

Chapter One

ELECTRIC CHARGES AND FIELDS

1.1 INTRODUCTION

All of us have the experience of seeing a spark or hearing a crackle when we take off our synthetic clothes or sweater, particularly in dry weather. This is almost inevitable with ladies garments like a polyester saree. Have you ever tried to find any explanation for this phenomenon? Another common example of electric discharge is the lightning that we see in the sky during thunderstorms. We also experience a sensation of an electric shock either while opening the door of a car or holding the iron bar of a bus after sliding from our seat. The reason for these experiences is discharge of electric charges through our body, which were accumulated due to rubbing of insulating surfaces. You might have also heard that this is due to generation of static electricity. This is precisely the topic we are going to discuss in this and the next chapter. Static means anything that does not move or change with time. *Electrostatics deals with the study of forces, fields and potentials arising from static charges.*

1.2 ELECTRIC CHARGE

Historically the credit of discovery of the fact that amber rubbed with wool or silk cloth attracts light objects goes to Thales of Miletus, Greece, around 600 BC. The name electricity is coined from the Greek word *elektron* meaning *amber*. Many such pairs of materials were known which

Physics

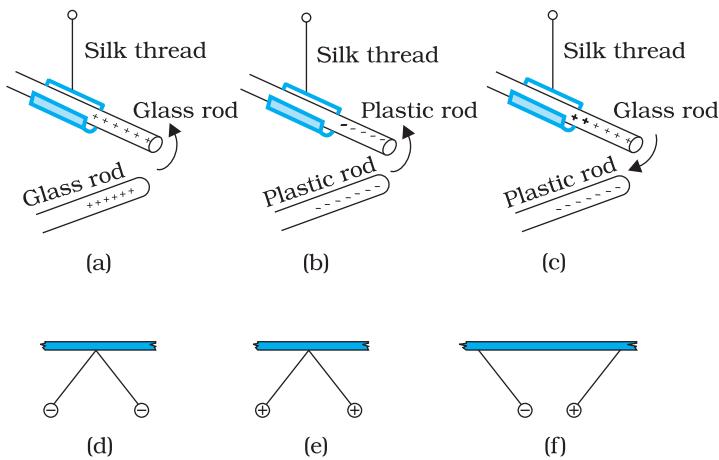


FIGURE 1.1 Rods and pith balls: like charges repel and unlike charges attract each other.

Interactive animation on simple electrostatic experiments:
<http://ephysics.physics.ucla.edu/travoltage/HTML/staticElectricity.htm>

PHYSICS

on rubbing could attract light objects like straw, pith balls and bits of papers. You can perform the following activity at home to experience such an effect. Cut out long thin strips of white paper and lightly iron them. Take them near a TV screen or computer monitor. You will see that the strips get attracted to the screen. In fact they remain stuck to the screen for a while.

It was observed that if two glass rods rubbed with wool or silk cloth are brought close to each other, they repel each other [Fig. 1.1(a)]. The two strands of wool or two pieces of silk cloth, with which the rods were rubbed, also repel each other. However, the glass rod and

wool attracted each other. Similarly, two plastic rods rubbed with cat's fur repelled each other [Fig. 1.1(b)] but attracted the fur. On the other hand, the plastic rod attracts the glass rod [Fig. 1.1(c)] and repels the silk or wool with which the glass rod is rubbed. The glass rod repels the fur.

If a plastic rod rubbed with fur is made to touch two small pith balls (now-a-days we can use polystyrene balls) suspended by silk or nylon thread, then the balls repel each other [Fig. 1.1(d)] and are also repelled by the rod. A similar effect is found if the pith balls are touched with a glass rod rubbed with silk [Fig. 1.1(e)]. A dramatic observation is that a pith ball touched with glass rod attracts another pith ball touched with plastic rod [Fig. 1.1(f)].

These seemingly simple facts were established from years of efforts and careful experiments and their analyses. It was concluded, after many careful studies by different scientists, that there were only two kinds of an entity which is called the *electric charge*. We say that the bodies like glass or plastic rods, silk, fur and pith balls are electrified. They acquire an electric charge on rubbing. The experiments on pith balls suggested that there are two kinds of electrification and we find that (i) *like charges repel* and (ii) *unlike charges attract* each other. The experiments also demonstrated that the charges are transferred from the rods to the pith balls on contact. It is said that the pith balls are electrified or are charged by contact. The property which differentiates the two kinds of charges is called the *polarity* of charge.

When a glass rod is rubbed with silk, the rod acquires one kind of charge and the silk acquires the second kind of charge. This is true for any pair of objects that are rubbed to be electrified. Now if the electrified glass rod is brought in contact with silk, with which it was rubbed, they no longer attract each other. They also do not attract or repel other light objects as they did on being electrified.

Thus, the charges acquired after rubbing are lost when the charged bodies are brought in contact. What can you conclude from these observations? It just tells us that unlike charges acquired by the objects

Electric Charges and Fields

neutralise or nullify each other's effect. Therefore the charges were named as *positive* and *negative* by the American scientist Benjamin Franklin. We know that when we add a positive number to a negative number of the same magnitude, the sum is zero. This might have been the philosophy in naming the charges as positive and negative. By convention, the charge on glass rod or cat's fur is called positive and that on plastic rod or silk is termed negative. If an object possesses an electric charge, it is said to be electrified or charged. When it has no charge it is said to be neutral.

UNIFICATION OF ELECTRICITY AND MAGNETISM

In olden days, electricity and magnetism were treated as separate subjects. Electricity dealt with charges on glass rods, cat's fur, batteries, lightning, etc., while magnetism described interactions of magnets, iron filings, compass needles, etc. In 1820 Danish scientist Oersted found that a compass needle is deflected by passing an electric current through a wire placed near the needle. Ampere and Faraday supported this observation by saying that electric charges in motion produce magnetic fields and moving magnets generate electricity. The unification was achieved when the Scottish physicist Maxwell and the Dutch physicist Lorentz put forward a theory where they showed the interdependence of these two subjects. This field is called *electromagnetism*. Most of the phenomena occurring around us can be described under electromagnetism. Virtually every force that we can think of like friction, chemical force between atoms holding the matter together, and even the forces describing processes occurring in cells of living organisms, have its origin in electromagnetic force. Electromagnetic force is one of the fundamental forces of nature.

Maxwell put forth four equations that play the same role in classical electromagnetism as Newton's equations of motion and gravitation law play in mechanics. He also argued that light is electromagnetic in nature and its speed can be found by making purely electric and magnetic measurements. He claimed that the science of optics is intimately related to that of electricity and magnetism.

The science of electricity and magnetism is the foundation for the modern technological civilisation. Electric power, telecommunication, radio and television, and a wide variety of the practical appliances used in daily life are based on the principles of this science. Although charged particles in motion exert both electric and magnetic forces, in the frame of reference where all the charges are at rest, the forces are purely electrical. You know that gravitational force is a long-range force. Its effect is felt even when the distance between the interacting particles is very large because the force decreases inversely as the square of the distance between the interacting bodies. We will learn in this chapter that electric force is also as pervasive and is in fact stronger than the gravitational force by several orders of magnitude (refer to Chapter 1 of Class XI Physics Textbook).

A simple apparatus to detect charge on a body is the *gold-leaf electroscope* [Fig. 1.2(a)]. It consists of a vertical metal rod housed in a box, with two thin gold leaves attached to its bottom end. When a charged object touches the metal knob at the top of the rod, charge flows on to the leaves and they diverge. The degree of divergence is an indicator of the amount of charge.

Physics

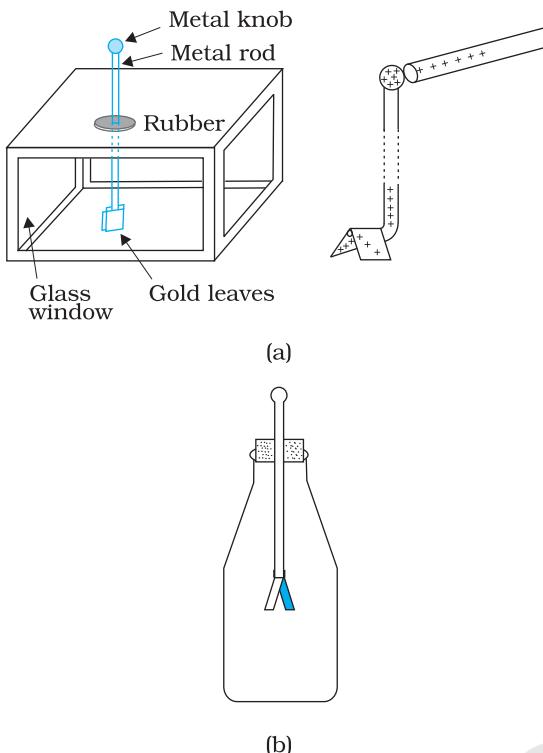


FIGURE 1.2 Electroscopes: (a) The gold leaf electroscope, (b) Schematics of a simple electroscope.

Students can make a simple electroscope as follows [Fig. 1.2(b)]: Take a thin aluminium curtain rod with ball ends fitted for hanging the curtain. Cut out a piece of length about 20 cm with the ball at one end and flatten the cut end. Take a large bottle that can hold this rod and a cork which will fit in the opening of the bottle. Make a hole in the cork sufficient to hold the curtain rod snugly. Slide the rod through the hole in the cork with the cut end on the lower side and ball end projecting above the cork. Fold a small, thin aluminium foil (about 6 cm in length) in the middle and attach it to the flattened end of the rod by cellulose tape. This forms the leaves of your electroscope. Fit the cork in the bottle with about 5 cm of the ball end projecting above the cork. A paper scale may be put inside the bottle in advance to measure the separation of leaves. The separation is a rough measure of the amount of charge on the electroscope.

To understand how the electroscope works, use the white paper strips we used for seeing the attraction of charged bodies. Fold the strips into half so that you make a mark of fold. Open the strip and iron it lightly with the mountain fold up, as shown in Fig. 1.3. Hold the strip by pinching it at the fold. You would notice that the two halves move apart.

This shows that the strip has acquired charge on ironing. When you fold it into half, both the halves have the same charge. Hence they repel each other. The same effect is seen in the leaf electroscope. On charging the curtain rod by touching the ball end with an electrified body, charge is transferred to the curtain rod and the attached aluminium foil. Both the halves of the foil get similar charge and therefore repel each other. The divergence in the leaves depends on the amount of charge on them. Let us first try to understand why material bodies acquire charge.

You know that all matter is made up of atoms and/or molecules. Although normally the materials are electrically neutral, they do contain charges; but their charges are exactly balanced. Forces that hold the molecules together, forces that hold atoms together in a solid, the adhesive force of glue, forces associated with surface tension, all are basically electrical in nature, arising from the forces between charged particles. Thus the electric force is all pervasive and it encompasses almost each and every field associated with our life. It is therefore essential that we learn more about such a force.

