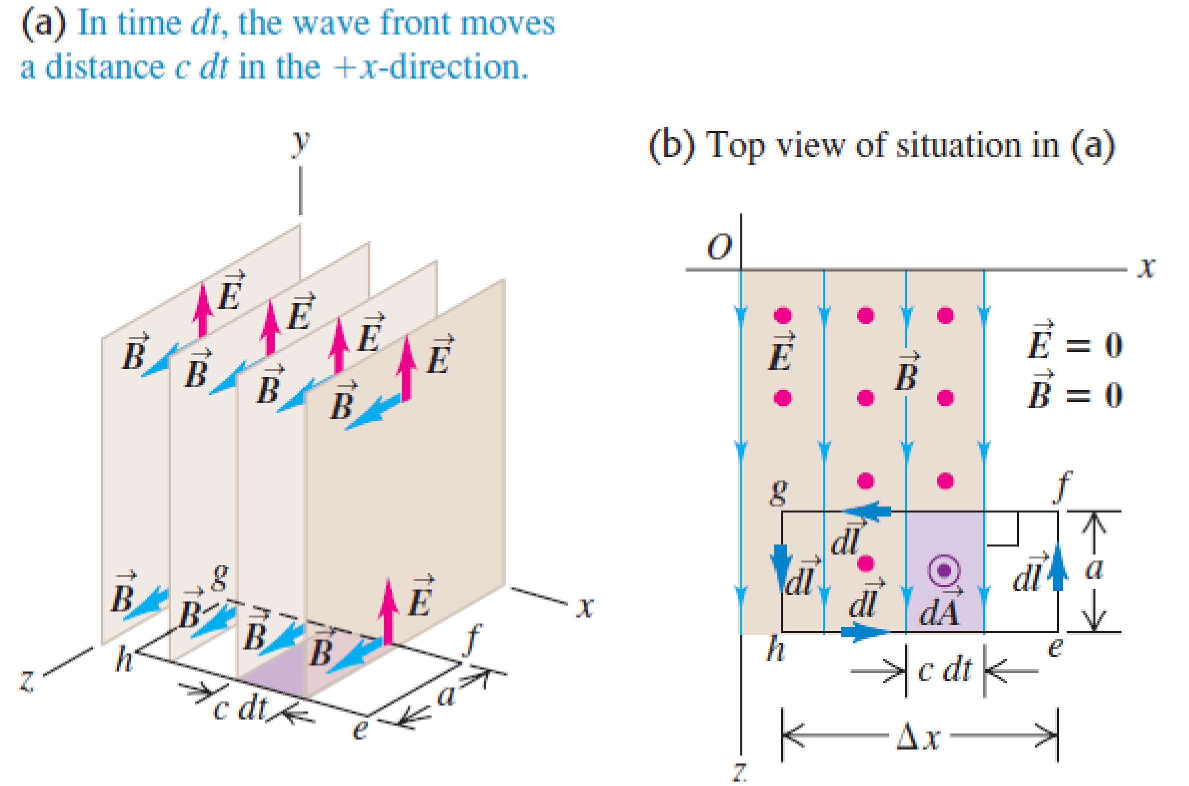
Finally, we carry out a similar calculation using Ampere's law, the remaining member of Maxwell's equations. There is no conduction current ($i_C = 0$) so Ampere's law is

$$\oint \vec{B} \cdot d\vec{l} = \mu_0 \varepsilon_0 \frac{d\Phi_E}{dt} \qquad (95)$$

To check whether our wave is consistent with Ampere's law, we move our rectangle so that it lies in the *xz*-plane, as shown in Fig. 52, and we again look at the situation at a time when the wave front has traveled partway through the rectangle. We take the vector area $d\vec{S}$ in the +*y*-direction, and so the right-hand rule requires that we integrate $\vec{B} \cdot d\vec{l}$ counterclockwise around the rectangle. The $\vec{B}$ field is zero at every point along side *ef*, and at each point on sides *fg* and *he* it is either zero or perpendicular to $d\vec{l}$. Only side *gh*, where $\vec{B}$ and $d\vec{l}$ are parallel, contributes to the integral, and we find

$$\oint \vec{B} \cdot d\vec{l} = Ba \qquad (96)$$

Hence, the left-hand side of Ampere's law, Eq. (95), is nonzero; the right-hand side must be nonzero as well. Thus $\vec{E}$ must have a (perpendicular to $\vec{B}$) so that the electric flux Φ_E through the rectangle and the time derivative $d\Phi_E/dt$ can be nonzero. We come to the same conclusion that we inferred from Faraday's law: In an electromagnetic wave, $\vec{E}$ and Φ_B must be mutually perpendicular.



(a) In time dt, the wave front moves a distance $c\,dt$ in the $+x$-direction.

(b) Top view of situation in (a)

Figure 52 - (a) Applying Ampere's law to a plane wave. (Compare to Fig. 51a.) (b) In a time *dt*, the electric flux through the rectangle in the *xz*-plane increases by an amount $d\Phi_E$ This increase equals the flux through the shaded rectangle with area *ac dt*; that is, $d\Phi_E = Eac\,dt$. Thus $d\Phi_E/dt = Eac$

In a time interval dt the electric flux Φ_E through the rectangle increases by $d\Phi_E = E(ac\,dt)$. Since we chose $d\vec{S}$ to be in the +*y*-direction, this flux change is positive; the rate of change of electric field is

$$\frac{d\Phi_E}{dt} = Eac \qquad (97)$$

Substituting Eqs. (96) and (97) into Ampere's law, Eq. (95), we find

$$\begin{gathered} Ba = \varepsilon_0\mu_0 Eac \\ B = \varepsilon_0\mu_0 E \end{gathered} \qquad (98)$$

Thus our assumed wave obeys Ampere's law only if $B, c,$ and E are related as in Eq. (98).

Our electromagnetic wave must obey *both* Ampere's law and Faraday's law, so Eqs. (94) and (98) must both be satisfied. This can happen only if $\mu_0\varepsilon_0 c = 1/c$ or

$$c = \frac{1}{\sqrt{\varepsilon_0 \mu_0}} \tag{99}$$

Inserting the numerical values of these quantities, we find

$$c = \frac{1}{\sqrt{8.85 \cdot 10^{-12} \times 4\pi \cdot 10^{-7}}} = 3 \cdot 10^8 \, m/s$$

Our assumed wave is consistent with all of Maxwell's equations, provided that the wave front moves with the speed given above, which you should recognize as the speed of light! Note that the *exact* value of is defined to be $299792458 \, m/s$ the modern value of is defined to agree with this when used in Eq. (99).

1.4.6 Sinusoidal electromagnetic waves

Sinusoidal electromagnetic waves are directly analogous to sinusoidal transverse mechanical waves on a stretched string. In a sinusoidal electromagnetic wave, $\vec{E}$ and $\vec{B}$ at any point in space are sinusoidal functions of time, and at any instant of time the *spatial* variation of the fields is also sinusoidal. Some sinusoidal electromagnetic waves are *plane waves;* the property that at any instant the fields are uniform over any plane perpendicular to the direction of propagation. The entire pattern travels in the direction of propagation with speed c. The directions of $\vec{E}$ and $\vec{B}$ are perpendicular to the direction of propagation (and to each other), so the wave is *transverse.* Electromagnetic waves produced by an oscillating point charge, shown in Fig. 47, are an example of sinusoidal waves that are *not* plane waves. But if we restrict our observations to a relatively small region of space at a sufficiently great distance from the source, even these waves are well approximated by plane waves (Fig. 53). In the same way, the curved surface of the (nearly) spherical earth appears flat to us because of our small size relative to the earth's radius. In this section we'll restrict our discussion to plane waves.

The frequency ν, the wavelength λ, and the speed of propagation c of any periodic wave are related by the usual wavelength–frequency relationship $c = \lambda\nu$. If the frequency ν is 10^8 Hz (100 MHz), typical of commercial FM radio broadcasts, the wavelength is

$$\lambda = \frac{3 \cdot 10^8 \, m/s}{10^8 \, Hz} = 3 \, m$$

Figure 48 shows the inverse proportionality between wavelength and frequency. Figure 54 shows a linearly polarized sinusoidal electromagnetic wave traveling in the +*x*-direction. The $\vec{E}$ and $\vec{B}$ vectors are shown for only a few points on the positive *x*-

axis. Note that the electric and magnetic fields oscillate in phase: $\vec{E}$ is maximum where $\vec{B}$ is maximum and $\vec{E}$ is zero where $\vec{B}$ is zero. Note also that where $\vec{E}$ is in the +*y*-direction, $\vec{B}$ is in the +*z*-direction; where $\vec{E}$ is in the –*y*-direction, $\vec{B}$ is in the –*z*-direction. At all points the vector product $\vec{E} \times \vec{B}$ is in the direction in which the wave is propagating (the +*x*-direction).

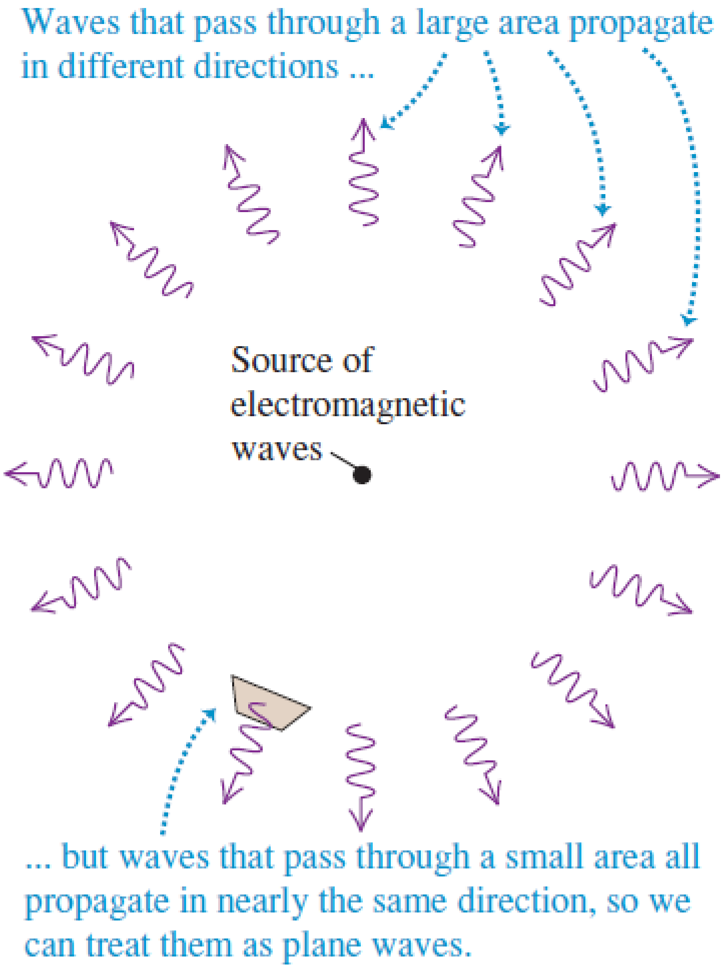

Waves that pass through a large area propagate in different directions ...
Source of electromagnetic waves
... but waves that pass through a small area all propagate in nearly the same direction, so we can treat them as plane waves.

Figure 53 - Waves passing through a small area at a sufficiently great distance from a source can be treated as plane waves

We can describe electromagnetic waves by means of *wave functions*. One form of the wave function for a transverse wave traveling in the +*x*-direction along a stretched string is:

$$y(x,t) = A\cos(kx - \omega t) \tag{100}$$

where $y(x,t)$ is the transverse displacement from its equilibrium position at time *t* of a point with coordinate *x* on the string. The quantity *A* is the maximum displacement, or *amplitude*, of the wave; ω is its *angular frequency*, equal to 2π times the frequency ν; and *k* is the *wave number*, equal to $2\pi/\lambda$, where λ is the wavelength.

Let $E_y(x,t)$ and $B_z(x,t)$ represent the instantaneous values of the *y*-component of $\vec{E}$ and the *z*-component of $\vec{B}$, respectively, in Fig. 54, and let E_{max} and B_{max} represent the maximum values, or *amplitudes*, of these fields. The wave functions for the wave are then

$$E_y(x,t) = E_{max}\cos(kx-\omega t)\,, B_z(x,t) = B_{max}\cos(kx-\omega t) \qquad (101)$$

We can also write the wave functions in vector form:

$$\begin{aligned}\vec{E}(x,t) &= \vec{j}E_{max}\cos(kx-\omega t),\\ \vec{B}(x,t) &= \vec{k}B_{max}\cos(kx-\omega t)\end{aligned} \qquad (102)$$

The sine curves in Fig. 54 represent instantaneous values of the electric and magnetic fields as functions of x at time $t=0$ - that is, $\vec{E}(x,t=0)$ and $\vec{B}(x,t=0)$. As time goes by, the wave travels to the right with speed c. Equations (101) and (102) show that at any point the sinusoidal oscillations of and are *in phase*. From Eq. (94) the amplitudes must be related by

$$E_{max} = cB_{max} \qquad (103)$$

These amplitude and phase relationships are also required for $E(x,t)$ and $B(x,t)$, which came from Faraday's law and Ampere's law, respectively.

Figure 54 shows the electric and magnetic fields of a wave traveling in the *negative x*-direction. At points where $\vec{E}$ is in the positive *y*-direction, $\vec{B}$ is in the *negative z*-direction; where $\vec{E}$ is in the negative *y*-direction, $\vec{B}$ is in the *positive z*-direction. The wave functions for this wave are

$$E_y(x,t) = E_{max}\cos(kx-\omega t)\,, B_z(x,t) = -B_{max}\cos(kx-\omega t) \qquad (104)$$

As with the wave traveling in the +*x*-direction, at any point the sinusoidal oscillations of the $\vec{E}$ and $\vec{B}$ fields are *in phase,* and the vector product $\vec{E}\times\vec{B}$ points in the direction of propagation.

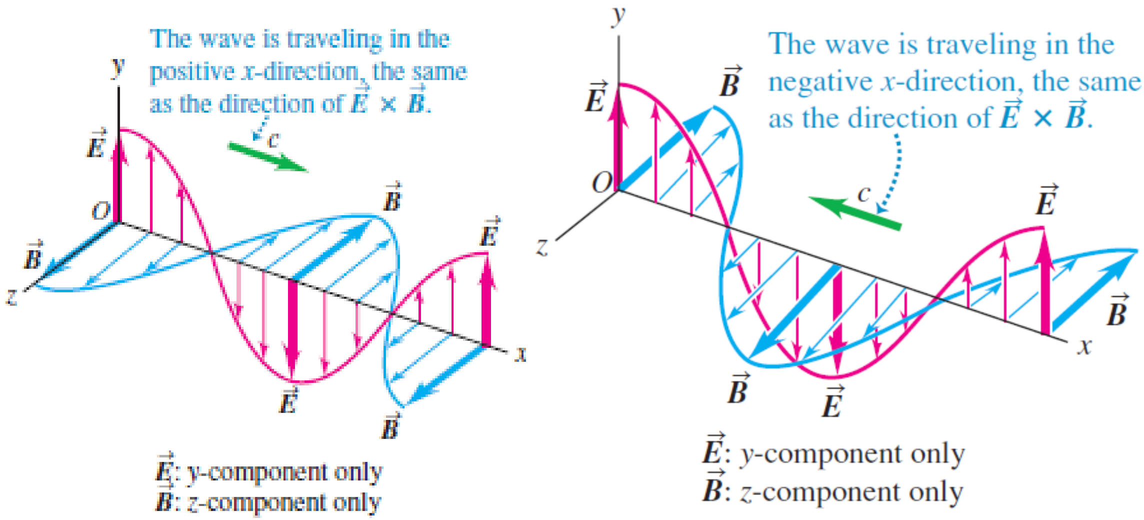

The wave is traveling in the positive x-direction, the same as the direction of $\vec{E} \times \vec{B}$.
c
y
$\vec{E}$
O
$\vec{B}$
z
$\vec{B}$
$\vec{E}$
$\vec{E}$
$\vec{B}$
x
$\vec{E}$: y-component only
$\vec{B}$: z-component only
y
$\vec{E}$
$\vec{B}$
The wave is traveling in the negative x-direction, the same as the direction of $\vec{E} \times \vec{B}$.
O
z
c
$\vec{E}$
$\vec{B}$
x
$\vec{B}$
$\vec{E}$
$\vec{E}$: y-component only
$\vec{B}$: z-component only

Figure 54 - Representation of one wavelength of a linearly polarized sinusoidal plane electromagnetic wave traveling in the negative x-direction at $t = 0$. The fields are shown only for points along the x-axis

The sinusoidal waves shown in Fig. 54 is both linearly polarized in the *y*-direction; the $\vec{E}$ field is always parallel to the *y*-axis.

1.4.7 Energy in electromagnetic waves

It is a familiar fact that energy is associated with electromagnetic waves; think of the energy in the sun’s radiation. Microwave ovens, radio transmitters, and lasers for eye surgery all make use of the energy that these waves carry. To understand how to utilize this energy, it’s helpful to derive detailed relationships for the energy in an electromagnetic wave.

We begin with the expressions derived early for the **energy densities** in electric and magnetic fields; we suggest you review those derivations now. Equations $u = \frac{1}{2}\varepsilon_0 E^2$ and $u = \frac{1}{2\mu_0}B^2$ show that in a region of empty space where $\vec{E}$ and $\vec{B}$ fields are present, the total energy density is given by

$$u = \frac{1}{2}\varepsilon_0 E^2 + \frac{1}{2\mu_0}B^2 \tag{105}$$

where ε_0 and μ_0 are, respectively, the permittivity and permeability of free space. For electromagnetic waves in vacuum, the magnitudes E and B are related by

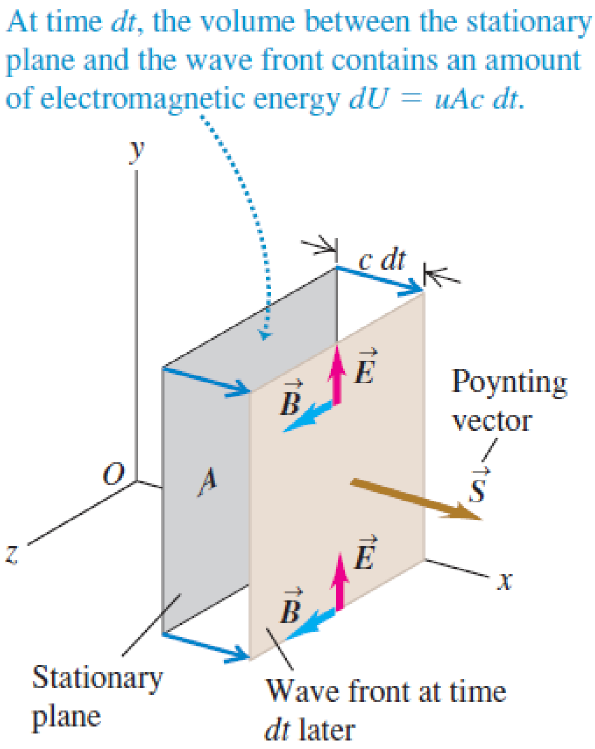
$$B = \frac{E}{c} = \sqrt{\varepsilon_0 \mu_0} E \tag{106}$$

Combining Eqs. (105) and (106), we can also express the energy density u in a simple electromagnetic wave in vacuum as

$$u = \frac{1}{2}E^2 + \frac{1}{2}\left(\sqrt{\varepsilon_0\mu_0}\right)^2 = \varepsilon_0 E^2 \qquad (107)$$

This shows that in vacuum, the energy density associated with the field $\vec{E}$ in our simple wave is equal to the energy density of the $\vec{B}$ field. In general, the electric field magnitude E is a function of position and time, as for the sinusoidal wave described by Eqs. (101); thus the energy density of an electromagnetic wave, given by Eq. (107), also depends in general on position and time.

Electromagnetic waves such as those we have described are *travelling* waves that transport energy from one region to another. We can describe this energy transfer in terms of energy transferred *per unit time per unit cross-sectional area,* or *power per unit area,* for an area perpendicular to the direction of wave travel.



At time *dt*, the volume between the stationary plane and the wave front contains an amount of electromagnetic energy $dU = uAc\,dt$.
y
$c\,dt$
$\vec{E}$
$\vec{B}$
Poynting vector
$\vec{S}$
O
A
z
$\vec{E}$
x
$\vec{B}$
Stationary plane
Wave front at time *dt* later

Figure 55 - A wave front at a time *dt* after it passes through the stationary plane with area *S*

To see how the energy flow is related to the fields, consider a stationary plane, perpendicular to the that coincides with the wave front at a certain time. In a time after this, the wave front moves a distance $dx = c\,dt$ to the right of the plane. Considering an area A on this stationary plane (Fig. 55), we note that the energy in the space to the right of this area must have passed through the area to reach the new location. The volume dV of the relevant region is the base area S times the length $c\,dt$, and the energy dU in this region is the energy density u times this volume:

$$dU = udV = (\varepsilon_0 E^2)(Sc\,dt) \qquad (108)$$

This energy passes through the area S in time dt. The energy flow per unit time per unit area, which we will call S is

$$S = \frac{1}{S}\frac{dU}{dt} = \varepsilon_0 E^2 \qquad (109)$$

Using Eqs. (94) and (99), we can derive the alternative forms

$$S = \frac{\varepsilon_0}{\sqrt{\varepsilon_0 \mu_0}} E^2 = \sqrt{\frac{\varepsilon_0}{\mu_0}} E^2 = \frac{EB}{\mu_0} \qquad (110)$$

We leave the derivation of Eq. (110) from Eq. (109) as an exercise for you. The units of S are energy per unit time per unit area, or power per unit area. The SI unit of S is $1\, J/s \cdot m^2$ or $1\, W/m^2$.

We can define a *vector* quantity that describes both the magnitude and direction of the energy flow rate:

$$\vec{S} = \frac{1}{\mu_0} \vec{E} \times \vec{B} \qquad (111)$$

The vector $\vec{S}$ is called the **Poynting vector;** it was introduced by the British physicist John Poynting (1852–1914). Its direction is in the direction of propagation of the wave. Since $\vec{E}$ and $\vec{B}$ are perpendicular, the magnitude of $\vec{S}$ is $S = \frac{EB}{\mu_0}$ from Eqs. (109) and (116) this is the energy flow per unit area and per unit time through a cross-sectional area perpendicular to the propagation direction. The total energy flow per unit time (power, P) out of any closed surface is the integral of $\vec{S}$ over the surface:

$$P = \oint \vec{S} \cdot d\vec{S} \qquad (112)$$

For the sinusoidal waves studied early, as well as for other more complex waves, the electric and magnetic fields at any point vary with time, so the Poynting vector at any point is also a function of time. Because the frequencies of typical electromagnetic waves are very high, the time variation of the Poynting vector is so rapid that it's most appropriate to look at its *average* value. The magnitude of the average value of $\vec{S}$ at a point is called the **intensity** of the radiation at that point. The SI unit of intensity is the same as for S, $1\, W/m^2$ (watt per square meter).

Let's work out the intensity of the sinusoidal wave described by Eqs. (102). We first substitute $\vec{E}$ and $\vec{B}$ into Eq. (111):

$$\vec{S}(x,t) = \frac{1}{\mu_0}\vec{E}(x,t) \times \vec{B}(x,t) = \tag{113}$$
$$= \frac{1}{\mu_0}[\vec{j}E_{max}\cos(kx-\omega t)] \times \left[\vec{k}B_{max}\cos(kx-\omega t)\right]$$

The vector product of the unit vectors is $\vec{j} \times \vec{k} = \vec{i}$ and $\cos^2(kx-\omega t)$ is never negative, so $\vec{S}(x,t)$ always points in the positive *x*-direction (the direction of wave propagation). The *x*-component of the Poynting vector is

$$S_x(x,t) = \frac{E_{max}B_{max}}{\mu_0}\cos^2(kx-\omega t) = \frac{E_{max}B_{max}}{2\mu_0}[1+\cos 2(kx-\omega t)] \tag{114}$$

The time average value of $\cos 2(kx-\omega t)$ is zero because at any point, it is positive during one half-cycle and negative during the other half. So the average value of the Poynting vector over a full cycle is $\vec{S}_{av} = \vec{i}S_{av}$ where

$$S_{av} = \frac{E_{max}B_{max}}{2\mu_0} \tag{115}$$

That is, the magnitude of the average value of $\vec{S}$ for a sinusoidal wave (the intensity I of the wave) is $1/2$ the maximum value. By using the relationships $E_{max} = cB_{max}$ and $\varepsilon_0\mu_0 = 1/c^2$ we can express the intensity in several equivalent
forms:

$$I = S_{av} = \frac{E_{max}B_{max}}{2\mu_0} = \frac{E_{max}^2}{2\mu_0 c} = \tag{116}$$
$$= \frac{1}{2}\sqrt{\frac{\varepsilon_0}{\mu_0}}E_{max}^2 = \frac{1}{2}\varepsilon_0 c E_{max}^2$$

We invite you to verify that these expressions are all equivalent.

For a wave traveling in the –*x*-direction, represented by Eqs. (104), the Poynting vector is in the –*x*-direction at every point, but its magnitude is the same as for a wave traveling in the +*x*-direction. Verifying these statements is left to you.

Discussion questions

1. By measuring the electric and magnetic fields at a point in space where there is an electromagnetic wave, can you determine the direction from which the wave came? Explain.
2. According to Ampere's law, is it possible to have both a conduction current and a displacement current at the same time? Is it possible for the effects of the two kinds of current to cancel each other exactly so that *no* magnetic field is produced? Explain.

3. Give several examples of electromagnetic waves that are encountered in everyday life. How are they all alike? How do they differ?
4. Sometimes neon signs located near a powerful radio station are seen to glow faintly at night, even though they are not turned on. What is happening?
5. Is polarization a property of all electromagnetic waves, or is it unique to visible light? Can sound waves be polarized? What fundamental distinction in wave properties is involved? Explain.
6. The light beam from a searchlight may have an electricfield magnitude of 1000V/m corresponding to a potential difference of 1500 V between the head and feet of a 1.5-m-tall person on whom the light shines. Does this cause the person to feel a strong electric shock? Why or why not?
7. For a certain sinusoidal wave of intensity *I*, the amplitude of the magnetic field is *B*. What would be the amplitude (in terms of *B*) in a similar wave of twice the intensity?
8. Most automobiles have vertical antennas for receiving radio broadcasts. Explain what this tells you about the direction of polarization of in the radio waves used in broadcasting.
9. If a light beam carries momentum, should a person holding a flashlight feel a recoil analogous to the recoil of a rifle when it is fired? Why is this recoil not actually observed?
10. A light source radiates a sinusoidal electromagnetic wave uniformly in all directions. This wave exerts an average pressure *p* on a perfectly reflecting surface a distance *R* away from it. What average pressure (in terms of *p*) would this wave exert on a perfectly absorbing surface that was twice as far from the source?
11. Does an electromagnetic *standing* wave have energy? Does it have momentum? Are your answers to these questions the same as for a *traveling* wave? Why or why not?
12. When driving on the upper level of the Bay Bridge, westbound from Oakland to San Francisco, you can easily pick up a number of radio stations on your car radio. But when driving eastbound on the lower level of the bridge, which has steel girders on either side to support the upper level, the radio reception is much worse. Why is there a difference?

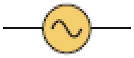
1.5 Alternating current

1.5.1 Phasors and alternating currents

During the 1880s in the United States there was a heated and acrimonious debate between two inventors over the best method of electric-power distribution. Thomas Edison favored direct current (dc)—that is, steady current that does not vary with time. George Westinghouse favored **alternating current (ac),** with sinusoidally varying voltages and currents. He argued that transformers (which we will study in this chapter) can be used to step the voltage up and down with ac but not with dc; low voltages are safer for consumer use, but high voltages and correspondingly low currents are best for long-distance power transmission to minimize i^2R losses in the cables.

Eventually, Westinghouse prevailed, and most present-day household and industrial power-distribution systems operate with alternating current. Any appliance that you plug into a wall outlet uses ac, and many battery-powered devices such as radios and cordless telephones make use of the dc supplied by the battery to create or amplify alternating currents. Circuits in modern communication equipment, including pagers and television, also make extensive use of ac.

To supply an alternating current to a circuit, a source of alternating emf or voltage is required. An example of such a source is a coil of wire rotating with constant angular velocity in a magnetic field. This develops a sinusoidal alternating emf and is the prototype of the commercial alternating-current generator or *alternator*.

We use the term **ac source** for any device that supplies a sinusoidally varying voltage (potential difference) u or i current The usual circuit-diagram symbol for an ac source is  .

A sinusoidal voltage might be described by a function such as

$$u = U\cos\omega t \qquad (31.1\ 17)$$

In this expression, u (lowercase) is the *instantaneous* potential difference; U (uppercase) is the maximum potential difference, which we call the **voltage amplitude;** and is the *angular frequency*, equal to 2π times the frequency ν (Fig. 56).

In the United States and Canada, commercial electric-power distribution systems always use a frequency of $\nu = 60\ Hz$, corresponding ; in much of the rest of the world, $\nu = 50\ Hz$ is used. Similarly, a sinusoidal current might be described as

$$i = I\cos\omega t \qquad (31.2\ 118)$$

where i (lowercase) is the instantaneous current and I (uppercase) is the maximum current or **current amplitude.**

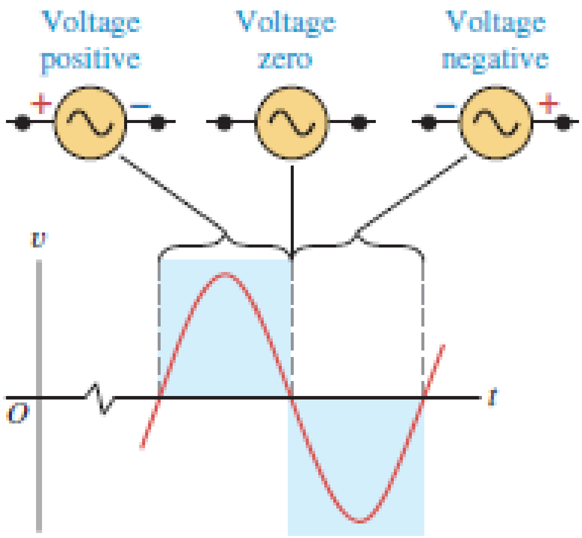


Figure 56 – The voltage across a sinusoidal ac source

To represent sinusoidally varying voltages and currents, we will use rotating vector diagrams similar to those we used in the study of simple harmonic motion. In these diagrams the instantaneous value of a quantity that varies sinusoidally with time is represented by the *projection* onto a horizontal axis of a vector with a length equal to the amplitude of the quantity. The vector rotates counterclockwise with constant angular speed ω. These rotating vectors are called **phasors,** and diagrams containing them are called **phasor diagrams.** Figure 57 shows a phasor diagram for the sinusoidal current described by Eq. (118). The projection of the phasor onto the horizontal axis at time t is $I\cos\omega t$; this is why we chose to use the cosine function rather than the sine in Eq. (118).

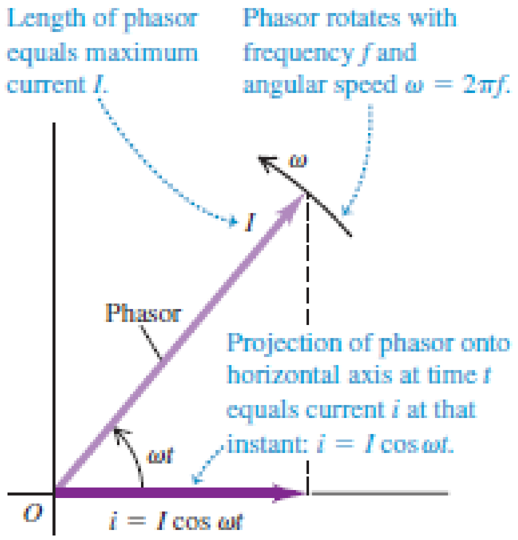

Length of phasor equals maximum current I.
Phasor rotates with frequency f and angular speed ω = 2πf.
ω
I
Phasor
Projection of phasor onto horizontal axis at time t equals current i at that instant: i = I cos ωt.
ωt
O
i = I cos ωt

Figure 57 – A phasor diagram

How do we measure a sinusoidally varying current? We used a d’Arsonval galvanometer to measure steady currents. But if we pass a *sinusoidal* current through a d’Arsonval meter, the torque on the moving coil varies sinusoidally, with one direction half the time and the opposite direction the other half. The needle may wiggle a little if the frequency is low enough, but its average deflection is zero. Hence a d’Arsonval meter by itself isn’t very useful for measuring alternating currents.

To get a measurable one-way current through the meter, we can use *diodes*. A diode is a device that conducts better in one direction than in the other; an ideal diode has zero resistance for one direction of current and infinite resistance for the other. Figure 58a shows one possible arrangement, called a *full-wave rectifier circuit*. The current through the galvanometer G is always upward, regardless of the direction of the current from the ac source (i.e., which part of the cycle the source is in). The graph in Fig. 58b shows the current through G: It pulsates but always has the same direction, and the average meter deflection is *not* zero.

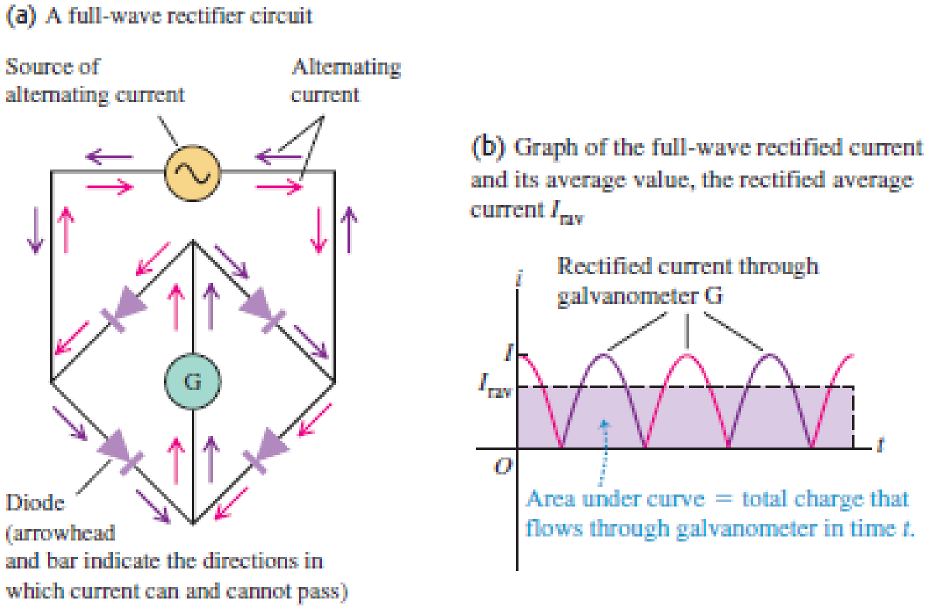

(a) A full-wave rectifier circuit
Source of alternating current
Alternating current
G
Diode (arrowhead and bar indicate the directions in which current can and cannot pass)
(b) Graph of the full-wave rectified current and its average value, the rectified average current I_{rav}
Rectified current through galvanometer G
i
I
I_{rav}
O
t
Area under curve = total charge that flows through galvanometer in time t.

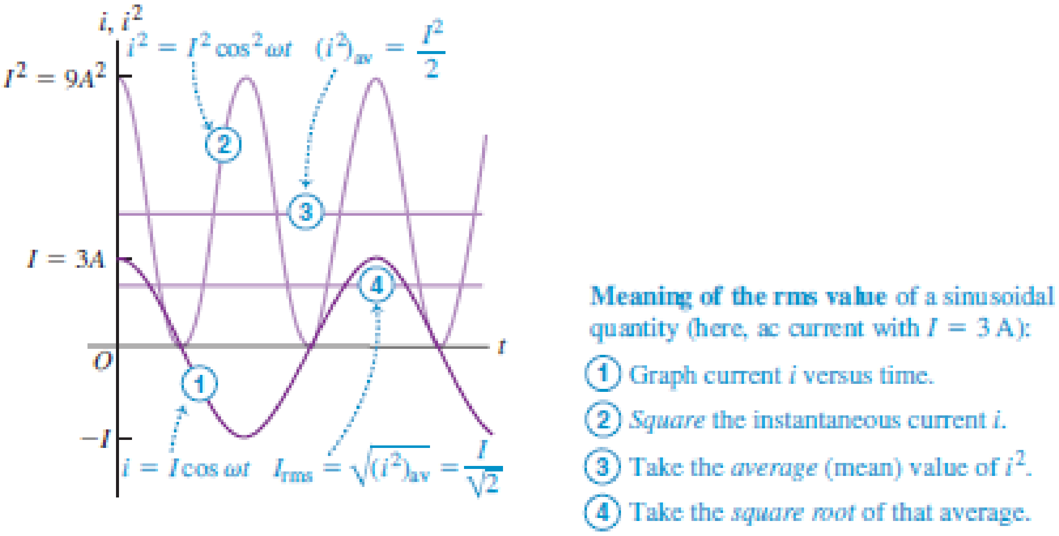
Figure 58 - (a) A full-wave rectifier circuit. (b) Graph of the resulting current through the galvanometer G.

The **rectified average current** I_{rav} is defined so that during any whole number of cycles, the total charge that flows is the same as though the current were constant with a value equal to I_{rav}. The notation I_{rav} and the name *rectified average* current emphasize that this is *not* the average of the original sinusoidal current. In Fig. 58b the total charge that flows in time *t* corresponds to the area under the curve of *i* versus *t;* this area must equal the rectangular area with height I_{rav}. We see that I_{rav} is less than the maximum current *I;* the two are related by

$$I_{rav} = \frac{2}{\pi} I = 0.637\, I \tag{119}$$

The galvanometer deflection is proportional to I_{rav}. The galvanometer scale can be calibrated to read *I* , I_{rav} or, most commonly, I_{rms}.

A more useful way to describe a quantity that can be either positive or negative is the *root-mean-square (rms) value.*We used rms values early in connection with the speeds of molecules in a gas. We *square* the instantaneous current i, take the *average* (mean) value of i^2, and finally take the *square root* of that average. This procedure defines the **root-mean-square current,** denoted as I_{rms} (Fig. 59). Even when i is negative, i^2 is always positive, so I_{rms} is never zero (unless i is zero at every instant).



Meaning of the rms value of a sinusoidal quantity (here, ac current with $I = 3$ A):

1 Graph current i versus time.

2 *Square* the instantaneous current i.

3 Take the *average* (mean) value of i^2.

4 Take the *square root* of that average.

Figure 59 - Calculating the root-mean-square (rms) value of an alternating current

Here’s how we obtain I_{rms} for a sinusoidal current, like that shown in Fig. 59. If the instantaneous current is given by $i = I \cos \omega t$, then

$$i^2 = I^2 \cos^2 \omega t \tag{120}$$

Using a double-angle formula from trigonometry,

$$\cos^2 \omega t = \frac{1}{2}(1 + \cos 2\omega t) \tag{121}$$

we find

$$i^2 = I^2 \frac{1}{2}(1 + \cos 2\omega t) = I^2 \frac{1}{2} + I^2 \frac{1}{2} \cos 2\omega t \tag{122}$$

The average of $\cos 2\omega t$ is zero because it is positive half the time and negative half the time. Thus the average of i^2 is simply $I^2/2$.The square root of this is I_{rms}:

$$I_{rms} = \frac{I}{\sqrt{2}} \tag{123}$$

In the same way, the root-mean-square value of a sinusoidal voltage with amplitude U is

$$U_{rms} = \frac{U}{\sqrt{2}} \tag{124}$$

We can convert a rectifying ammeter into a voltmeter by adding a series resistor, just as for the dc case discussed early. Meters used for ac voltage and current measurements are nearly always calibrated to read rms values, not maximum or rectified average. Voltages and currents in power distribution systems are always described in terms of their rms values. The usual household power supply, "120-volt ac," has an rms voltage of 120 V (Fig. 60). The voltage amplitude is

$$U = \sqrt{2}U_{rms} = \sqrt{2}(120\ V) = 170\ V \tag{125}$$

Figure 60 - This wall socket delivers a root-mean- square voltage of 120 V. Sixty times per second, the instantaneous voltage across its terminals varies from $\sqrt{2}(120\ V) = 170\ V$ to $-170\ V$ and back again

1.5.2 Resistance and reactance

In this section we will derive voltage–current relationships for individual circuit elements carrying a sinusoidal current. We'll consider resistors, inductors, and capacitors.