To electrify a neutral body, we need to add or remove one kind of charge. When we say that a body is charged, we always refer to this excess charge or deficit of charge. In solids, some of the electrons, being less tightly bound in the atom, are the charges which are transferred from one body to the other. A body can thus be charged positively by losing some of its electrons. Similarly, a body can be charged negatively

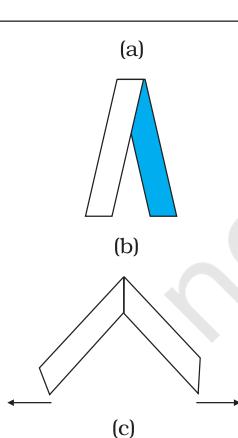


FIGURE 1.3 Paper strip experiment.

by gaining electrons. When we rub a glass rod with silk, some of the electrons from the rod are transferred to the silk cloth. Thus the rod gets positively charged and the silk gets negatively charged. No new charge is created in the process of rubbing. Also the number of electrons, that are transferred, is a very small fraction of the total number of electrons in the material body. Also only the less tightly bound electrons in a material body can be transferred from it to another by rubbing. Therefore, when a body is rubbed with another, the bodies get charged and that is why we have to stick to certain pairs of materials to notice charging on rubbing the bodies.

1.3 CONDUCTORS AND INSULATORS

A metal rod held in hand and rubbed with wool will not show any sign of being charged. However, if a metal rod with a wooden or plastic handle is rubbed without touching its metal part, it shows signs of charging. Suppose we connect one end of a copper wire to a neutral pith ball and the other end to a negatively charged plastic rod. We will find that the pith ball acquires a negative charge. If a similar experiment is repeated with a nylon thread or a rubber band, no transfer of charge will take place from the plastic rod to the pith ball. Why does the transfer of charge not take place from the rod to the ball?

Some substances readily allow passage of electricity through them, others do not. Those which allow electricity to pass through them easily are called *conductors*. They have electric charges (electrons) that are comparatively free to move inside the material. Metals, human and animal bodies and earth are conductors. Most of the non-metals like glass, porcelain, plastic, nylon, wood offer high resistance to the passage of electricity through them. They are called *insulators*. Most substances fall into one of the two classes stated above*.

When some charge is transferred to a conductor, it readily gets distributed over the entire surface of the conductor. In contrast, if some charge is put on an insulator, it stays at the same place. You will learn why this happens in the next chapter.

This property of the materials tells you why a nylon or plastic comb gets electrified on combing dry hair or on rubbing, but a metal article like spoon does not. The charges on metal leak through our body to the ground as both are conductors of electricity.

When we bring a charged body in contact with the earth, all the excess charge on the body disappears by causing a momentary current to pass to the ground through the connecting conductor (such as our body). This process of sharing the charges with the earth is called *grounding or earthing*. Earthing provides a safety measure for electrical circuits and appliances. A thick metal plate is buried deep into the earth and thick wires are drawn from this plate; these are used in buildings for the purpose of earthing near the mains supply. The electric wiring in our houses has three wires: live, neutral and earth. The first two carry

* There is a third category called *semiconductors*, which offer resistance to the movement of charges which is intermediate between the conductors and insulators.

Physics

electric current from the power station and the third is earthed by connecting it to the buried metal plate. Metallic bodies of the electric appliances such as electric iron, refrigerator, TV are connected to the earth wire. When any fault occurs or live wire touches the metallic body, the charge flows to the earth without damaging the appliance and without causing any injury to the humans; this would have otherwise been unavoidable since the human body is a conductor of electricity.

1.4 CHARGING BY INDUCTION

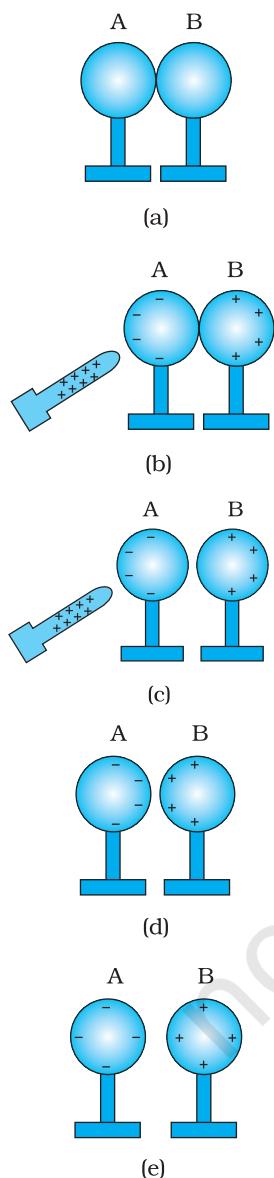


FIGURE 1.4 Charging by induction.

When we touch a pith ball with an electrified plastic rod, some of the negative charges on the rod are transferred to the pith ball and it also gets charged. Thus the pith ball is *charged by contact*. It is then repelled by the plastic rod but is attracted by a glass rod which is oppositely charged. However, why a electrified rod attracts light objects, is a question we have still left unanswered. Let us try to understand what could be happening by performing the following experiment.

- (i) Bring two metal spheres, A and B, supported on insulating stands, in contact as shown in Fig. 1.4(a).
- (ii) Bring a positively charged rod near one of the spheres, say A, taking care that it does not touch the sphere. The free electrons in the spheres are attracted towards the rod. This leaves an excess of positive charge on the rear surface of sphere B. Both kinds of charges are bound in the metal spheres and cannot escape. They, therefore, reside on the surfaces, as shown in Fig. 1.4(b). The left surface of sphere A, has an excess of negative charge and the right surface of sphere B, has an excess of positive charge. However, not all of the electrons in the spheres have accumulated on the left surface of A. As the negative charge starts building up at the left surface of A, other electrons are repelled by these. In a short time, equilibrium is reached under the action of force of attraction of the rod and the force of repulsion due to the accumulated charges. Fig. 1.4(b) shows the equilibrium situation. The process is called *induction of charge* and happens almost instantly. The accumulated charges remain on the surface, as shown, till the glass rod is held near the sphere. If the rod is removed, the charges are not acted by any outside force and they redistribute to their original neutral state.
- (iii) Separate the spheres by a small distance while the glass rod is still held near sphere A, as shown in Fig. 1.4(c). The two spheres are found to be oppositely charged and attract each other.
- (iv) Remove the rod. The charges on spheres rearrange themselves as shown in Fig. 1.4(d). Now, separate the spheres quite apart. The charges on them get uniformly distributed over them, as shown in Fig. 1.4(e).

In this process, the metal spheres will each be equal and oppositely charged. This is *charging by induction*. The positively charged glass rod does not lose any of its charge, contrary to the process of charging by contact.

When electrified rods are brought near light objects, a similar effect takes place. The rods induce opposite charges on the near surfaces of the objects and similar charges move to the farther side of the object.

Electric Charges and Fields

[This happens even when the light object is not a conductor. The mechanism for how this happens is explained later in Sections 1.10 and 2.10.] The centres of the two types of charges are slightly separated. We know that opposite charges attract while similar charges repel. However, the magnitude of force depends on the distance between the charges and in this case the force of attraction outweighs the force of repulsion. As a result the particles like bits of paper or pith balls, being light, are pulled towards the rods.

Example 1.1 How can you charge a metal sphere positively without touching it?

Solution Figure 1.5(a) shows an uncharged metallic sphere on an insulating metal stand. Bring a negatively charged rod close to the metallic sphere, as shown in Fig. 1.5(b). As the rod is brought close to the sphere, the free electrons in the sphere move away due to repulsion and start piling up at the farther end. The near end becomes positively charged due to deficit of electrons. This process of charge distribution stops when the net force on the free electrons inside the metal is zero. Connect the sphere to the ground by a conducting wire. The electrons will flow to the ground while the positive charges at the near end will remain held there due to the attractive force of the negative charges on the rod, as shown in Fig. 1.5(c). Disconnect the sphere from the ground. The positive charge continues to be held at the near end [Fig. 1.5(d)]. Remove the electrified rod. The positive charge will spread uniformly over the sphere as shown in Fig. 1.5(e).

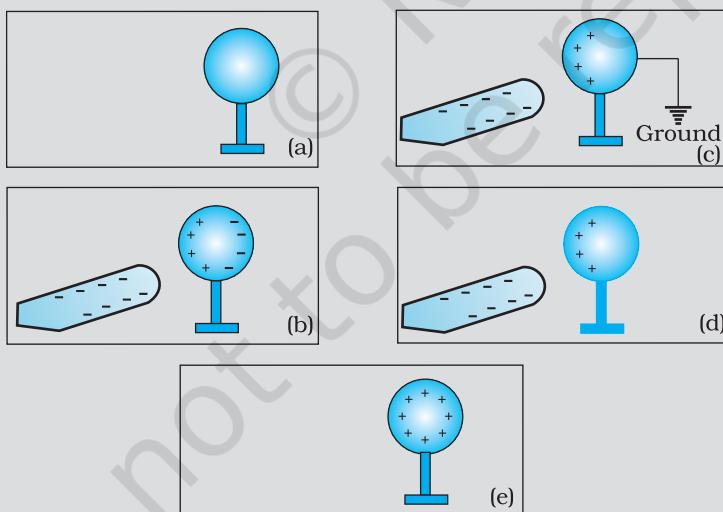


FIGURE 1.5

In this experiment, the metal sphere gets charged by the process of induction and the rod does not lose any of its charge.

Similar steps are involved in charging a metal sphere negatively by induction, by bringing a positively charged rod near it. In this case the electrons will flow from the ground to the sphere when the sphere is connected to the ground with a wire. Can you explain why?



Interactive animation on charging a two-sphere system by induction:
<http://www.physicsclassroom.com/mmedia/electrostatics/itsr.cfm>

1.5 BASIC PROPERTIES OF ELECTRIC CHARGE

We have seen that there are two types of charges, namely positive and negative and their effects tend to cancel each other. Here, we shall now describe some other properties of the electric charge.

If the sizes of charged bodies are very small as compared to the distances between them, we treat them as *point charges*. All the charge content of the body is assumed to be concentrated at one point in space.

1.5.1 Additivity of charges

We have not as yet given a quantitative definition of a charge; we shall follow it up in the next section. We shall tentatively assume that this can be done and proceed. If a system contains two point charges q_1 and q_2 , the total charge of the system is obtained simply by adding algebraically q_1 and q_2 , i.e., charges add up like real numbers or they are scalars like the mass of a body. If a system contains n charges $q_1, q_2, q_3, \dots, q_n$, then the total charge of the system is $q_1 + q_2 + q_3 + \dots + q_n$. Charge has magnitude but no direction, similar to the mass. However, there is one difference between mass and charge. Mass of a body is always positive whereas a charge can be either positive or negative. Proper signs have to be used while adding the charges in a system. For example, the total charge of a system containing five charges +1, +2, -3, +4 and -5, in some arbitrary unit, is $(+1) + (+2) + (-3) + (+4) + (-5) = -1$ in the same unit.