Resistor in an ac Circuit. First let's consider a resistor with resistance through which there is a sinusoidal current given by: $i = I\cos\omega t$. The positive direction of current is counterclockwise around the circuit, as in Fig. 61a. The current amplitude (maximum current) is *I*. From Ohm's law the instantaneous potential u_R of point *a* with respect to point *b* (that is, the instantaneous voltage across the resistor) is

$$u_R = (iR) \tag{126}$$

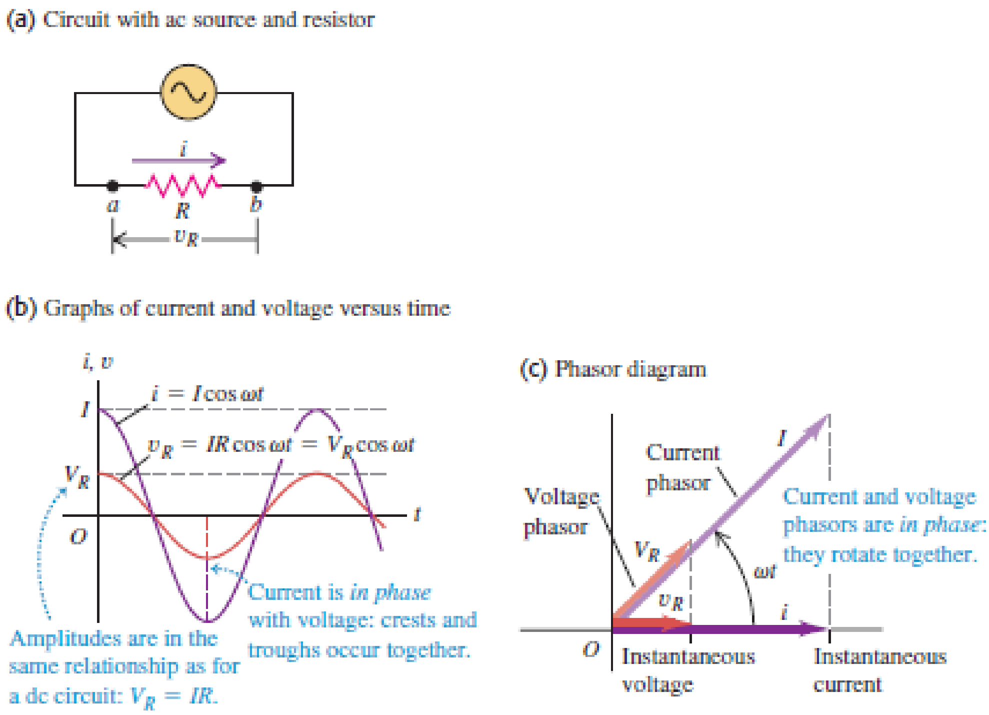
The maximum voltage U_R the *voltage amplitude,* is the coefficient of the cosine function:

$$U_R = IR \tag{127}$$

Hence we can also write

$$u_R = U_R \cos \omega t \tag{128}$$

The current and voltage are both proportional to so the current is *in phase* with the voltage. Equation (127) shows that the current and voltage amplitudes are related in the same way as in a dc circuit.


(a) Circuit with ac source and resistor

(b) Graphs of current and voltage versus time

(c) Phasor diagram

i
a
R
b
v_R
i, v
I
$i = I\cos\omega t$
$v_R = IR\cos\omega t = V_R\cos\omega t$
V_R
O
t
Current is *in phase* with voltage: crests and troughs occur together.
Amplitudes are in the same relationship as for a dc circuit: $V_R = IR$.
Current phasor
Voltage phasor
I
V_R
ωt
v_R
i
O
Current and voltage phasors are *in phase*: they rotate together.
Instantaneous voltage
Instantaneous current

Figure 61 – Resistance connected across an ac source

Figure 61b shows graphs of and as functions of time. The vertical scales for current and voltage are different, so the relative heights of the two curves are not significant. The corresponding phasor diagram is given in Fig. 61c. Because and are *in phase* and have the same frequency, the current and voltage phasors rotate together; they are parallel at each instant. Their projections on the horizontal axis represent the instantaneous current and voltage, respectively.

Inductor in an ac Circuit. Next, we replace the resistor in Fig. 61 with a pure inductor with self-inductance L and zero resistance (Fig. 62a). Again we assume that the current is $i = I\cos\omega t$ with the positive direction of current taken as counterclockwise around the circuit.

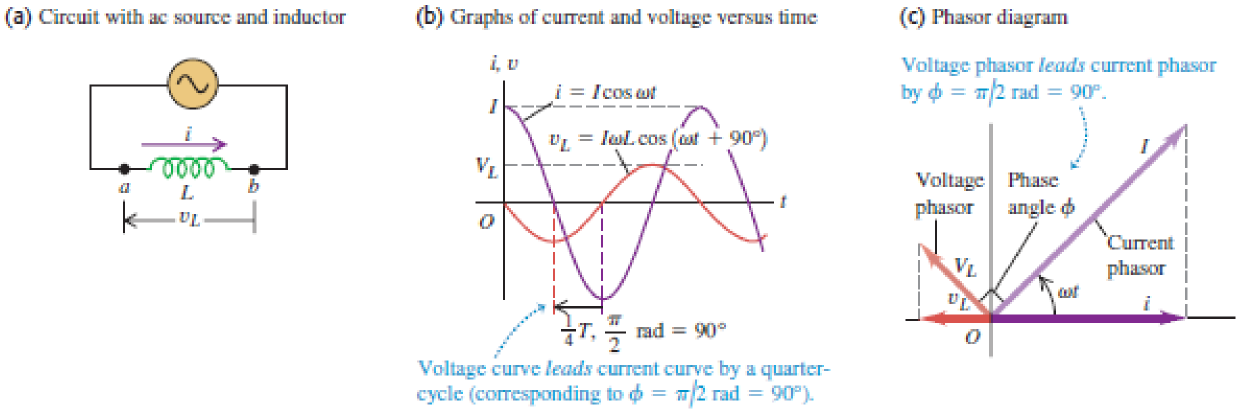

(a) Circuit with ac source and inductor
i
a
L
b
v_L
(b) Graphs of current and voltage versus time
i, v
I
$i = I\cos\omega t$
$v_L = I\omega L\cos(\omega t + 90°)$
V_L
O
t
$\frac{1}{4}T$, $\frac{\pi}{2}$ rad = 90°
Voltage curve *leads* current curve by a quarter-cycle (corresponding to $\phi = \pi/2$ rad = 90°).
(c) Phasor diagram
Voltage phasor *leads* current phasor by $\phi = \pi/2$ rad = 90°.
Voltage phasor
Phase angle ϕ
I
V_L
Current phasor
v_L
ωt
i
O

Figure 62– Inductance connected across an ac source

Although there is no resistance, there is a potential difference u_L between the inductor terminals a and b because the current varies with time, giving rise to a self-induced emf. The induced emf in the direction of is given by: $\mathcal{E} = -L\,di/dt$; however, the voltage is u_L *not* simply equal to $\mathcal{E}$. To see why, notice that if the current in the inductor is in the positive (counterclockwise) direction from a to b and is increasing, then di/dt is positive and the induced emf is directed to the left to oppose the increase in current; hence point a is at higher potential than is point b. Thus the potential of point a with respect to point b is positive and is given by $u_L = +L\frac{di}{dt}$, the *negative* of the induced emf. (You should convince yourself that this expression gives the correct sign of u_L in *all* cases, including i counterclockwise and decreasing, i clockwise and increasing, and i clockwise and decreasing). So we have

$$u_L = L\frac{di}{dt} = L\frac{d}{dt}(I\cos\omega t) = -I\omega L\sin\omega t \qquad (129)$$

The voltage u_L across the inductor at any instant is proportional to the *rate of change* of the current. The points of maximum voltage on the graph correspond to maximum steepness of the current curve, and the points of zero voltage are the points where the current curve instantaneously levels off at its maximum and minimum values (Fig. 62b). The voltage and current are “out of step” or *out of phase* by a quarter-cycle. Since the voltage peaks occur a quarter-cycle earlier than the current peaks, we say that the voltage *leads* the current by The phasor diagram in Fig. 62c also shows this relationship; the voltage phasor is ahead of the current phasor by 90°.

We can also obtain this phase relationship by rewriting Eq. (129) using the identity $\cos(A + 90) = -\sin A$:

$$u_L = I\omega L\cos(\omega t + 90) \qquad (130)$$

This result shows that the voltage can be viewed as a cosine function with a “head start” of 90° relative to the current.

As we have done in Eq. (130), we will usually describe the phase of the *voltage* relative to the *current*, not the reverse. Thus if the current i in a circuit is

$$i = I\cos\omega t \tag{131}$$

and the voltage of one point with respect to another is

$$u = U\cos(\omega t + \varphi) \tag{132}$$

we call φ the **phase angle;** it gives the phase of the *voltage* relative to the *current.* For a pure resistor, $\varphi = 0$, and for a pure inductor, $\varphi = 90°$.

From Eq. (129) or (130) the amplitude of the inductor voltage is

$$u_L = I\omega L \tag{133}$$

We define the **inductive reactance** X_L of an inductor as

$$X_L = \omega L \tag{134}$$

Using X_L, we can write Eq. (31.11) in a form similar to Eq. (127) for a resistor $(u_L = IX_L)$:

$$u_L = IX_L \tag{135}$$

Because X_L is the ratio of a voltage and a current, its SI unit is the ohm, the same as for resistance.

The inductive reactance X_L is really a description of the self-induced emf that opposes any change in the current through the inductor. From Eq. (135), for a given current amplitude I the voltage $u_R = +L\,di/dt$ across the inductor and the self-induced emf $\mathcal{E} = -L\,di/dt$ both have an amplitude U_L that is directly proportional to X_L. According to Eq. (134), the inductive reactance and self-induced emf increase with more rapid variation in current (that is, increasing angular frequency ω) and increasing inductance L.

If an oscillating voltage of a given amplitude U_R is applied across the inductor terminals, the resulting current will have a smaller amplitude I for larger values of X_L. Since X_L is proportional to frequency, a high-frequency voltage applied to the inductor gives only a small current, while a lower-frequency voltage of the same amplitude gives rise to a larger current. Inductors are used in some circuit applications, such as power supplies and radio-interference filters, to block high frequencies while permitting lower frequencies or dc to pass through. A circuit device that uses an inductor for this purpose is called a *low-pass filter*.

1.5.3 The L-R-C series circuit

Many ac circuits used in practical electronic systems involve resistance, inductive reactance, and capacitive reactance. Figure 62a shows a simple example: A series circuit containing a resistor, an inductor, a capacitor, and an ac source.

To analyze this and similar circuits, we will use a phasor diagram that includes the voltage and current phasors for each of the components. In this circuit, because of Kirchhoff's loop rule, the instantaneous *total* voltage u_{ab} across all three components is equal to the source voltage at that instant. We will show that the phasor representing this total voltage is the *vector sum* of the phasors for the individual voltages.

Figures 62b and 62c show complete phasor diagrams for the circuit of Fig. 62a. We assume that the source supplies a current i given by $i = I\cos\omega t$. Because the circuit elements are connected in series, the current at any instant is the same at every point in the circuit. Thus a *single phasor* I, with length proportional to the current amplitude, represents the current in *all* circuit elements.

We use the symbols $u_R, u_L,$ and u_C for the instantaneous voltages across $R, L,$ and C, and the symbols $U_R, U_L,$ and U_C for the maximum voltages. We denote the instantaneous and maximum *source* voltages by u and V. Then, in Fig. 62a, $u = u_{ad}, u_R = u_{ab}, u_L = u_{bc},$ and $u_C = u_{cd}$.

We have shown that the potential difference between the terminals of a resistor is *in phase* with the current in the resistor and that its maximum value U_R is given by Eq. (127):

$$U_R = IR \tag{136}$$

The phasor U_R in Fig. 62b, in phase with the current phasor I represents the voltage across the resistor. Its projection onto the horizontal axis at any instant gives the instantaneous potential difference u_R.

The voltage across an inductor *leads* the current by Its voltage amplitude is given by Eq. (135):

$$U_L = IX_L \tag{137}$$

The phasor U_L in Fig. 61b represents the voltage across the inductor, and its projection onto the horizontal axis at any instant equals u_L.

The voltage across a capacitor *lags* the current by 90. Its voltage amplitude is given by Eq.:

$$U_C = IX_C \tag{138}$$

The phasor U_C in Fig. 61b represents the voltage across the capacitor, and its projection onto the horizontal axis at any instant equals u_C.

The instantaneous potential difference between terminals a and d is equal at every instant to the (algebraic) sum of the potential differences u_R, u_L and u_C. That is, it equals the sum of the *projections* of the phasors U_R, U_L and U_C. But the sum of the projections of these phasors is equal to the *projection* of their *vector sum.* So the vector sum must be the phasor that represents the source voltage u and the instantaneous total voltage u_{ad} across the series of elements.

To form this vector sum, we first subtract the phasor U_C from the phasor U_L. (These two phasors always lie along the same line, with opposite directions.) This gives the phasor $U_L - U_C$. This is always at right angles to the phasor U_R so from the Pythagorean theorem the magnitude of the phasor U is

$$U = \sqrt{U_R^2 + (U_L - U_C)^2} = \sqrt{(IR)^2 + (IX_L - IX_C)^2} \qquad (139)$$

or

$$U = I\sqrt{R^2 + (X_L - X_C)^2}$$

We define the **impedance** Z of an ac circuit as the ratio of the voltage amplitude across the circuit to the current amplitude in the circuit. From Eq. (139) the impedance of the *L-R-C* series circuit is

$$Z = \sqrt{R^2 + (X_L - X_C)^2} \qquad (140)$$

so we can rewrite Eq. (139) as

$$U = IZ \qquad (141)$$

While Eq. (140) is valid only for an *L-R-C* series circuit, we can use Eq. (141) to define the impedance of *any* network of resistors, inductors, and capacitors as the ratio of the amplitude of the voltage across the network to the current amplitude. The SI unit of impedance is the ohm.

Equation (141) has a form similar to $U = IR$ with impedance Z in an ac circuit playing the role of resistance R in a dc circuit. Just as direct current tends to follow the path of least resistance, so alternating current tends to follow the path of lowest impedance. Note, however, that impedance is actually a function of $R, L,$ and C as well as of the angular frequency We can see this by substituting Eq. (134) for X_L and for X_C into Eq. (140), giving the following complete expression for Z for a series circuit:

$$Z = \sqrt{R^2 + (X_L - X_C)^2} = \sqrt{R^2 + [\omega L - 1/\omega C]^2} \qquad (142)$$

Hence for a given amplitude U of the source voltage applied to the circuit, the amplitude $I = U/Z$ of the resulting current will be different at different frequencies.

We'll explore this frequency dependence early. In the phasor diagram shown in Fig. 62b, the angle φ between the voltage and current phasors is the phase angle of the source voltage with respect to the current i; that is, it is the angle by which the source voltage leads the current.
From the diagram,

$$\tan\varphi = \frac{\omega L - 1/\omega C}{R} \tag{143}$$

If the current is $i = I\cos\omega t$ then the source voltage u is

$$u = U\cos(\omega t + \varphi) \tag{144}$$

Figure 62b shows the behavior of a circuit in which $X_L > X_C$. Figure 62c shows the behavior when $X_L < X_C$; the voltage phasor lies on the opposite side of the current phasor I and the voltage *lags* the current. In this case, $X_L - X_C$ is *negative,* $\tan\varphi$ is negative, and φ is a negative angle between 0 and 90. Since X_L and X_C depend on frequency, the phase angle depends on frequency as well.

All of the expressions that we've developed for an *L-R-C* series circuit are still valid if one of the circuit elements is missing. If the resistor is missing, we set $R = 0$; if the inductor is missing, we set $L = 0$. But if the capacitor is missing, we set $C = \infty$, corresponding to the absence of any potential difference $(u_C = q/C = 0)$ or any capacitive reactance $(X_L = 1/\omega C = 0)$.

In this entire discussion we have described magnitudes of voltages and currents in terms of their *maximum* values, the voltage and current *amplitudes.* But we remarked that these quantities are usually described in terms of rms values, not amplitudes. For any sinusoidally varying quantity, the rms value is always $1/\sqrt{2}$ times the amplitude. All the relationships between voltage and current that we have derived in this and the preceding sections are still valid if we use rms quantities throughout instead of amplitudes. For example, if we divide Eq. (141) by $\sqrt{2}$ we get

$$\frac{U}{\sqrt{2}} = \frac{I}{\sqrt{2}}Z \tag{145}$$

which we can rewrite as

$$U_{rms} = I_{rms}Z \tag{146}$$

We can translate Eqs. (127), (135), and (138) in exactly the same way.

We have considered only ac circuits in which an inductor, a resistor, and a capacitor are in series. You can do a similar analysis for an *L-R-C parallel* circuit; see.

Finally, we remark that in this section we have been describing the *steadystate* condition of a circuit, the state that exists after the circuit has been connected to the source for a long time. When the source is first connected, there may be additional voltages and currents, called *transients,* whose nature depends on the time in the cycle when the circuit is initially completed. A detailed analysis of transients is beyond our scope. They always die out after a sufficiently long time, and they do not affect the steady-state behavior of the circuit. But they can cause dangerous and damaging surges in power lines, which is why delicate electronic systems such as computers are often provided with power-line surge protectors.

1.5.4 Power in alternating current

Alternating currents play a central role in systems for distributing, converting, and using electrical energy, so it's important to look at power relationships in ac circuits. For an ac circuit with instantaneous current and current i amplitude I, we'll consider an element of that circuit across which the instantaneous potential difference is with voltage amplitude U. The instantaneous power p delivered to this circuit element is

$$p = ui \tag{147}$$

Let's first see what this means for individual circuit elements. We'll assume in each case that $i = I\cos\omega t$.

Power in a Resistor. Suppose first that the circuit element is a *pure resistor R* as in Fig. 61a; then $u = u_R$ and i are *in phase.*We obtain the graph representing p by multiplying the heights of the graphs of u and i in Fig. 61b at each instant. This graph is shown by the black curve in Fig. 63a. The product ui is always positive because u and i are always either both positive or both negative. Hence energy is supplied *to* the resistor at every instant for both directions of i, although the power is not constant.

The power curve for a pure resistor is symmetrical about a value equal to one-half its maximum value UI, so the *average power* P_{av} is

$$P_{av} = \frac{1}{2}UI \tag{148}$$

An equivalent expression is

$$P_{av} = \frac{U}{\sqrt{2}}\frac{I}{\sqrt{2}} = U_{rms}I_{rms} \tag{149}$$

Also, $U_{rms} = I_{rms}R$, so we can express by any of the equivalent forms

$$P_{av} = I_{rms}^2R = \frac{U_{rms}^2}{R} = U_{rms}I_{rms} \tag{150}$$

Note that the expressions in Eq. (150) have the same form as the corresponding relationships for a dc circuit. Also note that they are valid only for pure resistors, not for more complicated combinations of circuit elements.

Power in an Inductor. Next we connect the source to a pure inductor L, as in Fig. 61a. The voltage $u = u_L$ leads the current i by 90. When we multiply the curves of u and i the product ui is *negative* during the half of the cycle when u and i have *opposite* signs. The power curve, shown in Fig. 63b, is symmetrical about the horizontal axis; it is positive half the time and negative the other half, and the average power is zero. When p is positive, energy is being supplied to set up the magnetic field in the inductor; when is negative, the field is collapsing and the inductor is returning energy to the source. The net energy transfer over one cycle is zero.

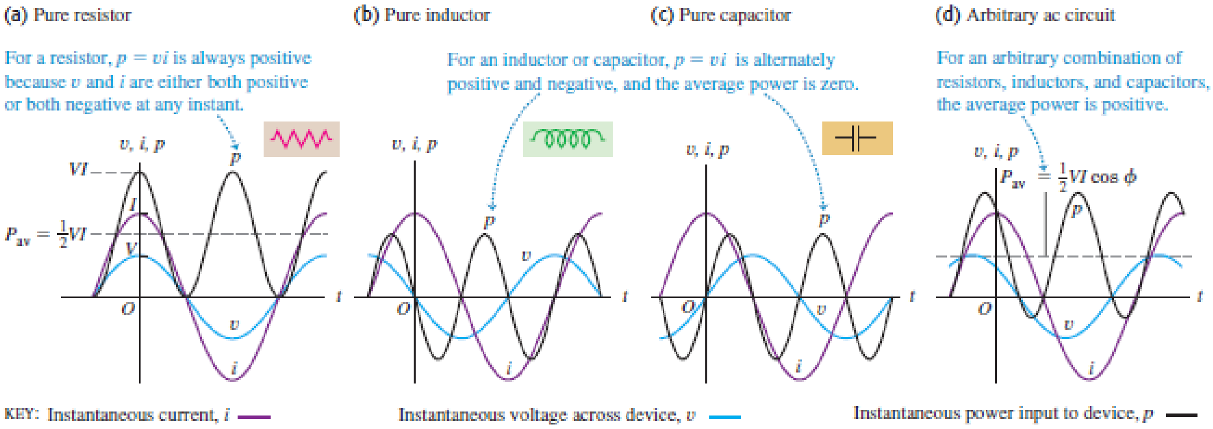


Figure 63 - Graphs of current, voltage, and power as function for (a) a pure resistor, (b) a pure inductor, (c) a pure capacitor, and (d) an arbitrary ac circuit that can have resistance, inductance, and capacitor

Power in a Capacitor. Finally, we connect the source to a pure capacitor as in Fig. 62a. The voltage $u = u_c$ lags the current by Figure 63c shows the power curve; the average power is again zero. Energy is supplied to charge the capacitor and is returned to the source when the capacitor discharges. The net energy transfer over one cycle is again zero.

Power in a General ac Circuit. In *any* ac circuit, with any combination of resistors, capacitors, and inductors, the voltage u across the entire circuit has some phase angle φ with respect to the current i. Then the instantaneous power is given by

$$p = ui = [U\cos(\omega t + \varphi)][I\cos\omega t] \tag{151}$$

The instantaneous power curve has the form shown in Fig. 63d. The area between the positive loops and the horizontal axis is greater than the area between the negative loops and the horizontal axis, and the average power is positive.

We can derive from Eq. (147) an expression for the *average* power P_{av} by using the identity for the cosine of the sum of two angles:

$$p = [U(\cos\omega t\cos\varphi - \sin\omega t\sin\varphi)][I\cos\omega t] = \\ = UI\cos\varphi\cos^2\omega t - UI\sin\varphi\cos\omega t\sin\omega t \tag{152}$$

From the discussion that led to Eq. (123), we see that the average value of $\cos^2\omega t$ (over one cycle) is $\frac{1}{2}$. The average value of $\cos\omega t\sin\omega t$ is zero because this product is equal to $\frac{1}{2}\sin 2\omega t$, whose average over a cycle is zero. So the average power P_{av} is

$$P_{av} = \frac{1}{2}UI\cos\varphi = U_{rms}I_{rms}\cos\varphi \tag{153}$$

When u and i are in phase, so $\varphi = 0$, the average power equals $\frac{1}{2}UI = U_{rms}I_{rms}$; when u and i are 90° out of phase, the average power is zero. In the general case, when u has a phase angle φ with respect to i, the average power equals $\frac{1}{2}I$ multiplied by $U\cos\varphi$, the component of the voltage phasor that is *in phase* with the current phasor. Figure 64 shows the general relationship of the current and voltage phasors. For the *L*-*R*-*C* series circuit, Figs. 62b and 62c show that $U\cos\varphi$ equals the voltage amplitude U_R for the resistor; hence Eq. (153) is the average power dissipated in the resistor. On average there is no energy flow into or out of the inductor or capacitor, so none of P_{av} goes into either of these circuit elements.

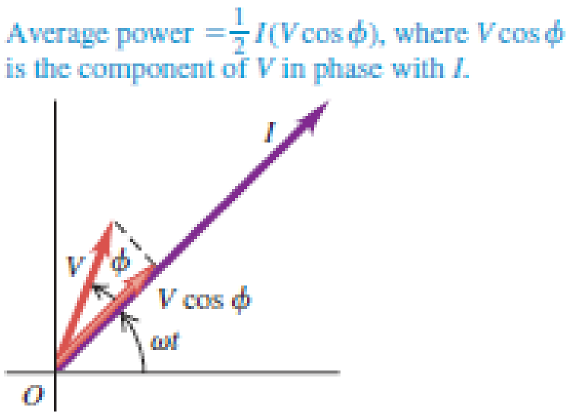

Average power = ½ I(V cos ϕ), where V cos ϕ
is the component of V in phase with I.
I
V
ϕ
V cos ϕ
ωt
O

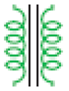
Figure 64 – Using phasors to calculate average power for an arbitary ac circuit

The factor $\cos\varphi$ is called the **power factor** of the circuit. For a pure resistance, $\varphi = 0, \cos\varphi = 1,$ and $P_{av} = U_{rms}I_{rms}$. For a pure inductor or capacitor, $\varphi = \pm 90°, \cos\varphi = 0$, and $P_{av} = 0$. For an *L*-*R*-*C* series circuit the power factor is equal to R/Z we leave the proof of this statement to you.

Alow power factor (large angle φ of lag or lead) is usually undesirable in power circuits. The reason is that for a given potential difference, a large current is needed to supply a given amount of power. This results in large i^2R losses in the transmission lines. Your electric power company may charge a higher rate to a client with a low power factor. Many types of ac machinery draw a *lagging* current; that is, the current drawn by the machinery lags the applied voltage. Hence the voltage leads the current, so $\varphi > 0$ and $\cos\varphi < 1$. The power factor can be corrected toward the ideal value of 1 by connecting a capacitor in parallel with the load. The current drawn by the capacitor *leads* the voltage (that is, the voltage across the capacitor lags the current), which compensates for the lagging current in the other branch of the circuit. The capacitor itself absorbs no net power from the line.

1.5.5 Transformers

One of the great advantages of ac over dc for electric-power distribution is that it is much easier to step voltage levels up and down with ac than with dc. For longdistance power transmission it is desirable to use as high a voltage and as small a current as possible; this reduces i^2R losses in the transmission lines, and smaller wires can be used, saving on material costs. Present-day transmission lines routinely operate at rms voltages of the order of 500 kV. On the other hand, safety considerations and insulation requirements dictate relatively low voltages in generating equipment and in household and industrial power distribution. The standard voltage for household wiring is 120 V in the United States and Canada and 240 V in many other countries. The necessary voltage conversion is accomplished by the use of **transformers.**

How Transformers Work. Figure 65 shows an idealized transformer. The key components of the transformer are two coils or *windings,* electrically insulated from each other but wound on the same core. The core is typically made of a material, such as iron, with a very large relative permeability K_m. This keeps the magnetic field lines due to a current in one winding almost completely within the core. Hence almost all of these field lines pass through the other winding, maximizing the *mutual inductance* of the two windings. The winding to which power is supplied is called the **primary;** the winding from which power is delivered is called the **secondary.** The circuit symbol for a transformer with an iron core, such as those used in power distribution systems, is 

Here's how a transformer works. The ac source causes an alternating current in the primary, which sets up an alternating flux in the core; this induces an emf in each winding, in accordance with Faraday's law. The induced emf in the secondary gives rise to an alternating current in the secondary, and this delivers energy to the device to which the secondary is connected. All currents and emfs have the same frequency as the ac source.

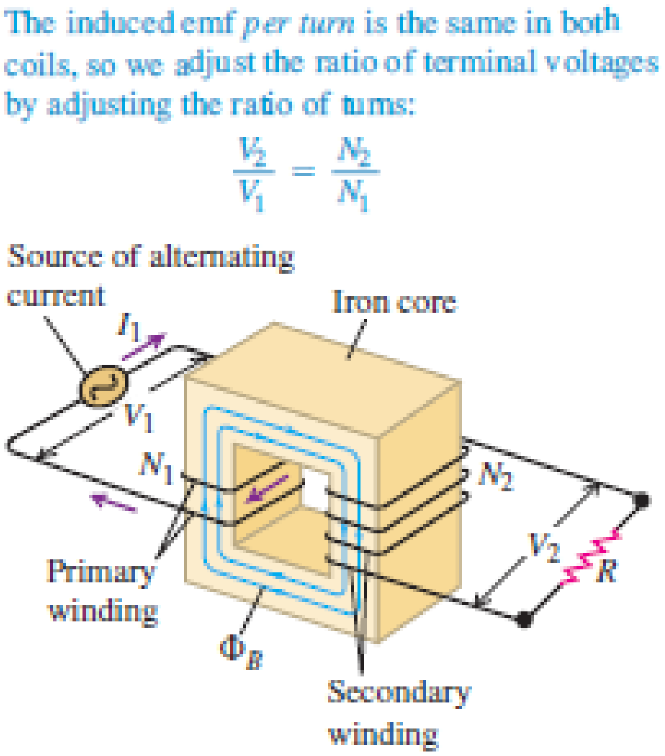

The induced emf *per turn* is the same in both coils, so we adjust the ratio of terminal voltages by adjusting the ratio of turns:
$\frac{V_2}{V_1} = \frac{N_2}{N_1}$
Source of alternating current
I_1
Iron core
V_1
N_1
N_2
V_2
R
Primary winding
Φ_B
Secondary winding

Figure 65 – Schematic diagram of an idealized step-up transformer. The primary is connected to an ac source; the secondary is connected to a device with resistance *R*

Let’s see how the voltage across the secondary can be made larger or smaller in amplitude than the voltage across the primary. We neglect the resistance of the windings and assume that all the magnetic field lines are confined to the iron core, so at any instant the magnetic flux Φ_B is the same in each turn of the primary and secondary windings. The primary winding has turns and the secondary winding has turns. When the magnetic flux changes because of changing currents in the two coils, the resulting induced emfs are

$$\mathcal{E}_1 = -N_1 \frac{d\Phi_B}{dt} \tag{154}$$

and

$$\mathcal{E}_2 = -N_2 \frac{d\Phi_B}{dt}$$

The flux *per turn* Φ_B is the same in both the primary and the secondary, so Eqs. (154) show that the induced emf *per turn* is the same in each. The ratio of the secondary emf $\mathcal{E}_2$ to the primary emf is $\mathcal{E}_1$ therefore equal at any instant to the ratio of secondary to primary turns:

$$\frac{\mathcal{E}_2}{\mathcal{E}_1} = \frac{N_2}{N_1} \tag{155}$$

Since $\mathcal{E}_1$ and $\mathcal{E}_2$ both oscillate with the same frequency as the ac source, Eq. (155) also gives the ratio of the amplitudes or of the rms values of the induced emfs. If the windings have zero resistance, the induced emfs $\mathcal{E}_1$ and $\mathcal{E}_2$ are equal to the terminal voltages across the primary and the secondary, respectively; hence

$$\frac{U_2}{U_1} = \frac{N_2}{N_1} \qquad (156)$$

where U_1 and U_2 are either the amplitudes or the rms values of the terminal voltages. By choosing the appropriate turns ratio $\frac{U_2}{U_1}$ we may obtain any desired secondary voltage from a given primary voltage. If $N_2 > N_1$ as in Fig. 65, then $U_2 > U_1$ and we have a *step-up* transformer; if $N_2 < N_1$, then $U_2 < U_1$ and we have a *step-down* transformer. At a power generating station, step-up transformers are used; the primary is connected to the power source and the secondary is connected to the transmission lines, giving the desired high voltage for transmission. Near the consumer, step-down transformers lower the voltage to a value suitable for use in home or industry (Fig. 66).

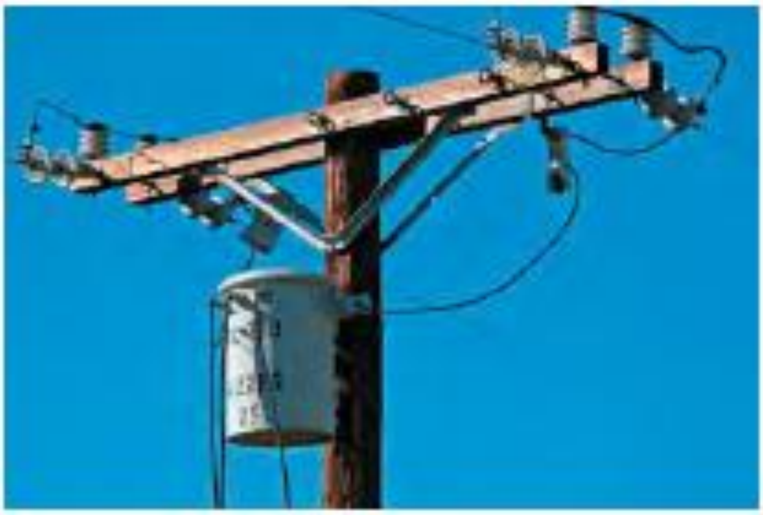


Figure 66 – The cylindrical can near the top of this power pole is a step-down transformer. It converts the high-voltage ac in the power lines to low-voltage (120 V) ac, which is then distributed to the surrounding homes and businesses

Even the relatively low voltage provided by a household wall socket is too high for many electronic devices, so a further step-down transformer is necessary. This is the role of an “ac adapter” such as those used to recharge a mobile phone or laptop computer from line voltage. Such adapters contain a step-down transformer that converts line voltage to a lower value, typically 3 to 12 volts, as well as diodes to convert alternating current to the direct current that small electronic devices require (Fig. 67).

Figure 67 - An ac adapter like this one converts household ac into low-voltage dc for use in electronic devices. It contains a step-down transformer to lower the voltage and diodes to rectify the output current

Energy Considerations for Transformers. If the secondary circuit is completed by a resistance then the amplitude or rms value of the current in the secondary circuit is $I_2 = U_2/R$. From energy considerations, the power delivered to the primary equals that taken out of the secondary (since there is no resistance in the windings), so

$$I_1 U_1 = I_2 U_2 \tag{157}$$

We can combine Eqs. (156) and (157) and the relationship $I_2 = U_2/R$ to eliminate U_2 and I_2 we obtain

$$\frac{U_1}{I_1} = \frac{R}{(N_2/N_1)^2} \tag{158}$$

This shows that when the secondary circuit is completed through a resistance R, the result is the same as if the *source* had been connected directly to a resistance equal to R divided by the square of the turns ratio, $(N_2/N_1)^2$. In other words, the transformer "transforms" not only voltages and currents, but resistances as well. More generally, we can regard a transformer as "transforming" the *impedance* of the network to which the secondary circuit is completed.

Equation (158) has many practical consequences. The power supplied by a source to a resistor depends on the resistances of both the resistor and the source. It can be shown that the power transfer is greatest when the two resistances are *equal.* The same principle applies in both dc and ac circuits. When a high-impedance ac source must be connected to a low-impedance circuit, such as an audio amplifier connected to a loudspeaker, the source impedance can be *matched* to that of the circuit by the use of a transformer with an appropriate turns ratio N_2/N_1.

Real transformers always have some energy losses. (That's why an ac adapter like the one shown in Fig. 67 feels warm to the touch after it's been in use for a while; the transformer is heated by the dissipated energy.) The windings have some resistance, leading to i^2R losses. There are also energy losses through hysteresis in the core. Hysteresis losses are minimized by the use of soft iron with a narrow hysteresis loop.

Another important mechanism for energy loss in a transformer core involves eddy currents. Consider a section *AA* through an iron transformer core (Fig. 68a). Since iron is a conductor, any such section can be pictured as several conducting circuits, one within the other (Fig. 64b). The flux through each of these circuits is continually changing, so eddy currents circulate in the entire volume of the core, with lines of flow that form planes perpendicular to the flux. These eddy currents are very undesirable; they waste energy through i^2R heating and themselves set up an opposing flux.

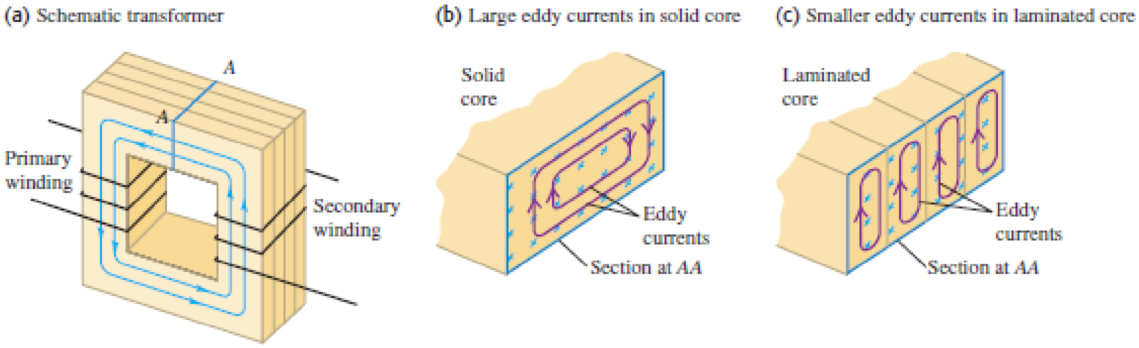


Figure 68 - (a) Primary and secondary windings in a transformer. (b) Eddy currents in the iron core, shown in the cross section at *AA*. (c) Using a laminated core reduces the eddy currents

The effects of eddy currents can be minimized by the use of a *laminated* core—that is, one built up of thin sheets or laminae. The large electrical surface resistance of each lamina, due either to a natural coating of oxide or to an insulating varnish, effectively confines the eddy currents to individual laminae (Fig. 68c). The possible eddy-current paths are narrower, the induced emf in each path is smaller, and the eddy currents are greatly reduced. The alternating magnetic field exerts forces on the current-carrying laminae that cause them to vibrate back and forth; this vibration causes the characteristic "hum" of an operating transformer. You can hear this same "hum" from the magnetic ballast of a fluorescent light fixture. Thanks to the use of soft iron cores and lamination, transformer efficiencies are usually well over 90%; in large installations they may reach 99%.

Discussion questions

1. Household electric power in most of western Europe is supplied at 240 V, rather than the 120 V that is standard in the United States and Canada. What are the advantages and disadvantages of each system?

2. The current in an ac power line changes direction 120 times per second, and its average value is zero. Explain how it is possible for power to be transmitted in such a system.
3. In an ac circuit, why is the average power for an inductor and a capacitor zero, but not for a resistor?
4. Fluorescent lights often use an inductor, called a ballast, to limit the current through the tubes. Why is it better to use an inductor rather than a resistor for this purpose?
5. Is it possible for the power factor of an *L-R-C* series ac circuit to be zero? Justify your answer on *physical* grounds.
6. In an *L-R-C* series circuit, can the instantaneous voltage across the capacitor exceed the source voltage at that same instant? Can this be true for the instantaneous voltage across the inductor? Across the resistor? Explain.
7. In an *L-R-C* series circuit, what are the phase angle and power factor $\cos\varphi$ when the resistance is much smaller than the inductive or capacitive reactance and the circuit is operated far from resonance? Explain.
8. When an *L-R-C* series circuit is connected across a 120-V ac line, the voltage rating of the capacitor may be exceeded even if it is rated at 200 or 400 V. How can this be?
9. A light bulb and a parallel-plate capacitor with air between the plates are connected in series to an ac source. What happens to the brightness of the bulb when a dielectric is inserted between the plates of the capacitor? Explain.
10. A coil of wire wrapped on a hollow tube and a light bulb are connected in series to an ac source. What happens to the brightness of the bulb when an iron rod is inserted in the tube?
11. A circuit consists of a light bulb, a capacitor, and an inductor connected in series to an ac source. What happens to the brightness of the bulb when the inductor is removed? When the inductor is left in the circuit but the capacitor is removed? Explain.
12. A circuit consists of a light bulb, a capacitor, and an inductor connected in series to an ac source. Is it possible for both the capacitor and the inductor to be removed and the brightness of the bulb to remain the same? Explain.
13. Can a transformer be used with dc? Explain. What happens if a transformer designed for 120-V ac is connected to a120-V dc line?
14. An ideal transformer has N_1 windings in the primary and N_2 windings in its secondary. If you double only the number of secondary windings, by what factor does (a) the voltage amplitude in the secondary change, and (b) the effective resistance of the secondary circuit change?
15. Some electrical appliances operate equally well on ac or dc, and others work only on ac or only on dc. Give examples of each, and explain the differences.

Topic 2 Optics

2.1 Geometrical optics

2.1.1 The nature of the light

Until the time of Isaac Newton (1642–1727), most scientists thought that light consisted of streams of particles (called *corpuscles*) emitted by light sources. Galileo and others tried (unsuccessfully) to measure the speed of light. Around 1665, evidence of *wave* properties of light began to be discovered. By the early 19th century, evidence that light is a wave had grown very persuasive.

In 1873, James Clerk Maxwell predicted the existence of electromagnetic waves and calculated their speed of propagation. This development, along with the experimental work of Heinrich Hertz starting in 1887, showed conclusively that light is indeed an electromagnetic wave.

The Two Personalities of Light. The wave picture of light is not the whole story, however. Several effects associated with emission and absorption of light reveal a particle aspect, in that the energy carried by light waves is packaged in discrete bundles called *photons* or *quanta.* These apparently contradictory wave and particle properties have been reconciled since 1930 with the development of quantum electrodynamics, a comprehensive theory that includes *both* wave and particle properties. The *propagation* of light is best described by a wave model, but understanding emission and absorption requires a particle approach.

Figure 69 - An electric heating element emits primarily infrared radiation. But if its temperature is high enough, it also emits a discernible amount of visible light

The fundamental sources of all electromagnetic radiation are electric charges in accelerated motion. All bodies emit electromagnetic radiation as a result of thermal motion of their molecules; this radiation, called *thermal radiation,* is a mixture of different wavelengths. At sufficiently high temperatures, all matter emits enough visible light to be self-luminous; a very hot body appears “red-hot” (Fig. 69) or “white-hot.” Thus hot matter in any form is a light source. Familiar examples are a candle flame, hot coals in a campfire, the coils in an electric room heater, and an incandescent lamp filament (which usually operates at a temperature of about 3000°C).

Light is also produced during electrical discharges through ionized gases. The bluish light of mercury-arc lamps, the orange-yellow of sodium-vapor lamps, and the various colors of “neon” signs are familiar. A variation of the mercury-arc lamp is the *fluorescent* lamp. This light source uses a material called a *phosphor* to convert the ultraviolet radiation from a mercury arc into visible light. This conversion makes fluorescent lamps more efficient than incandescent lamps in transforming electrical energy into light.

In most light sources, light is emitted independently by different atoms within the source; in a *laser,* by contrast, atoms are induced to emit light in a cooperative, coherent fashion. The result is a very narrow beam of radiation that can be enormously intense and that is much more nearly *monochromatic,* or singlefrequency, than light from any other source. Lasers are used by physicians for microsurgery, in a DVD or Blu-ray player to scan the information recorded on a video disc, in industry to cut through steel and to fuse high-melting-point materials, and in many other applications (Fig. 70).

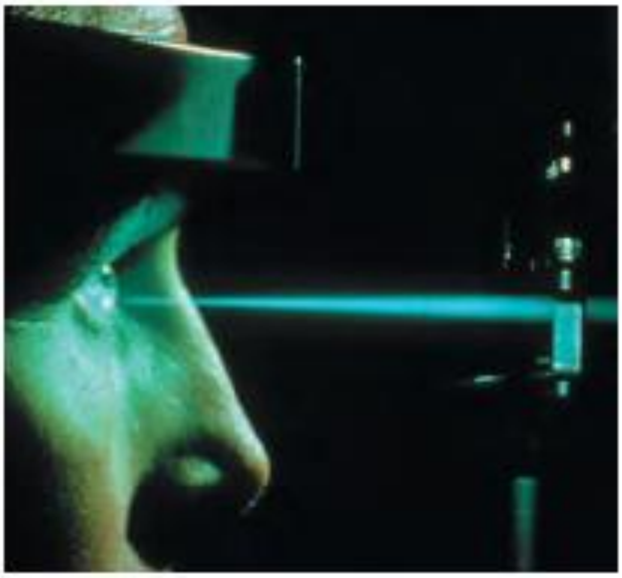


Figure 70 - Ophthalmic surgeons use lasers for repairing detached retinas and for cauterizing blood vessels in retinopathy. Pulses of blue-green light from an argon laser are ideal for this purpose, since they pass harmlessly through the transparent part of the eye but are absorbed by red pigments in the retina

No matter what its source, electromagnetic radiation travels in vacuum at the same speed. The speed of light in vacuum is defined to be

$$c = 2.99792458 \times 10^8 \ m/s$$

or $3 \times 10^8\ m/s$ to three significant figures. The duration of one second is defined by the cesium clock, so one meter is defined to be the distance that light travels in $1/299792458\ s$.

We often use the concept of a **wave front** to describe wave propagation. More generally, we define a wave front as *the locus of all adjacent points at which the phase of vibration of a physical quantity associated with the wave is the same.* That is, at any instant, all points on a wave front are at the same part of the cycle of their variation.

When we drop a pebble into a calm pool, the expanding circles formed by the wave crests, as well as the circles formed by the wave troughs between them, are wave fronts. Similarly, when sound waves spread out in still air from a pointlike source, or when electromagnetic radiation spreads out from a pointlike emitter, any spherical surface that is concentric with the source is a wave front, as shown in Fig. 71. In diagrams of wave motion we usually draw only parts of a few wave fronts, often choosing consecutive wave fronts that have the same phase and thus are one wavelength apart, such as crests of water waves. Similarly, a diagram for sound waves might show only the "pressure crests," the surfaces over which the pressure is maximum, and a diagram for electromagnetic waves might show only the "crests" on which the electric or magnetic field is maximum.

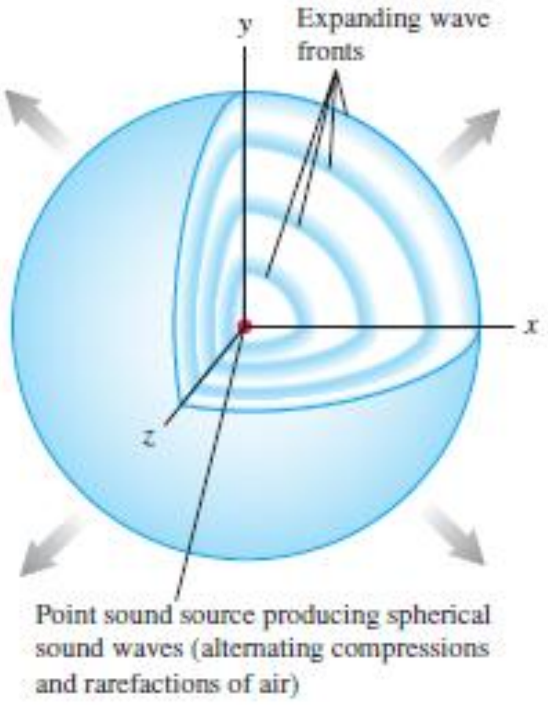

y
Expanding wave fronts
x
z
Point sound source producing spherical sound waves (alternating compressions and rarefactions of air)

Figure 71 - Spherical wave fronts of sound spread out uniformly in all directions from a point source in a motionless medium, such as still air, that has the same properties in all regions and in all directions. Electromagnetic waves in vacuum also spread out as shown here

We will often use diagrams that show the shapes of the wave fronts or their cross sections in some reference plane. For example, when electromagnetic waves are radiated by a small light source, we can represent the wave fronts as spherical surfaces concentric with the source or, as in Fig. 72a, by the circular intersections of these surfaces with the plane of the diagram. Far away from the source, where the radii of the spheres have become very large, a section of a spherical surface can be considered as a plane, and we have a *plane* wave like those discussed early (Fig, 72b).

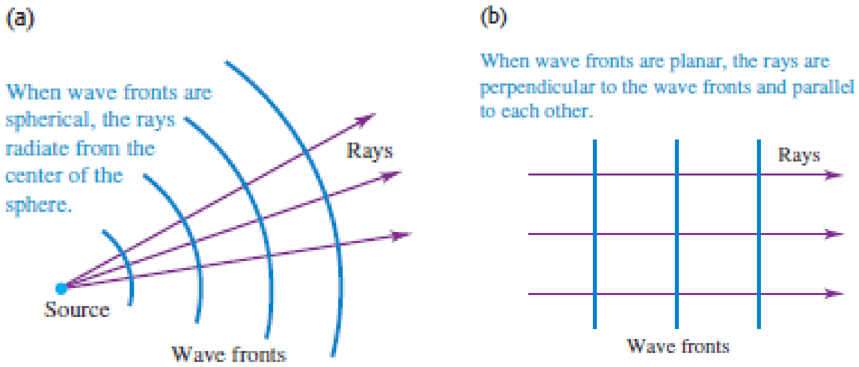

(a)
When wave fronts are spherical, the rays radiate from the center of the sphere.
Rays
Source
Wave fronts
(b)
When wave fronts are planar, the rays are perpendicular to the wave fronts and parallel to each other.
Rays
Wave fronts

Figure 72 – Wave fronts (blue) and rays (purple)

To describe the directions in which light propagates, it's often convenient to represent a light wave by **rays** rather than by wave fronts. Rays were used to describe light long before its wave nature was firmly established. In a particle theory of light, rays are the paths of the particles. From the wave viewpoint *a ray is an imaginary line along the direction of travel of the wave.* In Fig. 72a the rays are the radii of the spherical wave fronts, and in Fig. 72b they are straight lines perpendicular to the wave fronts. When waves travel in a homogeneous isotropic material (a material with the same properties in all regions and in all directions), the rays are always straight lines normal to the wave fronts. At a boundary surface between two materials, such as the surface of a glass plate in air, the wave speed and the direction of a ray may change, but the ray segments in the air and in the glass are straight lines.

The next several chapters will give you many opportunities to see the interplay of the ray, wave, and particle descriptions of light. The branch of optics for which the ray description is adequate is called **geometric optics;** the branch dealing specifically with wave behavior is called **physical optics.** This chapter and the following one are concerned mostly with geometric optics.

2.1.2 Reflection and refraction

In this section we'll use the *ray* model of light to explore two of the most important aspects of light propagation: **reflection** and **refraction.** When a light wave strikes a smooth interface separating two transparent materials (such as air and glass or water and glass), the wave is in general partly *reflected* and partly *refracted* (transmitted) into the second material, as shown in Fig. 73a. For example, when you look into a restaurant window from the street, you see a reflection of the street scene, but a person inside the restaurant can look out through the window at the same scene as light reaches him by refraction.

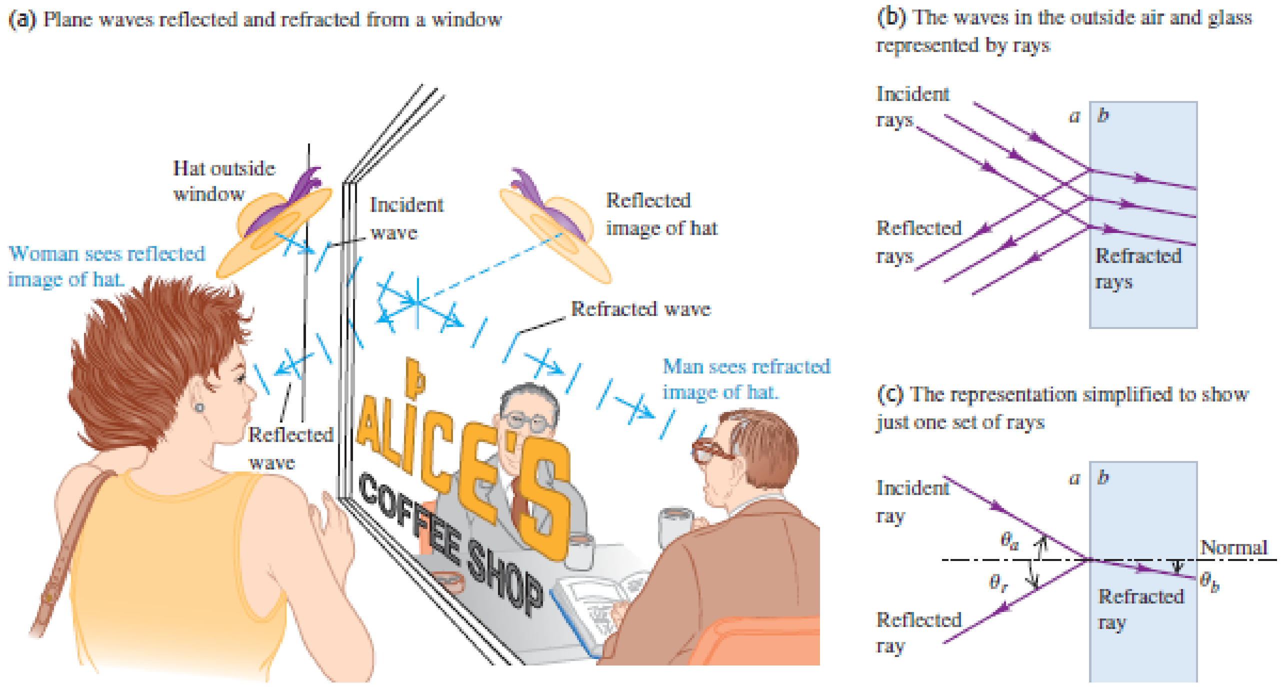

(a) Plane waves reflected and refracted from a window
Hat outside window
Woman sees reflected image of hat.
Incident wave
Reflected image of hat
Refracted wave
Man sees refracted image of hat.
Reflected wave
ALICE'S COFFEE SHOP
(b) The waves in the outside air and glass represented by rays
Incident rays
Reflected rays
Refracted rays
a
b
(c) The representation simplified to show just one set of rays
Incident ray
Reflected ray
Refracted ray
Normal
θ_a
θ_r
θ_b
a
b

Figure 73 - (a) A plane wave is in part reflected and in part refracted at the boundary between two media (in this case, air and glass). The light that reaches the inside of the coffee shop is refracted twice, once entering the glass and once exiting the glass. (b), (c) How light behaves at the interface between the air outside the coffee shop (material *a*) and the glass (material *b*). For the case shown here, material *b* has a larger index of refraction than material *a* ($n_b > n_a$) and the angle θ_b is smaller than θ_a

The segments of plane waves shown in Fig. 73a can be represented by bundles of rays forming *beams* of light (Fig. 73b). For simplicity we often draw only one ray in each beam (Fig. 73c). Representing these waves in terms of rays is the basis of geometric optics. We begin our study with the behavior of an individual ray.

We describe the directions of the incident, reflected, and refracted (transmitted) rays at a smooth interface between two optical materials in terms of the angles they make with the *normal* (perpendicular) to the surface at the point of incidence, as shown in Fig. 73c. If the interface is rough, both the transmitted light and the reflected light are scattered in various directions, and there is no single angle of transmission or reflection. Reflection at a definite angle from a very smooth surface is called **specular reflection** (from the Latin word for "mirror"); scattered reflection from a rough surface is called **diffuse reflection.** This distinction is shown in Fig. 74. Both kinds of reflection can occur with either transparent materials or *opaque* materials that do not transmit light. The vast majority of objects in your environment (including plants, other people, and this book) are visible to you because they reflect light in a diffuse manner from their surfaces. Our primary concern, however, will be with specular reflection from a very smooth surface such as highly polished glass or metal. Unless stated otherwise, when referring to "reflection" we will always mean *specular* reflection.

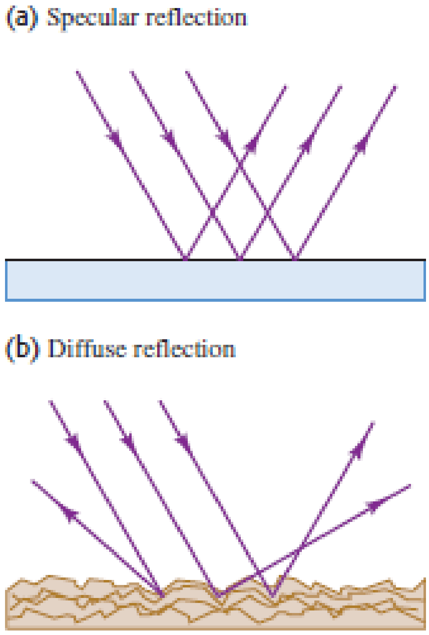


Figure 74 - Two types of reflection

The **index of refraction** of an optical material (also called the **refractive index**), denoted by plays a central role in geometric optics. It is the ratio of the speed of light in vacuum to the speed in the material:

$$n = \frac{c}{v} \tag{159}$$

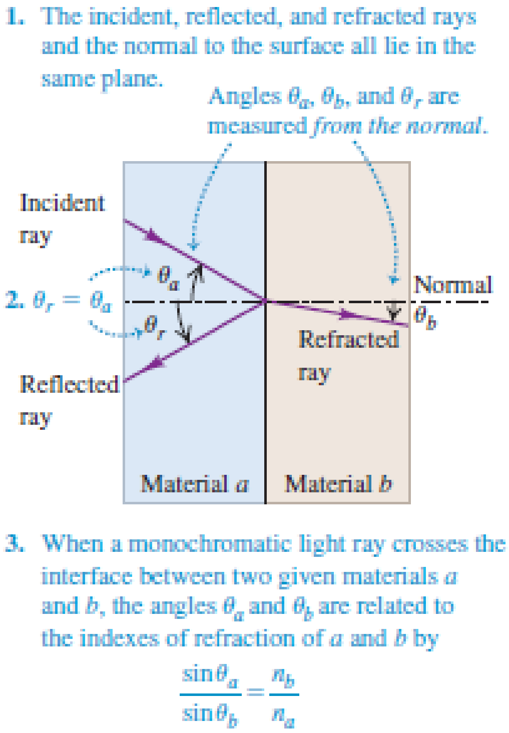
Light always travels *more slowly* in a material than in vacuum, so the value of in anything other than vacuum is always greater than unity. For vacuum, n = 1. Since n is a ratio of two speeds, it is a pure number without units.

Experimental studies of the directions of the incident, reflected, and refracted rays at a smooth interface between two optical materials lead to the following conclusions (Fig. 75):

1. **The incident, reflected, and refracted rays and the normal to the surface all lie in the same plane.** The plane of the three rays and the normal, called the **plane of incidence,** is perpendicular to the plane of the boundary surface between the two materials. We always draw ray diagrams so that the incident, reflected, and refracted rays are in the plane of the diagram.
2. **The angle of reflection is equal to the angle of incidence for all wavelengths and for any pair of materials.** That is, in Fig. 73c,

$$\theta_r = \theta_a \tag{160}$$

This relationship, together with the observation that the incident and reflected rays and the normal all lie in the same plane, is called the **law of reflection.**



1. The incident, reflected, and refracted rays and the normal to the surface all lie in the same plane.

Angles θ_a, θ_b, and θ_r are measured *from the normal.*

3. When a monochromatic light ray crosses the interface between two given materials a and b, the angles θ_a and θ_b are related to the indexes of refraction of a and b by

$$\frac{\sin\theta_a}{\sin\theta_b} = \frac{n_b}{n_a}$$

Figure 33.7 75– The laws of refraction and reflection

3. For monochromatic light and for a given pair of materials, a and b on opposite sides of the interface, **the ratio of the sines of the angles θ_a and θ_b, where both angles are measured from the normal to the surface, is equal to the inverse ratio of the two indexes of refraction:**

$$\frac{\sin\theta_a}{\sin\theta_b} = \frac{n_b}{n_a} \qquad (161)$$

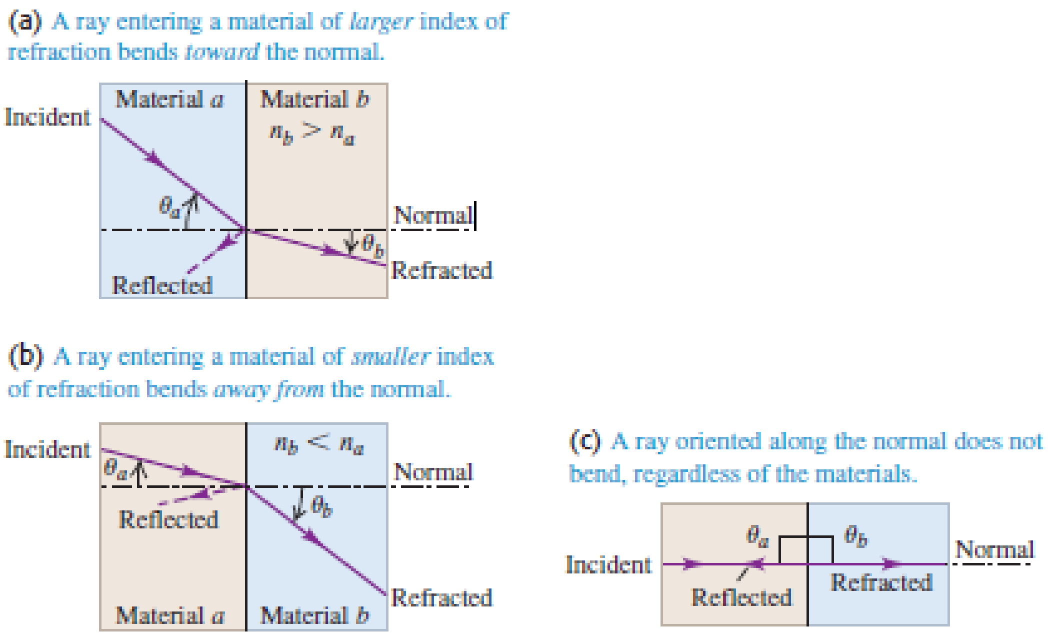
or

$$n_a \sin\theta_a = n_b \sin\theta_b \qquad (162)$$

This experimental result, together with the observation that the incident and refracted rays and the normal all lie in the same plane, is called the **law of refraction** or **Snell's law,** after the Dutch scientist Willebrord Snell (1591–1626). There is some doubt that Snell actually discovered it. The discovery that $n = \frac{c}{v}$ came much later.

While these results were first observed experimentally, they can be derived theoretically from a wave description of light.

Equations (161) and (162) show that when a ray passes from one material (a) into another material (b) having a larger index of refraction ($n_b > n_a$) and hence a slower wave speed, the angle θ_b with the normal is *smaller* in the second material than the angle θ_a in the first; hence the ray is bent *toward* the normal (Fig. 76a). When the second material has a *smaller* index of refraction than the first material ($n_b < n_a$) and hence a faster wave speed, the ray is bent *away from* the normal (Fig. 76b).

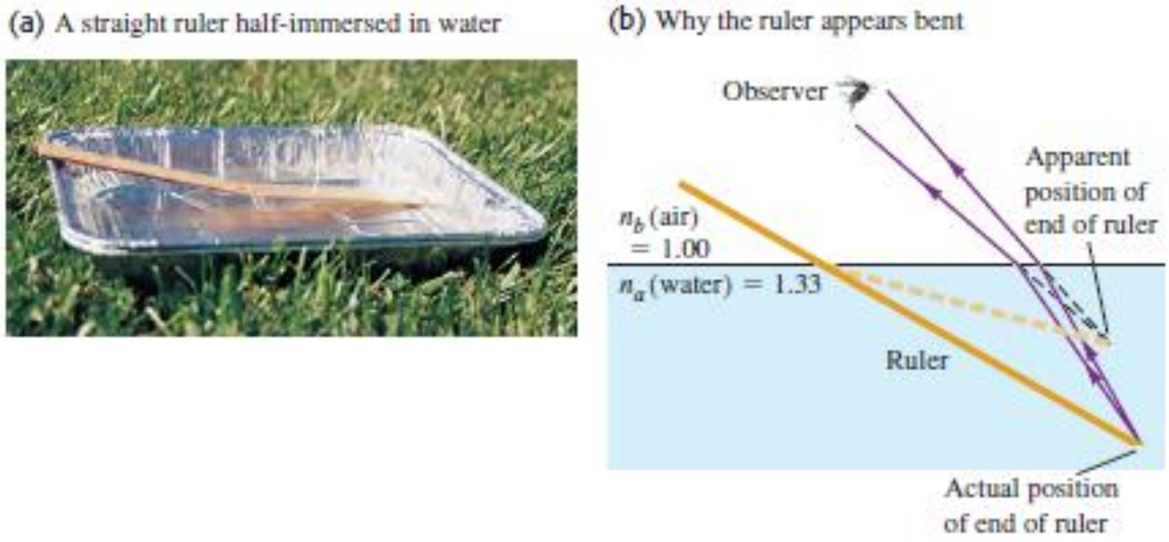

(a) A ray entering a material of *larger* index of refraction bends *toward* the normal.

(b) A ray entering a material of *smaller* index of refraction bends *away from* the normal.

(c) A ray oriented along the normal does not bend, regardless of the materials.

Figure 76 - Refraction and reflection in three cases. (a) Material b has a larger index of refraction than material a. (b) Material b has a smaller index of refraction than material a. (c) The incident light ray is normal to the interface between the materials

No matter what the materials on either side of the interface, in the case of *normal* incidence the transmitted ray is not bent at all (Fig. 76c). In this case $\theta_a = 0$ and $\sin \theta_a = 0$, so from Eq. (162) θ_b is also equal to zero, so the transmitted ray is also normal to the interface. Equation (160) shows that θ_r, too, is equal to zero, so the reflected ray travels back along the same path as the incident ray.


(a) A straight ruler half-immersed in water

(b) Why the ruler appears bent

Observer
Apparent position of end of ruler
n_b (air) = 1.00
n_a (water) = 1.33
Ruler
Actual position of end of ruler

Figure 77 - (a) This ruler is actually straight, but it appears to bend at the surface of the water. (b) Light rays from any submerged object bend away from the normal when they emerge into the air. As seen by an observer above the surface of the water, the object appears to be much closer to the surface than it actually is

The law of refraction explains why a partially submerged ruler or drinking straw appears bent; light rays coming from below the surface change in direction at the air–water interface, so the rays appear to be coming from a position above their actual point of origin (Fig. 77). A similar effect explains the appearance of the setting sun (Fig. 78).

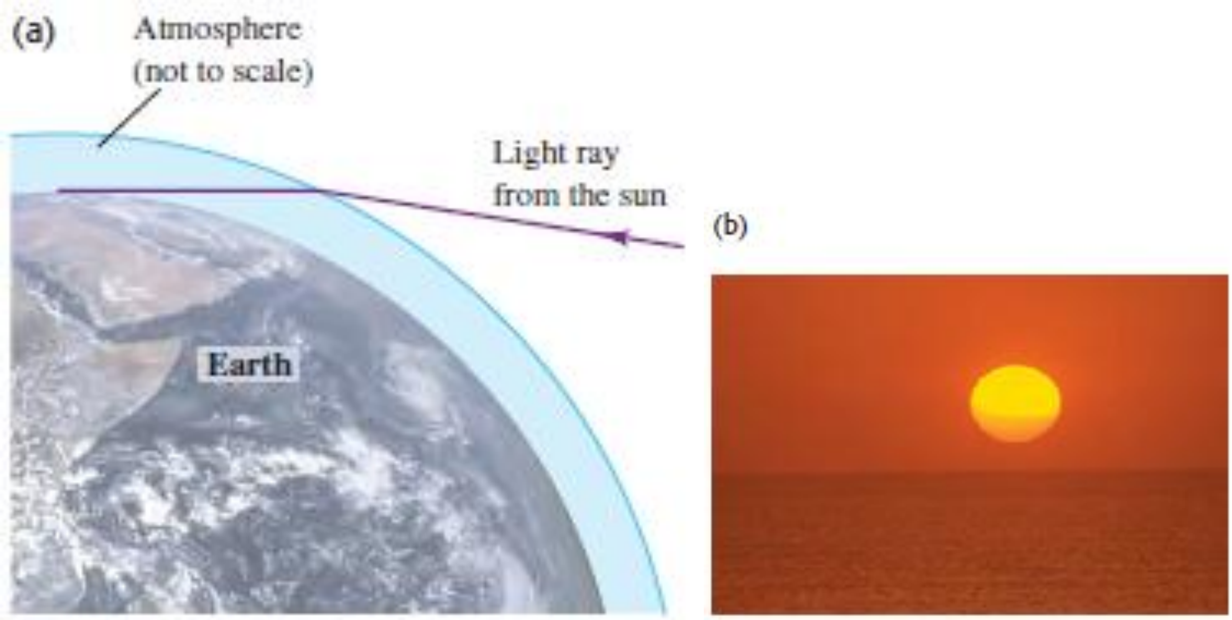


Figure 78 - (a) The index of refraction of air is slightly greater than 1, so light rays from the setting sun bend downward when they enter our atmosphere. (The effect is exaggerated in this figure.) (b) Stronger refraction occurs for light coming from the lower limb of the sun (the part that appears closest to the horizon), which passes through denser air in the lower atmosphere. As a result, the setting sun appears flattened vertically

An important special case is refraction that occurs at an interface between vacuum, for which the index of refraction is unity by definition, and a material. When a ray passes from vacuum into a material (b), so that $n_a = 1$ and $b > 1$ the ray is always bent *toward* the normal. When a ray passes from a material into vacuum, so that $n_a > 1$ and $n_b = 1$, the ray is always bent *away from* the normal.

The laws of reflection and refraction apply regardless of which side of the interface the incident ray comes from. If a ray of light approaches the interface in Fig. 76a or 76b from the right rather than from the left, there are again reflected and refracted rays; these two rays, the incident ray, and the normal to the surface again lie in the same plane. Furthermore, the path of a refracted ray is *reversible;* it follows the same path when going from b to a as when going from a to b [You can verify this using Eq. (162).] Since reflected and incident rays make the same angle with the normal, the path of a reflected ray is also reversible. That's why when you see someone's eyes in a mirror, they can also see you.

The *intensities* of the reflected and refracted rays depend on the angle of incidence, the two indexes of refraction, and the polarization (that is, the direction of the electric-field vector) of the incident ray. The fraction reflected is smallest at normal incidence ($\theta_a = 0℃$), where it is about 4% for an air–glass interface. This fraction increases with increasing angle of incidence to 100% at grazing incidence, when $\theta_a = 90$°C.

It's possible to use Maxwell's equations to predict the amplitude, intensity, phase, and polarization states of the reflected and refracted waves. Such an analysis is beyond our scope, however.

The index of refraction depends not only on the substance but also on the wavelength of the light. The dependence on wavelength is called *dispersion*. Indexes of refraction for several solids and liquids are given in Table 3 for a particular wavelength of yellow light.

The index of refraction of air at standard temperature and pressure is about 1.0003, and we will usually take it to be exactly unity. The index of refraction of a gas increases as its density increases. Most glasses used in optical instruments have indexes of refraction between about 1.5 and 2.0. A few substances have larger indexes; one example is diamond, with 2.417.

Table 3 – Index of refraction for yellow sodium light (wavelength 589 nm)

Substance	Index of refraction
Solids	
Ice	1.309
Fluorite	1.434
Polystyrene	1.49
Rock salt	1.544
Quartz	1.544
Zircon	1.923
Diamond	2.417
Fabulite	2.409
Rutile	2.62
Gases (typical values)	
Crown	1.52
Light flint	1.58
Medium flint	1.62
Dense flint	1.66
Lanthanum flint	1.8
Liquids at 20°C	
Methanol	1.329
Water	1.333
Ethanol	1.36
Carbon tetrachloride	1.46
Turpentine	1.472
Glycerine	1.473
Benzene	1.501
Carbon disulfide	1.628

We have discussed how the direction of a light ray changes when it passes from one material to another material with a different index of refraction. It's also important to see what happens to the *wave* characteristics of the light when this happens.

First, the frequency ν of the wave does not change when passing from one material to another. That is, the number of wave cycles arriving per unit time must equal the number leaving per unit time; this is a statement that the boundary surface cannot create or destroy waves.

Second, the wavelength λ of the wave *is* different in general in different materials. This is because in any material, $v = \lambda\nu$; since ν is the same in any material as in vacuum and v is always less than the wave speed c in vacuum, λ is also correspondingly reduced. Thus the wavelength λ of light in a material is *less than* the wavelength λ_0 of the same light in vacuum. From the above discussion, $\nu = c/\lambda_0 = v/\lambda$. Combining this with Eq. (159), $n = c/v$, we find

$$\lambda = \frac{\lambda_0}{n} \tag{163}$$

When a wave passes from one material into a second material with larger index of refraction, so that $n_b > n_a$, the wave speed decreases. The wavelength $\lambda_b = \lambda_0/n_b$ in the second material is then shorter than the wavelength $\lambda_a = \lambda_0/n_a$ in the first material. If instead the second material has a smaller index of refraction than the first material, so that then the wave speed increases. Then the wavelength λ_b in the second material is longer than the wavelength λ_a in the first material. This makes intuitive sense; the waves get "squeezed" (the wavelength gets shorter) if the wave speed decreases and get "stretched" (the wavelength gets longer) if the wave speed increases.

2.1.3 Plane and spherical surface

Before discussing what is meant by an image, we first need the concept of **object** as it is used in optics. By an *object* we mean anything from which light rays radiate. This light could be emitted by the object itself if it is *self-luminous,* like the glowing filament of a light bulb. Alternatively, the light could be emitted by another source (such as a lamp or the sun) and then reflected from the object; an example is the light you see coming from the pages of this book. Figure 79 shows light rays radiating in all directions from an object at a point P. For an observer to see this object directly, there must be no obstruction between the object and the observer's eyes. Note that light rays from the object reach the observer's left and right eyes at different angles; these differences are processed by the observer's brain to infer the *distance* from the observer to the object.

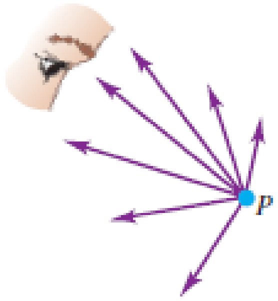


Figure 79 - Light rays radiate from a point object *P* in all directions

The object in Fig. 79 is a **point object** that has no physical extent. Real objects with length, width, and height are called **extended objects.** To start with, we'll consider only an idealized point object, since we can always think of an extended object as being made up of a very large number of point objects.

Suppose some of the rays from the object strike a smooth, plane reflecting surface (Fig. 80). This could be the surface of a material with a different index of refraction, which reflects part of the incident light, or a polished metal surface that reflects almost 100% of the light that strikes it. We will always draw the reflecting surface as a black line with a shaded area behind it, as in Fig. 80. Bathroom mirrors have a thin sheet of glass that lies in front of and protects the reflecting surface; we'll ignore the effects of this thin sheet.

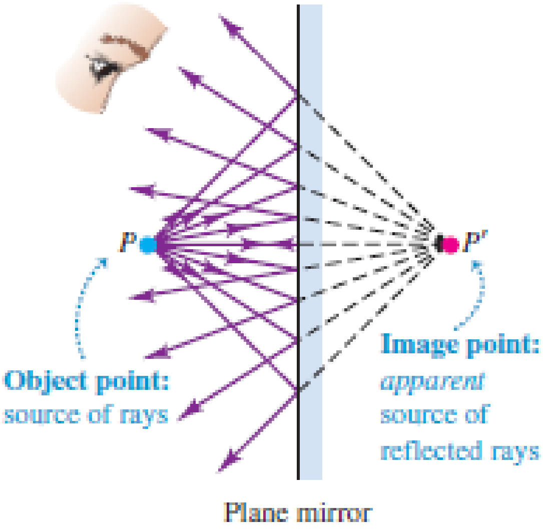


Figure 80 - Light rays from the object at point *P* are reflected from a plane mirror. The reflected rays entering the eye look as though they had come from image point *P'*

According to the law of reflection, all rays striking the surface are reflected at an angle from the normal equal to the angle of incidence. Since the surface is plane, the normal is in the same direction at all points on the surface, and we have *specular* reflection. After the rays are reflected, their directions are the same as though they had come from point P'. We call point *P* an *object point* and point the corresponding *image point,* and we say that the reflecting surface forms an **image** of point *P*. An observer who can see only the rays reflected from the surface, and who doesn't know that he's seeing a reflection, *thinks* that the rays originate from the image point P'. The image point is therefore a convenient way to describe the directions of the various reflected rays, just as the object point *P* describes the directions of the rays arriving at the surface *before* reflection.

If the surface in Fig. 80 were *not* smooth, the reflection would be *diffuse,* and rays reflected from different parts of the surface would go in uncorrelated directions (see Fig. 74b). In this case there would not be a definite image point from which all reflected rays seem to emanate. You can't see your reflection in the surface of a tarnished piece of metal because its surface is rough; polishing the metal smoothes the surface so that specular reflection occurs and a reflected image becomes visible.

An image is also formed by a plane *refracting* surface, as shown in Fig. 81. Rays coming from point *P* are refracted at the interface between two optical materials. When the angles of incidence are small, the final directions of the rays after refraction are the same as though they had come from point as shown, and again we call an *image point.* Early we described how this effect makes underwater objects appear closer to the surface than they really are (see Fig. 77).

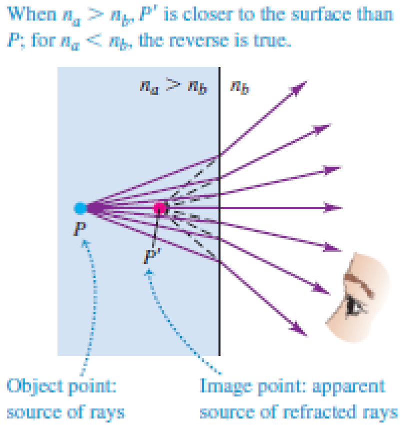

When $n_a > n_b$, P' is closer to the surface than P; for $n_a < n_b$, the reverse is true.
$n_a > n_b$
n_b
P
P'
Object point: source of rays
Image point: apparent source of refracted rays

Figure 81 - Light rays from the object at point *P* are refracted at the plane interface. The refracted rays entering the eye look as though they had come from image point *P*'

In both Figs. 80 and 81 the rays do not actually pass through the image point P'. Indeed, if the mirror in Fig. 80 is opaque, there is no light at all on its right side. If the outgoing rays don't actually pass through the image point, we call the image a **virtual image.** Later we will see cases in which the outgoing rays really *do* pass through an image point, and we will call the resulting image a **real image.** The images that are formed on a projection screen, on the photographic film in a camera, and on the retina of your eye are real images.

Image Formation by a Plane Mirror. Let's concentrate for now on images produced by *reflection;* we'll return to refraction later in the chapter. To find the precise location of the virtual image *P'* that a plane mirror forms of an object at *P*, we use the construction shown in Fig. 82. The figure shows two rays diverging from an object point *P* at a distance *s* to the left of a plane mirror. We call *s* the **object distance.** The ray *PV* is incident normally on the mirror (that is, it is perpendicular to the mirror surface), and it returns along its original path.

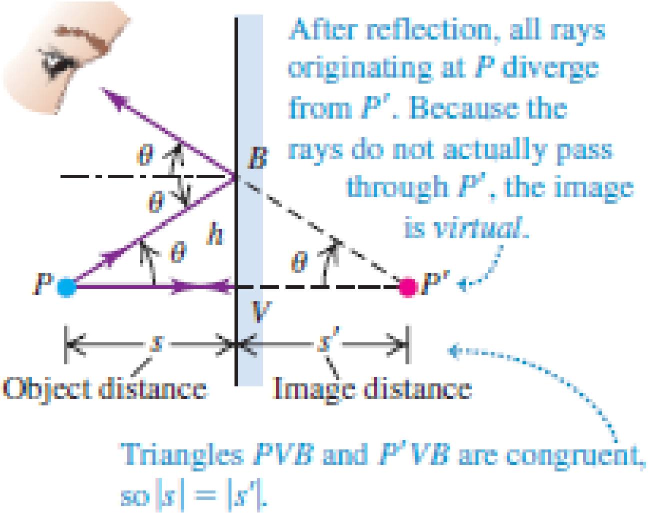

After reflection, all rays originating at *P* diverge from *P′*. Because the rays do not actually pass through *P′*, the image is *virtual*.
θ
θ
θ
θ
B
h
P
P′
V
s
s′
Object distance
Image distance
Triangles *PVB* and *P′VB* are congruent, so |*s*| = |*s′*|.

Figure 82 - Construction for determining the location of the image formed by a plane mirror. The image point *P’* is as far behind the mirror as the object point *P* is in front of it

The ray *PB* makes an angle θ with *PV*. It strikes the mirror at an angle of incidence θ and is reflected at an equal angle with the normal. When we extend the two reflected rays backward, they intersect at point *P’*, at a distance s’ behind the mirror. We call *s’* the **image distance.** The line between *P* and *P’* is perpendicular to the mirror. The two triangles *PVB* and *P’VB* are congruent, so *P* and *P’* are at equal distances from the mirror, *s’* and *s* and have equal magnitudes. The image point P’ is located exactly opposite the object point *P* as far *behind* the mirror as the object point is from the front of the mirror.

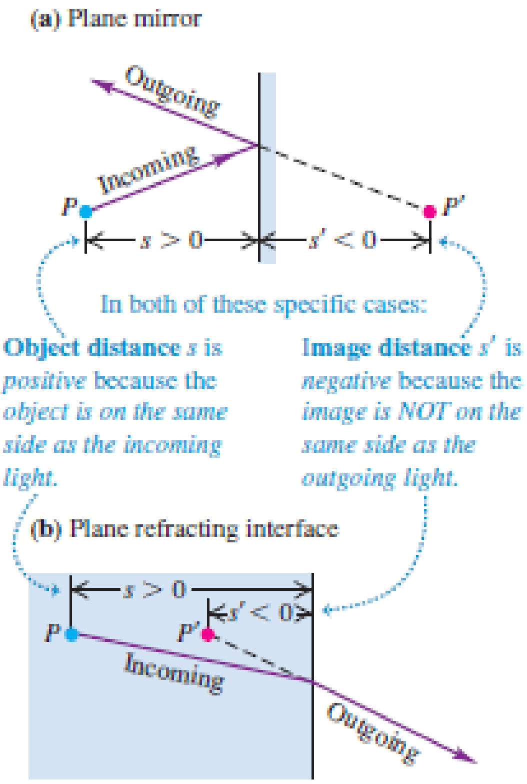
We can repeat the construction of Fig. 82 for each ray diverging from *P*. The directions of *all* the outgoing reflected rays are the same as though they had originated at point *P’*, confirming that *P’* is the *image* of *P*. No matter where the observer is located, she will always see the image at the point

Sign Rules. Before we go further, let’s introduce some general sign rules. These may seem unnecessarily complicated for the simple case of an image formed by a plane mirror, but we want to state the rules in a form that will be applicable to *all* the situations we will encounter later. These will include image formation by a plane or spherical reflecting or refracting surface, or by a pair of refracting surfaces forming a lens. Here are the rules:

1. **Sign rule for the object distance:** When the object is on the same side of the reflecting or refracting surface as the incoming light, the object distance *s* is positive; otherwise, it is negative.
2. **Sign rule for the image distance:** When the image is on the same side of the reflecting or refracting surface as the outgoing light, the image distance is positive; otherwise, it is negative.