1.5.2 Charge is conserved

We have already hinted to the fact that when bodies are charged by rubbing, there is transfer of electrons from one body to the other; no new charges are either created or destroyed. A picture of particles of electric charge enables us to understand the idea of conservation of charge. When we rub two bodies, what one body gains in charge the other body loses. Within an isolated system consisting of many charged bodies, due to interactions among the bodies, charges may get redistributed but it is found that *the total charge of the isolated system is always conserved*. Conservation of charge has been established experimentally.

It is not possible to create or destroy net charge carried by any isolated system although the charge carrying particles may be created or destroyed in a process. Sometimes nature creates charged particles: a neutron turns into a proton and an electron. The proton and electron thus created have equal and opposite charges and the total charge is zero before and after the creation.

1.5.3 Quantisation of charge

Experimentally it is established that all free charges are integral multiples of a basic unit of charge denoted by e . Thus charge q on a body is always given by

$$q = ne$$

Electric Charges and Fields

where n is any integer, positive or negative. This basic unit of charge is the charge that an electron or proton carries. By convention, the charge on an electron is taken to be negative; therefore charge on an electron is written as $-e$ and that on a proton as $+e$.

The fact that electric charge is always an integral multiple of e is termed as *quantisation of charge*. There are a large number of situations in physics where certain physical quantities are quantised. The quantisation of charge was first suggested by the experimental laws of electrolysis discovered by English experimentalist Faraday. It was experimentally demonstrated by Millikan in 1912.

In the International System (SI) of Units, a unit of charge is called a *coulomb* and is denoted by the symbol C. A coulomb is defined in terms of the unit of the electric current which you are going to learn in a subsequent chapter. In terms of this definition, one coulomb is the charge flowing through a wire in 1 s if the current is 1 A (ampere), (see Chapter 2 of Class XI, Physics Textbook , Part I). In this system, the value of the basic unit of charge is

$$e = 1.602192 \times 10^{-19} \text{ C}$$

Thus, there are about 6×10^{18} electrons in a charge of -1C . In electrostatics, charges of this large magnitude are seldom encountered and hence we use smaller units $1 \mu\text{C}$ (micro coulomb) = 10^{-6} C or 1 mC (milli coulomb) = 10^{-3} C .

If the protons and electrons are the only basic charges in the universe, all the observable charges have to be integral multiples of e . Thus, if a body contains n_1 electrons and n_2 protons, the total amount of charge on the body is $n_2 \times e + n_1 \times (-e) = (n_2 - n_1)e$. Since n_1 and n_2 are integers, their difference is also an integer. Thus the charge on any body is always an integral multiple of e and can be increased or decreased also in steps of e .

The step size e is, however, very small because at the macroscopic level, we deal with charges of a few μC . At this scale the fact that charge of a body can increase or decrease in units of e is not visible. The grainy nature of the charge is lost and it appears to be continuous.

This situation can be compared with the geometrical concepts of points and lines. A dotted line viewed from a distance appears continuous to us but is not continuous in reality. As many points very close to each other normally give an impression of a continuous line, many small charges taken together appear as a continuous charge distribution.

At the macroscopic level, one deals with charges that are enormous compared to the magnitude of charge e . Since $e = 1.6 \times 10^{-19} \text{ C}$, a charge of magnitude, say $1 \mu\text{C}$, contains something like 10^{13} times the electronic charge. At this scale, the fact that charge can increase or decrease only in units of e is not very different from saying that charge can take continuous values. Thus, at the macroscopic level, the quantisation of charge has no practical consequence and can be ignored. At the microscopic level, where the charges involved are of the order of a few tens or hundreds of e , i.e.,

they can be counted, they appear in discrete lumps and quantisation of charge cannot be ignored. It is the scale involved that is very important.

EXAMPLE 1.2

Example 1.2 If 10^9 electrons move out of a body to another body every second, how much time is required to get a total charge of 1 C on the other body?

Solution In one second 10^9 electrons move out of the body. Therefore the charge given out in one second is $1.6 \times 10^{-19} \times 10^9 \text{ C} = 1.6 \times 10^{-10} \text{ C}$. The time required to accumulate a charge of 1 C can then be estimated to be $1 \text{ C} \div (1.6 \times 10^{-10} \text{ C/s}) = 6.25 \times 10^9 \text{ s} = 6.25 \times 10^9 \div (365 \times 24 \times 3600) \text{ years} = 198 \text{ years}$. Thus to collect a charge of one coulomb, from a body from which 10^9 electrons move out every second, we will need approximately 200 years. One coulomb is, therefore, a very large unit for many practical purposes.

It is, however, also important to know what is roughly the number of electrons contained in a piece of one cubic centimetre of a material. A cubic piece of copper of side 1 cm contains about 2.5×10^{24} electrons.

EXAMPLE 1.3

Example 1.3 How much positive and negative charge is there in a cup of water?

Solution Let us assume that the mass of one cup of water is 250 g. The molecular mass of water is 18 g. Thus, one mole ($= 6.02 \times 10^{23}$ molecules) of water is 18 g. Therefore the number of molecules in one cup of water is $(250/18) \times 6.02 \times 10^{23}$.

Each molecule of water contains two hydrogen atoms and one oxygen atom, i.e., 10 electrons and 10 protons. Hence the total positive and total negative charge has the same magnitude. It is equal to $(250/18) \times 6.02 \times 10^{23} \times 10 \times 1.6 \times 10^{-19} \text{ C} = 1.34 \times 10^7 \text{ C}$.

1.6 COULOMB'S LAW

Coulomb's law is a quantitative statement about the force between two point charges. When the linear size of charged bodies are much smaller than the distance separating them, the size may be ignored and the charged bodies are treated as *point charges*. Coulomb measured the force between two point charges and found that it varied inversely as the square of the distance between the charges and was directly proportional to the product of the magnitude of the two charges and acted along the line joining the two charges. Thus, if two point charges q_1 , q_2 are separated by a distance r in vacuum, the magnitude of the force (\mathbf{F}) between them is given by

$$F = k \frac{|q_1 q_2|}{r^2} \quad (1.1)$$

How did Coulomb arrive at this law from his experiments? Coulomb used a torsion balance* for measuring the force between two charged metallic

* A torsion balance is a sensitive device to measure force. It was also used later by Cavendish to measure the very feeble gravitational force between two objects, to verify Newton's Law of Gravitation.

Electric Charges and Fields

spheres. When the separation between two spheres is much larger than the radius of each sphere, the charged spheres may be regarded as point charges. However, the charges on the spheres were unknown, to begin with. How then could he discover a relation like Eq. (1.1)? Coulomb thought of the following simple way: Suppose the charge on a metallic sphere is q . If the sphere is put in contact with an identical uncharged sphere, the charge will spread over the two spheres. By symmetry, the charge on each sphere will be $q/2^*$. Repeating this process, we can get charges $q/2$, $q/4$, etc. Coulomb varied the distance for a fixed pair of charges and measured the force for different separations. He then varied the charges in pairs, keeping the distance fixed for each pair. Comparing forces for different pairs of charges at different distances, Coulomb arrived at the relation, Eq. (1.1).

Coulomb's law, a simple mathematical statement, was initially experimentally arrived at in the manner described above. While the original experiments established it at a macroscopic scale, it has also been established down to subatomic level ($r \sim 10^{-10}$ m).

Coulomb discovered his law without knowing the *explicit* magnitude of the charge. In fact, it is the other way round: Coulomb's law can *now* be employed to furnish a definition for a unit of charge. In the relation, Eq. (1.1), k is so far arbitrary. We can choose any positive value of k . The choice of k determines the size of the unit of charge. In SI units, the value of k is about 9×10^9 . The unit of charge that results from this choice is called a coulomb which we defined earlier in Section 1.4. Putting this value of k in Eq. (1.1), we see that for $q_1 = q_2 = 1$ C, $r = 1$ m

$$F = 9 \times 10^9 \text{ N}$$

That is, 1 C is the charge that when placed at a distance of 1 m from another charge of the same magnitude *in vacuum* experiences an electrical force of repulsion of magnitude 9×10^9 N. One coulomb is evidently too big a unit to be used. In practice, in electrostatics, one uses smaller units like 1 mC or 1 μ C.

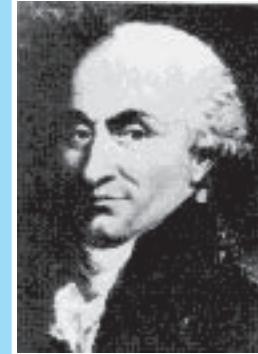
The constant k in Eq. (1.1) is usually put as $k = 1/4\pi\epsilon_0$ for later convenience, so that Coulomb's law is written as

$$F = \frac{1}{4\pi\epsilon_0} \frac{|q_1 q_2|}{r^2} \quad (1.2)$$

ϵ_0 is called the *permittivity of free space*. The value of ϵ_0 in SI units is

$$\epsilon_0 = 8.854 \times 10^{-12} \text{ C}^2 \text{ N}^{-1} \text{ m}^{-2}$$

* Implicit in this is the assumption of additivity of charges and conservation: two charges ($q/2$ each) add up to make a total charge q .



Charles Augustin de Coulomb (1736 – 1806)

Coulomb, a French physicist, began his career as a military engineer in the West Indies. In 1776, he returned to Paris and retired to a small estate to do his scientific research. He invented a torsion balance to measure the quantity of a force and used it for determination of forces of electric attraction or repulsion between small charged spheres. He thus arrived in 1785 at the inverse square law relation, now known as Coulomb's law. The law had been anticipated by Priestley and also by Cavendish earlier, though Cavendish never published his results. Coulomb also found the inverse square law of force between unlike and like magnetic poles.

CHARLES AUGUSTIN DE COULOMB (1736 – 1806)

Physics

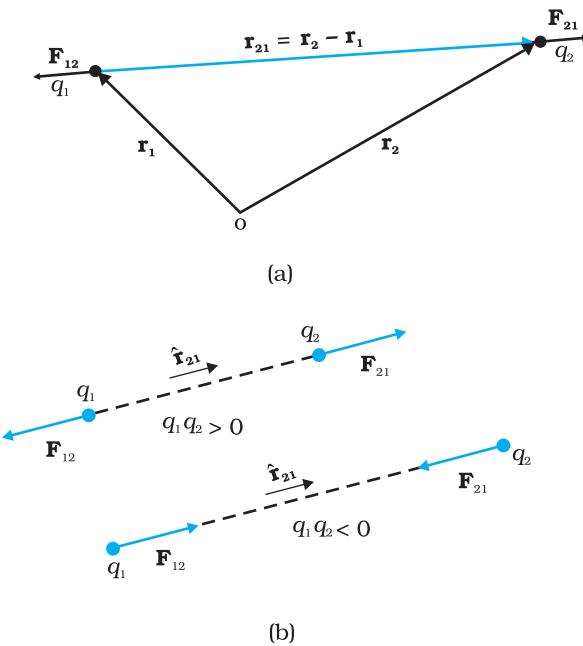


FIGURE 1.6 (a) Geometry and
(b) Forces between charges.