3. **Sign rule for the radius of curvature of a spherical surface:** When the center of curvature *C* is on the same side as the outgoing light, the radius of curvature is positive; otherwise, it is negative.

Figure 83 illustrates rules 1 and 2 for two different situations. For a mirror the incoming and outgoing sides are always the same; for example, in Figs. 80, 82, and 83a they are both on the left side. For the refracting surfaces in Figs. 81 and 83b the incoming and outgoing sides are on the left and right sides, respectively, of the interface between the two materials. (Note that other textbooks may use different rules.)


(**a**) Plane mirror

In both of these specific cases:

Object distance *s* is *positive* because the *object is on the same side as the incoming light.*

Image distance *s′* is *negative* because the *image is NOT on the same side as the outgoing light.*

(**b**) Plane refracting interface

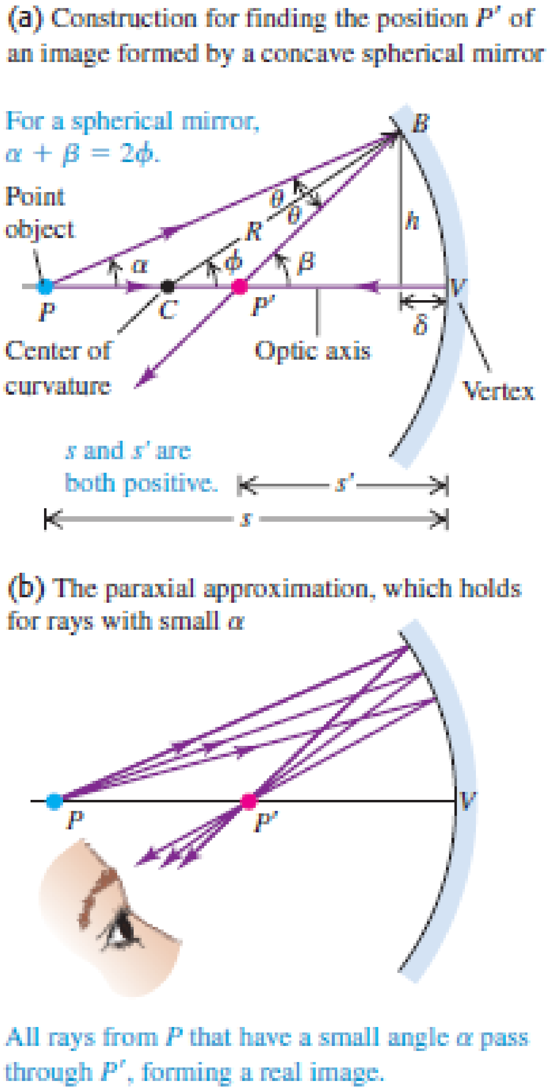
Figure 83 - For both of these situations, the object distance s is positive (rule 1) and the image distance s¿ is negative (rule 2)

In Figs. 82 and 83a the object distance *s* is *positive* because the object point *P* is on the incoming side (the left side) of the reflecting surface. The image distance *s'* is *negative* because the image point *P'* is *not* on the outgoing side (the left side) of the surface. The object and image distances s and *s'* are related simply by

$$s = -s' \qquad (164)$$

For a plane reflecting or refracting surface, the radius of curvature is infinite and not a particularly interesting or useful quantity; in these cases we really don't need sign rule 3. But this rule will be of great importance when we study image formation by *curved* reflecting and refracting surfaces later in the chapter.

Reflection at a Spherical Surface. A plane mirror produces an image that is the same size as the object. But there are many applications for mirrors in which the image and object must be of different sizes. A magnifying mirror used when applying makeup gives an image that is *larger* than the object, and surveillance mirrors (used in stores to help spot shoplifters) give an image that is *smaller* than the object. There are also applications of mirrors in which a *real* image is desired, so light rays do indeed pass through the image point A plane mirror by itself cannot perform any of these tasks. Instead, *curved* mirrors are used.


(a) Construction for finding the position P' of an image formed by a concave spherical mirror

For a spherical mirror, $\alpha + \beta = 2\phi$.

s and s' are both positive.

(b) The paraxial approximation, which holds for rays with small α

All rays from P that have a small angle α pass through P', forming a real image.

Figure 84 - (a) A concave spherical mirror forms a real image of a point object *P* on the mirror’s optic axis. (b) The eye sees some of the outgoing rays and perceives them as having come from *P*’

Image of a Point Object: Spherical Mirror. We’ll consider the special (and easily analyzed) case of image formation by a *spherical* mirror. Figure 84a shows a spherical mirror with radius of curvature *R*, with its concave side facing the incident light. The **center of curvature** of the surface (the center of the sphere of which the surface is a part) is at *C*, and the **vertex** of the mirror (the center of the mirror surface) is at *V*. The line *CV* is called the **optic axis.** Point *P* is an object point that lies on the optic axis; for the moment, we assume that the distance from *P* to *V* is greater than *R*.

Ray *PV*, passing through *C*, strikes the mirror normally and is reflected back on itself. Ray *PB*, at an angle α α with the axis, strikes the mirror at *B*, where the angles of incidence and reflection are θ. The reflected ray intersects the axis at point P'. We will show shortly that *all* rays from *P* intersect the axis at the *same* point as in Fig. 84b, provided that the angle is small. Point is therefore the *image* of object point *P'*. Unlike the reflected rays in Fig. 79, the reflected rays in Fig. 84b actually do intersect at point P' then diverge from P' *as if* they had originated at this point. Thus P' is a *real* image.

To see the usefulness of having a real image, suppose that the mirror is in a darkened room in which the only source of light is a self-luminous object at *P*. If you place a small piece of photographic film at P' all the rays of light coming from point *P* that reflect off the mirror will strike the same point P' on the film; when developed, the film will show a single bright spot, representing a sharply focused image of the object at point *P*. This principle is at the heart of most astronomical telescopes, which use large concave mirrors to make photographs of celestial objects. With a *plane* mirror like that in Fig. 80, placing a piece of film at the image point P' would be a waste of time; the light rays never actually pass through the image point, and the image can't be recorded on film. Real images are *essential* for photography.

Let's now find the location of the real image point P' in Fig. 84a and prove the assertion that all rays from *P* intersect at P' (provided that their angle with the optic axis is small). The object distance, measured from the vertex *V*, is *s*; the image distance, also measured from *V*, is s'. The signs of *s*, s' and the radius of curvature *R* are determined by the sign rules given early. The object point *P* is on the same side as the incident light, so according to sign rule 1, *s* is positive. The image point P' is on the same side as the reflected light, so according to sign rule 2, the image distance is also positive. The center of curvature *C* is on the same side as the reflected light, so according to sign rule 3, *R*, too, is positive; *R* is always positive when reflection occurs at the *concave* side of a surface
(Fig. 85).

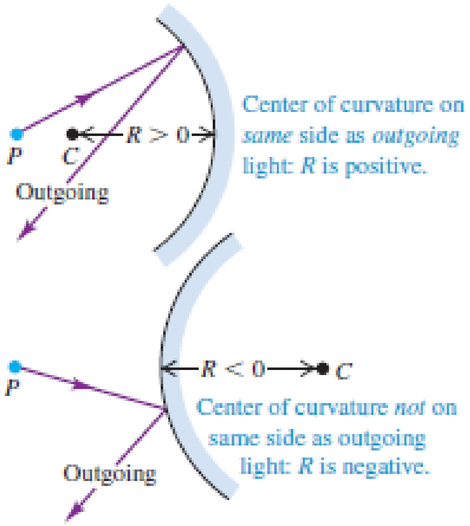


Figure 85 - The sign rule for the radius of a spherical mirror

We now use the following theorem from plane geometry: An exterior angle of a triangle equals the sum of the two opposite interior angles. Applying this theorem to triangles *PBC* and *P*'*BC* in Fig. 84a, we have Eliminating between these equations gives

$$\varphi = \alpha + \theta, \beta = \varphi + \theta \tag{165}$$

Eliminating θ between these equations gives

$$\alpha + \beta = 2\varphi \tag{166}$$

We may now compute the image distance s'. Let h represent the height of point *B* above the optic axis, and let δ represent the short distance from *V* to the foot of this vertical line. We now write expressions for the tangents of α, β and φ remembering that *s*, s' and *R* are all positive quantities:

$$\tan\alpha = \frac{h}{s-\delta}, \tan\beta = \frac{h}{s'-\delta}, \tan\varphi = \frac{h}{R-\delta} \tag{167}$$

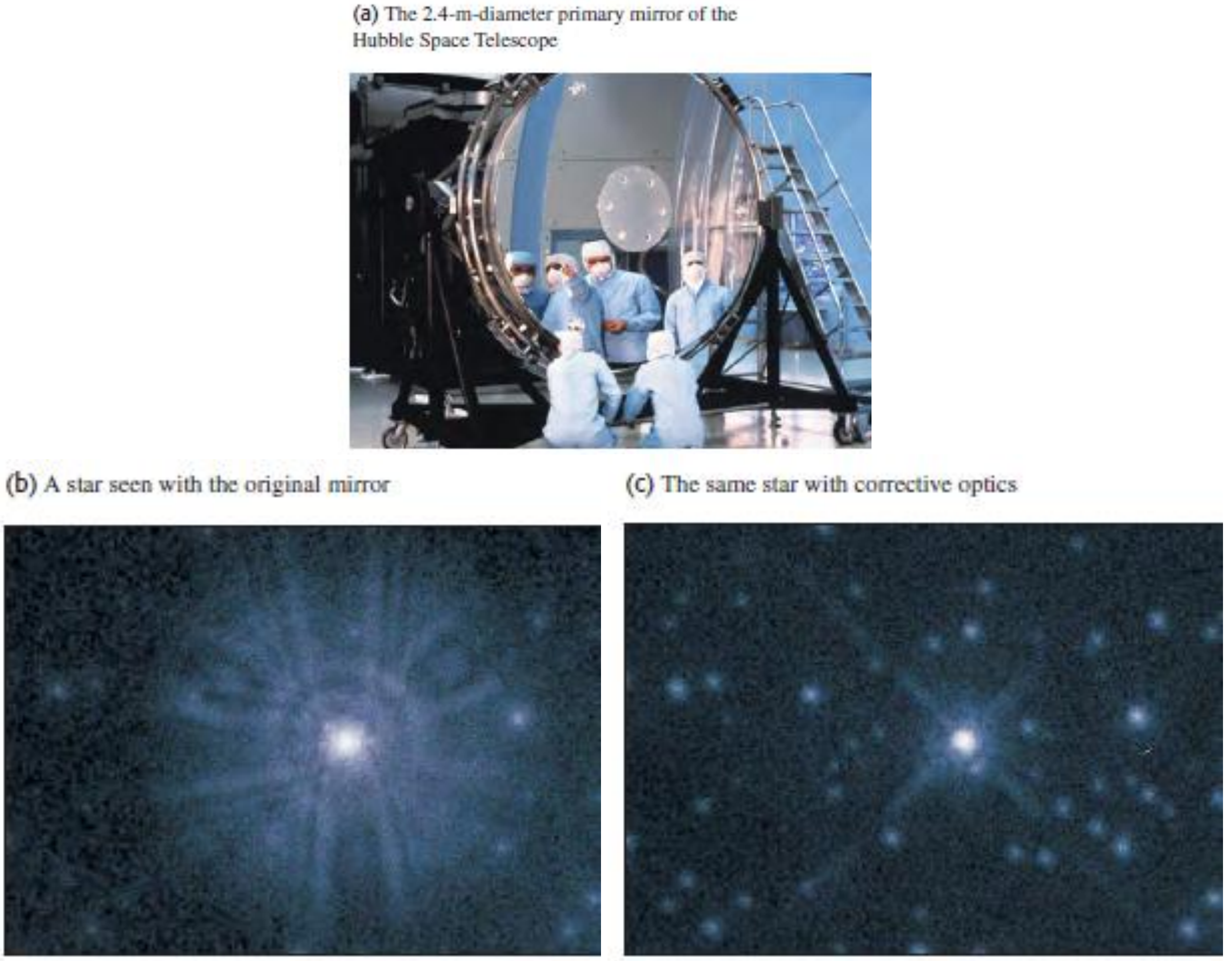
These trigonometric equations cannot be solved as simply as the corresponding algebraic equations for a plane mirror. However, *if the angle* α *is small*, the angles β and φ are also small. The tangent of an angle that is much less than one radian is nearly equal to the angle itself (measured in radians), so we can replace $\tan\alpha$ by α, and so on, in the equations above. Also, if α is small, we can neglect the distance δ compared with s', *s*, and *R*. So for small angles we have the following approximate relationships:

$$\alpha = \frac{h}{s}, \beta = \frac{h}{s'}, \varphi = \frac{h}{R} \tag{168}$$

Substituting these into Eq. (166) and dividing by h, we obtain a general relationship among s, s’ and R:

$$\frac{1}{s} + \frac{1}{s'} = \frac{2}{R} \tag{169}$$

This equation does not contain the angle α. Hence *all* rays from P that make sufficiently small angles with the axis intersect at P’ after they are reflected; this verifies our earlier assertion. Such rays, nearly parallel to the axis and close to it, are called **paraxial rays.** (The term **paraxial approximation** is often used for the approximations we have just described.) Since all such reflected light rays converge on the image point, a concave mirror is also called a *converging mirror*.



(a) The 2.4-m-diameter primary mirror of the Hubble Space Telescope

(b) A star seen with the original mirror

(c) The same star with corrective optics

Figure 86 - (a), (b) Soon after the Hubble Space Telescope (HST) was placed in orbit in 1990, it was discovered that the concave primary mirror (also called the *objective mirror*) was too shallow by about 1/50 the width of a human hair, leading to spherical aberration of the star’s image. (c) After corrective optics were installed in 1993, the effects of spherical aberration were almost completely eliminated.

Be sure you understand that Eq. (169), as well as many similar relationships that we will derive later in this chapter and the next, is only *approximately* correct. It results from a calculation containing approximations, and it is valid only for paraxial rays. If we increase the angle α that a ray makes with the optic axis, the point P' where the ray intersects the optic axis moves somewhat closer to the vertex than for a paraxial ray. As a result, a spherical mirror, unlike a plane mirror, does not form a precise point image of a point object; the image is "smeared out." This property of a spherical mirror is called **spherical aberration.** When the primary mirror of the Hubble Space Telescope (Fig. 86a) was manufactured, tiny errors were made in its shape that led to an unacceptable amount of spherical aberration (Fig. 86b). The performance of the telescope improved dramatically after the installation of corrective optics (Fig. 86c).

If the radius of curvature becomes infinite ($R = \infty$) the mirror becomes *plane,* and Eq. (169) reduces to Eq. (164) for a plane reflecting surface.

Focal Point and Focal Length. When the object point P is very far from the spherical mirror ($R = \infty$), the incoming rays are parallel. (The star shown in Fig. 86c is an example of such a distant object.) From Eq. (169) the image distance in this case is given by

$$\frac{1}{\infty}+\frac{1}{s'}=\frac{2}{R}, s' = \frac{R}{2} \tag{170}$$

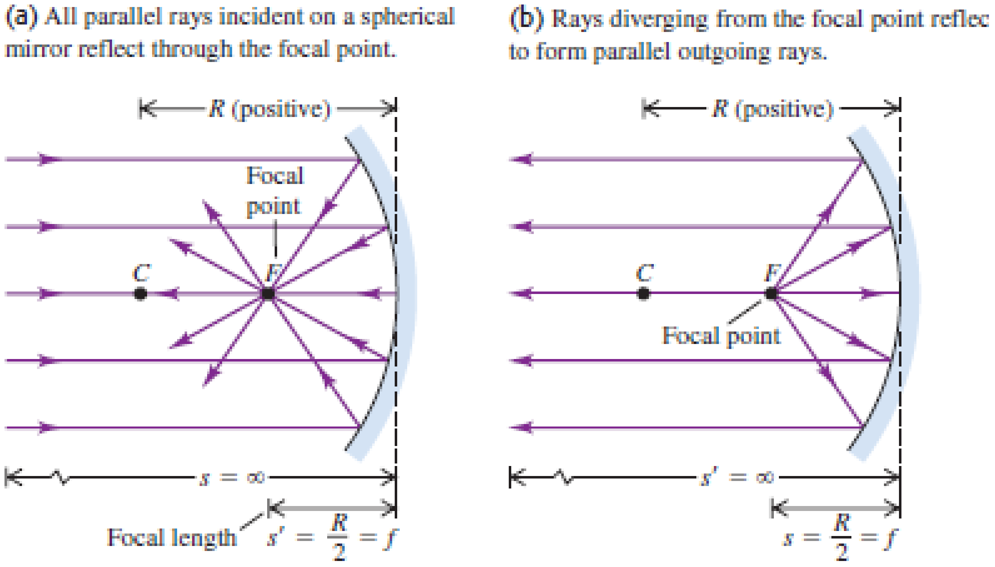
The situation is shown in Fig. 87a. The beam of incident parallel rays converges, after reflection from the mirror, to a point *F* at a distance R/2 from the vertex of the mirror. The point *F* at which the incident parallel rays converge is called the **focal point;** we say that these rays are brought to a focus. The distance from the vertex to the focal point, denoted by is called the **focal length.** We see that is related to the radius of curvature *R* by

$$f = \frac{R}{2} \tag{171}$$

The opposite situation is shown in Fig. 87b. Now the *object* is placed at the focal point *F*, so the object distance is $s = f = R/2$. The image distance s' is again given by Eq. (169):

$$\frac{2}{R}+\frac{1}{s'}=\frac{2}{R}, \quad \frac{1}{s'} = 0, \quad s' = \infty \tag{172}$$

With the object at the focal point, the reflected rays in Fig. 87b are parallel to the optic axis; they meet only at a point infinitely far from the mirror, so the image is at infinity.



(a) All parallel rays incident on a spherical mirror reflect through the focal point.

(b) Rays diverging from the focal point reflec to form parallel outgoing rays.

Figure 87 – The focal point and focal length of a concave mirror

Thus the focal point F of a spherical mirror has the properties that (1) any incoming ray parallel to the optic axis is reflected through the focal point and (2) any incoming ray that passes through the focal point is reflected parallel to the optic axis. For spherical mirrors these statements are true only for paraxial rays. For parabolic mirrors these statements are *exactly* true; this is why parabolic mirrors are preferred for astronomical telescopes. Spherical or parabolic mirrors are used in flashlights and headlights to form the light from the bulb into a parallel beam. Some solar-power plants use an array of plane mirrors to simulate an approximately spherical concave mirror; light from the sun is collected by the mirrors and directed to the focal point, where a steam boiler is placed.

We will usually express the relationship between object and image distances for a mirror, Eq. (169), in terms of the focal length f:

$$\frac{1}{s}+\frac{1}{s'}=\frac{1}{f} \qquad (173)$$

Convex Mirrors. In Fig. 88a the *convex* side of a spherical mirror faces the incident light. The center of curvature is on the side opposite to the outgoing rays; according to sign rule 3, R is negative (see Fig. 85). Ray PB is reflected, with the angles of incidence and reflection both equal to θ. The reflected ray, projected backward, intersects the axis at As with a concave mirror, *all* rays from P that are reflected by the mirror diverge from the same point P', provided that the angle α is small. Therefore P' is the image of P. The object distance s is positive, the image distance s' is negative, and the radius of curvature R is *negative* for a *convex* mirror.

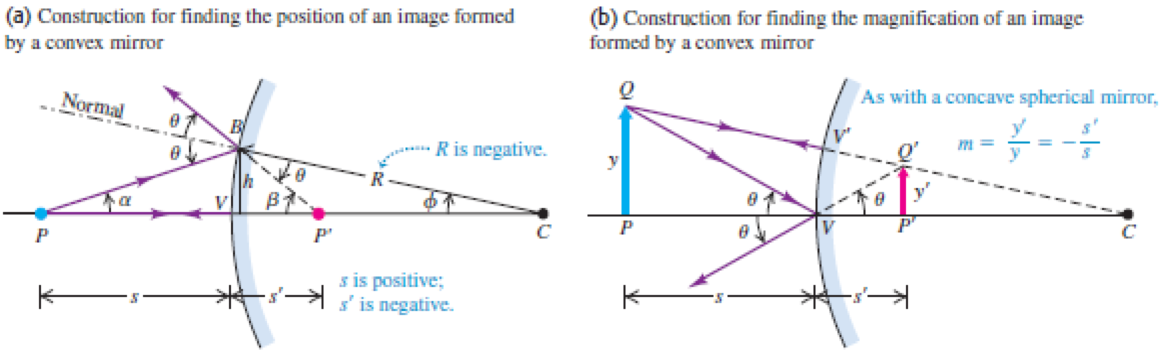

(a) Construction for finding the position of an image formed by a convex mirror
Normal
θ
θ
B
R is negative.
h
θ
R
α
V
β
φ
P
P′
C
s
s′
s is positive; s′ is negative.
(b) Construction for finding the magnification of an image formed by a convex mirror
Q
As with a concave spherical mirror,
V′
Q′
$m = \frac{y'}{y} = -\frac{s'}{s}$
y
θ
θ
y′
θ
P
V
P′
C
s
s′

Figure 88 – Image formation by a convex mirror

Figure 88b shows two rays diverging from the head of the arrow *PQ* and the virtual image P'Q' of this arrow. The same procedure that we used for a concave mirror can be used to show that for a convex mirror,

$$\frac{1}{s} + \frac{1}{s'} = \frac{2}{R} \tag{174}$$

and the lateral magnification is

$$m = \frac{y'}{y} = -\frac{s'}{s} \tag{175}$$

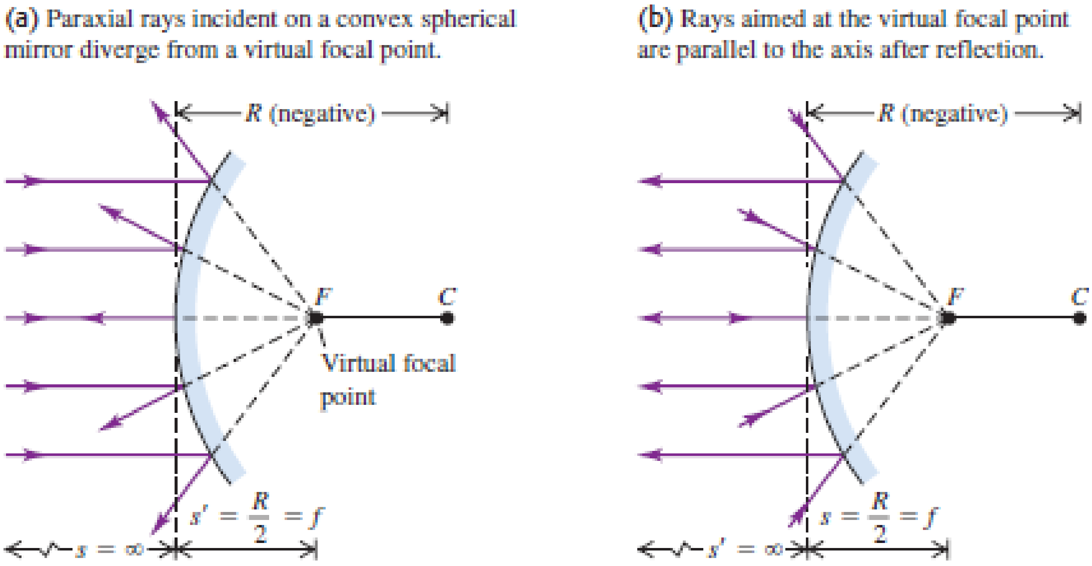

(a) Paraxial rays incident on a convex spherical mirror diverge from a virtual focal point.
R (negative)
F
C
Virtual focal point
$s' = \frac{R}{2} = f$
$s = \infty$
(b) Rays aimed at the virtual focal point are parallel to the axis after reflection.
R (negative)
F
C
$s = \frac{R}{2} = f$
$s' = \infty$

Figure 89 – The focal point and focal length of a convex mirror

These expressions are exactly the same as Eqs. (169) and (170) for a concave mirror. Thus when we use our sign rules consistently, Eqs. (169) and (170) are valid for both concave and convex mirrors.

When *R* is negative (convex mirror), incoming rays that are parallel to theoptic axis are not reflected through the focal point *F*. Instead, they diverge as though they had come from the point F at a distance f *behind* the mirror, as shown in Fig. 89a. In this case, f is the focal length, and F is called a *virtual focal point*. The corresponding image distance s' is negative, so both and R are negative, and Eq. (171), $f = R/2$, holds for convex as well as concave mirrors. In Fig. 89b the incoming rays are converging as though they would meet at the virtual focal point F, and they are reflected parallel to the optic axis.

In summary, Eqs. (169) through (174), the basic relationships for image formation by a spherical mirror, are valid for both concave and convex mirrors, provided that we use the sign rules consistently.

Graphical Methods for Mirrors. We used Eqs. (173) and (174) to find the position and size of the image formed by a mirror. We can also determine the properties of the image by a simple *graphical* method. This method consists of finding the point of intersection of a few particular rays that diverge from a point of the object (such as point Q in Fig. 90) and are reflected by the mirror. Then (neglecting aberrations) *all* rays from this object point that strike the mirror will intersect at the same point. For this construction we always choose an object point that is *not* on the optic axis. Four rays that we can usually draw easily are shown in Fig. 90. These are called **principal rays.**

1. *A ray parallel to the axis,* after reflection, passes through the focal point F of a concave mirror or appears to come from the (virtual) focal point of a convex mirror.

2. *A ray through (or proceeding toward) the focal point F* is reflected parallel to the axis.

3. *A ray along the radius* through or away from the center of curvature C intersects the surface normally and is reflected back along its original path. 4. *A ray to the vertex V* is reflected forming equal angles with the optic axis.

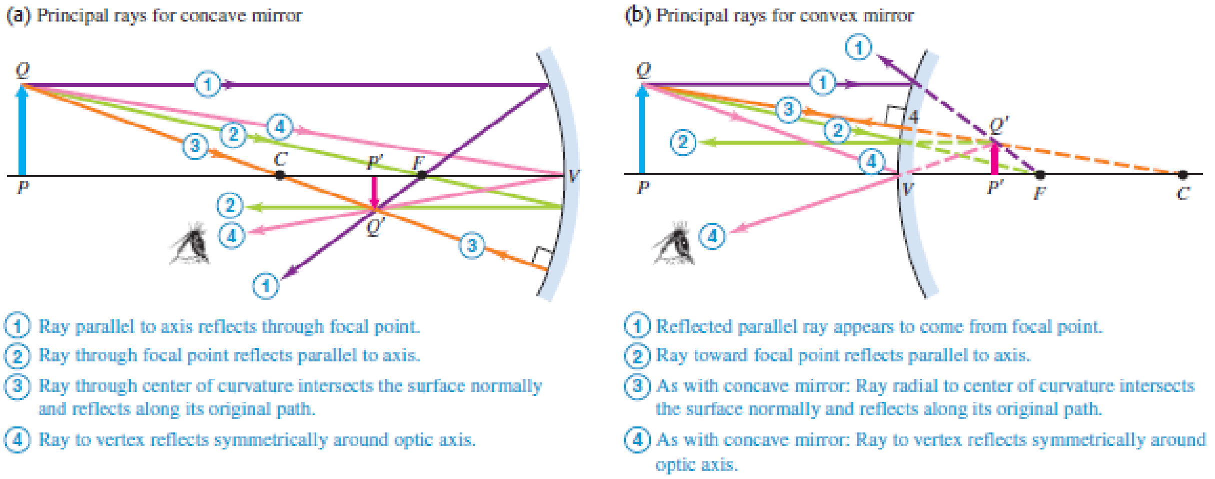

(a) Principal rays for concave mirror
Q
P
C
P'
F
V
Q'
1
2
3
4
(b) Principal rays for convex mirror
Q
P
V
Q'
P'
F
C
1
2
3
4
1 Ray parallel to axis reflects through focal point.
2 Ray through focal point reflects parallel to axis.
3 Ray through center of curvature intersects the surface normally and reflects along its original path.
4 Ray to vertex reflects symmetrically around optic axis.
1 Reflected parallel ray appears to come from focal point.
2 Ray toward focal point reflects parallel to axis.
3 As with concave mirror: Ray radial to center of curvature intersects the surface normally and reflects along its original path.
4 As with concave mirror: Ray to vertex reflects symmetrically around optic axis.

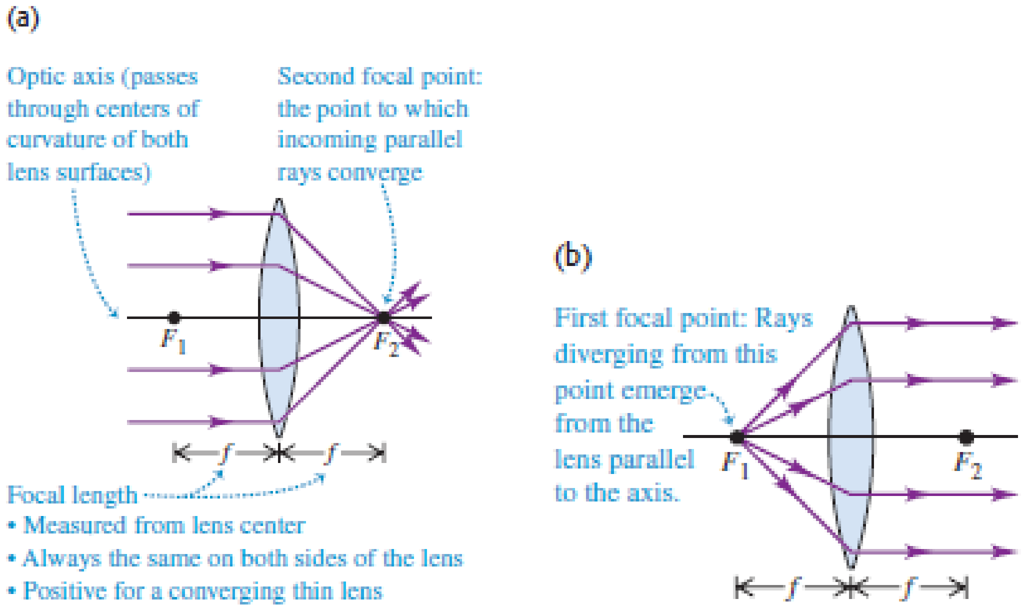
Figure 90 - The graphical method of locating an image formed by a spherical mirror. The colours of the rays are for identification only; they do not refer to specific colours of light

Once we have found the position of the image point by means of the intersection of any two of these principal rays we can draw the path of any other ray from the object point to the same image point.

2.1.4 Thin lenses

The most familiar and widely used optical device (after the plane mirror) is the *lens.* A lens is an optical system with two refracting surfaces. The simplest lens has two *spherical* surfaces close enough together that we can neglect the distance between them (the thickness of the lens); we call this a **thin lens.** If you wear eyeglasses or contact lenses while reading, you are viewing these words through a pair of thin lenses. We can analyze thin lenses in detail using the results for refraction by a single spherical surface. However, we postpone this analysis until later in the section so that we can first discuss the properties of thin lenses.

Properties of a Lens. A lens of the shape shown in Fig. 91 has the property that when a beam of rays parallel to the axis passes through the lens, the rays converge to a point F_2 (Fig. 91a) and form a real image at that point. Such a lens is called a **converging lens.** Similarly, rays passing through point emerge from the lens as a beam of parallel rays (Fig. 91b). The points F_1 and F_2 are called the first and second *focal points,* and the distance (measured from the center of the lens) is called the *focal length.* Note the similarities between the two focal points of a converging lens and the single focal point of a concave mirror (see Fig. 87). As for a concave mirror, the focal length of a converging lens is defined to be a *positive* quantity, and such a lens is also called a *positive lens.*



Optic axis (passes through centers of curvature of both lens surfaces)

Second focal point: the point to which incoming parallel rays converge

Focal length
- Measured from lens center
- Always the same on both sides of the lens
- Positive for a converging thin lens

First focal point: Rays diverging from this point emerge from the lens parallel to the axis.

Figure 91 - F_1 and F_2 are the first and second focal points of a converging thin lens. The numerical value of f is positive

The central horizontal line in Fig. 91 is called the *optic axis*, as with spherical mirrors. The centers of curvature of the two spherical surfaces lie on and define the optic axis. The two focal lengths in Fig. 91, both labelled *f*, *are always equal* for a thin lens, even when the two sides have different curvatures. We will derive this somewhat surprising result later in the section, when we derive the relationship of *f* to the index of refraction of the lens and the radii of curvature of its surfaces.

Image of an Extended Object: Converging Lens. Like a concave mirror, a converging lens can form an image of an extended object. Figure 92 shows how to find the position and lateral magnification of an image made by a thin converging lens. Using the same notation and sign rules as before, we let *s* and *s'* be the object and image distances, respectively, and let *y* and *y'* be the object and image heights. Ray *QA*, parallel to the optic axis before refraction, passes through the second focal point F_2 after refraction. Ray *QOQ'* passes undeflected straight through the center of the lens because at the center the two surfaces are parallel and (we have assumed) very close together. There is refraction where the ray enters and leaves the material but no net change in direction.

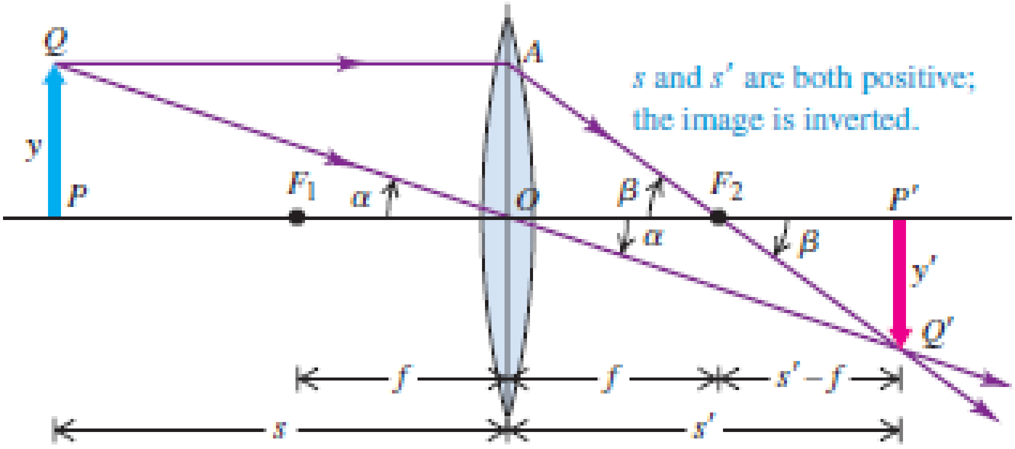


Figure 92 - Construction used to find image position for a thin lens. To emphasize that the lens is assumed to be very thin, the ray *QAQ'* is shown as bent at the midplane of the lens rather than at the two surfaces and ray *QOQ'* is shown as a straight line

The two angles labeled α in Fig. 92 are equal. Therefore the two right triangles *PQO* and *P'Q'O'* are *similar*, and ratios of corresponding sides are equal. Thus

$$\frac{y}{s} = -\frac{y'}{y}$$

or

$$\frac{y'}{y} = -\frac{s'}{s} \tag{176}$$

(The reason for the negative sign is that the image is below the optic axis and *y'* is negative.) Also, the two angles labeled β are equal, and the two right triangles OAF_1 and $P'Q'F_2$ are similar, so

$$\frac{y}{f} = -\frac{y'}{s' - f}$$

or

$$\frac{y'}{y} = -\frac{s' - f}{f} \tag{177}$$

We now equate Eqs. (176) and (177), divide by *s'* and rearrange to obtain

$$\frac{1}{s} + \frac{1}{s'} = \frac{1}{f} \tag{178}$$

This analysis also gives the lateral magnification $m = y/y'$ for the lens

$$m = -\frac{s}{s'} \tag{179}$$

The negative sign tells us that when s and *s'* are both positive, as in Fig. 92, the image is *inverted*, and *y* and *y'* have opposite signs.

Equations (178) and (179) are the basic equations for thin lenses. They are *exactly* the same as the corresponding equations for spherical mirrors, Eqs. (173) and (174). As we will see, the same sign rules that we used for spherical mirrors are also applicable to lenses. In particular, consider a lens with a positive focal length (a converging lens). When an object is outside the first focal point F_1 of this lens (that is, when $s > f$), the image distance is positive (that is, the image is on the same side as the outgoing rays); this image is real and inverted, as in Fig. 92. An object placed inside the first focal point of a converging lens, so that $s < f$, produces an image with a negative value of *s'*; this image is located on the same side of the lens as the object and is virtual, erect, and larger than the object. You can verify these statements algebraically using Eqs. (178) and (179); we'll also verify them in the next section, using graphical methods analogous to those introduced for mirrors.

Diverging Lenses. So far we have been discussing *converging* lenses. Figure 93 shows a **diverging lens;** the beam of parallel rays incident on this lens *diverges* after refraction. The focal length of a diverging lens is a negative quantity, and the lens is also called a *negative lens*. The focal points of a negative lens are reversed, relative to those of a positive lens. The second focal point, F_2, of a negative lens is the point from which rays that are originally parallel to the axis *appear to diverge* after refraction, as in Fig. 93a. Incident rays converging toward the first focal point F_1 as in Fig. 93b, emerge from the lens parallel to its axis. You can see that a diverging lens has the same relationship to a converging lens as a convex mirror has to a concave mirror.

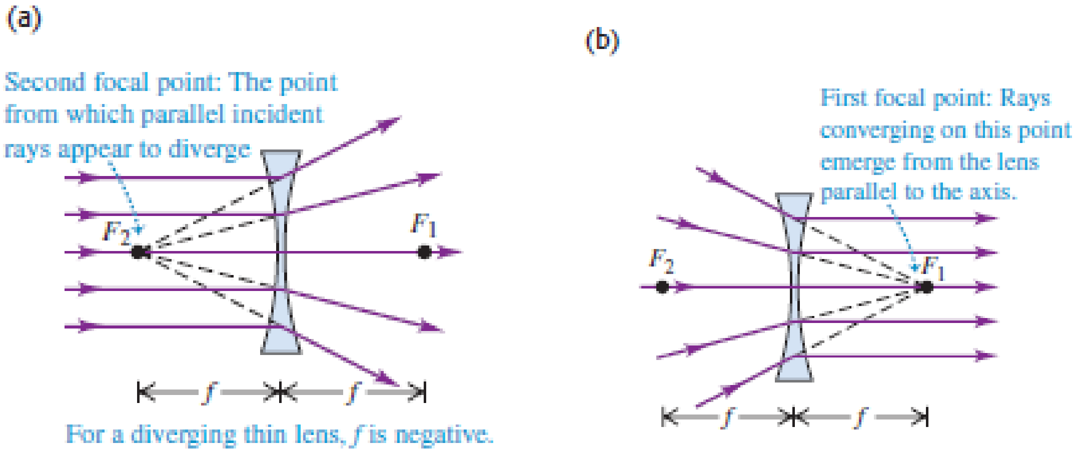

(a)
Second focal point: The point from which parallel incident rays appear to diverge
F_2
F_1
f
f
For a diverging thin lens, f is negative.
(b)
First focal point: Rays converging on this point emerge from the lens parallel to the axis.
F_2
F_1
f
f

Figure 93 - F_2 and F_1 are the second and first focal points of a diverging thin lens, respectively. The numerical value of is negative

Equations (178) and (179) apply to *both* positive and negative lenses. Figure 94 shows various types of lenses, both converging and diverging. Here's an important observation: *Any lens that is thicker at its center than at its edges is a converging lens with positive ; and any lens that is thicker at its edges than at its center is a diverging lens with negative* (provided that the lens has a greater index of refraction than the surrounding material). We can prove this using the *lensmaker's equation*, which it is our next task to derive.