Coulomb's force law between two point charges \$q_1\$ and \$q_2\$ located at \$\mathbf{r}_1\$ and \$\mathbf{r}_2\$ is then expressed as

$$\mathbf{F}_{21} = \frac{1}{4\pi\epsilon_0} \frac{q_1 q_2}{r_{21}^2} \hat{\mathbf{r}}_{21} \quad (1.3)$$

Some remarks on Eq. (1.3) are relevant:

- Equation (1.3) is valid for any sign of \$q_1\$ and \$q_2\$ whether positive or negative. If \$q_1\$ and \$q_2\$ are of the same sign (either both positive or both negative), \$\mathbf{F}_{21}\$ is along \$\hat{\mathbf{r}}_{21}\$, which denotes repulsion, as it should be for like charges. If \$q_1\$ and \$q_2\$ are of opposite signs, \$\mathbf{F}_{21}\$ is along \$-\hat{\mathbf{r}}_{21}\$ (\$=\hat{\mathbf{r}}_{12}\$), which denotes attraction, as expected for unlike charges. Thus, we do not have to write separate equations for the cases of like and unlike charges. Equation (1.3) takes care of both cases correctly [Fig. 1.6(b)].
- The force \$\mathbf{F}_{12}\$ on charge \$q_1\$ due to charge \$q_2\$, is obtained from Eq. (1.3), by simply interchanging 1 and 2, i.e.,

$$\mathbf{F}_{12} = \frac{1}{4\pi\epsilon_0} \frac{q_1 q_2}{r_{12}^2} \hat{\mathbf{r}}_{12} = -\mathbf{F}_{21}$$

Thus, Coulomb's law agrees with the Newton's third law.

- Coulomb's law [Eq. (1.3)] gives the force between two charges \$q_1\$ and \$q_2\$ in vacuum. If the charges are placed in matter or the intervening space has matter, the situation gets complicated due to the presence of charged constituents of matter. We shall consider electrostatics in matter in the next chapter.

Since force is a vector, it is better to write Coulomb's law in the vector notation. Let the position vectors of charges \$q_1\$ and \$q_2\$ be \$\mathbf{r}_1\$ and \$\mathbf{r}_2\$ respectively [see Fig. 1.6(a)]. We denote force on \$q_1\$ due to \$q_2\$ by \$\mathbf{F}_{12}\$ and force on \$q_2\$ due to \$q_1\$ by \$\mathbf{F}_{21}\$. The two point charges \$q_1\$ and \$q_2\$ have been numbered 1 and 2 for convenience and the vector leading from 1 to 2 is denoted by \$\mathbf{r}_{21}\$:

$$\mathbf{r}_{21} = \mathbf{r}_2 - \mathbf{r}_1$$

In the same way, the vector leading from 2 to 1 is denoted by \$\mathbf{r}_{12}\$:

$$\mathbf{r}_{12} = \mathbf{r}_1 - \mathbf{r}_2 = -\mathbf{r}_{21}$$

The magnitude of the vectors \$\mathbf{r}_{21}\$ and \$\mathbf{r}_{12}\$ is denoted by \$r_{21}\$ and \$r_{12}\$, respectively (\$r_{12} = r_{21}\$). The direction of a vector is specified by a unit vector along the vector. To denote the direction from 1 to 2 (or from 2 to 1), we define the unit vectors:

$$\hat{\mathbf{r}}_{21} = \frac{\mathbf{r}_{21}}{r_{21}}, \quad \hat{\mathbf{r}}_{12} = \frac{\mathbf{r}_{12}}{r_{12}}, \quad \hat{\mathbf{r}}_{21} = \hat{\mathbf{r}}_{12}$$

Electric Charges and Fields

Example 1.4 Coulomb's law for electrostatic force between two point charges and Newton's law for gravitational force between two stationary point masses, both have inverse-square dependence on the distance between the charges/masses. (a) Compare the strength of these forces by determining the ratio of their magnitudes (i) for an electron and a proton and (ii) for two protons. (b) Estimate the accelerations of electron and proton due to the electrical force of their mutual attraction when they are 1 Å ($= 10^{-10}$ m) apart? ($m_p = 1.67 \times 10^{-27}$ kg, $m_e = 9.11 \times 10^{-31}$ kg)

Solution

(a) (i) The electric force between an electron and a proton at a distance r apart is:

$$F_e = -\frac{1}{4\pi\epsilon_0} \frac{e^2}{r^2}$$

where the negative sign indicates that the force is attractive. The corresponding gravitational force (always attractive) is:

$$F_G = -G \frac{m_p m_e}{r^2}$$

where m_p and m_e are the masses of a proton and an electron respectively.

$$\left| \frac{F_e}{F_G} \right| = \frac{e^2}{4\pi\epsilon_0 G m_p m_e} = 2.4 \times 10^{39}$$

(ii) On similar lines, the ratio of the magnitudes of electric force to the gravitational force between two protons at a distance r apart is :

$$\left| \frac{F_e}{F_G} \right| = \frac{e^2}{4\pi\epsilon_0 G m_p m_p} = 1.3 \times 10^{36}$$

However, it may be mentioned here that the signs of the two forces are different. For two protons, the gravitational force is attractive in nature and the Coulomb force is repulsive. The actual values of these forces between two protons inside a nucleus (distance between two protons is $\sim 10^{-15}$ m inside a nucleus) are $F_e \sim 230$ N whereas $F_G \sim 1.9 \times 10^{-34}$ N.

The (dimensionless) ratio of the two forces shows that electrical forces are enormously stronger than the gravitational forces.

(b) The electric force \mathbf{F} exerted by a proton on an electron is same in magnitude to the force exerted by an electron on a proton; however the masses of an electron and a proton are different. Thus, the magnitude of force is

$$|\mathbf{F}| = \frac{1}{4\pi\epsilon_0} \frac{e^2}{r^2} = 8.987 \times 10^9 \text{ Nm}^2/\text{C}^2 \times (1.6 \times 10^{-19}\text{C})^2 / (10^{-10}\text{m})^2 \\ = 2.3 \times 10^{-8} \text{ N}$$

Using Newton's second law of motion, $F = ma$, the acceleration that an electron will undergo is

$$a = 2.3 \times 10^{-8} \text{ N} / 9.11 \times 10^{-31} \text{ kg} = 2.5 \times 10^{22} \text{ m/s}^2$$

Comparing this with the value of acceleration due to gravity, we can conclude that the effect of gravitational field is negligible on the motion of electron and it undergoes very large accelerations under the action of Coulomb force due to a proton.

The value for acceleration of the proton is

$$2.3 \times 10^{-8} \text{ N} / 1.67 \times 10^{-27} \text{ kg} = 1.4 \times 10^{19} \text{ m/s}^2$$



Interactive animation on Coulomb's law:
http://webphysics.davidson.edu/physlet_resources/bu_semester2/menu_semester2.html

Example 1.5 A charged metallic sphere A is suspended by a nylon thread. Another charged metallic sphere B held by an insulating handle is brought close to A such that the distance between their centres is 10 cm, as shown in Fig. 1.7(a). The resulting repulsion of A is noted (for example, by shining a beam of light and measuring the deflection of its shadow on a screen). Spheres A and B are touched by uncharged spheres C and D respectively, as shown in Fig. 1.7(b). C and D are then removed and B is brought closer to A to a distance of 5.0 cm between their centres, as shown in Fig. 1.7(c). What is the expected repulsion of A on the basis of Coulomb's law? Spheres A and C and spheres B and D have identical sizes. Ignore the sizes of A and B in comparison to the separation between their centres.

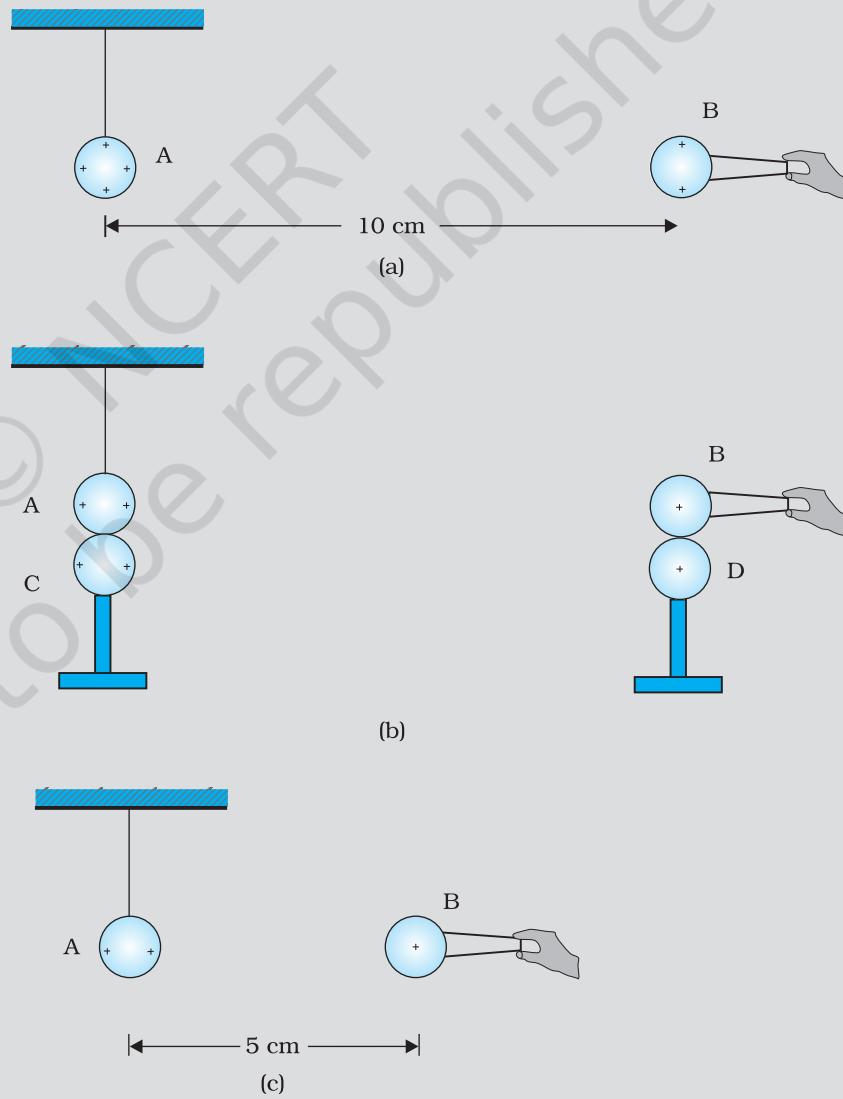


FIGURE 1.7

Solution Let the original charge on sphere A be q and that on B be q' . At a distance r between their centres, the magnitude of the electrostatic force on each is given by

$$F = \frac{1}{4\pi\epsilon_0} \frac{qq'}{r^2}$$

neglecting the sizes of spheres A and B in comparison to r . When an identical but uncharged sphere C touches A, the charges redistribute on A and C and, by symmetry, each sphere carries a charge $q/2$. Similarly, after D touches B, the redistributed charge on each is $q'/2$. Now, if the separation between A and B is halved, the magnitude of the electrostatic force on each is

$$F' = \frac{1}{4\pi\epsilon_0} \frac{(q/2)(q'/2)}{(r/2)^2} = \frac{1}{4\pi\epsilon_0} \frac{(qq')}{r^2} = F$$

Thus the electrostatic force on A, due to B, remains unaltered.

EXAMPLE 1.5

1.7 FORCES BETWEEN MULTIPLE CHARGES

The mutual electric force between two charges is given by Coulomb's law. How to calculate the force on a charge where there are not one but several charges around? Consider a system of n stationary charges $q_1, q_2, q_3, \dots, q_n$ in vacuum. What is the force on q_1 due to q_2, q_3, \dots, q_n ? Coulomb's law is not enough to answer this question. Recall that forces of mechanical origin add according to the parallelogram law of addition. Is the same true for forces of electrostatic origin?

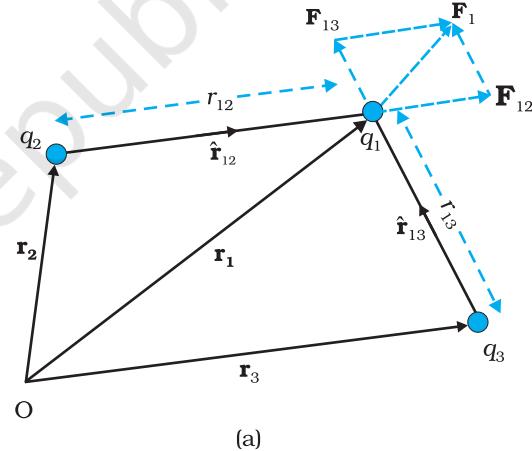
Experimentally it is verified that *force on any charge due to a number of other charges is the vector sum of all the forces on that charge due to the other charges, taken one at a time. The individual forces are unaffected due to the presence of other charges.* This is termed as the *principle of superposition*.

To better understand the concept, consider a system of three charges q_1, q_2 and q_3 , as shown in Fig. 1.8(a). The force on one charge, say q_1 , due to two other charges q_2, q_3 can therefore be obtained by performing a vector addition of the forces due to each one of these charges. Thus, if the force on q_1 due to q_2 is denoted by \mathbf{F}_{12} , \mathbf{F}_{12} is given by Eq. (1.3) even though other charges are present.

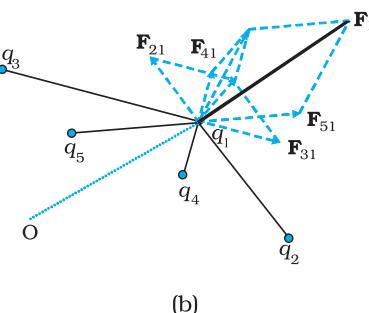
$$\text{Thus, } \mathbf{F}_{12} = \frac{1}{4\pi\epsilon_0} \frac{q_1 q_2}{r_{12}^2} \hat{\mathbf{r}}_{12}$$

In the same way, the force on q_1 due to q_3 , denoted by \mathbf{F}_{13} , is given by

$$\mathbf{F}_{13} = \frac{1}{4\pi\epsilon_0} \frac{q_1 q_3}{r_{13}^2} \hat{\mathbf{r}}_{13}$$



(a)



(b)

FIGURE 1.8 A system of (a) three charges (b) multiple charges.

Physics

which again is the Coulomb force on q_1 due to q_3 , even though other charge q_2 is present.

Thus the total force \mathbf{F}_1 on q_1 due to the two charges q_2 and q_3 is given as

$$\mathbf{F}_1 = \mathbf{F}_{12} + \mathbf{F}_{13} = \frac{1}{4\pi\epsilon_0} \frac{q_1 q_2}{r_{12}^2} \hat{\mathbf{r}}_{12} + \frac{1}{4\pi\epsilon_0} \frac{q_1 q_3}{r_{13}^2} \hat{\mathbf{r}}_{13} \quad (1.4)$$

The above calculation of force can be generalised to a system of charges more than three, as shown in Fig. 1.8(b).

The principle of superposition says that in a system of charges q_1, q_2, \dots, q_n , the force on q_1 due to q_2 is the same as given by Coulomb's law, i.e., it is unaffected by the presence of the other charges q_3, q_4, \dots, q_n . The total force \mathbf{F}_1 on the charge q_1 , due to all other charges, is then given by the vector sum of the forces $\mathbf{F}_{12}, \mathbf{F}_{13}, \dots, \mathbf{F}_{1n}$:

i.e.,

$$\begin{aligned} \mathbf{F}_1 &= \mathbf{F}_{12} + \mathbf{F}_{13} + \dots + \mathbf{F}_{1n} = \frac{1}{4\pi\epsilon_0} \left[\frac{q_1 q_2}{r_{12}^2} \hat{\mathbf{r}}_{12} + \frac{q_1 q_3}{r_{13}^2} \hat{\mathbf{r}}_{13} + \dots + \frac{q_1 q_n}{r_{1n}^2} \hat{\mathbf{r}}_{1n} \right] \\ &= \frac{q_1}{4\pi\epsilon_0} \sum_{i=2}^n \frac{q_i}{r_{1i}^2} \hat{\mathbf{r}}_{1i} \end{aligned} \quad (1.5)$$

The vector sum is obtained as usual by the parallelogram law of addition of vectors. All of electrostatics is basically a consequence of Coulomb's law and the superposition principle.

Example 1.6 Consider three charges q_1, q_2, q_3 each equal to q at the vertices of an equilateral triangle of side l . What is the force on a charge Q (with the same sign as q) placed at the centroid of the triangle, as shown in Fig. 1.9?

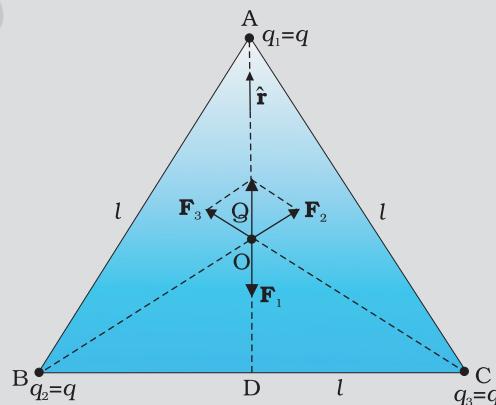


FIGURE 1.9

Solution In the given equilateral triangle ABC of sides of length l , if we draw a perpendicular AD to the side BC,

$AD = AC \cos 30^\circ = (\sqrt{3}/2) l$ and the distance AO of the centroid O from A is $(2/3) AD = (1/\sqrt{3}) l$. By symmetry $AO = BO = CO$.

Thus,

$$\text{Force } \mathbf{F}_1 \text{ on } Q \text{ due to charge } q \text{ at A} = \frac{3}{4\pi\epsilon_0} \frac{Qq}{l^2} \text{ along AO}$$

$$\text{Force } \mathbf{F}_2 \text{ on } Q \text{ due to charge } q \text{ at B} = \frac{3}{4\pi\epsilon_0} \frac{Qq}{l^2} \text{ along BO}$$

$$\text{Force } \mathbf{F}_3 \text{ on } Q \text{ due to charge } q \text{ at C} = \frac{3}{4\pi\epsilon_0} \frac{Qq}{l^2} \text{ along CO}$$

The resultant of forces \mathbf{F}_2 and \mathbf{F}_3 is $\frac{3}{4\pi\epsilon_0} \frac{Qq}{l^2}$ along OA, by the parallelogram law. Therefore, the total force on $Q = \frac{3}{4\pi\epsilon_0} \frac{Qq}{l^2} (\hat{\mathbf{r}} - \hat{\mathbf{r}})$

$= 0$, where $\hat{\mathbf{r}}$ is the unit vector along OA.

It is clear also by symmetry that the three forces will sum to zero. Suppose that the resultant force was non-zero but in some direction. Consider what would happen if the system was rotated through 60° about O.

Example 1.7 Consider the charges q , q , and $-q$ placed at the vertices of an equilateral triangle, as shown in Fig. 1.10. What is the force on each charge?

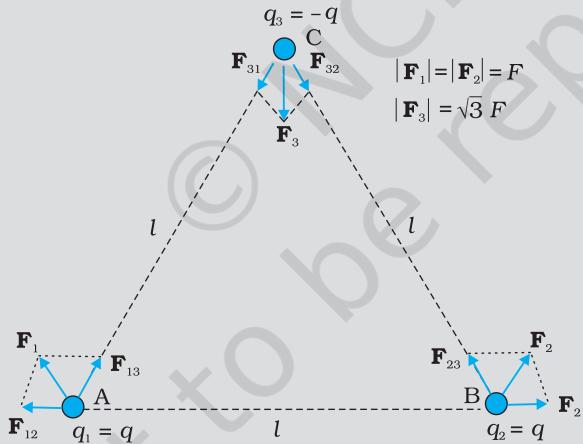


FIGURE 1.10

Solution The forces acting on charge q at A due to charges q at B and $-q$ at C are \mathbf{F}_{12} along BA and \mathbf{F}_{13} along AC respectively, as shown in Fig. 1.10. By the parallelogram law, the total force \mathbf{F}_1 on the charge q at A is given by

$$\mathbf{F}_1 = F \hat{\mathbf{r}}_1 \text{ where } \hat{\mathbf{r}}_1 \text{ is a unit vector along BC.}$$

The force of attraction or repulsion for each pair of charges has the

$$\text{same magnitude } F = \frac{q^2}{4\pi\epsilon_0 l^2}$$

The total force \mathbf{F}_2 on charge q at B is thus $\mathbf{F}_2 = F \hat{\mathbf{r}}_2$, where $\hat{\mathbf{r}}_2$ is a unit vector along AC.

EXAMPLE 1.6

EXAMPLE 1.7

EXAMPLE 1.7

Similarly the total force on charge $-q$ at C is $\mathbf{F}_3 = \sqrt{3} F \hat{\mathbf{n}}$, where $\hat{\mathbf{n}}$ is the unit vector along the direction bisecting the $\angle BCA$. It is interesting to see that the sum of the forces on the three charges is zero, i.e.,

$$\mathbf{F}_1 + \mathbf{F}_2 + \mathbf{F}_3 = 0$$

The result is not at all surprising. It follows straight from the fact that Coulomb's law is consistent with Newton's third law. The proof is left to you as an exercise.