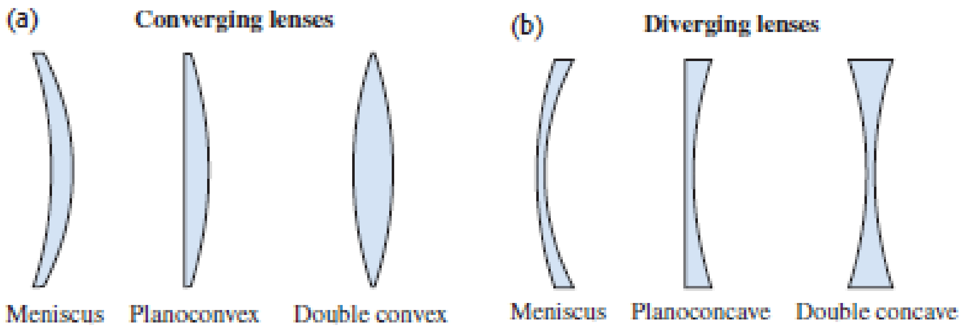


Figure 94 – Various types of lenses

The Lensmaker's Equation. We'll now derive Eq. (178) in more detail and at the same time derive the *lensmaker's equation,* which is a relationship among the focal length the index of refraction *n* of the lens, and the radii of curvature R_1 and R_2 of the lens surfaces. We use the principle that an image formed by one reflecting or refracting surface can serve as the object for a second reflecting or refracting surface.

We begin with the somewhat more general problem of two spherical interfaces separating three materials with indexes of refraction n_a, n_b and n_c as shown in Fig. 95. The object and image distances for the first surface are s_1 and s_1', and those for the second surface are s_2 and s_2'. We assume that the lens is thin, so that the distance *t* between the two surfaces is small in comparison with the object and image distances and can therefore be neglected. This is usually the case with eyeglass lenses. Then s_2 and s_1' have the same magnitude but opposite sign. For example, if the first image is on the outgoing side of the first surface, $s_1{}'$ is positive. But when viewed as an object for the second surface, the first image is *not* on the incoming side of that surface. So we can say that $s_2 = -s_1{}'$.

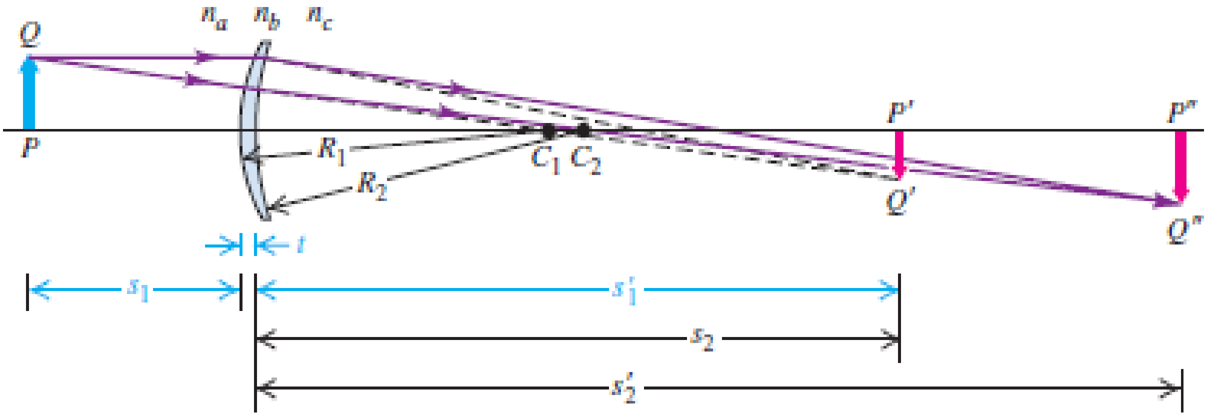


Figure 95 - The image formed by the first surface of a lens serves as the object for the second surface. The distances $s_1{}'$ and s_2 are taken to be equal; this is a good approximation if the lens thickness t is small.

We need to use the single-surface equation, twice, once for each surface. The two resulting equations are

$$\begin{gathered}\frac{n_a}{s_1}+\frac{n_b}{s_1'}=\frac{n_b-n_a}{R_1}\\ \frac{n_b}{s_2}+\frac{n_c}{s_2'}=\frac{n_c-n_b}{R_2}\end{gathered}\tag{180}$$

Ordinarily, the first and third materials are air or vacuum, so we set $n_a = n_c = 1$. The second index n_b is that of the lens, which we can call simply n. Substituting these values and the relationship $s_2 = -s_1{}'$ we get

$$\begin{gathered}\frac{1}{s_1}+\frac{n}{s_1'}=\frac{n-1}{R_1}\\ -\frac{n}{s_2}+\frac{1}{s_2'}=\frac{1-n}{R_2}\end{gathered}\tag{181}$$

To get a relationship between the initial object position s_1 and the final image position $s_2{}'$.n we add these two equations. This eliminates the term $n/s_1{}'$ and we obtain

$$\frac{1}{s_1}+\frac{1}{s_2'}=(n-1)\left(\frac{1}{R_1}-\frac{1}{R_2}\right)\tag{182}$$

Finally, thinking of the lens as a single unit, we call the object distance simply s instead of s_1', and we call the final image distance s' instead of s_2'. Making these substitutions, we have

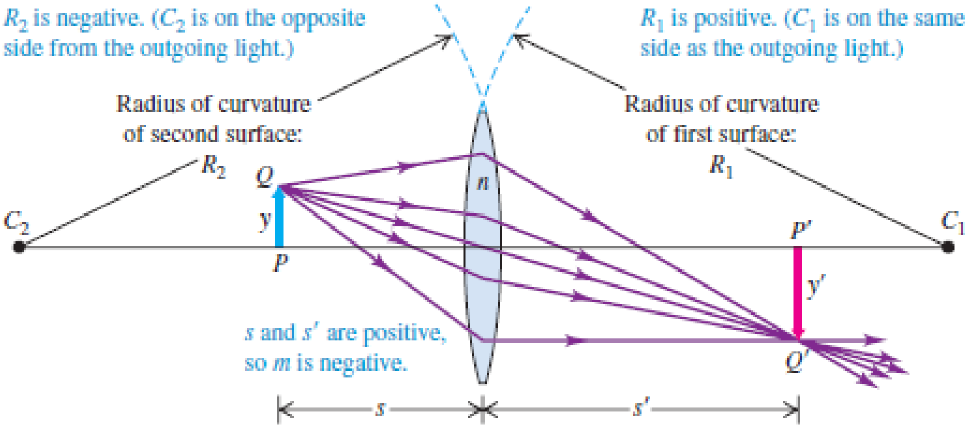
$$\frac{1}{s}+\frac{1}{s'}=(n-1)\left(\frac{1}{R_1}-\frac{1}{R_2}\right) \tag{183}$$

Now we compare this with the other thin-lens equation, Eq. (178). We see that the object and image distances *s* and *s'* appear in exactly the same places in both equations and that the focal length is given by

$$\frac{1}{f}=(n-1)\left(\frac{1}{R_1}-\frac{1}{R_2}\right) \tag{184}$$

This is the **lensmaker's equation.** In the process of rederiving the relationship among object distance, image distance, and focal length for a thin lens, we have also derived an expression for the focal length f of a lens in terms of its index of refraction *n* and the radii of curvature R_1 and R_2 of its surfaces. This can be used to show that all the lenses in Fig. 94a are converging lenses with positive focal lengths and that all the lenses in Fig. 94b are diverging lenses with negative focal lengths.

We use all our sign rules with Eqs. (183) and (184). For example, in Fig. 96, *s*, *s'* and R_1 are positive, but R_2 is negative.



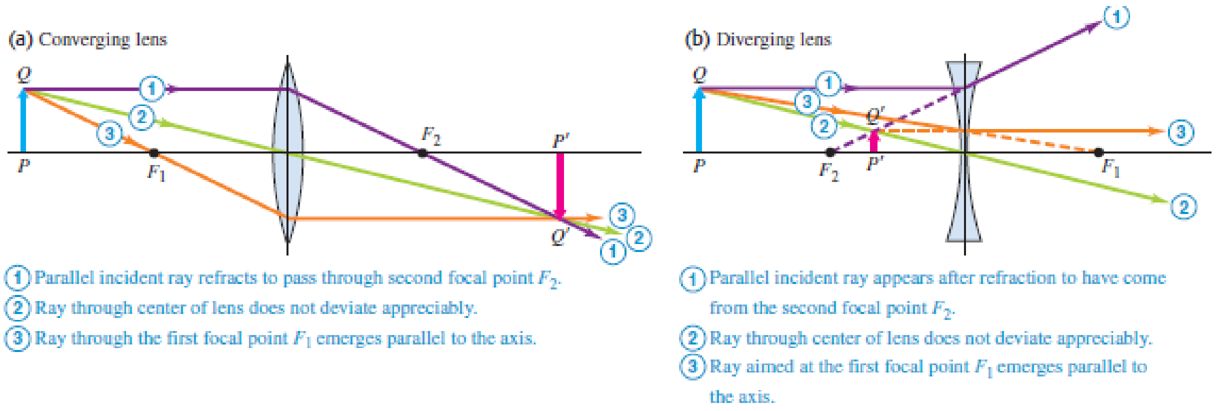
R_2 is negative. (C_2 is on the opposite side from the outgoing light.)
R_1 is positive. (C_1 is on the same side as the outgoing light.)
Radius of curvature of second surface: R_2
Radius of curvature of first surface: R_1
C_2
Q
y
P
n
P'
C_1
y'
Q'
s and s' are positive, so m is negative.
s
s'

Figure 96 - A converging thin lens with a positive focal length *f*

It is not hard to generalize Eq. (184) to the situation in which the lens is immersed in a material with an index of refraction greater than unity. We invite you to work out the lensmaker's equation for this more general situation.

We stress that the paraxial approximation is indeed an approximation! Rays that are at sufficiently large angles to the optic axis of a spherical lens will not be brought to the same focus as paraxial rays; this is the same problem of spherical aberration that plagues spherical *mirrors*. To avoid this and other limitations of thin spherical lenses, lenses of more complicated shape are used in precision optical instruments.

Graphical Methods for Lenses. We can determine the position and size of an image formed by a thin lens by using a graphical method very similar to the one we used in early for spherical mirrors. Again we draw a few special rays called *principal rays* that diverge from a point of the object that is *not* on the optic axis. The intersection of these rays, after they pass through the lens, determines the position and size of the image. In using this graphical method, we will consider the entire deviation of a ray as occurring at the midplane of the lens, as shown in Fig. 97. This is consistent with the assumption that the distance between the lens surfaces is negligible.



(a)
1 Parallel incident ray refracts to pass through second focal point F_2.
2 Ray through center of lens does not deviate appreciably.
3 Ray through the first focal point F_1 emerges parallel to the axis.

(b)
1 Parallel incident ray appears after refraction to have come from the second focal point F_2.
2 Ray through center of lens does not deviate appreciably.
3 Ray aimed at the first focal point F_1 emerges parallel to the axis.

Figure 97 - The graphical method of locating an image formed by a thin lens. The colours of the rays are for identification only; they do not refer to specific colours of light

The three principal rays whose paths are usually easy to trace for lenses are shown in Fig. 97:

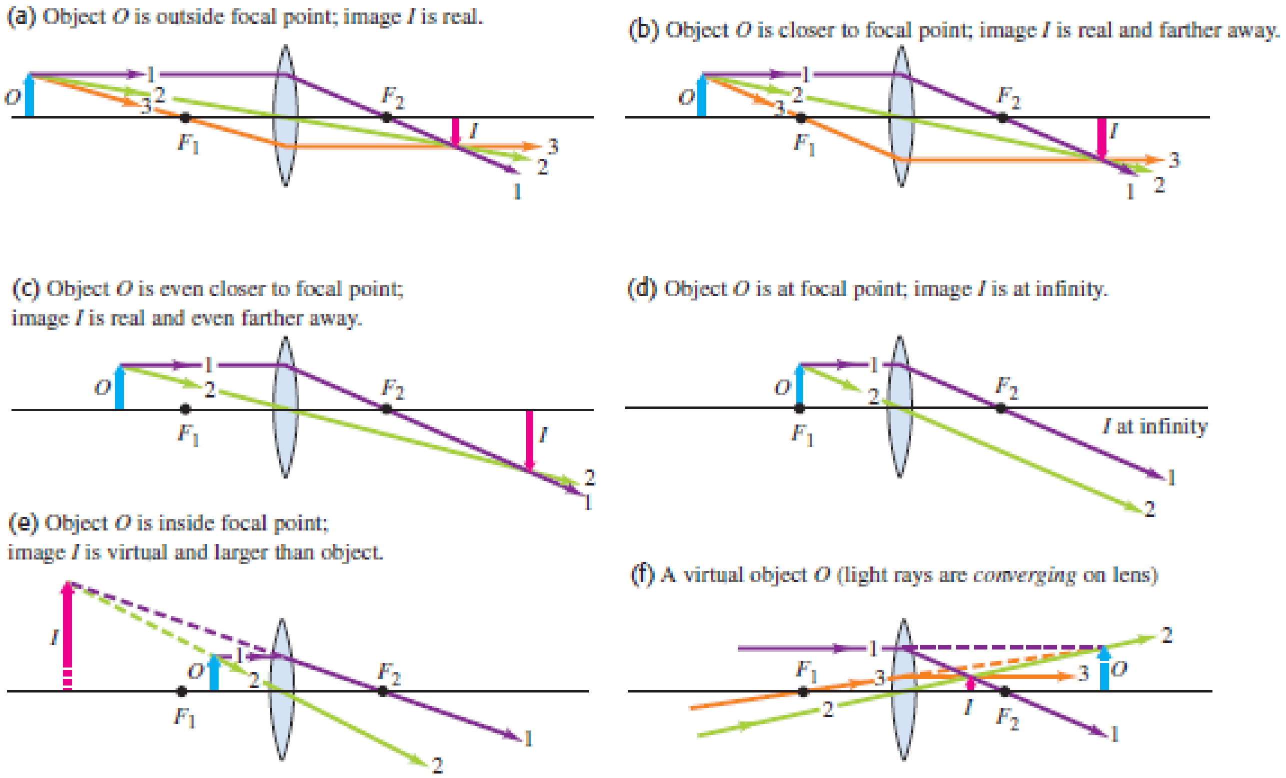
1. *A ray parallel to the axis* emerges from the lens in a direction that passes through the second focal point F_2 of a converging lens, or appears to come from the second focal point of a diverging lens.

2. *A ray through the center of the lens* is not appreciably deviated; at the center of the lens the two surfaces are parallel, so this ray emerges at essentially the same angle at which it enters and along essentially the same line.

3. *A ray through (or proceeding toward) the first focal point F_1* emerges parallel to the axis.

When the image is real, the position of the image point is determined by the intersection of any two rays 1, 2, and 3 (Fig. 97a). When the image is virtual, we extend the diverging outgoing rays backward to their intersection point to find the image point (Fig. 97b).

Figure 98 shows principal-ray diagrams for a converging lens for several object distances. We suggest you study each of these diagrams very carefully, comparing each numbered ray with the above description.


(a) Object *O* is outside focal point; image *I* is real.

(b) Object *O* is closer to focal point; image *I* is real and farther away.

(c) Object *O* is even closer to focal point; image *I* is real and even farther away.

(d) Object *O* is at focal point; image *I* is at infinity.

(e) Object *O* is inside focal point; image *I* is virtual and larger than object.

(f) A virtual object *O* (light rays are *converging* on lens)

Figure 98 – Formation of images by a thin converging lens for various object distances. The principal rays are numbered

Parts (a), (b), and (c) of Fig. 98 help explain what happens in focusing a camera. For a photograph to be in sharp focus, the film must be at the position of the real image made by the camera’s lens. The image distance increases as the object is brought closer, so the film is moved farther behind the lens (i.e., the lens is moved farther in front of the film). In Fig. 98d the object is at the focal point; ray 3 can’t be drawn because it doesn’t pass through the lens. In Fig. 98e the object distance is less than the focal length. The outgoing rays are divergent, and the image is *virtual;* its position is located by extending the outgoing rays backward, so the image distance *s’* is negative. Note also that the image is erect and larger than the object. Figure 98f corresponds to a *virtual object*. The incoming rays do not diverge from a real object, but are *converging* as though they would meet at the tip of the virtual object *O* on the right side; the object distance *s* is negative in this case. The image is real and is located between the lens and the second focal point. This situation can arise if the rays that strike the lens in Fig. 98f emerge from another converging lens (not shown) to the left of the figure.

Discussion questions

1. A spherical mirror is cut in half horizontally. Will an image be formed by the bottom half of the mirror? If so, where will the image be formed?
2. The laws of optics also apply to electromagnetic waves invisible to the eye. A satellite TV dish is used to detect radio waves coming from orbiting satellites. Why is a curved reflecting surface (a “dish”) used? The dish is always

concave, never convex; why? The actual radio receiver is placed on an arm and suspended in front of the dish. How far in front of the dish should it be placed?

3. Explain why the focal length of a *plane* mirror is infinite, and explain what it means for the focal point to be at infinity.
4. If a spherical mirror is immersed in water, does its focal length change? Explain.
5. For what range of object positions does a concave spherical mirror form a real image? What about a convex spherical mirror?
6. When a room has mirrors on two opposite walls, an infinite series of reflections can be seen. Discuss this phenomenon in terms of images. Why do the distant images appear fainter?
7. For a spherical mirror, if $s=f$, then $s'=\infty$ and the lateral magnification is infinite. Does this make sense? If so, what does it mean?
8. You may have noticed a small convex mirror next to your bank's ATM. Why is this mirror convex, as opposed to flat or concave? What considerations determine its radius of curvature?
9. A student claims that she can start a fire on a sunny day using just the sun's rays and a concave mirror. How is this done? Is the concept of image relevant? Can she do the same thing with a convex mirror? Explain.
10. A person looks at his reflection in the concave side of a shiny spoon. Is it right side up or inverted? Does it matter how far his face is from the spoon? What if he looks in the convex side? (Try this yourself!)
11. The bottom of the passenger-side mirror on your car notes, "Objects in mirror are closer than they appear." Is this true? Why?
12. How could you very quickly make an approximate measurement of the focal length of a converging lens? Could the same method be applied if you wished to use a diverging lens? Explain.
13. The focal length of a simple lens depends on the color (wavelength) of light passing through it. Why? Is it possible for a lens to have a positive focal length for some colors and negative for others? Explain.
14. When a converging lens is immersed in water, does its focal length increase or decrease in comparison with the value in air? Explain.
15. A spherical air bubble in water can function as a lens. Is it a converging or diverging lens? How is its focal length related to its radius?
16. Can an image formed by one reflecting or refracting surface serve as an object for a second reflection or refraction? Does it matter whether the first image is real or virtual? Explain.
17. If a piece of photographic film is placed at the location of a real image, the film will record the image. Can this be done with a virtual image? How might one record a virtual image?
18. You've entered a survival contest that will include building a crude telescope. You are given a large box of lenses. Which two lenses do you pick? How do you quickly identify them?

19. You can’t see clearly underwater with the naked eye, but you *can* if you wear a face mask or goggles (with air between your eyes and the mask or goggles). Why is there a difference? Could you instead wear eyeglasses (with water between your eyes and the eyeglasses) in order to see underwater? If so, should the lenses be converging or diverging? Explain.
20. You take a lens and mask it so that light can pass through only the bottom half of the lens. How does the image formed by the masked lens compare to the image formed before masking?

2.2 Interference

2.2.1 Interference and coherent sources

As we discussed early, the term **interference** refers to any situation in which two or more waves overlap in space. When this occurs, the total wave at any point at any instant of time is governed by the **principle of superposition.**. This principle also applies to electromagnetic waves and is the most important principle in all of physical optics. The principle of superposition states:

When two or more waves overlap, the resultant displacement at any point and at any instant is found by adding the instantaneous displacements that would be produced at the point by the individual waves if each were present alone.

(In some special situations, such as electromagnetic waves propagating in a crystal, this principle may not apply. A discussion of these is beyond our scope.)

We use the term “displacement” in a general sense. With waves on the surface of a liquid, we mean the actual displacement of the surface above or below its normal level. With sound waves, the term refers to the excess or deficiency of pressure. For electromagnetic waves, we usually mean a specific component of electric or magnetic field.

Interference in Two or Three Dimensions. We have already discussed one important case of interference, in which two identical waves propagating in opposite directions combine to produce a *standing wave.* We saw this in Chapters 15 and 16 for transverse waves on a string and for longitudinal waves in a fluid filling a pipe; we described the same phenomenon for electromagnetic waves early. In all of these cases the waves propagated along only a single axis: along a string, along the length of a fluid-filled pipe, or along the propagation direction of an electromagnetic plane wave. But light waves can (and do) travel in *two* or *three* dimensions, as can any kind of wave that propagates in a two- or three-dimensional medium. In this section we’ll see what happens when we combine waves that spread out in two or three dimensions from a pair of identical wave sources.

Interference effects are most easily seen when we combine *sinusoidal* waves with a single frequency f and wavelength λ. Figure 99 shows a “snapshot” of a *single* source S_1 of sinusoidal waves and some of the wave fronts produced by this source. The figure shows only the wave fronts corresponding to wave *crests,* so the spacing between successive wave fronts is one wavelength. The material surrounding S_1 is uniform, so the wave speed is the same in all directions, and there is no refraction (and hence no bending of the wave fronts). If the waves are two-dimensional, like waves on the surface of a liquid, the circles in Fig. 99 represent circular wave fronts; if the waves propagate in three dimensions, the circles represent spherical wave fronts spreading away from S_1.

In optics, sinusoidal waves are characteristic of **monochromatic light** (light of a single colour). While it’s fairly easy to make water waves or sound waves of a single frequency, common sources of light *do not* emit monochromatic (single-frequency) light. For example, incandescent light bulbs and flames emit a continuous distribution of wavelengths. By far the most nearly monochromatic source that is available at present is the *laser.* An example is the helium–neon laser, which emits red light at 632.8 nm with a wavelength range of the order of or about one part in 10^9. As we analyze interference and diffraction effects in this chapter and the next, we will assume that we are working with monochromatic waves (unless we explicitly state otherwise).

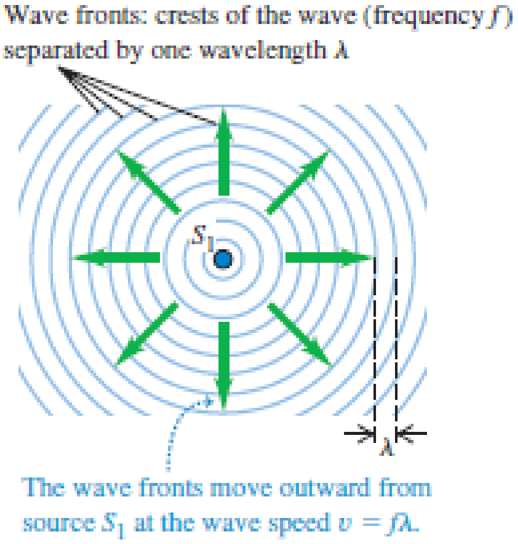
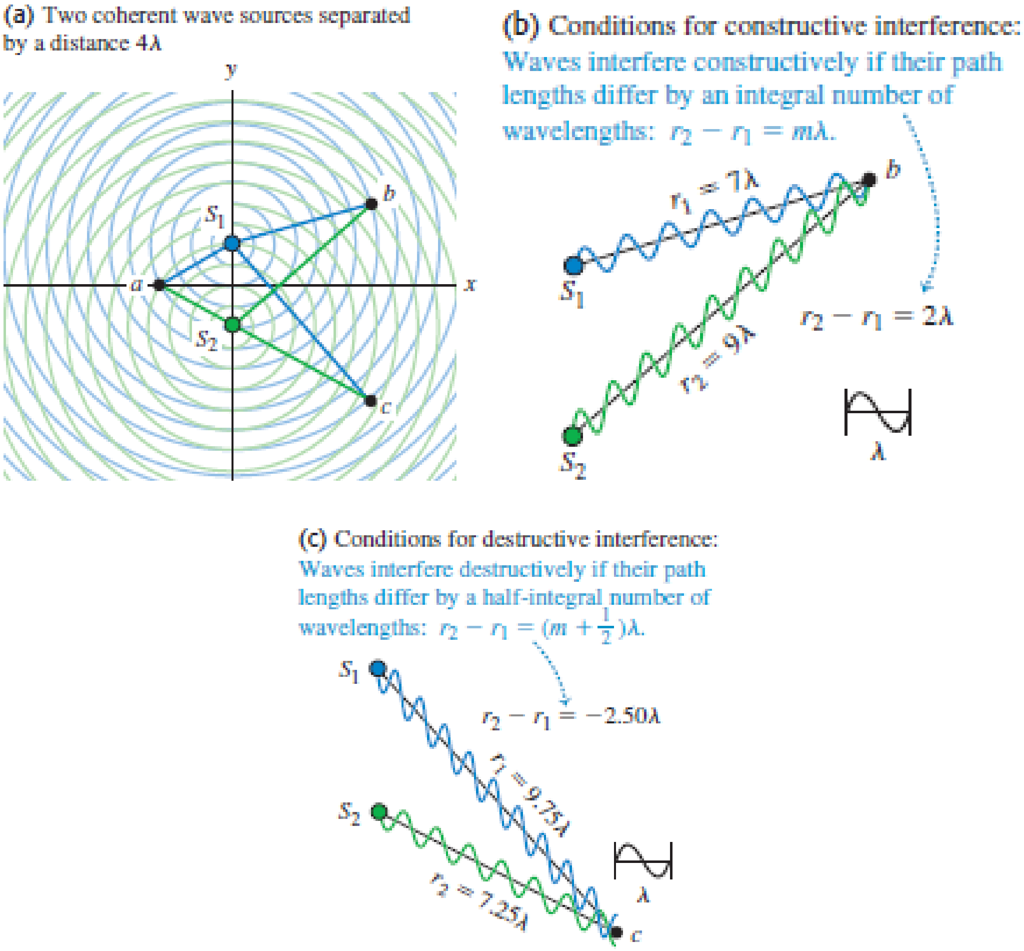

Wave fronts: crests of the wave (frequency f) separated by one wavelength λ
S_1
λ
The wave fronts move outward from source S_1 at the wave speed $v = f\lambda$.

Figure 99 - A “snapshot” of sinusoidal waves of frequency f and wavelength spreading out from source S_1 in all directions

Constructive and Destructive Interference. Two identical sources of monochromatic waves, and are shown in Fig. 100a. The two sources produce waves of the same amplitude and the same wavelength l. In addition, the two sources are permanently *in phase;* they vibrate in unison. They might be two loudspeakers driven by the same amplifier, two radio antennas powered by the same transmitter, or two small slits in an opaque screen, illuminated by the same monochromatic light source. We will see that if there were not a constant phase relationship between the two sources, the phenomena we are about to discuss would not occur. Two monochromatic sources of the same frequency and with a constant phase relationship (not necessarily in phase) are said to be **coherent.** We also use the term *coherent* *waves* (or, for light waves, *coherent light*) to refer to the waves emitted by two such sources.

If the waves emitted by the two coherent sources are *transverse*, like electromagnetic waves, then we will also assume that the wave disturbances produced by both sources have the same *polarization* (that is, they lie along the same line). For example, the sources S_1 and S_2 in Fig. 100a could be two radio antennas in the form of long rods oriented parallel to the *z*-axis (perpendicular to the plane of the figure); at any point in the *xy*-plane the waves produced by both antennas have $\vec{E}$ fields with only a *z*-component. Then we need only a single scalar function to describe each wave; this makes the analysis much easier.

We position the two sources of equal amplitude, equal wavelength, and (if the waves are transverse) the same polarization along the *y*-axis in Fig. 100a, equidistant from the origin. Consider a point *a* on the *y*-axis. From symmetry the two distances from S_1 to *a* and from S_2 to *a* are *equal;* waves from the two sources thus require equal times to travel to *a*. Hence waves that leave S_1 and S_2 in phase arrive at *a* in phase. The two waves add, and the total amplitude at *a* is *twice* the amplitude of each individual wave. This is true for *any* point on the *x*-axis.


(a) Two coherent wave sources separated by a distance 4λ

(b) Conditions for constructive interference: Waves interfere constructively if their path lengths differ by an integral number of wavelengths: $r_2 - r_1 = m\lambda$.

(c) Conditions for destructive interference: Waves interfere destructively if their path lengths differ by a half-integral number of wavelengths: $r_2 - r_1 = (m + \frac{1}{2})\lambda$.

Figure 100 - (a) A “snapshot” of sinusoidal waves spreading out from two coherent sources S_1 and S_2. Constructive interference occurs at point a (equidistant from the two sources) and (b) at point *b* (c) Destructive interference occurs at point *c*.

Similarly, the distance from S_2 to point b is exactly two wavelengths *greater* than the distance from S_2 to b. A wave crest from S_1 arrives at b exactly two cycles earlier than a crest emitted at the same time from S_1, and again the two waves arrive in phase. As at point a, the total amplitude is the sum of the amplitudes of the waves from S_1 and S_2.

In general, when waves from two or more sources arrive at a point *in phase*, they reinforce each other: The amplitude of the resultant wave is the *sum* of the amplitudes of the individual waves. This is called **constructive interference** (Fig. 100b). Let the distance from S_1 to any point *P* be r_1, and let the distance from S_2 to *P* be r_2. For constructive interference to occur at *P*, the path difference $r_2 - r_1$ for the two sources must be an integral multiple of the wavelength λ:

$$r_2 - r_1 = m\lambda, \qquad (m = 0, \pm 1, \pm 2, \pm 3 \ldots) \tag{185}$$

In Fig. 100a, points *a* and *b* satisfy Eq. (185) with $m = 0$ and $m = +2$ respectively.

Something different occurs at point *c* in Fig. 100a. At this point, the path difference $r_2 - r_1 = 2.5\lambda$, which is a *half-integral* number of wavelengths. Waves from the two sources arrive at point exactly a half-cycle out of phase. A crest of one wave arrives at the same time as a crest in the opposite direction (a “trough”) of the other wave (Fig. 100c). The resultant amplitude is the *difference* between the two individual amplitudes. If the individual amplitudes are equal, then the total amplitude is *zero*! This cancellation or partial cancellation of the individual waves is called **destructive interference.** The condition for destructive interference in the situation shown in Fig. 100a is

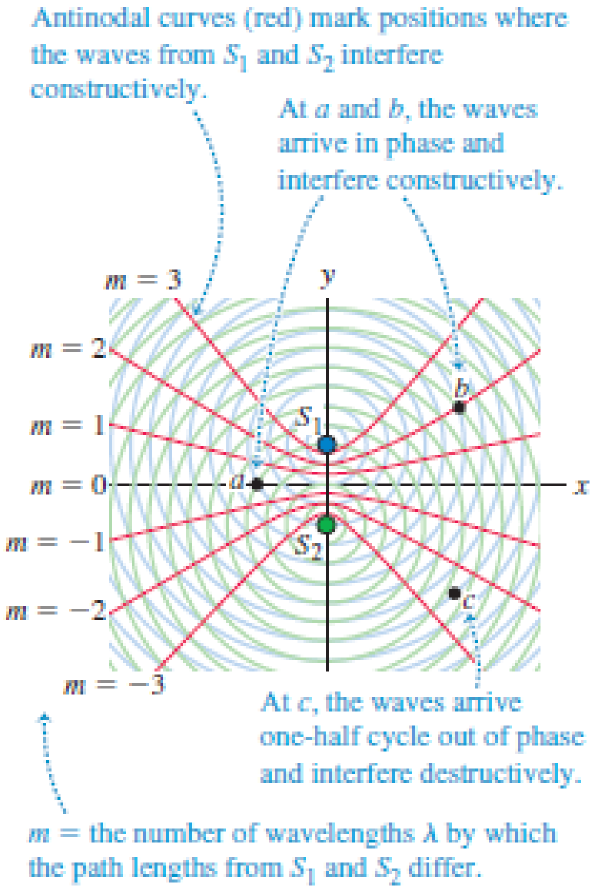
$$r_2 - r_1 = \left(m + \frac{1}{2}\right)\lambda, \qquad (m = 0, \pm 1, \pm 2, \pm 3 \ldots) \tag{186}$$

The path difference at point *c* in Fig. 100a satisfies Eq. (186) with $m = -3$.

Figure 101 shows the same situation as in Fig. 100a, but with red curves that show all positions where *constructive* interference occurs. On each curve, the path difference $r_2 - r_1$ is equal to an integer times the wavelength, as in Eq. (185). These curves are called **antinodal curves.** In a standing wave formed by interference between waves propagating in opposite directions, the antinodes are points at which the amplitude is maximum; likewise, the wave amplitude in the situation of Fig. 101 is maximum along the antinodal curves. Not shown in Fig. 101 are the **nodal curves,** which are the curves that show where *destructive* interference occurs in accordance with Eq. (186); these are analogous to the *nodes* in a standing-wave pattern. A nodal curve lies between each two adjacent antinodal curves in Fig. 101; one such curve, corresponding to $r_2 - r_1 = -2.5\lambda$ passes through point

In some cases, such as two loudspeakers or two radio-transmitter antennas, the interference pattern is three-dimensional. Think of rotating the colour curves of Fig. 101 around the *y*-axis; then maximum constructive interference occurs at all points on the resulting surfaces of revolution.

For Eqs. (185) and (186) to hold, the two sources must have the same wavelength and must *always* be in phase. These conditions are rather easy to satisfy for sound waves. But with *light* waves there is no practical way to achieve a constant phase relationship (coherence) with two independent sources. This is because of the way light is emitted. In ordinary light sources, atoms gain excess energy by thermal agitation or by impact with accelerated electrons. Such an "excited" atom begins to radiate energy and continues until it has lost all the energy it can, typically in a time of the order of $10^{-8}\,s$. The many atoms in a source ordinarily radiate in an unsynchronized and random phase relationship, and the light that is emitted from *two* such sources has no definite phase relationship.



Antinodal curves (red) mark positions where the waves from S_1 and S_2 interfere constructively.

At a and b, the waves arrive in phase and interfere constructively.

At c, the waves arrive one-half cycle out of phase and interfere destructively.

m = the number of wavelengths λ by which the path lengths from S_1 and S_2 differ.

Figure 101 - The same as Fig. 100a, but with red antinodal curves (curves of maximum amplitude) superimposed. All points on each curve satisfy Eq. (185) with the value of m shown. The nodal curves (not shown) lie between each adjacent pair of antinodal curves.

However, the light from a single source can be split so that parts of it emerge from two or more regions of space, forming two or more *secondary sources.* Then any random phase change in the source affects these secondary sources equally and does not change their *relative* phase.

The distinguishing feature of light from a *laser* is that the emission of light from many atoms is synchronized in frequency and phase. As a result, the random phase changes mentioned above occur much less frequently. Definite phase relationships are preserved over correspondingly much greater lengths in the beam, and laser light is much more coherent than ordinary light.

2.2.2 Two-sources interference of the light

The interference pattern produced by two coherent sources of *water* waves of the same wavelength can be readily seen in a ripple tank with a shallow layer of water (Fig. 102). This pattern is not directly visible when the interference is between *light* waves, since light traveling in a uniform medium cannot be seen. (A shaft of afternoon sunlight in a room is made visible by scattering from airborne dust particles.)

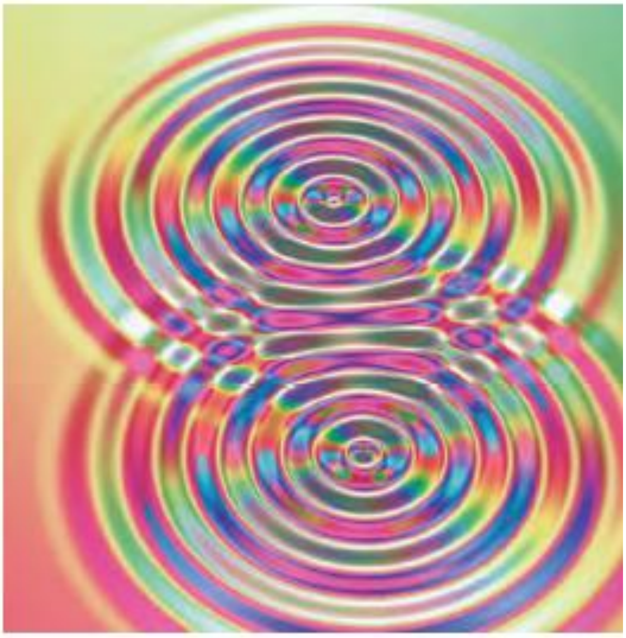


Figure 102 - The concepts of constructive interference and destructive interference apply to these water waves as well as to light waves and sound waves

One of the earliest quantitative experiments to reveal the interference of light from two sources was performed in 1800 by the English scientist Thomas Young. We will refer back to this experiment several times in this and later chapters, so it's important to understand it in detail. Young's apparatus is shown in perspective in Fig. 103a. A light source (not shown) emits monochromatic light; however, this light is not suitable for use in an interference experiment because emissions from different parts of an ordinary source are not synchronized. To remedy this, the light is directed at a screen with a narrow slit $S_0, 1\ \mu m$ or so wide. The light emerging from the slit originated from only a small region of the light source; thus slit behaves more nearly like the idealized source shown in Fig. 99. (In modern versions of the experiment, a laser is used as a source of coherent light, and the slit S_0 isn't needed.) The light from slit S_0 falls on a screen with two other narrow slits S_1 and S_2, each $1\ \mu m$ or so wide and a few tens or hundreds of micrometers apart. Cylindrical wave fronts spread out from slit S_0 and reach slits S_1 and S_2 *in phase* because they travel equal distances from S_0. The waves *emerging* from slits S_1 and S_2 are therefore also always in phase, so S_1 and S_2 are *coherent* sources. The interference of waves from S_1 and S_2 produces a pattern in space like that to the right of the sources in Figs. 100a and 101.

To visualize the interference pattern, a screen is placed so that the light from S_1 and S_2 falls on it (Fig. 103b). The screen will be most brightly illuminated at points P, where the light waves from the slits interfere constructively, and will be darkest at points where the interference is destructive.

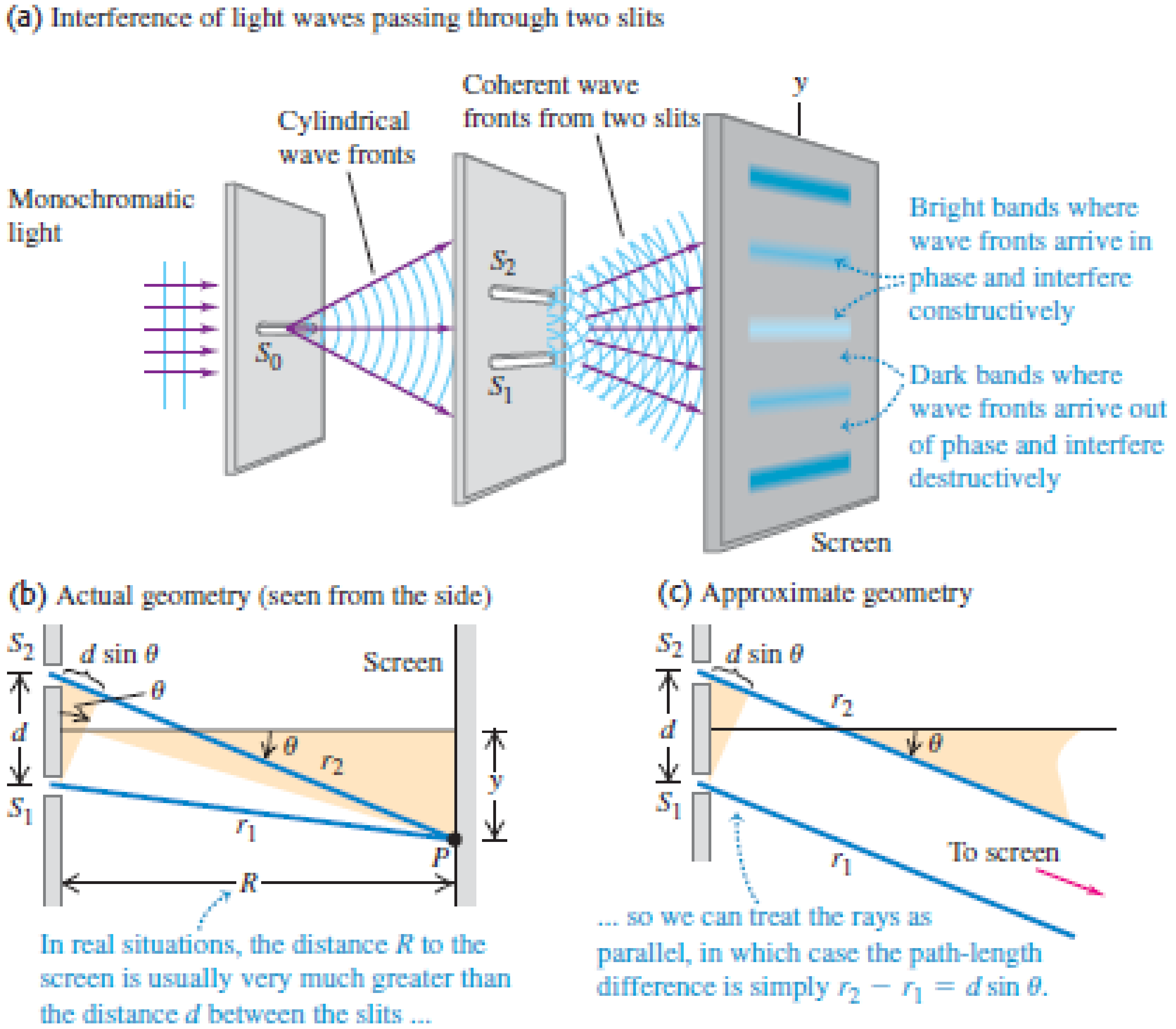

(a) Interference of light waves passing through two slits
Coherent wave fronts from two slits
y
Cylindrical wave fronts
Monochromatic light
S_2
S_0
S_1
Bright bands where wave fronts arrive in phase and interfere constructively
Dark bands where wave fronts arrive out of phase and interfere destructively
Screen
(b) Actual geometry (seen from the side)
S_2
$d \sin \theta$
θ
Screen
d
θ
r_2
y
S_1
r_1
P
R
In real situations, the distance R to the screen is usually very much greater than the distance d between the slits ...
(c) Approximate geometry
S_2
$d \sin \theta$
r_2
d
θ
S_1
To screen
r_1
... so we can treat the rays as parallel, in which case the path-length difference is simply $r_2 - r_1 = d \sin \theta$.