1.8 ELECTRIC FIELD

Let us consider a point charge Q placed in vacuum, at the origin O. If we place another point charge q at a point P, where $\mathbf{OP} = \mathbf{r}$, then the charge Q will exert a force on q as per Coulomb's law. We may ask the question: If charge q is removed, then what is left in the surrounding? Is there nothing? If there is nothing at the point P, then how does a force act when we place the charge q at P. In order to answer such questions, the early scientists introduced the concept of *field*. According to this, we say that the charge Q produces an electric field everywhere in the surrounding. When another charge q is brought at some point P, the field there acts on it and produces a force. The electric field produced by the charge Q at a point \mathbf{r} is given as

$$\mathbf{E}(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} \frac{Q}{r^2} \hat{\mathbf{r}} = \frac{1}{4\pi\epsilon_0} \frac{Q}{r^2} \hat{\mathbf{r}} \quad (1.6)$$

where $\hat{\mathbf{r}} = \mathbf{r}/r$, is a unit vector from the origin to the point \mathbf{r} . Thus, Eq.(1.6) specifies the value of the electric field for each value of the position vector \mathbf{r} . The word "field" signifies how some distributed quantity (which could be a scalar or a vector) varies with position. The effect of the charge has been incorporated in the existence of the electric field. We obtain the force \mathbf{F} exerted by a charge Q on a charge q , as

$$\mathbf{F} = \frac{1}{4\pi\epsilon_0} \frac{Qq}{r^2} \hat{\mathbf{r}} \quad (1.7)$$

Note that the charge q also exerts an equal and opposite force on the charge Q . The electrostatic force between the charges Q and q can be looked upon as an interaction between charge q and the electric field of Q and *vice versa*. If we denote the position of charge q by the vector \mathbf{r} , it experiences a force \mathbf{F} equal to the charge q multiplied by the electric field \mathbf{E} at the location of q . Thus,

$$\mathbf{F}(\mathbf{r}) = q \mathbf{E}(\mathbf{r}) \quad (1.8)$$

Equation (1.8) defines the SI unit of electric field as N/C*.

Some important remarks may be made here:

- (i) From Eq. (1.8), we can infer that if q is unity, the electric field due to a charge Q is numerically equal to the force exerted by it. Thus, the *electric field due to a charge Q at a point in space may be defined as the force that a unit positive charge would experience if placed*

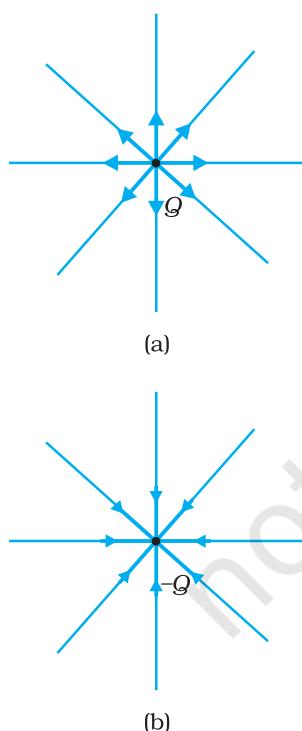


FIGURE 1.11 Electric field (a) due to a charge Q , (b) due to a charge $-Q$.

* An alternate unit V/m will be introduced in the next chapter.

at that point. The charge Q , which is producing the electric field, is called a *source charge* and the charge q , which tests the effect of a source charge, is called a *test charge*. Note that the source charge Q must remain at its original location. However, if a charge q is brought at any point around Q , Q itself is bound to experience an electrical force due to q and will tend to move. A way out of this difficulty is to make q negligibly small. The force \mathbf{F} is then negligibly small but the ratio \mathbf{F}/q is finite and defines the electric field:

$$\mathbf{E} = \lim_{q \rightarrow 0} \left(\frac{\mathbf{F}}{q} \right) \quad (1.9)$$

A practical way to get around the problem (of keeping Q undisturbed in the presence of q) is to hold Q to its location by unspecified forces! This may look strange but actually this is what happens in practice. When we are considering the electric force on a test charge q due to a charged planar sheet (Section 1.15), the charges on the sheet are held to their locations by the forces due to the unspecified charged constituents inside the sheet.

- (ii) Note that the electric field \mathbf{E} due to Q , though defined operationally in terms of some test charge q , is independent of q . This is because \mathbf{F} is proportional to q , so the ratio \mathbf{F}/q does not depend on q . The force \mathbf{F} on the charge q due to the charge Q depends on the particular location of charge q which may take any value in the space around the charge Q . Thus, the electric field \mathbf{E} due to Q is also dependent on the space coordinate \mathbf{r} . For different positions of the charge q all over the space, we get different values of electric field \mathbf{E} . The field exists at every point in three-dimensional space.
- (iii) For a positive charge, the electric field will be directed radially outwards from the charge. On the other hand, if the source charge is negative, the electric field vector, at each point, points radially inwards.
- (iv) Since the magnitude of the force \mathbf{F} on charge q due to charge Q depends only on the distance r of the charge q from charge Q , the magnitude of the electric field \mathbf{E} will also depend only on the distance r . Thus at equal distances from the charge Q , the magnitude of its electric field \mathbf{E} is same. The magnitude of electric field \mathbf{E} due to a point charge is thus same on a sphere with the point charge at its centre; in other words, it has a spherical symmetry.

1.8.1 Electric field due to a system of charges

Consider a system of charges q_1, q_2, \dots, q_n with position vectors $\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n$ relative to some origin O. Like the electric field at a point in space due to a single charge, electric field at a point in space due to the system of charges is defined to be the force experienced by a unit test charge placed at that point, without disturbing the original positions of charges q_1, q_2, \dots, q_n . We can use Coulomb's law and the superposition principle to determine this field at a point P denoted by position vector \mathbf{r} .

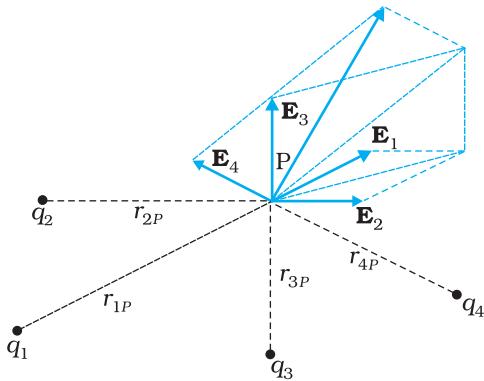


FIGURE 1.12 Electric field at a point due to a system of charges is the vector sum of the electric fields at the point due to individual charges.

Electric field \mathbf{E}_1 at \mathbf{r} due to q_1 at \mathbf{r}_1 is given by

$$\mathbf{E}_1 = \frac{1}{4\pi\epsilon_0} \frac{q_1}{r_{1P}^2} \hat{\mathbf{r}}_{1P}$$

where $\hat{\mathbf{r}}_{1P}$ is a unit vector in the direction from q_1 to P , and r_{1P} is the distance between q_1 and P .

In the same manner, electric field \mathbf{E}_2 at \mathbf{r} due to q_2 at \mathbf{r}_2 is

$$\mathbf{E}_2 = \frac{1}{4\pi\epsilon_0} \frac{q_2}{r_{2P}^2} \hat{\mathbf{r}}_{2P}$$

where $\hat{\mathbf{r}}_{2P}$ is a unit vector in the direction from q_2 to P and r_{2P} is the distance between q_2 and P . Similar expressions hold good for fields \mathbf{E}_3 , \mathbf{E}_4 , ..., \mathbf{E}_n due to charges q_3 , q_4 , ..., q_n .

By the superposition principle, the electric field \mathbf{E} at \mathbf{r} due to the system of charges is (as shown in Fig. 1.12)

$$\begin{aligned} \mathbf{E}(\mathbf{r}) &= \mathbf{E}_1(\mathbf{r}) + \mathbf{E}_2(\mathbf{r}) + \dots + \mathbf{E}_n(\mathbf{r}) \\ &= \frac{1}{4\pi\epsilon_0} \frac{q_1}{r_{1P}^2} \hat{\mathbf{r}}_{1P} + \frac{1}{4\pi\epsilon_0} \frac{q_2}{r_{2P}^2} \hat{\mathbf{r}}_{2P} + \dots + \frac{1}{4\pi\epsilon_0} \frac{q_n}{r_{nP}^2} \hat{\mathbf{r}}_{nP} \\ \mathbf{E}(\mathbf{r}) &= \frac{1}{4\pi\epsilon_0} \sum_{i=1}^n \frac{q_i}{r_{iP}^2} \hat{\mathbf{r}}_{iP} \end{aligned} \quad (1.10)$$

\mathbf{E} is a vector quantity that varies from one point to another point in space and is determined from the positions of the source charges.

1.8.2 Physical significance of electric field

You may wonder why the notion of electric field has been introduced here at all. After all, for any system of charges, the measurable quantity is the force on a charge which can be directly determined using Coulomb's law and the superposition principle [Eq. (1.5)]. Why then introduce this intermediate quantity called the electric field?

For electrostatics, the concept of electric field is convenient, but not really necessary. Electric field is an elegant way of characterising the electrical environment of a system of charges. Electric field at a point in the space around a system of charges tells you the force a unit positive test charge would experience if placed at that point (without disturbing the system). Electric field is a characteristic of the system of charges and is independent of the test charge that you place at a point to determine the field. The term *field* in physics generally refers to a quantity that is defined at every point in space and may vary from point to point. Electric field is a vector field, since force is a vector quantity.

The true physical significance of the concept of electric field, however, emerges only when we go beyond electrostatics and deal with time-dependent electromagnetic phenomena. Suppose we consider the force between two distant charges q_1 , q_2 in accelerated motion. Now the greatest speed with which a signal or information can go from one point to another is c , the speed of light. Thus, the effect of any motion of q_1 on q_2 cannot

Electric Charges and Fields

arise instantaneously. There will be some time delay between the effect (force on q_2) and the cause (motion of q_1). It is precisely here that the notion of electric field (strictly, electromagnetic field) is natural and very useful. *The field picture is this: the accelerated motion of charge q_1 produces electromagnetic waves, which then propagate with the speed c , reach q_2 and cause a force on q_2 .* The notion of field elegantly accounts for the time delay. Thus, even though electric and magnetic fields can be detected only by their effects (forces) on charges, they are regarded as physical entities, not merely mathematical constructs. They have an *independent dynamics* of their own, i.e., they evolve according to laws of their own. They can also transport energy. Thus, a source of time-dependent electromagnetic fields, turned on briefly and switched off, leaves behind propagating electromagnetic fields transporting energy. The concept of field was first introduced by Faraday and is now among the central concepts in physics.

Example 1.8 An electron falls through a distance of 1.5 cm in a uniform electric field of magnitude $2.0 \times 10^4 \text{ N C}^{-1}$ [Fig. 1.13(a)]. The direction of the field is reversed keeping its magnitude unchanged and a proton falls through the same distance [Fig. 1.13(b)]. Compute the time of fall in each case. Contrast the situation with that of ‘free fall under gravity’.