Figure 103 - (a) Young's experiment to show interference of light passing through two slits. A pattern of bright and dark areas appears on the screen (see Fig. 104). (b) Geometrical analysis of Young's experiment. For the case shown, $r_2 > r_1$ and both y and θ are positive. If point P is on the other side of the screen's center, $r_2 < r_1$ and both y and θ are negative. (c) Approximate geometry when the distance R to the screen is much greater than the distance d between the slits

To simplify the analysis of Young's experiment, we assume that the distance R from the slits to the screen is so large in comparison to the distance d between the slits that the lines from S_1 and S_2 to P are very nearly parallel, as in Fig. 103c. This is usually the case for experiments with light; the slit separation is typically a few millimeters, while the screen may be a meter or more away. The differencein path length is then given by

$$r_2 - r_1 = d \sin\theta \tag{187}$$

where θ is the angle between a line from slits to screen (shown in blue in Fig. 103c) and the normal to the plane of the slits (shown as a thin black line).

Constructive and Destructive Two-Slit Interference. We found early that constructive interference (reinforcement) occurs at points where the path difference is an integral number of wavelengths, $m\lambda$, where $m = 0, \pm 1, \pm 2, \pm 3$ .... So the bright regions on the screen in Fig. 103a occur at angles θ for which

$$d \sin\theta = m\lambda, \quad (m = 0, \pm 1, \pm 2, \pm 3 \ldots.) \tag{188}$$

Similarly, destructive interference (cancellation) occurs, forming dark regions on the screen, at points for which the path difference is a half-integral number of wavelengths, :

$$d \sin\theta = \left(m + \frac{1}{2}\right)\lambda, \quad (m = 0, \pm 1, \pm 2, \pm 3 \ldots.) \tag{189}$$

Thus the pattern on the screen of Figs. 103a and 103b is a succession of bright and dark bands, or **interference fringes,** parallel to the slits S_1 and S_2. A photograph of such a pattern is shown in Fig. 104. The center of the pattern is a bright band corresponding to $m = 0$ in Eq. (188); this point on the screen is equidistant from the two slits.

We can derive an expression for the positions of the centers of the bright bands on the screen. In Fig. 103b, is measured from the center of the pattern, corresponding to the distance from the center of Fig. 104. Let y_m be the distance from the center of the pattern (θ=0) to the center of the bright band. Let θ_m be the corresponding value of θ; then

$$y_m = R \tan\theta_m \tag{190}$$

In experiments such as this, the distances y_m are often much smaller than the distance R from the slits to the screen. Hence θ_m is very small, $\tan\theta_m$ is very nearly equal to $\sin\theta_m$, and

$$y_m = R \sin\theta_m \tag{191}$$

Combining this with Eq. (188), we find that *for small angles only,*

$$y_m = R\frac{m\lambda}{d} \tag{192}$$

We can measure R and d, as well as the positions y_m of the bright fringes, so this experiment provides a direct measurement of the wavelength λ. Young's experiment was in fact the first direct measurement of wavelengths of light.

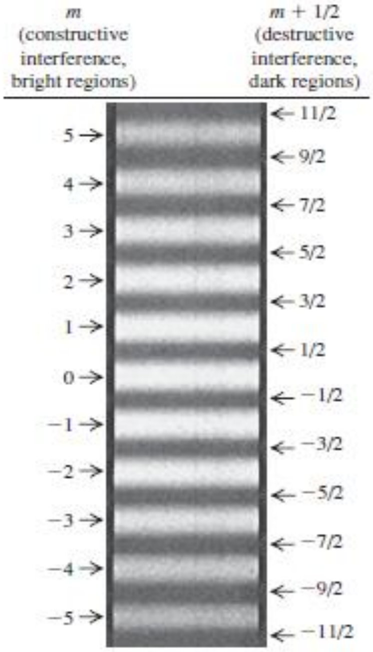


Figure 104 – Photograph of interference fingers produced on a screen in Young's double slit experiment

The distance between adjacent bright bands in the pattern is *inversely* proportional to the distance *d* between the slits. The closer together the slits are, the more the pattern spreads out. When the slits are far apart, the bands in the pattern are closer together.

While we have described the experiment that Young performed with visible light, the results given in Eqs. (188) and (189) are valid for *any* type of wave, provided that the resultant wave from two coherent sources is detected at a point that is far away in comparison to the separation *d*.

Discussion questions

1. A two-slit interference experiment is set up, and the fringes are displayed on a screen. Then the whole apparatus is immersed in the nearest swimming pool. How does the fringe pattern change?
2. Could an experiment similar to Young's two-slit experiment be performed with sound? How might this be carried out? Does it matter that sound waves are longitudinal and electromagnetic waves are transverse? Explain.
3. Monochromatic coherent light passing through two thin slits is viewed on a distant screen. Are the bright fringes equally spaced on the screen? If so, why? If not, which ones are closest to being equally spaced?

4. In a two-slit interference pattern on a distant screen, are thebright fringes midway between the dark fringes? Is this ever a good approximation?
5. Would the headlights of a distant car form a two-source interference pattern? If so, how might it be observed? If not, why not?
6. Could the Young two-slit interference experiment be performed with gamma rays? If not, why not? If so, discuss differences in the experimental design compared to the experiment with visible light.
7. Coherent red light illuminates two narrow slits that are 25 cm apart. Will a two-slit interference pattern be observed when the lightfrom the slits falls on a screen? Explain.
8. Coherent light with wavelength λ falls on two narrow slits separated by a distance *d*. If *d* is less than some minimum value, no dark fringes are observed. Explain. In terms λ, of what is this minimum value of *d*.
9. In using the superposition principle to calculate intensities in interference patterns, could you add the intensities of the waves instead of their amplitudes? Explain.
10. A glass windowpane with a thin film of water on it reflects less than when it is perfectly dry. Why?
11. A *very* thin soap film (n=1.33) whose thickness is much less than a wavelength of visible light, looks black; it appears to reflect no light at all. Why? By contrast, an equally thin layer of soapy water (n=1.333) on glass (n=1.5) appears quite shiny. Why is there a difference?
12. Interference can occur in thin films. Why is it important that the films be *thin*? Why don’t you get these effects with a relatively *thick* film? Where should you put the dividing line between “thin” and “thick”? Explain your reasoning.
13. Monochromatic light is directed at normal incidence on a thin film. There is destructive interference for the reflected light, so the intensity of the reflected light is very low. What happened to the energy of the incident light?
14. When a thin oil film spreads out on a puddle of water, the thinnest part of the film looks dark in the resulting interference pattern. What does this tell you about the relative magnitudes of the refractive indexes of oil and water?

2.3 Diffraction

2.3.1 Frensel and Fraunhofer diffraction

Everyone is used to the idea that sound bends around corners. If sound didn’t behave this way, you couldn’t hear a police siren that’s out of sight around a corner or the speech of a person whose back is turned to you. What may surprise you (and certainly surprised many scientists of the early 19th century) is that *light* can bend around corners as well. When light from a point source falls on a straightedge and casts a shadow, the edge of the shadow is never perfectly sharp. Some light appears in the area that we expect to be in the shadow, and we find alternating bright and dark fringes in the illuminated area. In general, light emerging from apertures doesn’t behave precisely according to the predictions of the straight-line ray model of geometric optics.

Light emerging from arrays of apertures also forms patterns whose character depends on the colour of the light and the size and spacing of the apertures. Examples of this effect include the colours of iridescent butterflies and the "rainbow" you see reflected from the surface of a compact disc. We'll explore similar effects with x rays that are used to study the atomic structure of solids and liquids. Finally, we'll look at the physics of a *hologram,* a special kind of interference pattern recorded on photographic film and reproduced. When properly illuminated, it forms a three-dimensional image of the original object.

According to geometric optics, when an opaque object is placed between a point light source and a screen, as in Fig. 105, the shadow of the object forms a perfectly sharp line. No light at all strikes the screen at points within the shadow, and the area outside the shadow is illuminated nearly uniformly. But the *wave* nature of light causes effects that can't be understood with geometric optics. An important class of such effects occurs when light strikes a barrier that has an aperture or an edge. The interference patterns formed in such a situation are grouped under the heading **diffraction.**

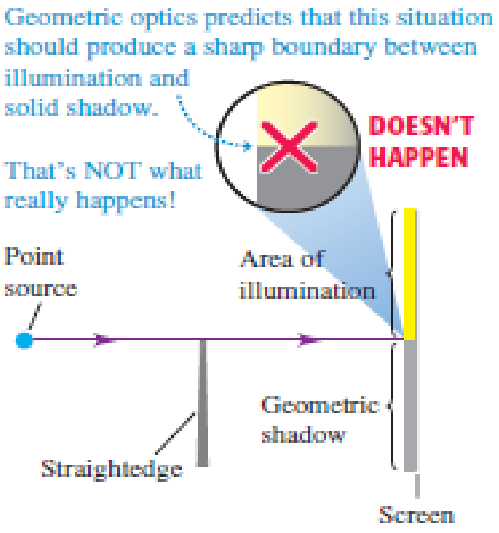

Geometric optics predicts that this situation should produce a sharp boundary between illumination and solid shadow.
DOESN'T HAPPEN
That's NOT what really happens!
Point source
Area of illumination
Geometric shadow
Straightedge
Screen

Figure 105 – A point source of light illuminates a straightedge

Figure 106 shows an example of diffraction. The photograph in Fig. 106a was made by placing a razor blade halfway between a pinhole, illuminated by monochromatic light, and a photographic film. The film recorded the shadow cast by the blade. Figure 106b is an enlargement of a region near the shadow of the right edge of the blade. The position of the *geometric* shadow line is indicated by arrows. The area outside the geometric shadow is bordered by alternating bright and dark bands. There is some light in the shadow region, although this is not very visible in the photograph. The first bright band in Fig. 106b, just to the right of the geometric shadow, is considerably brighter than in the region of uniform illumination to the extreme right. This simple experiment gives us some idea of the richness and complexity of what might seem to be a simple idea, the casting of a shadow by an opaque object.

We don't often observe diffraction patterns such as Fig. 106 in everyday life because most ordinary light sources are neither monochromatic nor point sources. If we use a white frosted light bulb instead of a point source to illuminate the razor blade in Fig. 106, each wavelength of the light from every point of the bulb forms its own diffraction pattern, but the patterns overlap so much that we can't see any individual pattern.

We can analyze diffraction patterns using Huygens's principle. This principle states that we can consider every point of a wave front as a source of secondary wavelets. These spread out in all directions with a speed equal to the speed of propagation of the wave. The position of the wave front at any later time is the *envelope* of the secondary wavelets at that time. To find the resultant displacement at any point, we combine all the individual displacements produced by these secondary waves, using the superposition principle and taking into account their amplitudes and relative phases.

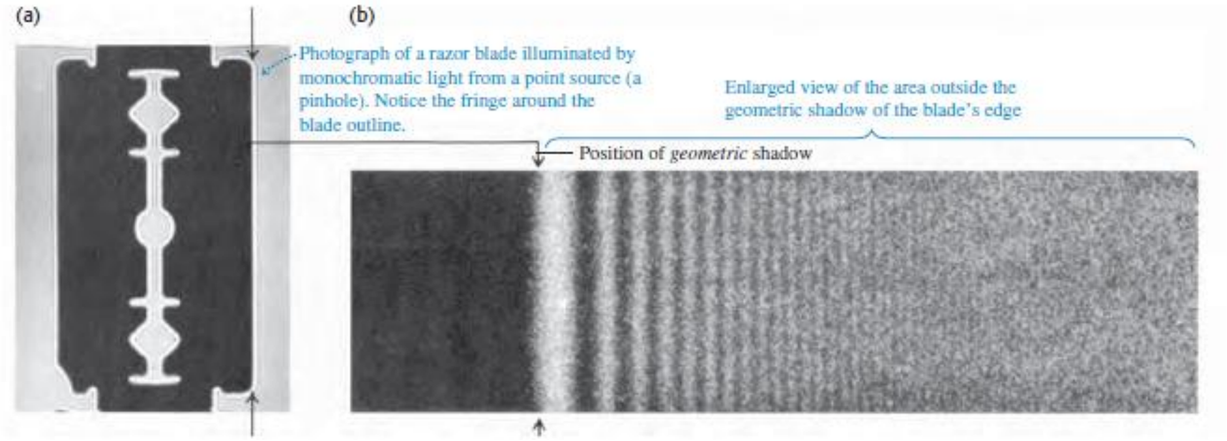

(a)
(b)
Photograph of a razor blade illuminated by monochromatic light from a point source (a pinhole). Notice the fringe around the blade outline.
Enlarged view of the area outside the geometric shadow of the blade's edge
Position of *geometric* shadow

Figure 106 – An example of diffraction

In Fig. 105, both the point source and the screen are relatively close to the obstacle forming the diffraction pattern. This situation is described as *near-field diffraction* or **Fresnel diffraction,** pronounced "Freh-nell" (after the French scientist Augustin Jean Fresnel, 1788–1827). By contrast, we use the term **Fraunhofer diffraction** (after the German physicist Joseph von Fraunhofer, 1787–1826) for situations in which the source, obstacle, and screen are far enough apart that we can consider all lines from the source to the obstacle to be parallel, and can likewise consider all lines from the obstacle to a given point on the screen to be parallel. We will restrict the following discussion to Fraunhofer diffraction, which is usually simpler to analyze in detail than Fresnel diffraction.

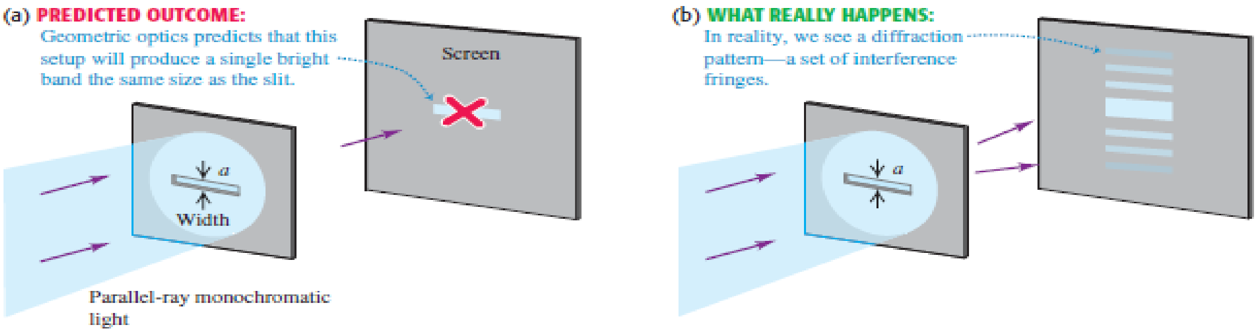
Diffraction is sometimes described as "the bending of light around an obstacle." But the process that causes diffraction is present in the propagation of *every* wave. When part of the wave is cut off by some obstacle, we observe diffraction effects that result from interference of the remaining parts of the wave fronts. Optical instruments typically use only a limited portion of a wave; for example, a telescope uses only the part of a wave that is admitted by its objective lens or mirror. Thus diffraction plays a role in nearly all optical phenomena.

Finally, we emphasize that there is no fundamental distinction between *interference* and *diffraction.* We used the term *interference* for effects involving waves from a small number of sources, usually two. *Diffraction* usually involves a *continuous* distribution of Huygens's wavelets across the area of an aperture, or a very large number of sources or apertures. But both interference and diffraction are consequences of superposition and Huygens's principle.

2.3.2 Diffraction from a single slit

In this section we'll discuss the diffraction pattern formed by plane-wave (parallelray) monochromatic light when it emerges from a long, narrow slit, as shown in Fig. 107. We call the narrow dimension the *width*, even though in this figure it is a vertical dimension.

According to geometric optics, the transmitted beam should have the same cross section as the slit, as in Fig. 107a. What is *actually* observed is the pattern shown in Fig. 107b. The beam spreads out vertically after passing through the slit. The diffraction pattern consists of a central bright band, which may be much broader than the width of the slit, bordered by alternating dark and bright bands with rapidly decreasing intensity. About 85% of the power in the transmitted beam is in the central bright band, whose width is *inversely* proportional to the width of the slit. In general, the smaller the width of the slit, the broader the entire diffraction pattern. (The *horizontal* spreading of the beam in Fig. 107b is negligible because the horizontal dimension of the slit is relatively large.) You can observe a similar diffraction pattern by looking at a point source, such as a distant street light, through a narrow slit formed between your two thumbs held in front of your eye; the retina of your eye corresponds to the screen.



(a) PREDICTED OUTCOME:
Geometric optics predicts that this setup will produce a single bright band the same size as the slit.

(b) WHAT REALLY HAPPENS:
In reality, we see a diffraction pattern—a set of interference fringes.

Figure 107 - (a) The "shadow" of a horizontal slit as incorrectly predicted by geometric optics. (b) A horizontal slit actually produces a diffraction pattern. The slit width has been greatly exaggerated

Single-Slit Diffraction: Locating the Dark Fringes. [Figure 108 shows a side view of the same setup; the long sides of the slit are perpendicular to the figure, and plane waves are incident on the slit from the left. According to Huygens's principle, each element of area of the slit opening can be considered as a source of secondary waves. In particular, imagine dividing the slit into several narrow strips of equal width, parallel to the long edges and perpendicular to the page. Figure 108a shows two such strips. Cylindrical secondary wavelets, shown in cross section, spread out from each strip.

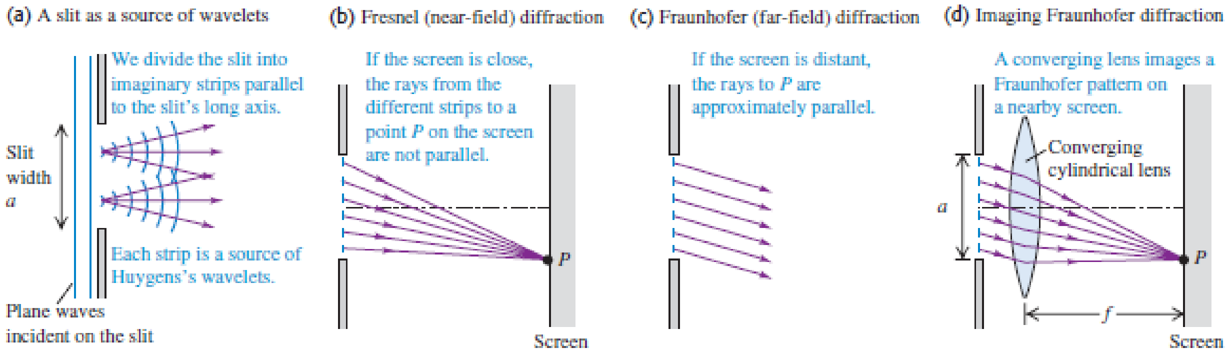

(a) A slit as a source of wavelets
We divide the slit into imaginary strips parallel to the slit's long axis.
Slit width a
Each strip is a source of Huygens's wavelets.
Plane waves incident on the slit
(b) Fresnel (near-field) diffraction
If the screen is close, the rays from the different strips to a point P on the screen are not parallel.
P
Screen
(c) Fraunhofer (far-field) diffraction
If the screen is distant, the rays to P are approximately parallel.
(d) Imaging Fraunhofer diffraction
A converging lens images a Fraunhofer pattern on a nearby screen.
Converging cylindrical lens
a
P
f
Screen

Figure 108 - Diffraction by a single rectangular slit. The long sides of the slit are perpendicular to the figure

In Fig. 108b a screen is placed to the right of the slit. We can calculate the resultant intensity at a point P on the screen by adding the contributions from the individual wavelets, taking proper account of their various phases and amplitudes. It's easiest to do this calculation if we assume that the screen is far enough away that all the rays from various parts of the slit to a particular point P on the screen are parallel, as in Fig. 108c. An equivalent situation is Fig. 108d, in which the rays to the lens are parallel and the lens forms a reduced image of the same pattern that would be formed on an infinitely distant screen without the lens. We might expect that the various light paths through the lens would introduce additional phase shifts, but in fact it can be shown that all the paths have *equal* phase shifts, so this is not a problem.

The situation of Fig. 108b is Fresnel diffraction; those in Figs. 108c and 108d, where the outgoing rays are considered parallel, are Fraunhofer diffraction. We can derive quite simply the most important characteristics of the Fraunhofer diffraction pattern from a single slit. First consider two narrow strips, one just below the top edge of the drawing of the slit and one at its center, shown in end view in Fig. 109. The difference in path length to point P is $(a/2)\sin\theta$, where a is the slit width and θ is the angle between the perpendicular to the slit and a line from the center of the slit to P. Suppose this path difference happens to be equal to $\lambda/2$; then light from these two strips arrives at point P with a halfcycle phase difference, and cancellation occurs.

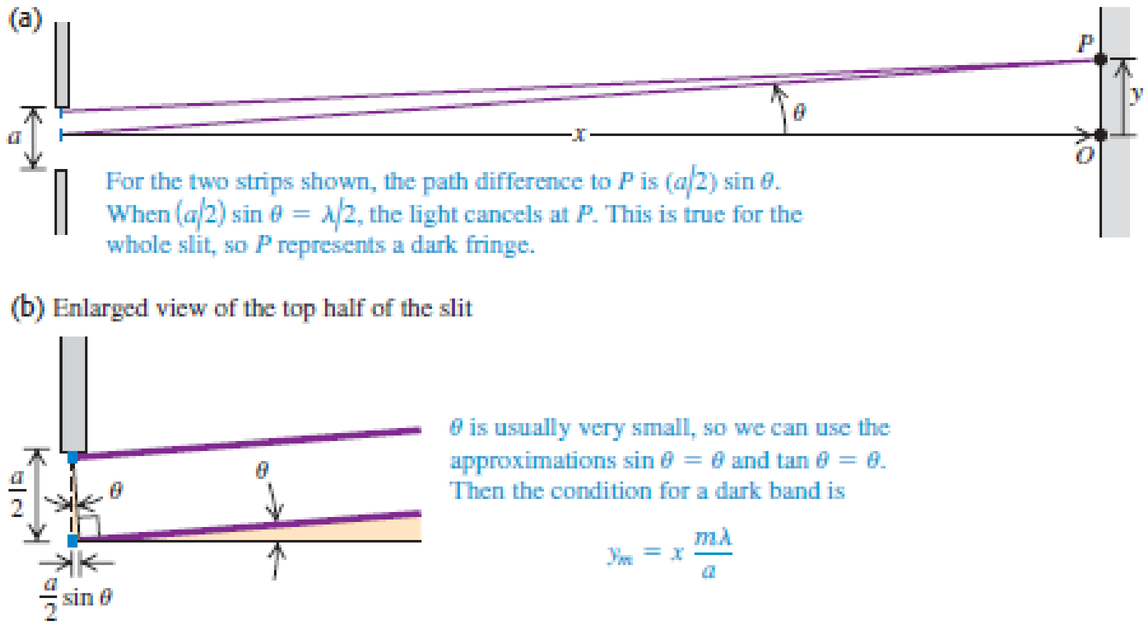
Similarly, light from two strips immediately *below* the two in the figure also arrives at P a half-cycle out of phase. In fact, the light from *every* strip in the top half of the slit cancels out the light from a corresponding strip in the bottom half. Hence the combined light from the entire slit completely cancels at P, giving a dark fringe in the interference pattern. A dark fringe occurs whenever

$$\frac{a}{2}\sin\theta = \pm\frac{\lambda}{2}$$

or (193)

$$\sin\theta = \pm\frac{\lambda}{a}$$

The plus-or-minus sign in Eq. (193) says that there are symmetric dark fringes above and below point O in Fig. 109a. The upper fringe $(\theta > 0)$occurs at a point P where light from the bottom half of the slit travels $\lambda/2$ farther to P than does light from the top half; the lower fringe $(\theta < 0)$ occurs where light from the *top* half travels $\lambda/2$ farther than light from the *bottom* half.



For the two strips shown, the path difference to P is $(a/2)\sin\theta$. When $(a/2)\sin\theta = \lambda/2$, the light cancels at P. This is true for the whole slit, so P represents a dark fringe.

θ is usually very small, so we can use the approximations $\sin\theta = \theta$ and $\tan\theta = \theta$. Then the condition for a dark band is

$$y_m = x\frac{m\lambda}{a}$$

Figure 109 - Side view of a horizontal slit. When the distance to the screen is much greater than the slit width the rays from a distance apart may be considered parallel

We may also divide the screen into quarters, sixths, and so on, and use the above argument to show that a dark fringe occurs whenever $\sin\theta = \pm 2\frac{\lambda}{a}, \pm 3\frac{\lambda}{a}$, and so on. Thus the condition for a *dark* fringe is

$$\sin\theta = \frac{m\lambda}{a}, \quad (m = 0, \pm 1, \pm 2, \pm 3 \ldots) \tag{194}$$

For example, if the slit width is equal to ten wavelengths $(a = 10\lambda)$ dark fringes occur at $\theta = \pm\frac{1}{10}, \pm\frac{1}{10}, \pm\frac{3}{10}, \ldots\ldots$ Between the dark fringes are bright fringes. We also note that $\sin\theta = 0$ corresponds to a *bright* band; in this case, light from the entire slit arrives at P in phase. Thus it would be wrong to put $m = 0$ in Eq. (194). The central bright fringe is wider than the other bright fringes, as Fig. 107b shows. In the small-angle approximation that we will use below, it is exactly *twice* as wide.

With light, the wavelength is of the order of $500\,nm = 5 \times 10^{-7}m$. This is often much smaller than the slit width a typical slit width is $10^{-2}cm = 10^{-4}m$. Therefore the values of θ in Eq. (194) are often so small that the approximation $\sin\theta \approx \theta$ (where θ is in radians) is a very good one. In that case we can rewrite this equation as

$$\theta = \frac{m\lambda}{a}, \quad (m = 0, \pm1, \pm2, \pm3 \dots) \tag{195}$$

where θ is in *radians.* Also, if the distance from slit to screen is x, as in Fig. 109a, and the vertical distance of the mth dark band from the center of the pattern is y_m, then $\tan\theta = y_m/x$. For small θ we may also approximate $\tan\theta$ by θ (in radians), and we then find

$$y_m = x\frac{m\lambda}{a} \tag{196}$$

Figure 110 is a photograph of a single-slit diffraction pattern with the $m = \pm1, \pm2,$ and ±3 minima labeled.

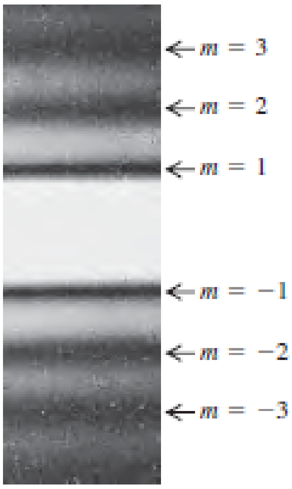


Figure 110 - Photograph of the Fraunhofer diffraction pattern of a single horizontal slit

Discussion questions

1. Why can we readily observe diffraction effects for sound waves and water waves, but not for light? Is this because light travels so much faster than these other waves? Explain.
2. What is the difference between Fresnel and Fraunhofer diffraction? Are they different *physical* processes? Explain.
3. You use a lens of diameter *D* and light of wavelength λ and frequency *f* to form an image of two closely spaced and distant objects. Which of the following will increase the resolving power? (a) Use a lens with a smaller diameter; (b)

use light of higher frequency; (c) use light of longer wavelength. In each case justify your answer.

4. Light of wavelength λ and frequency passes *f* through a single slit of width . The diffraction pattern is observed on a screen a distance from the slit. Which of the following will *decrease* the width of the central maximum? (a) Decrease the slit width; (b) decrease the frequency of the light; (c) decrease the wavelength of the light; (d) decrease the distance of the screen from the slit. In each case justify your answer.
5. In a diffraction experiment with waves of wavelength λ there will be *no* intensity minima (that is, no dark fringes) if the slit width is small enough. What is the maximum slit width for which this occurs? Explain your answer.
6. The predominant sound waves used in human speech have wavelengths in the range from 1.0 to 3.0 meters. Using the ideas of diffraction, explain how it is possible to hear a person's voice even when he is facing away from you.
7. In single-slit diffraction, what is when θ=0. In view of your answer, why is the single-slit intensity *not* equal to zero at the center?
8. A rainbow ordinarily shows a range of colors. But if the water droplets that form the rainbow are small enough, the rainbow will appear white. Explain why, using diffraction ideas. How small do you think the raindrops would have to be for this to occur?
9. Some loudspeaker horns for outdoor concerts (at which the entire audience is seated on the ground) are wider vertically than horizontally. Use diffraction ideas to explain why this is more efficient at spreading the sound uniformly over the audience than either a square speaker horn or a horn that is wider horizontally than vertically. Would this still be the case if the audience were seated at different elevations, as in an amphitheater? Why or why not?
10. Information is stored on an audio compact disc, CD-ROM, or DVD disc in a series of pits on the disc. These pits are scanned by a laser beam. An important limitation on the amount of information that can be stored on such a disc is the width of the laserbeam. Explain why this should be, and explain how using a shorter-wavelength laser allows more information to be stored on a disc of the same size.
11. With which color of light can the Hubble Space Telescope see finer detail in a distant astronomical object: red, blue, or ultraviolet? Explain your answer.
12. Could x-ray diffraction effects with crystals be observed by using visible light instead of x rays? Why or why not?
13. Why is a diffraction grating better than a two-slit setup for measuring wavelengths of light?
14. One sometimes sees rows of evenly spaced radio antenna towers. A student remarked that these act like diffraction gratings. What did she mean? Why would one want them to act like a diffraction grating?
15. If a hologram is made using 600-nm light and then viewed with 500-nm light, how will the images look compared to those observed when viewed with 600-nm light? Explain.

16. A hologram is made using 600-nm light and then viewed by using white light from an incandescent bulb. What will be seen? Explain.
17. Ordinary photographic film reverses black and white, in the sense that the most brightly illuminated areas become blackest upon development (hence the term *negative*). Suppose a hologram negative is viewed directly, without making a positive transparency. How will the resulting images differ from those obtained with the positive? Explain.

Topic 3 Modern physics

3.1 Relativity

3.1.1 Invariance of physical laws

When the year 1905 began, Albert Einstein was an unknown 25-year old clerk in the Swiss patent office. By the end of that amazing year he had published three papers of extraordinary importance. One was an analysis of Brownian motion; a second (for which he was awarded the Nobel Prize) was on the photoelectric effect. In the third, Einstein introduced his **special theory of relativity,** proposing drastic revisions in the Newtonian concepts of space and time.

The special theory of relativity has made wide-ranging changes in our understanding of nature, but Einstein based it on just two simple postulates. One states that the laws of physics are the same in all inertial frames of reference; the other states that the speed of light in vacuum is the same in all inertial frames. These innocent-sounding propositions have far-reaching implications. Here are three:(1) Events that are simultaneous for one observer may not be simultaneous for another. (2) When two observers moving relative to each other measure a time interval or a length, they may not get the same results. (3) For the conservation principles for momentum and energy to be valid in all inertial systems, Newton's second law and the equations for momentum and kinetic energy have to be revised.

Relativity has important consequences in *all* areas of physics, including electromagnetism, atomic and nuclear physics, and high-energy physics. Although many of the results derived in this chapter may run counter to your intuition, the theory is in solid agreement with experimental observations.

Let's take a look at the two postulates that make up the special theory of relativity. Both postulates describe what is seen by an observer in an *inertial frame of reference*. The theory is "special" in the sense that it applies to observers in such special reference frames.

Einstein's first postulate, called the **principle of relativity,** states: **The laws of physics are the same in every inertial frame of reference.** If the laws differed, that difference could distinguish one inertial frame from the others or make one frame somehow more "correct" than another. Here are two examples. Suppose you watch two children playing catch with a ball while the three of you are aboard a train moving with constant velocity. Your observations of the motion *of the ball,* no matter how carefully done, can't tell you how fast (or whether) the train is moving. This is because Newton's laws of motion are the same in every inertial frame. Another example is the electromotive force (emf) induced in a coil of wire by a nearby moving permanent magnet. In the frame of reference in which the *coil* is stationary (Fig. 111a), the moving magnet causes a change of magnetic flux through the coil, and this induces an emf. In a different frame of reference in which the *magnet* is stationary (Fig. 111b), the motion of the coil through a magnetic field induces the emf. According to the principle of relativity, both of these frames of reference are equally valid. Hence the same emf must be induced in both situations shown in Fig. 111. As we saw, this is indeed the case, so Faraday's law is consistent with the principle of relativity. Indeed, *all* of the laws of electromagnetism are the same in every inertial frame of reference.

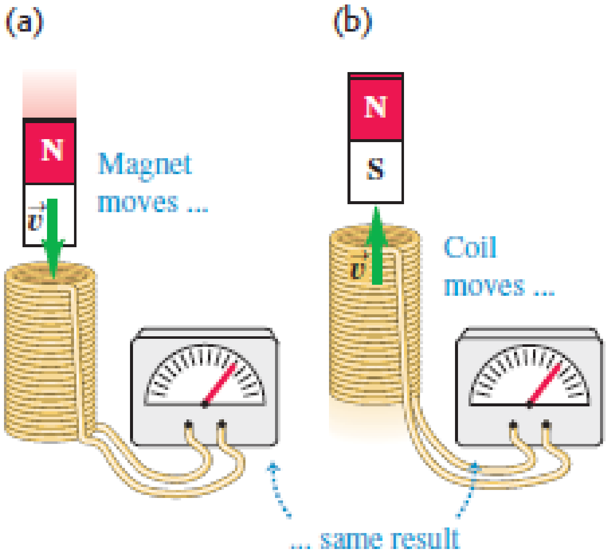


Figure 111 - The same emf is induced in the coil whether (a) the magnet moves relative to the coil or (b) the coil moves relative to the magnet

Equally significant is the prediction of the speed of electromagnetic radiation, derived from Maxwell's equations. According to this analysis, light and all other electromagnetic waves travel in vacuum with a constant speed, now defined to equal exactly $299792458\, m/s$. (We often use the approximate value $\theta c = 3 \times 10^8 m/s$, which is within one part in 1000 of the exact value.) As we will see, the speed of light in vacuum plays a central role in the theory of relativity.

During the 19th century, most physicists believed that light travelled through a hypothetical medium called the *ether*, just as sound waves travel through air. If so, the speed of light measured by observers would depend on their motion relative to the ether and would therefore be different in different directions. The Michelson-Morley experiment, was an effort to detect motion of the earth relative to the ether. Einstein's conceptual leap was to recognize that if Maxwell's equations are valid in all inertial frames, then the speed of light in vacuum should also be the same in all frames and in all directions. In fact, Michelson and Morley detected *no* ether motion across the earth, and the ether concept has been discarded. Although Einstein may not have known about this negative result, it supported his bold hypothesis of the constancy of the speed of light in vacuum.

Einstein's second postulate states: The speed of light in vacuum is the same in all inertial frames of reference and is independent of the motion of the source.

Let’s think about what this means. Suppose two observers measure the speed of light in vacuum. One is at rest with respect to the light source, and the other is moving away from it. Both are in inertial frames of reference. According to the principle of relativity, the two observers must obtain the same result, despite the fact that one is moving with respect to the other.

If this seems too easy, consider the following situation. A spacecraft moving past the earth at fires a missile straight ahead with a speed of 2000 m/s (relative to the spacecraft) (Fig. 112). What is the missile’s speed relative to the earth? Simple, you say; this is an elementary problem in relative velocity. The correct answer, according to Newtonian mechanics, is 3000 m/s. But now suppose the spacecraft turns on a searchlight, pointing in the same direction in which the missile was fired. An observer on the spacecraft measures the speed of light emitted by the searchlight and obtains the value According to Einstein’s second postulate, the motion of the light after it has left the source cannot depend on the motion of the source. So the observer on earth who measures the speed of this same light must also obtain the value c, *not* $\theta + 1000\, m/s$. This result contradicts our elementary notion of relative velocities, and it may not appear to agree with common sense. But “common sense” is intuition based on everyday experience, and this does not usually include measurements of the speed of light.

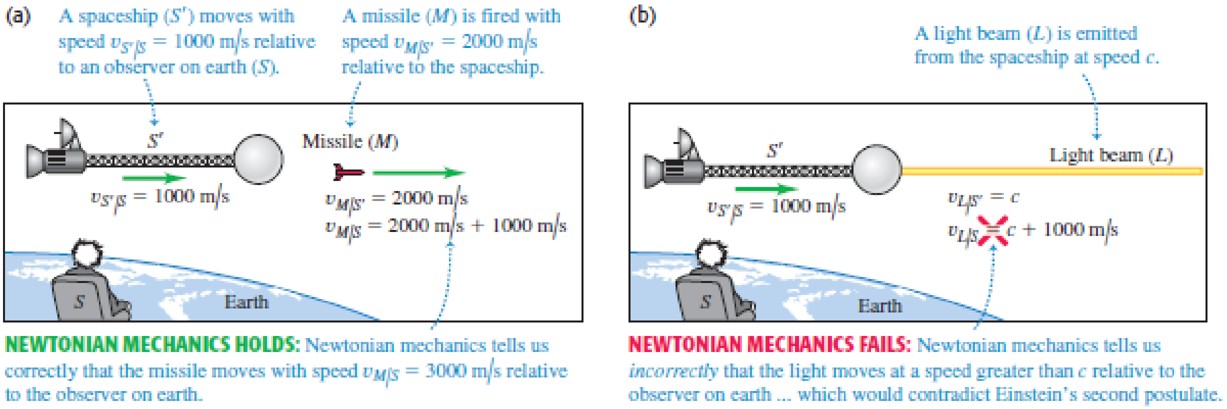

(a)
A spaceship (S') moves with speed $v_{S'/S} = 1000$ m/s relative to an observer on earth (S).
A missile (M) is fired with speed $v_{M/S'} = 2000$ m/s relative to the spaceship.
S'
Missile (M)
$v_{S'/S} = 1000$ m/s
$v_{M/S'} = 2000$ m/s
$v_{M/S} = 2000$ m/s + 1000 m/s
S
Earth
NEWTONIAN MECHANICS HOLDS: Newtonian mechanics tells us correctly that the missile moves with speed $v_{M/S} = 3000$ m/s relative to the observer on earth.
(b)
A light beam (L) is emitted from the spaceship at speed c.
S'
Light beam (L)
$v_{S'/S} = 1000$ m/s
$v_{L/S'} = c$
$v_{L/S} = c + 1000$ m/s
S
Earth
NEWTONIAN MECHANICS FAILS: Newtonian mechanics tells us *incorrectly* that the light moves at a speed greater than c relative to the observer on earth ... which would contradict Einstein’s second postulate.

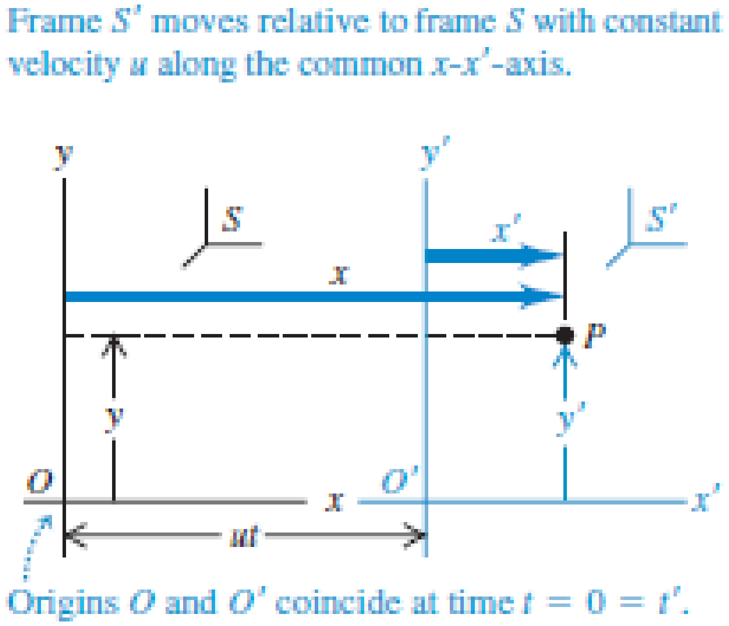
Figure 112 - (a) Newtonian mechanics makes correct predictions about relatively slow-moving objects; (b) it makes incorrect predictions about the behavior of light

Einstein’s second postulate immediately implies the following result:

It is impossible for an inertial observer to travel at *c*, the speed of light in vacuum.

We can prove this by showing that travel at implies a logical contradiction. Suppose that the spacecraft S’ in Fig. 112b is moving at the speed of light relative to an observer on the earth, so that $v_{S'/S} = c$. If the spacecraft turns on a headlight, the second postulate now asserts that the earth observer S measures the headlight beam to be also moving at c.Thus this observer measures that the headlight beam and the spacecraft move together and are always at the same point in space. But Einstein’s second postulate also asserts that the headlight beam moves at a speed relative to the spacecraft, so they *cannot* be at the same point in space. This contradictory result can be avoided only if it is impossible for an inertial observer, such as a passenger on the spacecraft, to move at c. As we go through our discussion of relativity, you may find yourself asking the question Einstein asked himself as a 16-year-old student, "What would I see if I were travelling at the speed of light?" Einstein realized only years later that his question's basic flaw was that he could *not* travel at

The Galilean Coordinate Transformation. Let's restate this argument symbolically, using two inertial frames of reference, labelled S for the observer on earth and S' for the moving spacecraft, as shown in Fig. 113. To keep things as simple as possible, we have omitted the *z*-axis. The *x*-axis of the two frames lie along the same line, but the origin O' of frame S' moves relative to the origin O of frame S with constant velocity u along the common *x*-*x'*-axis. We on earth set our clocks so that the two origins coincide at time *t* = 0, so their separation at a later time *t* is *ut*.



Frame S' moves relative to frame S with constant velocity u along the common x-x'-axis.
y
y'
S
S'
x'
x
P
y
y'
O
O'
x
x'
ut
Origins O and O' coincide at time $t = 0 = t'$.

Figure 113 – The position of particle P can be described by the coordinates x and y in frame of references S or by x' and y' in frame S'

Now think about how we describe the motion of a particle P. This might be an exploratory vehicle launched from the spacecraft or a pulse of light from a laser. We can describe the *position* of this particle by using the earth coordinates (x, y, z) in S or the spacecraft coordinates (x', y', z') in S'. Figure 113 shows that these are simply related by

$$x = x' + ut, \quad y = y', \quad z = z' \tag{197}$$

These equations, based on the familiar Newtonian notions of space and time, are called the **Galilean coordinate transformation.**

If particle S moves in the *x*-direction, its instantaneous velocity v_x as measured by an observer stationary in S is $v_x = dx/dt$. Its velocity $v_x{}'$ as measured by an observer stationary in S' is $v_x' = dx'/dt$. We can derive a relationship between v_x and $v_x{}'$ by taking the derivative with respect to of the first of Eqs. (197):

$$\frac{dx}{dt} = \frac{dx'}{dt} + u \qquad (198)$$

Now dx/dt is the velocity v_x measured in S, and dx'/dt is the velocity $v_x{}'$ measured in S', so we get the *Galilean velocity transformation* for one-dimensional motion:

$$v_x = v_x' + u \qquad (199)$$

Now here's the fundamental problem. Applied to the speed of light in vacuum, Eq. (194) says that $c = c' + u$. Einstein's second postulate, supported subsequently by a wealth of experimental evidence, says that $c = c'$. This is a genuine inconsistency, not an illusion, and it demands resolution. If we accept this postulate, we are forced to conclude that Eqs. (197) and (199) *cannot* be precisely correct, despite our convincing derivation. These equations have to be modified to bring them into harmony with this principle.

The resolution involves some very fundamental modifications in our kinematic concepts. The first idea to be changed is the seemingly obvious assumption that the observers in frames S and S' use the same *time scale,* formally stated as $t = t'$. Alas, we are about to show that this everyday assumption cannot be correct; the two observers *must* have different time scales. We must define the velocity v' in frame S' as $v' = dx'/dt'$, not as dx'/dt; the two quantities are not the same. The difficulty lies in the concept of *simultaneity,* which is our next topic. A careful analysis of simultaneity will help us develop the appropriate modifications of our notions about space and time.

3.1.2 Relativity of time intervals

We can derive a quantitative relationship between time intervals in different coordinate systems. To do this, let's consider another thought experiment. As before, a frame of reference S' moves along the common $x - x'$-axis with constant speed u relative to a frame S. u must be less than the speed of light c. Mavis, who is riding along with frame S', measures the time interval between two events that occur at the *same* point in space. Event 1 is when a flash of light from a light source leaves O'. Event 2 is when the flash returns to O', having been reflected from a mirror a distance d away, as shown in Fig. 114a. We label the time interval Δt_0 using the subscript zero as a reminder that the apparatus is at rest, with zero velocity, in frame S'. The flash of light moves a total distance 2*d*, so the time interval is

$$\Delta t_0 = \frac{2d}{c} \qquad (200)$$

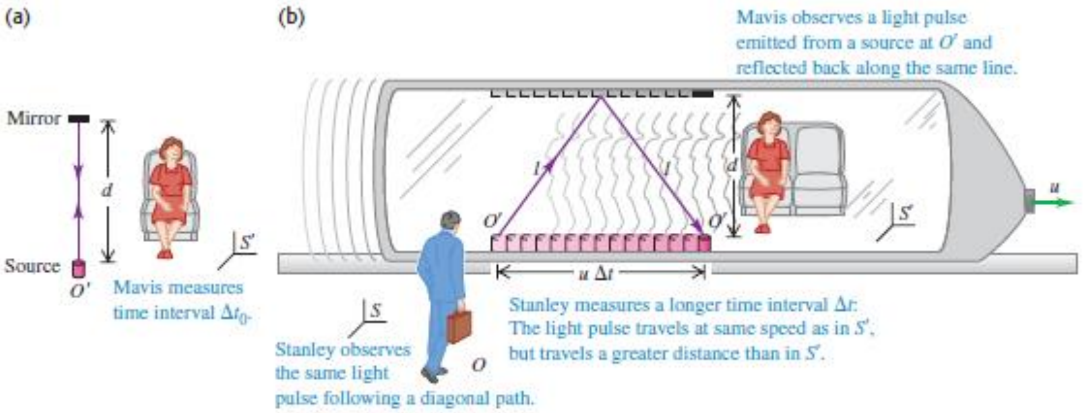

(a)
(b)
Mavis observes a light pulse emitted from a source at O′ and reflected back along the same line.
Mirror
d
Source
O′
Mavis measures time interval Δt₀.
S′
l
l
d
O′
O′
u
S′
u Δt
S
Stanley measures a longer time interval Δt: The light pulse travels at same speed as in S′, but travels a greater distance than in S′.
Stanley observes the same light pulse following a diagonal path.
O

Figure 114 - (a) Mavis, in frame of reference S', observes a light pulse emitted from a source at O' and reflected back along the same line. (b) How Stanley (in frame of reference S) and Mavis observe the same light pulse. The positions of O' at the times of departure and return of the pulse are shown

The round-trip time measured by Stanley in frame is a different interval Δt; in his frame of reference the two events occur at *different* points in space. During the time Δt, the source moves relative to S a distance $u\Delta t$ (Fig. 114b). In S' the round-trip distance is $2d$ perpendicular to the relative velocity, but the round-trip distance in S is the longer distance $2l$, where

$$l = \sqrt{d^2 + \left(\frac{u\Delta t}{2}\right)^2} \tag{201}$$

In writing this expression, we have assumed that both observers measure the same distance d. We will justify this assumption in the next section. The speed of light is the same for both observers, so the round-trip time measured in S is

$$\Delta t = \frac{2l}{c} = \frac{2}{c}\sqrt{d^2 + \left(\frac{u\Delta t}{2}\right)^2} \tag{202}$$

We would like to have a relationship between Δt and that is independent of d To get this, we solve Eq. (200) for d and substitute the result into Eq. (202), obtaining

$$\Delta t = \frac{2}{c}\sqrt{\left(\frac{c\Delta t_0}{2}\right)^2 + \left(\frac{u\Delta t}{2}\right)^2} \tag{203}$$

Now we square this and solve for Δt the result is

$$\Delta t = \frac{\Delta t_0}{\sqrt{1 - u^2/c^2}} \tag{204}$$

Since the quantity $\sqrt{1-u^2/c^2}$ is less than 1, Δt is greater than Δt_0: Thus Stanley measures a *longer* round-trip time for the light pulse than does Mavis.

Time Dilation. We may generalize this important result. In a particular frame of reference, suppose that two events occur at the same point in space. The time interval between these events, as measured by an observer at rest in this same frame (which we call the *rest frame* of this observer), is Δt_0. Then an observer in a second frame moving with constant speed relative to the rest frame will measure the time interval to be Δt, where

$$\Delta t = \frac{\Delta t_0}{\sqrt{1-u^2/c^2}} \tag{205}$$

We recall that no inertial observer can travel at $u = c$ and we note that $\sqrt{1-u^2/c^2}$ is imaginary for $u > c$. Thus Eq. (205) gives sensible results only when $u < c$. The denominator of Eq. (205) is always smaller than 1, so Δt is always *larger* than Δt_0. Thus we call this effect **time dilation.**

Think of an old-fashioned pendulum clock that has one second between ticks, as measured by Mavis in the clock's rest frame; this is Δt_0. If the clock's rest frame is moving relative to Stanley, he measures a time between ticks Δt that is longer than one second. In brief, *observers measure any clock to run slow if it moves relative to them* (Fig. 115). Note that this conclusion is a direct result of the fact that the speed of light in vacuum is the same in both frames of reference.

Figure 115 - This image shows an exploding star, called a *supernova,* within a distant galaxy. The brightness of a typical supernova decays at a certain rate. But supernovae that are moving away from us at a substantial fraction of the speed of light decay more slowly, in accordance with Eq. (205). The decaying supernova is a moving "clock" that runs slow

The quantity $1/\sqrt{1-u^2/c^2}$ in Eq. (205) appears so often in relativity that it is given its own symbol γ (the Greek letter gamma):

$$\gamma = \frac{1}{\sqrt{1-u^2/c^2}} \tag{206}$$

In terms of this symbol, we can express the time dilation formula, Eq. (205), as

$$\Delta t = \gamma \Delta t_0 \tag{207}$$

As a further simplification, u/c is sometimes given the symbol β (the Greek letter beta); then $\gamma = 1/\sqrt{1-\beta^2}$.

Figure 116 shows a graph of γ as a function of the relative speed u of two frames of reference. When u is very small compared to $c, u^2/c^2$ is much smaller than 1 and γ is very nearly *equal* to 1. In that limit, Eqs. (205) and (207) approach the Newtonian relationship $\Delta t = \Delta t_0$, corresponding to the same time interval in all frames of reference.

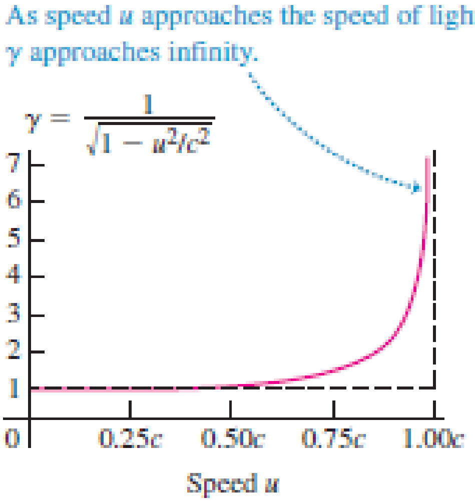
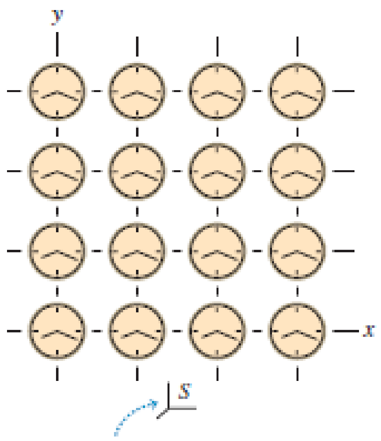


Figure 116 - The quantity $\gamma = 1/\sqrt{1-\beta^2}$ as a function of the relative speed of two frames of reference

If the relative speed u is great enough that γ is appreciably greater than 1, the speed is said to be *relativistic;* if the difference between γ and 1 is negligibly small, the speed u is called *nonrelativistic*. Thus $u = 6 \times 10^7 \frac{m}{s} = 0.2c$ (for which $\gamma = 1.02$) is a relativistic speed, but $u = 6 \times 10^4 \frac{m}{s} = 0.0002c$ (for which $\gamma = 1.00000002$) is a nonrelativistic speed.

Proper Time. There is only one frame of reference in which a clock is at rest, and there are infinitely many in which it is moving. Therefore the time interval measured between two events (such as two ticks of the clock) that occur at the same point in a particular frame is a more fundamental quantity than the interval between events at different points. We use the term **proper time** to describe the time interval between two events that occur *at the same point.*

In thought experiments, it's often helpful to imagine many observers with synchronized clocks at rest at various points in a particular frame of reference. We can picture a frame of reference as a coordinate grid with lots of synchronized clocks distributed around it, as suggested by Fig. 117. Only when a clock is moving relative to a given frame of reference do we have to watch for ambiguities of synchronization or simultaneity.



The grid is three dimensional; identical planes of clocks lie in front of and behind the page, connected by grid lines perpendicular to the page.

Figure 117 - A frame of reference pictured as a coordinate system with a grid of synchronized clocks

Throughout this chapter we will frequently use phrases like "Stanley *observes* that Mavis passes the point $x = 5\,m, y = 0, z = 0$ at time 2 s." This means that Stanley is using a grid of clocks in his frame of reference, like the grid shown in Fig. 117, to record the time of an event. We could restate the phrase as "When Mavis passes the point at $x = 5\,m, y = 0, z = 0$, the clock at that location in Stanley's frame of reference reads 2 s." We will avoid using phrases like "Stanley *sees* that Mavis is a certain point at a certain time," because there is a time delay for light to travel to Stanley's eye from the position of an event.

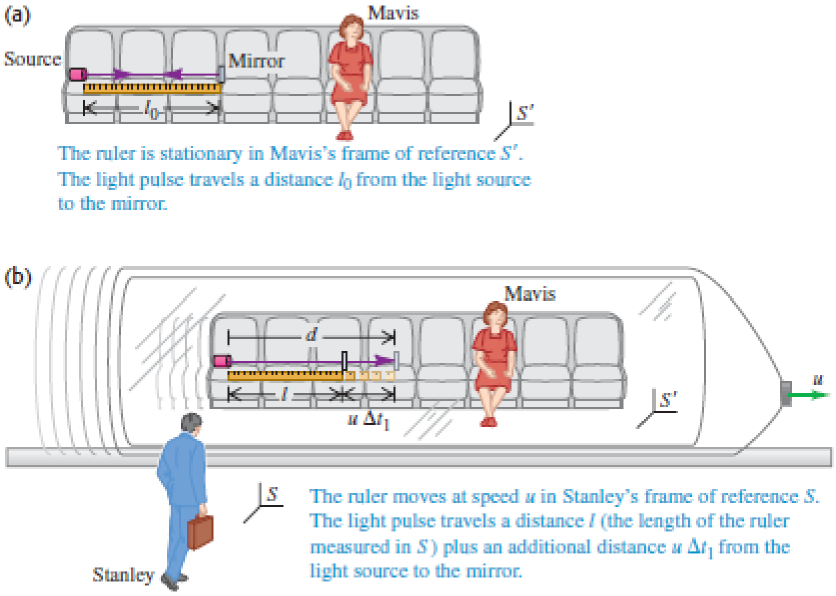
3.1.3 Relativity of length

Not only does the time interval between two events depend on the observer's frame of reference, but the *distance* between two points may also depend on the observer's frame of reference. The concept of simultaneity is involved. Suppose you want to measure the length of a moving car. One way is to have two assistants make marks on the pavement at the positions of the front and rear bumpers. Then you measure the distance between the marks. But your assistants have to make their marks *at the same time*. If one marks the position of the front bumper at one time and the other marks the position of the rear bumper half a second later, you won't get the car's true length. Since we've learned that simultaneity isn't an absolute concept, we have to proceed with caution.

Lengths Parallel to the Relative Motion. To develop a relationship between lengths that are measured parallel to the direction of motion in various coordinate systems, we consider another thought experiment. We attach a light source to one end of a ruler and a mirror to the other end. The ruler is at rest in reference frame S', and its length in this frame is l_0 (Fig. 118a). Then the time Δt_0 required for a light pulse to make the round trip from source to mirror and back is

$$\Delta t_0 = \frac{2l_0}{c} \tag{208}$$

This is a proper time interval because departure and return occur at the same point in S'.



The ruler is stationary in Mavis's frame of reference S'. The light pulse travels a distance l_0 from the light source to the mirror.

The ruler moves at speed u in Stanley's frame of reference S. The light pulse travels a distance l (the length of the ruler measured in S) plus an additional distance $u\,\Delta t_1$ from the light source to the mirror.

Figure 118 - (a) A ruler is at rest in Mavis's frame S'. A light pulse is emitted from a source at one end of the ruler, reflected by a mirror at the other end, and returned to the source position. (b) Motion of the light pulse as measured in Stanley's frame S

In reference frame S the ruler is moving to the right with speed u during this travel of the light pulse (Fig. 118b). The length of the ruler in S is l, and the time of travel from source to mirror, as measured in S, is Δt_1. During this interval the ruler, with source and mirror attached, moves a distance $u\Delta t_1$. The total length of path d from source to mirror is not but rather

$$d = l + u\Delta t_1 \tag{209}$$

The light pulse travels with speed so it is also true that

$$d = c\Delta t_1 \tag{210}$$

Combining Eqs. (209) and (210) to eliminate d we find

$$c\Delta t_1 = l + u\Delta t_1$$

or

$$\Delta t_1 = \frac{l}{c-u} \tag{211}$$

(Dividing the distance l by $c-u$ does *not* mean that light travels with speed $c-u$, but rather that the distance the pulse travels in S is greater than l).

In the same way we can show that the time Δt_2 for the return trip from mirror to source is

$$\Delta t_2 = \frac{l}{c+u} \tag{212}$$

The *total* time $\Delta t = \Delta t_1 + \Delta t_2$ for the round trip, as measured in S, is

$$\Delta t = \frac{l}{c-u} + \frac{l}{c+u} = \frac{2l}{c(1-u^2/c^2)} \tag{213}$$

We also know that Δt and Δt_0 are related by Eq. (205) because Δt_0 is a proper time in S'. Thus Eq. (209) for the round-trip time in the rest frame S' of the ruler becomes

$$\Delta t\sqrt{1-\frac{u^2}{c^2}} = \frac{l_0}{\gamma} \tag{214}$$

Finally, combining Eqs. (213) and (214) to eliminate Δt and simplifying, we obtain

$$l = l_0\sqrt{1-\frac{u^2}{c^2}} = \frac{l_0}{\gamma} \tag{215}$$

We have used the quantity $\gamma = 1/\sqrt{1-\frac{u^2}{c^2}}$ defined in Eq. (206). Thus the length measured in S, in which the ruler is moving, is *shorter* than the length l_0 measured in its frame S'.

A length measured in the frame in which the body is at rest (the rest frame of the body) is called a **proper length;** thus l_0 is a proper length in S', and the length measured in any other frame moving relative to S' is *less than* l_0. This effect is called **length contraction.**

When u is very small in comparison to c, γ approaches 1. Thus in the limit of small speeds we approach the Newtonian relationship $l = l_0$. This and the corresponding result for time dilation show that Eqs. (197), the Galilean coordinate transformation, are usually sufficiently accurate for relative speeds much smaller than c. If u is a reasonable fraction of c, however, the quantity $\sqrt{1-\frac{u^2}{c^2}}$ can be appreciably less than 1. Then can be substantially smaller than l_0, and the effects of length contraction can be substantial (Fig. 119).

Figure 119 - The speed at which electrons traverse the 3-km beam line of the LAC National Accelerator Laboratory is slower than c by less than 1 cm/s. As measured in the reference frame of such an electron, the beam line (which extends from the top to the bottom of this photograph) is only about 15 cm long!

3.1.4. Relativistic work and energy

When we developed the relationship between work and kinetic energy we used Newton's laws of motion. When we generalize these laws according to the principle of relativity, we need a corresponding generalization of the equation for kinetic energy.

Relativistic Kinetic Energy. We use the work–energy theorem, beginning with the definition of work. When the net force and displacement are in the same direction, the work done by that force is $A = \int Fdx$. We substitute the expression for relativistic force $F = \frac{m}{(1-u^2/c^2)^{3/2}} a$, the applicable relativistic version of Newton's second law. In moving a particle of rest mass from point x_1 to point x_2,

$$A = \int_{x_1}^{x_2} Fdx = \int_{x_1}^{x_2} \frac{madx}{(1-u^2/c^2)^{3/2}} \tag{216}$$

To derive the generalized expression for kinetic energy K as a function of speed v, we would like to convert this to an integral on v. To do this, first remember that the kinetic energy of a particle equals the net work done on it in moving it from rest to the speed v: $K = A$. Thus we let the speeds be zero at point x_1 and v at point x_2. So as not to confuse the variable of integration with the final speed, we change v to v_x in Eq. 216. That is, v_x is the varying x-component of the velocity of the particle as the net force accelerates it from rest to a speed v. We also realize that dx and dv_x are the infinitesimal changes in x and v_x, respectively, in the time interval dt. Because $v_x = dx/dt$ and $a = dv_x/dt$, we can rewrite adx in Eq. (216) as

$$adx = \frac{dv_x}{dt} dx = dx \frac{dv_x}{dt} = \frac{dx}{dt} v_x = v_x dx \tag{217}$$

Making these substitutions gives us

$$K = A = \int_0^v \frac{mv_x dx}{(1-u^2/c^2)^{3/2}} \tag{218}$$

We can evaluate this integral by a simple change of variable; the final result is

$$K = \frac{mc^2}{\sqrt{1-v^2/c^2}} - mc^2 = (\gamma - 1)mc^2 \tag{219}$$

As v approaches c, the kinetic energy approaches infinity. If Eq. (219) is correct, it must also approach the Newtonian expression $K = \frac{1}{2}mv^2$ when v is much smaller than c (Fig. 120).

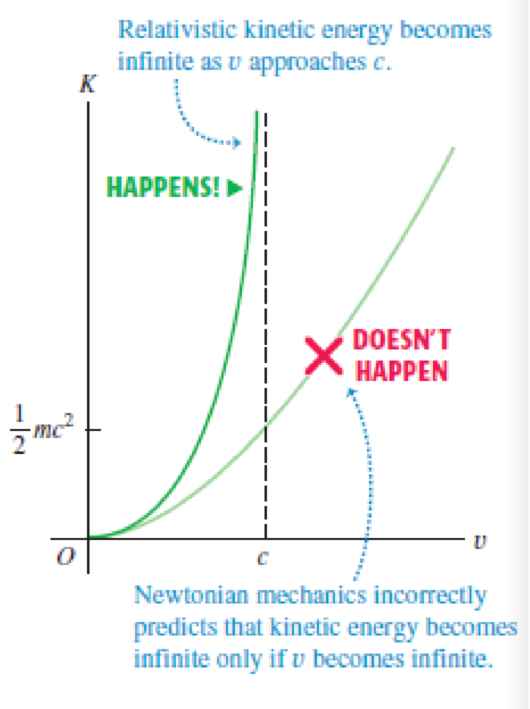


Figure 120 - Graph of the kinetic energy of a particle of rest mass m as a function of speed v. Also shown is the Newtonian prediction, which gives correct results only at speeds much less than c.

When v is much smaller than c, all the terms in the series in binominal theorem except the first are negligibly small, and we obtain the Newtonian expression $\frac{1}{2}mv^2$.

Rest Energy. Equation (219) for the kinetic energy of a moving particle includes a term $mc^2/\sqrt{1-v^2/c^2}$ that depends on the motion and a second energy term mc^2 that is independent of the motion. It seems that the kinetic energy of a particle is the difference between some **total energy** E and an energy mc^2 that it has even when it is at rest. Thus we can rewrite Eq. (219) as

$$E = K + mc^2 = \frac{mc^2}{\sqrt{1-\frac{v^2}{c^2}}} = \gamma mc^2 \tag{220}$$

For a particle at rest ($K = 0$), we see that $E = mc^2$. The energy mc^2 associated with rest mass m rather than motion is called the **rest energy** of the particle.

There is in fact direct experimental evidence that rest energy really does exist. The simplest example is the decay of a neutral *pion*. This is an unstable subatomic particle of rest mass m_π; when it decays, it disappears and electromagnetic radiation appears. If a neutral pion has no kinetic energy before its decay, the total energy of the radiation after its decay is found to equal exactly $m_\pi c^2$. In many other fundamental particle transformations the sum of the rest masses of the particles changes. In every case there is a corresponding energy change, consistent with the assumption of a rest energy mc^2 associated with a rest mass m. Historically, the principles of conservation of mass and of energy developed quite independently. The theory of relativity shows that they are actually two special cases of a single broader conservation principle, the *principle of conservation of mass and energy.* In some physical phenomena, neither the sum of the rest masses of the particles nor the total energy other than rest energy is separately conserved, but there is a more general conservation principle: In an isolated system, when the sum of the rest masses changes, there is always a change in $1/c^2$ times the total energy other than the rest energy. This change is equal in magnitude but opposite in sign to the change in the sum of the rest masses.

This more general mass-energy conservation law is the fundamental principle involved in the generation of power through nuclear reactions. When a uranium nucleus undergoes fission in a nuclear reactor, the sum of the rest masses of the resulting fragments is *less than* the rest mass of the parent nucleus. An amount of energy is released that equals the mass decrease multiplied by c^2. Most of this energy can be used to produce steam to operate turbines for electric power generators.

We can also relate the total energy E of a particle (kinetic energy plus rest energy) directly to its momentum by combining Eq. $\vec{p} = \frac{m\vec{v}}{\sqrt{1-v^2/c^2}}$ for relativistic momentum and Eq. (220) for total energy to eliminate the particle's velocity. The simplest procedure is to rewrite these equations in the following forms:

$$\left(\frac{E}{mc^2}\right)^2 = \frac{1}{1 - v^2/c^2} \tag{221}$$

and

$$\left(\frac{p}{mc}\right)^2 = \frac{v^2/c^2}{1 - v^2/c^2} \tag{222}$$

Subtracting the second of these from the first and rearranging, we find

$$E^2 = (mc^2)^2 + (pc)^2 \tag{223}$$

Again we see that for a particle at rest $p = 0, E = mc^2$.

Equation (223) also suggests that a particle may have energy and momentum even when it has no rest mass. In such a case, $m = 0$ and

$$E = pc \tag{224}$$

In fact, zero rest mass particles do exist. Such particles always travel at the speed of light in vacuum. One example is the *photon,* the quantum of electromagnetic radiation. Photons are emitted and absorbed during changes of state of an atomic or nuclear system when the energy and momentum of the system change.

Discussion questions

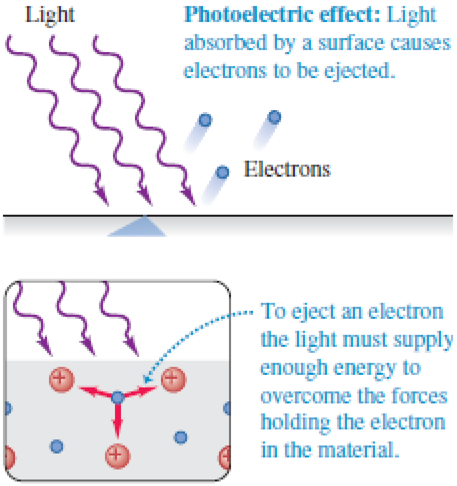
1. You are standing on a train platform watching a high-speed train pass by. A light inside one of the train cars is turned on and then a little later it is turned off. (a) Who can measure the proper time interval for the duration of the light: you or a passenger on the train? (b) Who can measure the proper length of the train car: you or a passenger on the train? (c) Who can measure the proper length of a sign attached to a post on the train platform: you or a passenger on the train? In each case explain your answer.
2. If simultaneity is not an absolute concept, does that mean that we must discard the concept of causality? If event *A* is to *cause* event *B*, must occur first. Is it possible that in some frames *A* appears to be the cause of *B* and in others appears to be the cause of *A*? Explain.
3. What do you think would be different in everyday life if the speed of light were 10 m/s instead of $3*10^8$ m/s?
4. The average life span in the United States is about 70 years. Does this mean that it is impossible for an average person to travel a distance greater than 70 light-years away from the earth? (A light-year is the distance light travels in a year.) Explain.
5. You are holding an elliptical serving platter. How would you need to travel for the serving platter to appear round to another observer?
6. Two events occur at the same space point in a particular inertial frame of reference and are simultaneous in that frame. Is it possible that they may not be simultaneous in a different inertial frame? Explain.
7. A high-speed train passes a train platform. Larry is a passenger on the train, Adam is standing on the train platform, and David is riding a bicycle toward the platform in the same direction as the train is traveling. Compare the length of a train car as measured by Larry, Adam, and David.
8. The theory of relativity sets an upper limit on the speed that a particle can have. Are there also limits on the energy and momentum of a particle? Explain.
9. A student asserts that a material particle must always have a speed slower than that of light, and a massless particle must always move at exactly the speed of light. Is she correct? If so, how do massless particles such as photons and neutrinos acquire this speed? Can’t they start from rest and accelerate? Explain.
10. The speed of light relative to still water is If the water is $2,25*10^8$ m/s moving past us, the speed of light we measure depends on the speed of the water. Do these facts violate Einstein’s second postulate? Explain.
11. When a monochromatic light source moves toward an observer, its wavelength appears to be shorter than the value measured when the source is at rest. Does this contradict the hypothesis that the speed of light is the same for all observers? Explain.
12. In principle, does a hot gas have more mass than the same gas when it is cold? Explain. In practice, would this be a measurable effect? Explain.

13. Why do you think the development of Newtonian mechanics preceded the more refined relativistic mechanics by so many years?

3.2 Light waves behaving as particles and particles behaving as waves

3.2.1 Photoelectric effect

A phenomenon that gives insight into the nature of light is the **photoelectric effect,** in which a material emits electrons from its surface when illuminated (Fig. 121). To escape from the surface, an electron must absorb enough energy from the incident light to overcome the attraction of positive ions in the material. These attractions constitute a potential-energy barrier; the light supplies the “kick” that enables the electron to escape.



Photoelectric effect: Light absorbed by a surface causes electrons to be ejected.

To eject an electron the light must supply enough energy to overcome the forces holding the electron in the material.

Figure 121 – The photoelectric effect

The photoelectric effect has a number of applications. Digital cameras and night-vision scopes use it to convert light energy into an electric signal that is reconstructed into an image. On the moon, sunlight striking the surface causes surface dust to eject electrons, leaving the dust particles with a positive charge. The mutual electric repulsion of these charged dust particles causes them to rise above the moon’s surface, a phenomenon that was observed from lunar orbit by the Apollo astronauts.

Early we explored the wave model of light, which Maxwell formulated two decades before the photoelectric effect was observed. Is the photoelectric effect consistent with this model? Two conducting electrodes are enclosed in an evacuated glass tube and connected by a battery, and the cathode is illuminated. Depending on the potential difference V_{AC} between the two electrodes, electrons emitted by the illuminated cathode (called *photoelectrons*) may travel across to the anode, producing a *photocurrent* in the external circuit. (The tube is evacuated to a pressure of 0.01 Pa or less to minimize collisions between the electrons and gas molecules.)

The illuminated cathode emits photoelectrons with various kinetic energies. If the electric field points toward the cathode, all the electrons are accelerated toward the anode and contribute to the photocurrent. But by reversing the field and adjusting its strength, we can prevent the less energetic electrons from reaching the anode. In fact, we can determine the maximum kinetic energy K_{max} of the emitted electrons by making the potential of the anode relative to the cathode, V_{AC}, just negative enough so that the current stops. This occurs for $V_{AC} = -V_0$, where V_0 is called the **stopping potential**. As an electron moves from the cathode to the anode, the potential decreases by V_0 and negative work $-eV_0$ is done on the (negatively charged) electron. The most energetic electron leaves the cathode with kinetic energy $K_{max} = \frac{1}{2}mv_{max}^2$ and has zero kinetic energy at the anode. Using the work–energy theorem, we have

$$A_{tot} = -eV_0 = \Delta K = 0 - K_{max} \qquad (225)$$
$$K_{max} = \frac{1}{2}mv_{max}^2 = eV_0$$

Hence by measuring the stopping potential V_0, we can determine the maximum kinetic energy with which electrons leave the cathode. (We are ignoring any effects due to differences in the materials of the cathode and anode.)

The experimental results proved to be very different from these predictions. Here is what was found in the years between 1877 and 1905:

Experimental Result 1: *The photocurrent depends on the light frequency*. For a given material, monochromatic light with a frequency below a minimum threshold frequency produces no photocurrent, regardless of intensity. For most metals the threshold frequency is in the ultraviolet (corresponding to wavelengths between 200 and 300 nm), but for other materials like potassium oxide and cesium oxide it is in the visible spectrum (between 380 and 750 nm).

Experimental Result 2: There is *no measurable time delay* between when the light is turned on and when the cathode emits photoelectrons (assuming the frequency of the light exceeds the threshold frequency). This is true no matter how faint the light is.

Experimental Result 3: **The stopping potential does not depend on intensity, but does depend on frequency**. Figure 122 shows graphs of photocurrent as a function of potential difference V_{AC} for light of a given frequency and two different intensities. The reverse potential difference $-V_0$ needed to reduce the current to zero is the same for both intensities. The only effect of increasing the intensity is to increase the number of electrons per second and hence the photocurrent i. (The curves level off when V_{AC} is large and positive because at that point all the emitted electrons are being collected by the anode.) If the intensity is held constant but the frequency is increased, the stopping potential also increases. In other words, the greater the light frequency, the higher the energy of the ejected photoelectrons.

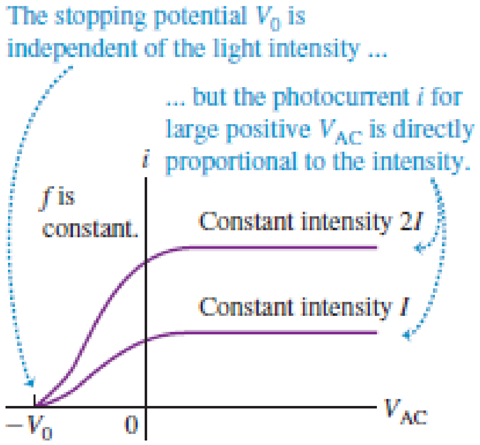


Figure 122 - Photocurrent i for a constant light frequency f as a function of the potential of the anode with respect to the cathode

These results directly contradict Maxwell's description of light as an electromagnetic wave. A solution to this dilemma was provided by Albert Einstein in 1905. His proposal involved nothing less than a new picture of the nature of light.

Einstein made the radical postulate that a beam of light consists of small packages of energy called **photons** or *quanta*. This postulate was an extension of an idea developed five years earlier by Max Planck to explain the properties of blackbody radiation. In Einstein's picture, the energy E of an individual photon is equal to a constant h times the photon frequency ν. From the relationship $\nu = c/\lambda$ for electromagnetic waves in vacuum, we have

$$E = h\nu = \frac{hc}{\lambda} \qquad (226)$$

where h is a universal constant called **Planck's constant.** The numerical value of this constant, to the accuracy known at present, is

$$h = 6.62606896(33) \times 10^{-34}\, J \cdot s$$

In Einstein's picture, an individual photon arriving at the surface in Fig. 121a is absorbed by a single electron. This energy transfer is an all-or-nothing process, in contrast to the continuous transfer of energy in the wave theory of light; the electron gets all of the photon's energy or none at all. The electron can escape from the surface only if the energy it acquires is greater than the work function A_w. Thus photoelectrons will be ejected only if $h\nu > A_w$, or $\nu > A_w/h$. Einstein's postulate therefore explains why the photoelectric effect occurs only for frequencies greater than a minimum threshold frequency. This postulate is also consistent with the observation that greater intensity causes a greater photocurrent(Fig. 122). Greater intensity at a particular frequency means a greater number of photons per second absorbed, and thus a greater number of electronsemitted per second and a greater photocurrent.

Einstein’s postulate also explains why there is no delay between illumination and the emission of photoelectrons. As soon as photons of sufficient energy strike the surface, electrons can absorb them and be ejected.

Finally, Einstein’s postulate explains why the stopping potential for a given surface depends only on the light frequency. Recall that A_w is the *minimum* energy needed to remove an electron from the surface. Einstein applied conservation of energy to find that the *maximum* kinetic energy $K_{max} = \frac{1}{2}mv_{max}^2$ for an emitted electron is the energy $h\nu$ gained from a photon minus the work function A_w:

$$K_{max} = \frac{1}{2}mv_{max}^2 = h\nu - A_w \tag{227}$$

Substituting $K_{max} = eV_0$ from Eq. (225), we find

$$eV_0 = h\nu - A_w \tag{228}$$

Equation (228) shows that the stopping potential increases with increasing frequency ν. The intensity doesn’t appear in Eq. (228), so V_0 is independent of intensity. As a check of Eq. (228), we can measure the stopping potential V_0 for each of several values of frequency ν for a given cathode material (Fig. 123). A graph of V_0 as a function of ν turns out to be a straight line, verifying Eq. (228), and from such a graph we can determine both the work function A_w for the material and the value of the quantity h/e. After the electron charge $-e$ was measured by Robert Millikan in 1909, Planck’s constant h could also be determined from these measurements.

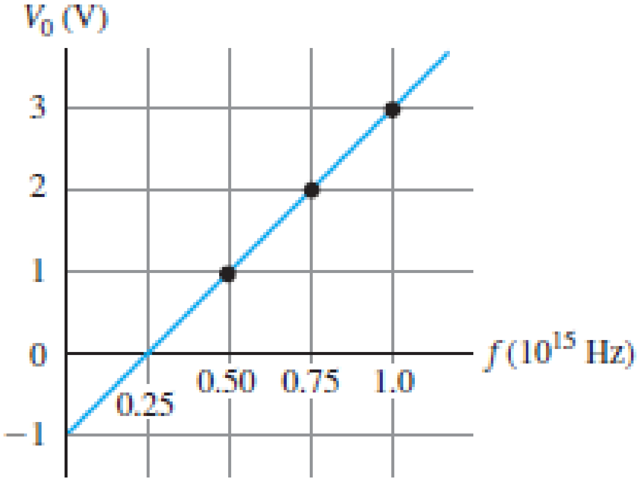


Figure 123 - Stopping potential as a function of frequency for a particular cathode material

Electron energies and work functions are usually expressed in electron volts (eV). To four significant figures,

$$1\ eV = 1.602 \times 10^{-19} J$$

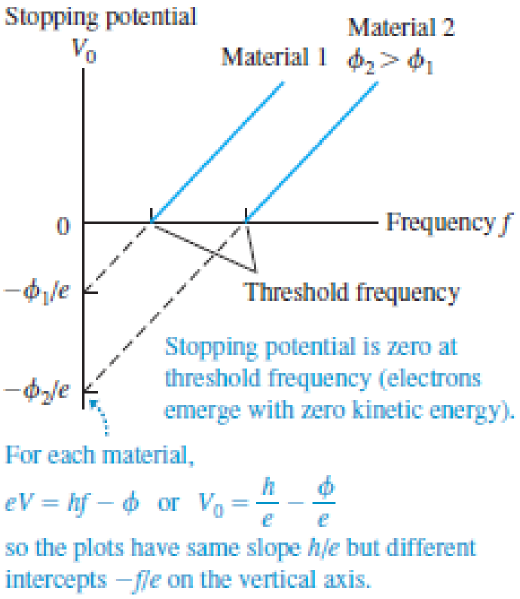
To this accuracy, Planck's constant is

$$h = 6.626 \times 10^{-34} J \cdot s = 4.13610^{-15}\ eV \cdot s$$

Table 4 lists the work functions of several elements. These values are approximate because they are very sensitive to surface impurities. The greater the work function, the higher the minimum frequency needed to emit photoelectrons (Fig. 124).