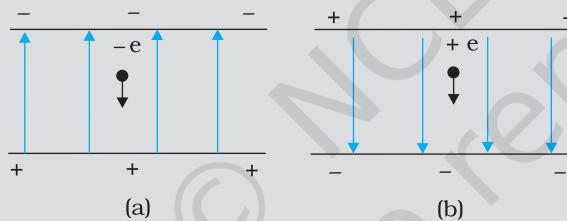


FIGURE 1.13

Solution In Fig. 1.13(a) the field is upward, so the negatively charged electron experiences a downward force of magnitude eE where E is the magnitude of the electric field. The acceleration of the electron is

$$a_e = eE/m_e$$

where m_e is the mass of the electron.

Starting from rest, the time required by the electron to fall through a

$$\text{distance } h \text{ is given by } t_e = \sqrt{\frac{2h}{a_e}} = \sqrt{\frac{2hm_e}{eE}}$$

For $e = 1.6 \times 10^{-19} \text{ C}$, $m_e = 9.11 \times 10^{-31} \text{ kg}$,

$$E = 2.0 \times 10^4 \text{ N C}^{-1}$$

$$h = 1.5 \times 10^{-2} \text{ m}$$

$$t_e = 2.9 \times 10^{-9} \text{ s}$$

In Fig. 1.13 (b), the field is downward, and the positively charged proton experiences a downward force of magnitude eE . The acceleration of the proton is

$$a_p = eE/m_p$$

where m_p is the mass of the proton; $m_p = 1.67 \times 10^{-27} \text{ kg}$. The time of fall for the proton is

EXAMPLE 1.8

EXAMPLE 1.8

$$t_p = \sqrt{\frac{2h}{a_p}} = \sqrt{\frac{2hm_p}{eE}} = 1.3 \times 10^{-7} \text{ s}$$

Thus, the heavier particle (proton) takes a greater time to fall through the same distance. This is in basic contrast to the situation of 'free fall under gravity' where the time of fall is independent of the mass of the body. Note that in this example we have ignored the acceleration due to gravity in calculating the time of fall. To see if this is justified, let us calculate the acceleration of the proton in the given electric field:

$$\begin{aligned} a_p &= \frac{eE}{m_p} \\ &= \frac{(1.6 \times 10^{-19} \text{ C}) \times (2.0 \times 10^4 \text{ N C}^{-1})}{1.67 \times 10^{-27} \text{ kg}} \\ &= 1.9 \times 10^{12} \text{ m s}^{-2} \end{aligned}$$

which is enormous compared to the value of g (9.8 m s^{-2}), the acceleration due to gravity. The acceleration of the electron is even greater. Thus, the effect of acceleration due to gravity can be ignored in this example.

Example 1.9 Two point charges q_1 and q_2 , of magnitude $+10^{-8} \text{ C}$ and -10^{-8} C , respectively, are placed 0.1 m apart. Calculate the electric fields at points A, B and C shown in Fig. 1.14.

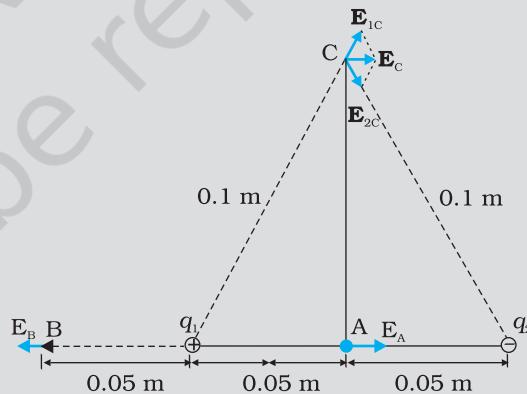


FIGURE 1.14

Solution The electric field vector \mathbf{E}_{1A} at A due to the positive charge q_1 points towards the right and has a magnitude

$$E_{1A} = \frac{(9 \times 10^9 \text{ N m}^2 \text{ C}^{-2}) \times (10^{-8} \text{ C})}{(0.05 \text{ m})^2} = 3.6 \times 10^4 \text{ N C}^{-1}$$

The electric field vector \mathbf{E}_{2A} at A due to the negative charge q_2 points towards the right and has the same magnitude. Hence the magnitude of the total electric field E_A at A is

$$E_A = E_{1A} + E_{2A} = 7.2 \times 10^4 \text{ N C}^{-1}$$

\mathbf{E}_A is directed toward the right.

EXAMPLE 1.9

The electric field vector \mathbf{E}_{1B} at B due to the positive charge q_1 points towards the left and has a magnitude

$$E_{1B} = \frac{(9 \times 10^9 \text{ Nm}^2 \text{ C}^{-2}) \times (10^{-8} \text{ C})}{(0.05 \text{ m})^2} = 3.6 \times 10^4 \text{ N C}^{-1}$$

The electric field vector \mathbf{E}_{2B} at B due to the negative charge q_2 points towards the right and has a magnitude

$$E_{2B} = \frac{(9 \times 10^9 \text{ Nm}^2 \text{ C}^{-2}) \times (10^{-8} \text{ C})}{(0.15 \text{ m})^2} = 4 \times 10^3 \text{ N C}^{-1}$$

The magnitude of the total electric field at B is

$$E_B = E_{1B} - E_{2B} = 3.2 \times 10^4 \text{ N C}^{-1}$$

\mathbf{E}_B is directed towards the left.

The magnitude of each electric field vector at point C, due to charge q_1 and q_2 is

$$E_{1C} = E_{2C} = \frac{(9 \times 10^9 \text{ Nm}^2 \text{ C}^{-2}) \times (10^{-8} \text{ C})}{(0.10 \text{ m})^2} = 9 \times 10^3 \text{ N C}^{-1}$$

The directions in which these two vectors point are indicated in Fig. 1.14. The resultant of these two vectors is

$$E_C = E_1 \cos \frac{\pi}{3} + E_2 \cos \frac{\pi}{3} = 9 \times 10^3 \text{ N C}^{-1}$$

\mathbf{E}_C points towards the right.

EXAMPLE 1.9

1.9 ELECTRIC FIELD LINES

We have studied electric field in the last section. It is a vector quantity and can be represented as we represent vectors. Let us try to represent \mathbf{E} due to a point charge pictorially. Let the point charge be placed at the origin. Draw vectors pointing along the direction of the electric field with their lengths proportional to the strength of the field at each point. Since the magnitude of electric field at a point decreases inversely as the square of the distance of that point from the charge, the vector gets shorter as one goes away from the origin, always pointing radially outward. Figure 1.15 shows such a picture. In this figure, each arrow indicates the electric field, i.e., the force acting on a unit positive charge, placed at the tail of that arrow. Connect the arrows pointing in one direction and the resulting figure represents a field line. We thus get many field lines, all pointing outwards from the point charge. Have we lost the information about the strength or magnitude of the field now, because it was contained in the length of the arrow? No. Now the magnitude of the field is represented by the density of field lines. \mathbf{E} is strong near the charge, so the density of field lines is more near the charge and the lines are closer. Away from the charge, the field gets weaker and the density of field lines is less, resulting in well-separated lines.

Another person may draw more lines. But the number of lines is not important. In fact, an infinite number of lines can be drawn in any region.

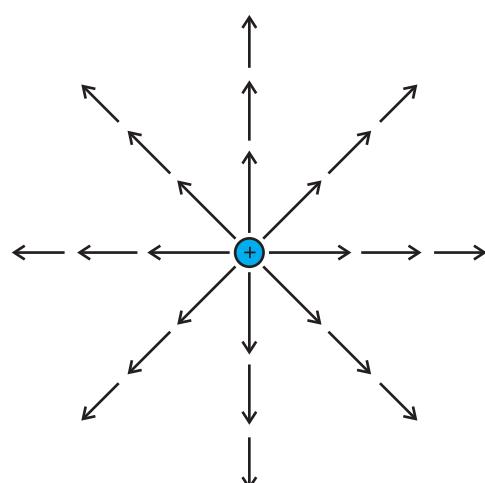


FIGURE 1.15 Field of a point charge.

Physics

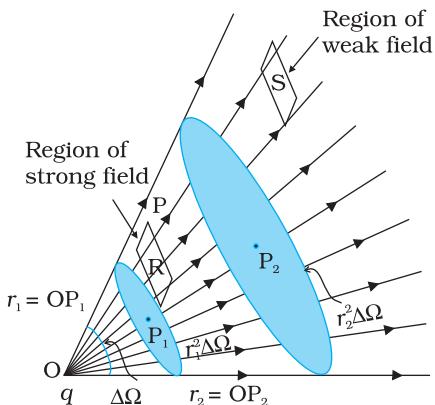


FIGURE 1.16 Dependence of electric field strength on the distance and its relation to the number of field lines.

It is the relative density of lines in different regions which is important.

We draw the figure on the plane of paper, i.e., in two-dimensions but we live in three-dimensions. So if one wishes to estimate the density of field lines, one has to consider the number of lines per unit cross-sectional area, perpendicular to the lines. Since the electric field decreases as the square of the distance from a point charge and the area enclosing the charge increases as the square of the distance, the number of field lines crossing the enclosing area remains constant, whatever may be the distance of the area from the charge.

We started by saying that the field lines carry information about the direction of electric field at different points in space. Having drawn a certain set of field lines, the relative density (i.e., closeness) of the field lines at different points indicates the relative strength of electric field at those points. The field lines crowd where the field is strong and are spaced apart where it is weak. Figure 1.16 shows a set of field lines. We

can imagine two equal and small elements of area placed at points R and S normal to the field lines there. The number of field lines in our picture cutting the area elements is proportional to the magnitude of field at these points. The picture shows that the field at R is stronger than at S.

To understand the dependence of the field lines on the area, or rather the *solid angle* subtended by an area element, let us try to relate the area with the solid angle, a generalization of angle to three dimensions. Recall how a (plane) angle is defined in two-dimensions. Let a small transverse line element Δl be placed at a distance r from a point O. Then the angle subtended by Δl at O can be approximated as $\Delta\theta = \Delta l/r$. Likewise, in three-dimensions the solid angle* subtended by a small perpendicular plane area ΔS , at a distance r , can be written as $\Delta\Omega = \Delta S/r^2$. We know that in a given solid angle the number of radial field lines is the same. In Fig. 1.16, for two points P_1 and P_2 at distances r_1 and r_2 from the charge, the element of area subtending the solid angle $\Delta\Omega$ is $r_1^2 \Delta\Omega$ at P_1 and an element of area $r_2^2 \Delta\Omega$ at P_2 , respectively. The number of lines (say n) cutting these area elements are the same. The number of field lines, cutting unit area element is therefore $n/(r_1^2 \Delta\Omega)$ at P_1 and $n/(r_2^2 \Delta\Omega)$ at P_2 , respectively. Since n and $\Delta\Omega$ are common, the strength of the field clearly has a $1/r^2$ dependence.

The picture of field lines was invented by Faraday to develop an intuitive non-mathematical way of visualizing electric fields around charged configurations. Faraday called them *lines of force*. This term is somewhat misleading, especially in case of magnetic fields. The more appropriate term is *field lines* (electric or magnetic) that we have adopted in this book.