Table 4 – Work functions of several elements

Element	Work function (eV)
Aluminium	4.3
Carbon	5
Copper	4.7
Gold	5.1
Nickel	5.1
Silicon	4.8
Silver	4.3
Sodium	2.7



Stopping potential is zero at threshold frequency (electrons emerge with zero kinetic energy).

For each material,

$eV = hf - \phi$ or $V_0 = \frac{h}{e} - \frac{\phi}{e}$

so the plots have same slope h/e but different intercepts $-f/e$ on the vertical axis.

Figure 124 - Stopping potential as a function of frequency for two cathode materials having different work functions f.

The photon picture explains a number of other phenomena in which light is absorbed. One example is a *suntan*, which is caused when the energy in sunlight triggers a chemical reaction in skin cells that leads to increased production of the pigment melanin. This reaction can occur only if a specific molecule in the cell absorbs a certain minimum amount of energy. A short-wavelength ultraviolet photon has enough energy to trigger the reaction, but a longer-wavelength visible-light photon does not. Hence ultraviolet light causes tanning, while visible light cannot.

3.2.2 Compton's effect

The final aspect of light that we must test against Einstein's photon model is its behavior after the light is produced and before it is eventually absorbed. We can do this by considering the scattering of light.

In the photon model we imagine the scattering process as a

collision of two particles, the incident photon and an electron that is initially at rest (Fig. 125a). The incident photon would give up part of its energy and momentum to the electron, which recoils as a result of this impact. The scattered photon that remains can fly off at a variety of angles with respect to the incident direction, but it has less energy and less momentum than the incident photon (Fig. 125b). The energy and momentum of a photon are given by $E = h\nu = \frac{hc}{\lambda}$ (Eq. 226) and $p = \frac{h\nu}{c} = \frac{h}{\lambda}$. Therefore, in the photon model, the scattered light has a lower frequency ν and longer wavelength l than the incident light.

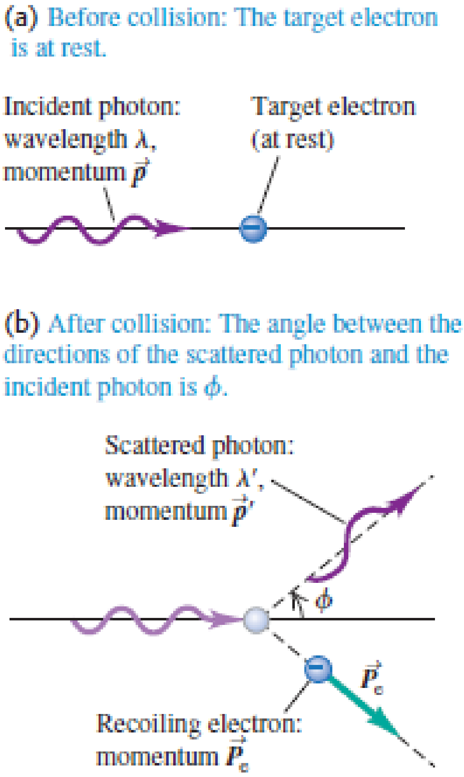

(a) Before collision: The target electron is at rest.
Incident photon: wavelength λ, momentum $\vec{p}$
Target electron (at rest)
(b) After collision: The angle between the directions of the scattered photon and the incident photon is ϕ.
Scattered photon: wavelength λ', momentum $\vec{p}'$
ϕ
$\vec{P}_e$
Recoiling electron: momentum $\vec{P}_e$

Figure 125 – The photon model of scattering by an electron

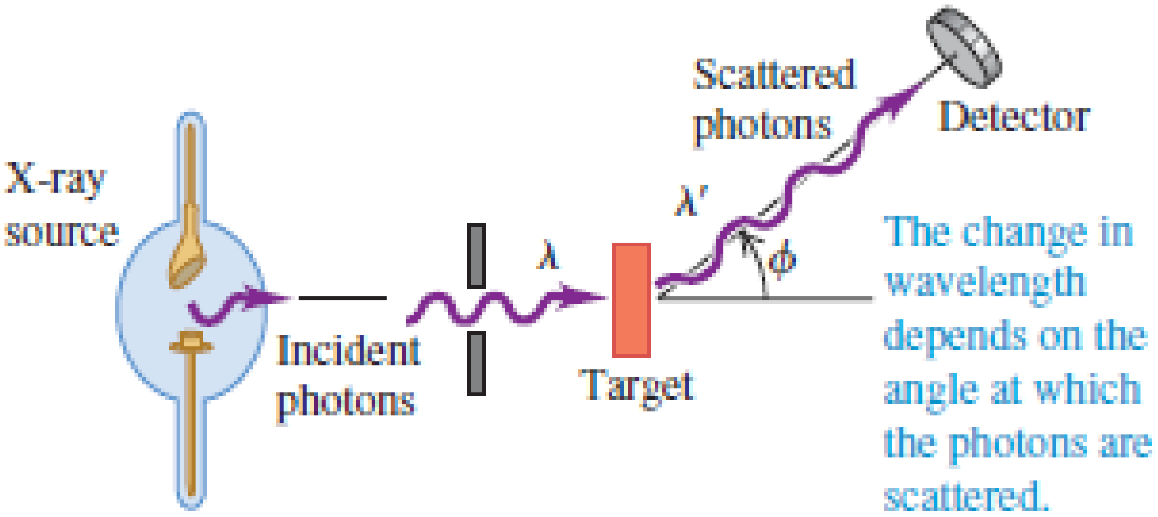
The definitive experiment that tested these predictions of the wave and photon models was carried out in 1922 by the American physicist Arthur H. Compton. In his experiment Compton aimed a beam of x rays at a solid target and measured the wavelength of the radiation scattered from the target (Fig. 126). He discovered that some of the scattered radiation has smaller frequency (longer wavelength) than the incident radiation and that the change in wavelength depends on the angle through which the radiation is scattered. This is precisely what the photon model predicts for light scattered from electrons in the target, a process that is now called **Compton scattering**.

Specifically, if the scattered radiation emerges at an angle θ with respect to the incident direction, as shown in Fig. 126, and if λ and λ' are the wavelengths of the incident and scattered radiation, respectively, Compton found that

$$\lambda' - \lambda = \frac{h}{mc} = (1 - \cos\theta) \tag{229}$$

where m is the electron rest mass. In other words, λ' is greater than λ. The quantity h/mc that appears in Eq. (229) has units of length. Its numerical value is

$$\frac{h}{mc} = \frac{6.626 \times 10^{-34}\, J \cdot s}{(9.109 \times 10^{-31} kg)(2.299 \times 10^{8} m/s)} = 2.426 \times 10^{-12}\, m$$



The change in wavelength depends on the angle at which the photons are scattered.

Figure 126 – A Compton-effect experiment

Compton showed that Einstein’s photon theory, combined with the principles of conservation of energy and conservation of momentum, provides a beautifully clear explanation of his experimental results. We outline the derivation below. The electron recoil energy may be in the relativistic range, so we have to use the relativistic energy–momentum relationships, Eqs. (223) and (224). The incident photon has momentum $\vec{p}$ with magnitude p and energy pc. The scattered photon has momentum $\vec{p}'$ with magnitude and energy *pc’*. The electron is initially at rest, so its initial momentum is zero and its initial energy is its rest energy mc^2. The final electron momentum $\vec{P}_e$ has magnitude P_e, and the final electron energy is given by

$E^2 = (mc^2)^2 + (P_e c)^2$. Then energy conservation gives us the relationship $pc + mc^2 = p'c + E_e$.
Rearranging, we find

$$(pc - p'c + mc^2)^2 = E_e^2 = (mc^2)^2 + (P_e c)^2 \tag{230}$$

We can eliminate the electron momentum $\vec{P}_e$ from Eq. (230) by using momentum conservation. From Fig. 127 we see that $\vec{p} = \vec{p}' + \vec{P}_e$, or

$$\vec{P}_e = \vec{p} - \vec{p}' \tag{231}$$

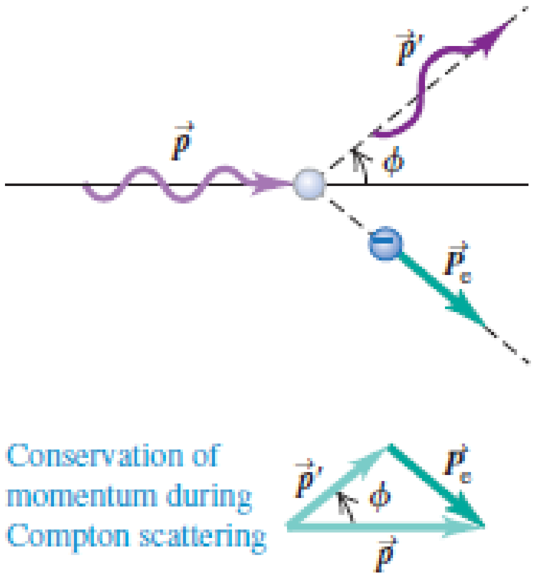


Figure 127 – Vector diagram showing conservation of momentum in Compton scattering

By taking the scalar product of each side of Eq. (231) with itself, we find

$$P_e^2 = p^2 + p'^2 - 2pp' \cos\theta \tag{232}$$

We now substitute this expression for P_e^2 into Eq. (230) and multiply out the left side. We divide out a common factor c^2; several terms cancel, and when the resulting equation is divided through by (pp') the result is

$$\frac{mc}{p'} - \frac{mc}{p} = 1 - \cos\theta \tag{233}$$

Finally, we substitute $p' = h/\lambda'$ and $p = h/\lambda$, then multiply by h/mc to obtain Eq. (229).

When the wavelengths of x rays scattered at a certain angle are measured, the curve of intensity per unit wavelength as a function of wavelength has two peaks

(Fig. 128). The longer-wavelength peak represents Compton scattering. The shorter-wavelength peak, labeled is at the wavelength of the incident x-rays and corresponds to x-ray scattering from tightly bound electrons. In such scattering processes the entire atom must recoil, so the m in Eq. (229) is the mass of the entire atom rather than of a single electron. The resulting wavelength shifts are negligible.

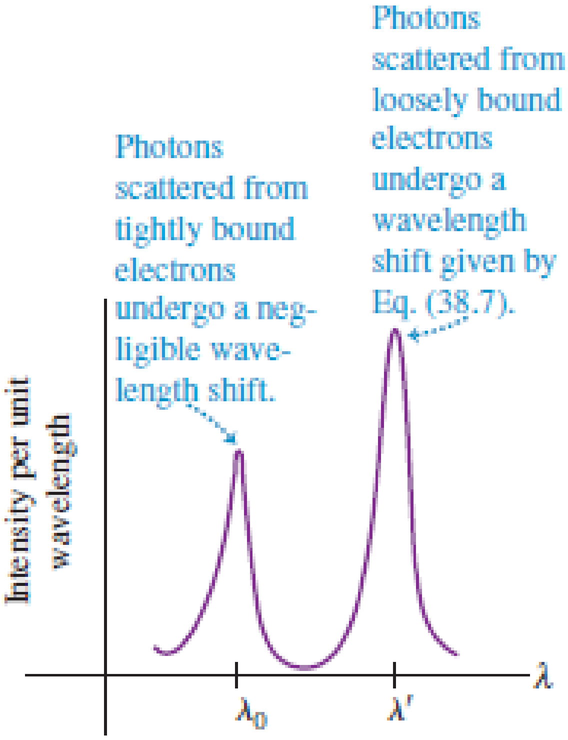


Figure 128- Intensity as a function of wavelength for photons scattered at an angle of 135° in a Compton-scattering experiment

Discussion questions

1. In what ways do photons resemble other particles such as electrons? In what ways do they differ? Do photons have mass? Do they have electric charge? Can they be accelerated? What mechanical properties do they have?
2. There is a certain probability that a single electron may simultaneously absorb *two* identical photons from a high-intensity laser.
3. According to the photon model, light carries its energy in packets called quanta or photons. Why then don't we see a series of flashes when we look at things?
4. Would you expect effects due to the photon nature of light to be generally more important at the low-frequency end of the electromagnetic spectrum (radio waves) or at the high-frequency end (x rays and gamma rays)? Why?
5. During the photoelectric effect, light knocks electrons out of metals. So why don't the metals in your home lose their electrons when you turn on the lights?
6. Most black-and-white photographic film (with the exception of some special-purpose films) is less sensitive to red light than blue light and has almost no sensitivity to infrared. How can these properties be understood on the basis of photons?

7. Human skin is relatively insensitive to visible light, but ultraviolet radiation can cause severe burns. Does this have anything to do with photon energies? Explain.
8. In a photoelectric-effect experiment, the photocurrent i for large positive values of V_{AC} has the same value no matter what thelight frequency f (provided that f is higher than the threshold frequency f_0) Explain why.
9. In an experiment involving the photoelectric effect, if the intensity of the incident light (having frequency higher than the threshold frequency) is reduced by a factor of 10 without changing anything else, which (if any) of the following statements about this process will be true? (a) The number of photoelectrons will most likely be reduced by a factor of 10. (b) The maximum kinetic energy of the ejected photoelectrons will most likely be reduced by a factor of 10. (c) The maximum speed of the ejected photoelectrons will most likely be reduced by a factor of 10. (d) The maximum speed of the ejected photoelectrons will most likely be reduced by a factor of $\sqrt{10}$ (e) The time for the first photoelectron to be ejected will be increased by a factor of 10.
10. The materials called *phosphors* that coat the inside of a fluorescent lamp convert ultraviolet radiation (from the mercuryvapor discharge inside the tube) into visible light. Could one also make a phosphor that converts visible light to ultraviolet? Explain.
11. In a photoelectric-effect experiment, which of the following will increase the maximum kinetic energy of the photoelectrons? (a) Use light of greater intensity; (b) use light of higher frequency; (c) use light of longer wavelength; (d) use a metal surface with a larger work function. In each case justify your answer.
12. A photon of frequency f undergoes Compton scattering from an electron at rest and scatters through an angle The frequency of the scattered photon is How is related to Does your answer depend on Explain.
13. Can Compton scattering occur with protons as well as electrons? For example, suppose a beam of x rays is directed at a target of liquid hydrogen. (Recall that the nucleus of hydrogen consists of a single proton.) Compared to Compton scattering with electrons, what similarities and differences would you expect? Explain.
14. Why must engineers and scientists shield against x-ray production in high-voltage equipment?
15. In attempting to reconcile the wave and particle models of light, some people have suggested that the photon rides up and down on the crests and troughs of the electromagnetic wave. What things are *wrong* with this description?
16. If a proton and an electron have the same speed, which has the longer de Broglie wavelength? Explain.
17. If a proton and an electron have the same kinetic energy, which has the longer de Broglie wavelength? Explain.
18. Does a photon have a de Broglie wavelength? If so, how is it related to the wavelength of the associated electromagnetic wave? Explain.

19. When an electron beam goes through a very small hole, it produces a diffraction pattern on a screen, just like that of light. Does this mean that an electron spreads out as it goes through the hole? What does this pattern mean?
20. Galaxies tend to be strong emitters of photons Lyman-α (from the n to n=1 transition in atomic hydrogen). But the intergalactic medium—the very thin gas between the galaxies— tends to *absorb* photons. What can you infer from these observations about the temperature in these two environments?Explain.
21. The emission of a photon by an isolated atom is a recoil process in which momentum is conserved. Thus Eq. (39.5) should include a recoil kinetic energy for the atom. Why is this energynegligible in that equation?
22. How might the energy levels of an atom be measured directly—that is, without recourse to analysis of spectra?
23. Elements in the gaseous state emit line spectra with welldefined wavelengths. But hot solid bodies always emit a continuous spectrum—that is, a continuous smear of wavelengths. Can you account for this difference?
24. As a body is heated to a very high temperature and becomes self-luminous, the apparent color of the emitted radiation shifts from red to yellow and finally to blue as the temperature increases. Why does the color shift? What other changes in the character of the radiation occur?
25. The peak-intensity wavelength of red dwarf stars, which have surface temperatures around 3000 K, is about 1000 nm, which is beyond the visible spectrum. So why are we able to see these stars, and why do they appear red?
26. Why go through the expense of building an electron microscope for studying very small objects such as organic molecules? Why not just use extremely short electromagnetic waves, which are much cheaper to generate?
27. Which has more total energy: a hydrogen atom with an electron in a high shell (large *n*) or in a low shell (small *n*)? Which is moving faster: the high-shell electron or the low-shell electron? Is there a contradiction here? Explain.
28. Does the uncertainty principle have anything to do with marksmanship? That is, is the accuracy with which a bullet can be aimed at a target limited by the uncertainty principle? Explain.
29. Suppose a two-slit interference experiment is carried out using an electron beam. Would the same interference pattern result if one slit at a time is uncovered instead of both at once? If not, why not? Doesn’t each electron go through one slit or the other? Or does every electron go through both slits? Discuss the latter possibility in light of the principle of complementarity.
30. Laser light results from transitions from long-lived metastable states. Why is it more monochromatic than ordinary light?
31. Could an electron-diffraction experiment be carried out using three or four slits? Using a grating with many slits? What sort of results would you expect with a grating? Would the uncertainty principle be violated? Explain.
32. Why can an electron microscope have greater magnification than an ordinary microscope?

33. When you check the air pressure in a tire, a little air always escapes; the process of making the measurement changes the quantity being measured. Think of other examples of measurements that change or disturb the quantity being measured.

3.3 Nuclei physics

3.3.1 Properties of nuclei

During the past century, applications of nuclear physics have had enormous effects on humankind, some beneficial, some catastrophic. Many people have strong opinions about applications such as bombs and reactors. Ideally, those opinions should be based on understanding, not on prejudice or emotion, and we hope this chapter will help you to reach that ideal.

Every atom contains at its center an extremely dense, positively charged *nucleus*, which is much smaller than the overall size of the atom but contains most of its total mass. We will look at several important general properties of nuclei and of the nuclear force that holds them together. The stability or instability of a particular nucleus is determined by the competition between the attractive nuclear force among the protons and neutrons and the repulsive electrical interactions among the protons. Unstable nuclei *decay*, transforming themselves spontaneously into other nuclei by a variety of processes. Nuclear reactions can also be induced by impact on a nucleus of a particle or another nucleus. Two classes of reactions of special interest are *fission* and *fusion*. We could not survive without the energy released by one nearby fusion reactor, our sun.

As we described, Rutherford found that the nucleus is tens of thousands of times smaller in radius than the atom itself. Since Rutherford's initial experiments, many additional scattering experiments have been performed, using high-energy protons, electrons, and neutrons as well as alpha particles (helium-4 nuclei). These experiments show that we can model a nucleus as a sphere with a radius R that depends on the total number of nucleons (neutrons and protons) in the nucleus. This number is called the **nucleon number** A. The radii of most nuclei are represented quite well by the equation

$$R = R_0 A^{1/3} \tag{234}$$

where R_0 is an experimentally determined constant:

$$R_0 = 1.2 \times 10^{-15}\, m = 1.2\, fm$$

The nucleon number A in Eq. (234) is also called the **mass number** because it is the nearest whole number to the mass of the nucleus measured in unified atomic mass units (u). (The proton mass and the neutron mass are both approximately 1 u.) The best current conversion factor is

$$1\,u = 1.660538782(83) \times 10^{-27} kg$$

Note that when we speak of the masses of nuclei and particles, we mean their rest masses.

The building blocks of the nucleus are the proton and the neutron. In a neutral atom, the nucleus is surrounded by one electron for every proton in the nucleus. The masses of these particles are

$$\text{Proton: } m_p = 1.007276\,u = 1.672622 \times 10^{-27} kg$$
$$\text{Neutron: } m_p = 1.008665\,u = 1.674927 \times 10^{-27} kg$$
$$\text{Electron: } m_p = 0.00054858\,u = 9.10938 \times 10^{-31} kg$$

The number of protons in a nucleus is the **atomic number** *Z*. The number of neutrons is the **neutron number** *N*. The nucleon number or mass number *A* is the sum of the number of protons *Z* and the number of neutrons *N*:

$$A = Z + N \qquad (235)$$

A single nuclear species having specific values of both Z and *N* is called a *nuclide*. The electron structure of an atom, which is responsible for its chemical properties, is determined by the charge Ze of the nucleus. The table shows some nuclides that have the same *Z* but different *N*. These nuclides are called **isotopes** of that element; they have different masses because they have different numbers of neutrons in their nuclei. A familiar example is chlorine (Cl, *Z*=17). About 76% of chlorine nuclei have *N*=18; the other 24% have *N*=20. Different isotopes of an element usually have slightly different physical properties such as melting and boiling temperatures and diffusion rates. The two common isotopes of uranium with *A*=235 and 238 are usually separated industrially by taking advantage of the different diffusion rates of gaseous uranium hexafluoride (UF_7) containing the two isotopes.

The symbol of the element, with a pre-subscript equal to Z and a pre-superscript equal to the mass number *A*. The general format for an element El is ${}^{A}_{Z}El$. The isotopes of chlorine mentioned above, with *A*=35 and 37, are written ${}^{35}_{17}Cl$ and ${}^{37}_{17}Cl$ and pronounced "chlorine-35" and "chlorine-37," respectively. This name of the element determines the atomic number *Z*, so the pre-subscript *Z* is sometimes omitted, as in ${}^{37}Cl$.

3.3.2 Radioactivity

Among about 2500 known nuclides, fewer than 300 are stable. The others are unstable structures that decay to form other nuclides by emitting particles and electromagnetic radiation, a process called **radioactivity.** The time scale of these decay processes ranges from a small fraction of a microsecond to billions of years. The *stable* nuclides are shown by dots on the graph in Fig. 129, where the neutron number *N* and proton number (or atomic number) *Z* for each nuclide are plotted. Such a chart is called a *Segru chart*, after its inventor, the Italian-American physicist Emilio Segre (1905–1989).

Each blue line perpendicular to the line *N=Z* represents a specific value of the mass number *A=Z+N*. Most lines of constant *A* pass through only one or two stable nuclides; that is, there is usually a very narrow range of stability for a given mass number. The lines at *A=20*, *A=40*, *A=60*, and *A=80* are examples. In four cases these lines pass through *three* stable nuclides—namely, at *A=96*, *124*, *130*, and *136*.

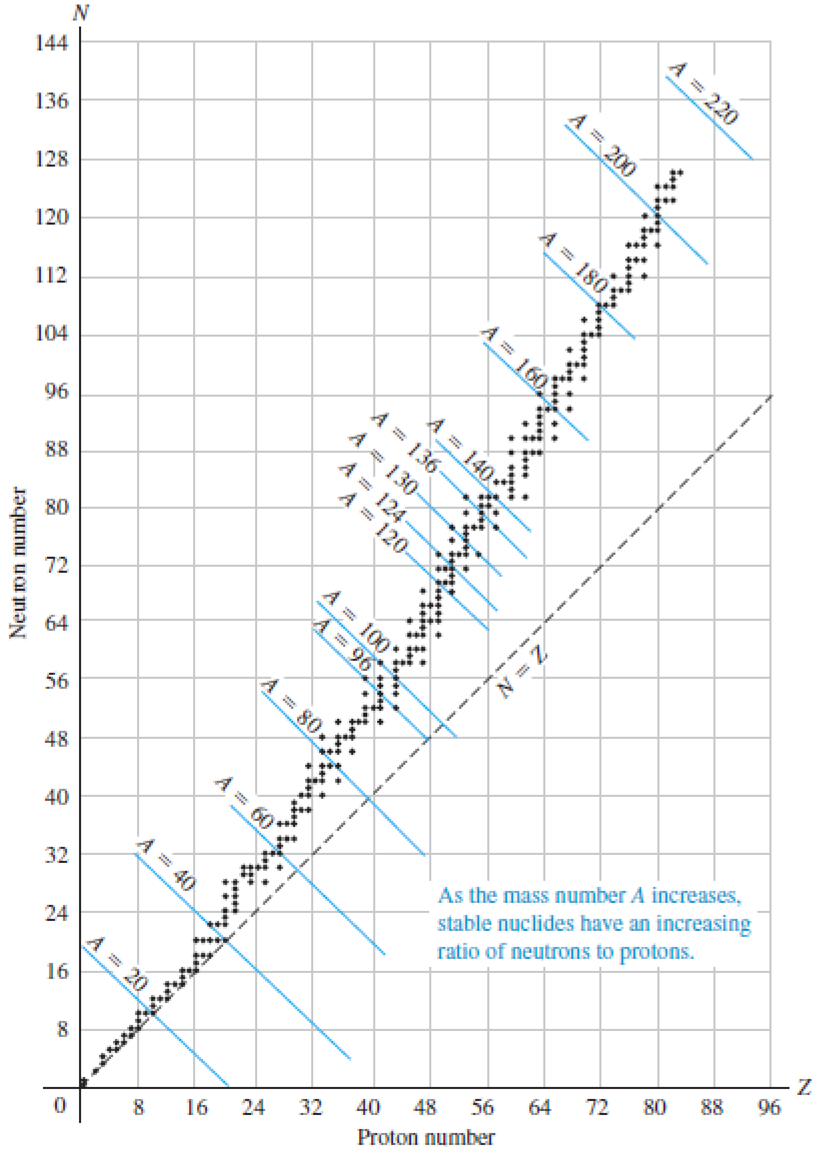


Figure 129 - Segrè chart showing neutron number and proton number for stable nuclides

Four stable nuclides have both odd *Z* and odd *N*:

$$^{2}_{1}H,\ \ ^{6}_{3}Li,\ \ ^{10}_{5}B\ ,\ \ ^{14}_{7}N$$

These are called *odd-odd nuclides*. The absence of other odd-odd nuclides shows the influence of pairing. Also, there is *no* stable nuclide with *A=5* or *A=8*. The doubly magic ${}^{4}_{2}He$ nucleus, with a pair of protons and a pair of neutrons, has no interest in accepting a fifth particle into its structure. Collections of eight nucleons decay to smaller nuclides, with ${}^{8}_{4}Be$ a nucleus immediately splitting into two ${}^{4}_{2}He$ nuclei.

The points on the Segrи chart representing stable nuclides define a rather narrow stability region. For low mass numbers, the numbers of protons and neutrons are approximately equal, $N \approx Z$. The ratio N/Z increases gradually with *A*, up to about 1.6 at large mass numbers, because of the increasing influence of the electrical repulsion of the protons. Points to the right of the stability region represent nuclides that have too many protons relative to neutrons to be stable. In these cases, repulsion wins, and the nucleus comes apart. To the left are nuclides with too many neutrons relative to protons. In these cases the energy associated with the neutrons is out of balance with that associated with the protons, and the nuclides decay in a process that converts neutrons to protons. The graph also shows that no nuclide with $A > 209$ or $Z > 83$ is stable. A nucleus is unstable if it is too big. Note that there is no stable nuclide with $Z = 43$ (technetium) or 61 (promethium).

Alpha Decay. Nearly 90% of the 2500 known nuclides are *radioactive;* they are not stable but decay into other nuclides. When unstable nuclides decay into different nuclides, they usually emit alpha (α) or beta (β) particles. An **alpha particle** is a ${}^{4}_{2}He$ nucleus, two protons and two neutrons bound together, with total spin zero. Alpha emission occurs principally with nuclei that are too large to be stable. When a nucleus emits an alpha particle, its *N* and *Z* values each decrease by 2 and *A* decreases by 4, moving it closer to stable territory on the Segre chart.

A familiar example of an alpha emitter is radium, ${}^{226}_{88}Ra$ (Fig. 130a). The speed of the emitted alpha particle, determined from the curvature of its path in a transverse magnetic field, is about $1.52 \times 10^{7}\ m/s$. This speed, although large, is only 5% of the speed of light, so we can use the nonrelativistic kinetic-energy expression $K = \frac{1}{2}mv^{2}$:

$$K = \frac{1}{2}(6.64 \times 10^{-27} kg)(1.52 \times 10^{7}\ m/s) = 7.67 \times 10^{-13} J = 4.79\ MeV$$

Alpha particles are always emitted with definite kinetic energies, determined by conservation of momentum and energy. Because of their charge and mass, alpha particles can travel only several centimeters in air, or a few tenths or hundredths of a millimeter through solids, before they are brought to rest by collisions.

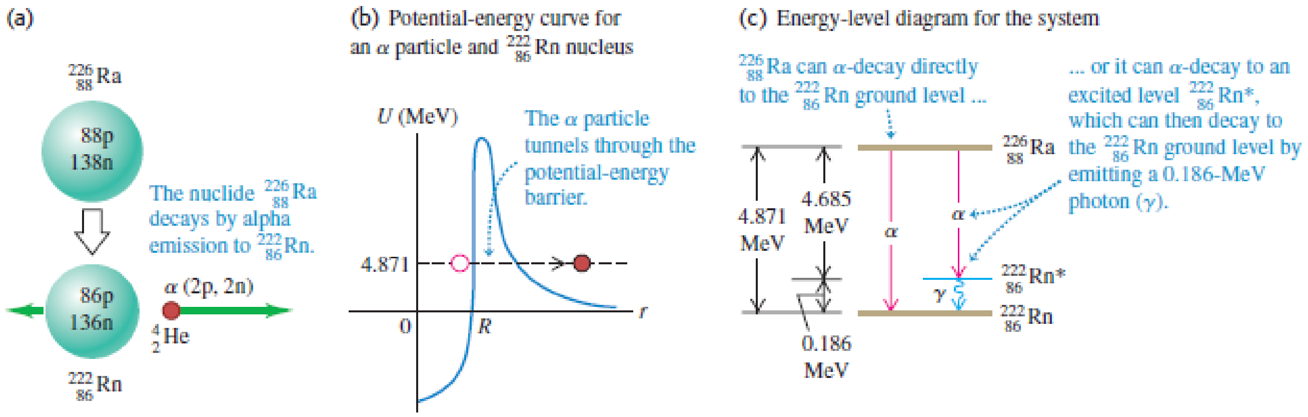

(a)
$^{226}_{88}$Ra
88p
138n
The nuclide $^{226}_{88}$Ra decays by alpha emission to $^{222}_{86}$Rn.
86p
136n
α (2p, 2n)
$^{4}_{2}$He
$^{222}_{86}$Rn
(b) Potential-energy curve for an α particle and $^{222}_{86}$Rn nucleus
U (MeV)
The α particle tunnels through the potential-energy barrier.
4.871
0
R
r
(c) Energy-level diagram for the system
$^{226}_{88}$Ra can α-decay directly to the $^{222}_{86}$Rn ground level ...
... or it can α-decay to an excited level $^{222}_{86}$Rn*, which can then decay to the $^{222}_{86}$Rn ground level by emitting a 0.186-MeV photon (γ).
4.871 MeV
4.685 MeV
α
α
γ
$^{226}_{88}$Ra
$^{222}_{86}$Rn*
$^{222}_{86}$Rn
0.186 MeV

Figure 130 – Alpha decay of the unstable radium nuclide $^{226}_{88}Ra$

Some nuclei can spontaneously decay by emission of particles because energy is released in their alpha decay. You can use conservation of mass-energy to show that

> **alpha decay is possible whenever the mass of the original neutral atom is greater than the sum of the masses of the final neutral atom and the neutral helium-4 atom.**

In alpha decay, the α particle tunnels through a potential-energy barrier, as Fig. 130b shows.

Beta Decay. There are three different simple types of beta decay: *beta-minus*, *beta-plus*, and *electron capture*. A **beta-minus particle** (β^-) is an electron. It's not obvious how a nucleus can emit an electron if there aren't any electrons in the nucleus. Emission of a β^- involves *transformation* of a neutron into a proton, an electron, and a third particle called an *antineutrino*. In fact, if you freed a neutron from a nucleus, it would decay into a proton, an electron, and an antineutrino in an average time of about 15 minutes.

Beta particles can be identified and their speeds can be measured with techniques that are similar to the Thomson experiments. The speeds of beta particles range up to 0.9995 of the speed of light, so their motion is highly relativistic. They are emitted with a continuous spectrum of energies. This would not be possible if the only two particles were the β^- and the recoiling nucleus, since energy and momentum conservation would then require a definite speed for the β^-. Thus there must be a third particle involved. From conservation of charge, it must be neutral, and from conservation of angular momentum, it must be a spin ½ particle.

This third particle is an antineutrino, the *antiparticle* of a **neutrino**. The symbol for a neutrino is ν_e (the Greek letter nu). Both the neutrino and the antineutrino have zero charge and zero (or very small) mass and therefore produce very little observable effect when passing through matter. Both evaded detection until 1953, when Frederick Reines and Clyde Cowan succeeded in observing the antineutrino directly. We now know that there are at least three varieties of neutrinos, each with its corresponding antineutrino; one is associated with beta decay and the other two are associated with the decay of two unstable particles, the muon and the tau particle. The antineutrino that is emitted in β^- decay is denoted as $\bar{\nu}_e$. The basic process of β^- decay is

$$n \rightarrow p + \beta^- + \bar{\nu}_e \qquad (236)$$

Beta-minus decay usually occurs with nuclides for which the neutron-to-proton ratio *N/*Z is too large for stability. In β^- decay, N decreases by 1, Z increases by 1, and *A* doesn't change. You can use conservation of mass-energy to show that

beta-minus decay can occur whenever the mass of the original neutral atom is larger than that of the final atom.

We have noted that β^- decay occurs with nuclides that have too large a neutron-to-proton ratio *N/Z*. Nuclides for which *N/Z* is too small for stability can emit a *positron*, the electron's antiparticle, which is identical to the electron but with positive charge. The basic process, called *beta-plus decay* (β^+) is

$$p \rightarrow n + \beta^+ + \nu_e \qquad (237)$$

where is β^+ a positron and ν_e is the electron neutrino.

Beta-plus decay can occur whenever the mass of the original neutral atom is at least two electron masses larger than that of the final atom.

You can show this using conservation of mass-energy.

The third type of beta decay is *electron capture*. There are a few nuclides for which β^+ emission is not energetically possible but in which an orbital electron (usually in the *K* shell) can combine with a proton in the nucleus to form a neutron and a neutrino. The neutron remains in the nucleus and the neutrino is emitted. The basic process is

$$p + \beta^- \rightarrow n + \nu_e \qquad (238)$$

You can use conservation of mass-energy to show that

electron capture can occur whenever the mass of the original neutral atom is larger than that of the final atom.

In all types of beta decay, **A** remains constant. However, in beta-plus decay and electron capture, *N* increases by 1 and Z decreases by 1 as the neutron–proton ratio increases toward a more stable value. The reaction of Eq. (238) also helps to explain the formation of a neutron star.

3.3.3 Nuclei reactions

In the preceding sections we studied the decay of unstable nuclei, especially spontaneous emission of an α or β particle, sometimes followed by γ emission. Nothing needs to be done to initiate this decay, and nothing can be done to control it. This section examines some nuclear reactions, rearrangements of nuclear components that result from a bombardment by a particle rather than a spontaneous natural process. Rutherford suggested in 1919 that a massive particle with sufficient kinetic energy might be able to penetrate a nucleus. The result would be either a new nucleus with greater atomic number and mass number or a decay of the original nucleus. Rutherford bombarded nitrogen ($^{14}_{7}N$) with particles and obtained an oxygen ($^{17}_{8}O$) nucleus and a proton:

$$^{4}_{2}He + ^{14}_{7}N \rightarrow ^{17}_{8}O + ^{1}_{1}H \qquad (239)$$

Rutherford used alpha particles from naturally radioactive sources. Early we'll describe some of the particle accelerators that are now used to initiate nuclear reactions.

Nuclear reactions are subject to several conservation laws. The classical conservation principles for charge, momentum, angular momentum, and energy (including rest energies) are obeyed in all nuclear reactions. An additional conservation law, not anticipated by classical physics, is conservation of the total number of nucleons. The numbers of protons and neutrons need not be conserved separately; in β decay, neutrons and protons change into one another.

When two nuclei interact, charge conservation requires that the sum of the initial atomic numbers must equal the sum of the final atomic numbers. Because of conservation of nucleon number, the sum of the initial mass numbers must also equal the sum of the final mass numbers. In general, these are not elastic collisions, and the total initial mass does not equal the total final mass.

The difference between the masses before and after the reaction corresponds to the reaction energy, according to the mass–energy relationship $E = mc^2$. If initial particles A and B interact to produce final particles C and D, the reaction energy Q is defined as

$$Q = (M_A + M_B - M_C - M_D)c^2 \qquad (240)$$

To balance the electrons, we use the neutral atomic masses in Eq. (240). That is, we use the mass of $^{1}_{1}H$ for a proton, $^{2}_{1}H$ for a deuteron, $^{4}_{2}He$ for an α particle, and so on. When Q is positive, the total mass decreases and the total kinetic energy increases. Such a reaction is called an exoergic reaction. When Q is negative, the mass increases and the kinetic energy decreases, and the reaction is called an endoergic reaction. The terms exothermal and endothermal, borrowed from chemistry, are also used. In an endoergic reaction the reaction cannot occur at all unless the initial kinetic energy in the center-of-mass reference frame is at least as great as $|Q|$. That is, there is a threshold energy, the minimum kinetic energy to make an endoergic reaction go.

Ordinarily, the endoergic reaction of part (b) of Example 43.11 would be produced by bombarding stationary $^{14}_{7}N$ nuclei with alpha particles from an accelerator. In this case an alpha's kinetic energy must be greater than 1.192 MeV. If all the alpha's kinetic energy went solely to increasing the rest energy, the final kinetic energy would be zero, and momentum would not be conserved. When a particle with mass m and kinetic energy K collides with a stationary particle with mass M, the total kinetic energy K_{cm} in the center-of-mass coordinate system (the energy available to cause reactions) is

$$K_{cm} = \frac{M}{M+m} K \tag{241}$$

This expression assumes that the kinetic energies of the particles and nuclei are much less than their rest energies.

Discussion questions

1. Can a hydrogen atom emit x rays? If so, how? If not, why not?
2. Neutrons have a magnetic dipole moment and can undergo spin flips by absorbing electromagnetic radiation. Why, then, are protons rather than neutrons used in MRI of body tissues?
3. Why aren't the masses of all nuclei integer multiples of the mass of a single nucleon?
4. What are the six known elements for which *Z* is a magic number? Discuss what properties these elements have as a consequence of their special values of *Z*.
5. Heavy, unstable nuclei usually decay by emitting an α or particle. Why don't they usually emit a single proton or neutron?
6. The only two stable nuclides with more protons than neutrons ${}^{1}_{1}H$ are and ${}^{3}_{2}He$. Why is Z>N so uncommon?
7. Compared to particles with the same energy, particles can much more easily penetrate through matter. Why is this?
8. In a nuclear decay equation, why can we represent an electron as ${}^{0}_{-1}\beta^{-}$?. What are the equivalent representations for a positron, a neutrino, and an antineutrino?
9. Why is the alpha, beta, or gamma decay of an unstable nucleus unaffected by the *chemical* situation of the atom, such as the nature of the molecule or solid in which it is bound? The chemical situation of the atom can, however, have an effect on the half-life in electron capture. Why is this?
10. In the process of *internal conversion*, a nucleus decays from an excited state to a ground state by giving the excitation energy directly to an atomic electron rather than emitting a gamma-ray photon. Why can this process also produce x-ray photons?
11. **T**he activity of atmospheric carbon *before* 1900 was given. Discuss why this activity may have changed since 1900.
12. One problem in radiocarbon dating of biological samples, especially very old ones, is that they can easily be contaminated with modern biological material during the measurement process. What effect would such contamination have on the estimated age? Why is such contamination a more serious problem for samples of older material than for samples of younger material?

13. The most common radium isotope found on earth, has a half-life of about 1600 years. If the earth was formed well over 10^9 years ago, why is there any radium left now?
14. Fission reactions occur only for nuclei with large nucleon numbers, while exoergic fusion reactions occur only for nuclei with small nucleon numbers. Why is this?
15. When a large nucleus splits during nuclear fission, the daughter nuclei of the fission fly apart with enormous kineticenergy. Why does this happen?
16. As stars age, they use up their supply of hydrogen and eventually begin producing energy by a reaction that involves the fusion of three helium nuclei to form a carbon nucleus. Would you expect the interiors of these old stars to be hotter or cooler than the interiors of younger stars? Explain.

Reference

1 Hugh D. Young, Roger A. Freedman – University physics with modern physics, 13-th edition, 2012, 1598
2 Benjamin Crowell - Simple Nature. An Introduction to Physics for Engineering and Physical Science Students, 2001-2008, p.452
3 Robert G. Brown - Introductory Physics I, Elementary Mechanics, 2007
James H. Dann Basic Physics II, p.371
4 FHSST Authors - The Free High School Science Texts: A Textbook for High School Students Studying Physics, 2005, 528