Electric field lines are thus a way of pictorially mapping the electric field around a configuration of charges. An electric field line is, in general,

* Solid angle is a measure of a cone. Consider the intersection of the given cone with a sphere of radius R . The solid angle $\Delta\Omega$ of the cone is defined to be equal to $\Delta S/R^2$, where ΔS is the area on the sphere cut out by the cone.

a curve drawn in such a way that the tangent to it at each point is in the direction of the net field at that point. An arrow on the curve is obviously necessary to specify the direction of electric field from the two possible directions indicated by a tangent to the curve. A field line is a space curve, i.e., a curve in three dimensions.

Figure 1.17 shows the field lines around some simple charge configurations. As mentioned earlier, the field lines are in 3-dimensional space, though the figure shows them only in a plane. The field lines of a single positive charge are radially outward while those of a single negative charge are radially inward. The field lines around a system of two positive charges (q, q) give a vivid pictorial description of their mutual repulsion, while those around the configuration of two equal and opposite charges ($q, -q$), a dipole, show clearly the mutual attraction between the charges. The field lines follow some important general properties:

- Field lines start from positive charges and end at negative charges. If there is a single charge, they may start or end at infinity.
- In a charge-free region, electric field lines can be taken to be continuous curves without any breaks.
- Two field lines can never cross each other. (If they did, the field at the point of intersection will not have a unique direction, which is absurd.)
- Electrostatic field lines do not form any closed loops. This follows from the conservative nature of electric field (Chapter 2).

1.10 ELECTRIC FLUX

Consider flow of a liquid with velocity \mathbf{v} , through a small flat surface dS , in a direction normal to the surface. The rate of flow of liquid is given by the volume crossing the area per unit time $v dS$ and represents the flux of liquid flowing across the plane. If the normal to the surface is not parallel to the direction of flow of liquid, i.e., to \mathbf{v} , but makes an angle θ with it, the projected area in a plane perpendicular to \mathbf{v} is $v dS \cos \theta$. Therefore the flux going out of the surface dS is $\mathbf{v} \cdot \hat{\mathbf{n}} dS$.

For the case of the electric field, we define an analogous quantity and call it *electric flux*.

We should however note that there is no flow of a physically observable quantity unlike the case of liquid flow.

In the picture of electric field lines described above, we saw that the number of field lines crossing a unit area, placed normal to the field at a point is a measure of the strength of electric field at that point. This means that if

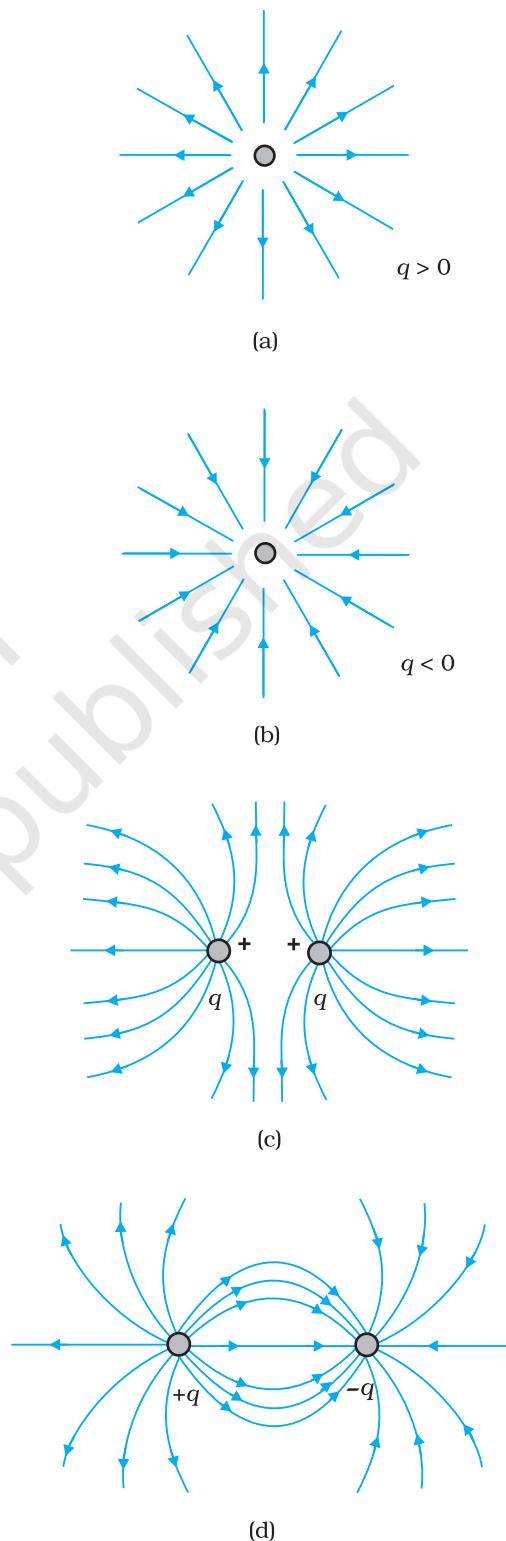


FIGURE 1.17 Field lines due to some simple charge configurations.

Physics

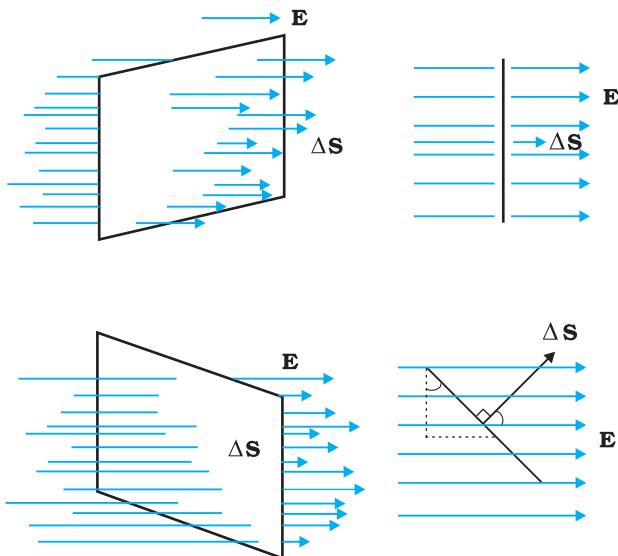


FIGURE 1.18 Dependence of flux on the inclination θ between \mathbf{E} and $\hat{\mathbf{n}}$.

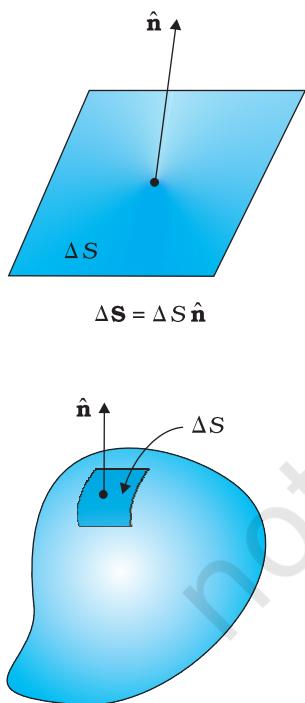


FIGURE 1.19
Convention for defining normal $\hat{\mathbf{n}}$ and $\Delta \mathbf{S}$.

we place a small planar element of area ΔS normal to \mathbf{E} at a point, the number of field lines crossing it is proportional* to $E \Delta S$. Now suppose we tilt the area element by angle θ . Clearly, the number of field lines crossing the area element will be smaller. The projection of the area element normal to E is $\Delta S \cos \theta$. Thus, the number of field lines crossing ΔS is proportional to $E \Delta S \cos \theta$. When $\theta = 90^\circ$, field lines will be parallel to ΔS and will not cross it at all (Fig. 1.18).

The orientation of area element and not merely its magnitude is important in many contexts. For example, in a stream, the amount of water flowing through a ring will naturally depend on how you hold the ring. If you hold it normal to the flow, maximum water will flow through it than if you hold it with some other orientation. This shows that an area element should be treated as a vector. It has a

magnitude and also a direction. How to specify the direction of a planar area? Clearly, the normal to the plane specifies the orientation of the plane. Thus the direction of a planar area vector is along its normal.

How to associate a vector to the area of a curved surface? We imagine dividing the surface into a large number of very small area elements. Each small area element may be treated as planar and a vector associated with it, as explained before.

Notice one ambiguity here. The direction of an area element is along its normal. But a normal can point in two directions. Which direction do we choose as the direction of the vector associated with the area element? This problem is resolved by some convention appropriate to the given context. For the case of a closed surface, this convention is very simple. The vector associated with every area element of a closed surface is taken to be in the direction of the *outward* normal. This is the convention used in Fig. 1.19. Thus, the area element vector $\Delta \mathbf{S}$ at a point on a closed surface equals $\Delta S \hat{\mathbf{n}}$ where ΔS is the magnitude of the area element and $\hat{\mathbf{n}}$ is a unit vector in the direction of outward normal at that point.

We now come to the definition of electric flux. Electric flux $\Delta\phi$ through an area element $\Delta \mathbf{S}$ is defined by

$$\Delta\phi = \mathbf{E} \cdot \Delta \mathbf{S} = E \Delta S \cos \theta \quad (1.11)$$

which, as seen before, is proportional to the number of field lines cutting the area element. The angle θ here is the angle between \mathbf{E} and $\Delta \mathbf{S}$. For a closed surface, with the convention stated already, θ is the angle between \mathbf{E} and the outward normal to the area element. Notice we could look at the expression $E \Delta S \cos \theta$ in two ways: $E (\Delta S \cos \theta)$ i.e., E times the

* It will not be proper to say that the number of field lines is equal to $E \Delta S$. The number of field lines is after all, a matter of how many field lines we choose to draw. What is physically significant is the relative number of field lines crossing a given area at different points.

projection of area normal to \mathbf{E} , or $E_{\perp} \Delta S$, i.e., component of \mathbf{E} along the normal to the area element times the magnitude of the area element. The unit of electric flux is N C⁻¹ m².

The basic definition of electric flux given by Eq. (1.11) can be used, in principle, to calculate the total flux through any given surface. All we have to do is to divide the surface into small area elements, calculate the flux at each element and add them up. Thus, the total flux ϕ through a surface S is

$$\phi \simeq \sum \mathbf{E} \cdot \Delta \mathbf{S} \quad (1.12)$$

The approximation sign is put because the electric field \mathbf{E} is taken to be constant over the small area element. This is mathematically exact only when you take the limit $\Delta S \rightarrow 0$ and the sum in Eq. (1.12) is written as an integral.

1.11 ELECTRIC DIPOLE

An electric dipole is a pair of equal and opposite point charges q and $-q$, separated by a distance $2a$. The line connecting the two charges defines a direction in space. By convention, the direction from $-q$ to q is said to be the direction of the dipole. The mid-point of locations of $-q$ and q is called the centre of the dipole.

The total charge of the electric dipole is obviously zero. This does not mean that the field of the electric dipole is zero. Since the charge q and $-q$ are separated by some distance, the electric fields due to them, when added, do not exactly cancel out. However, at distances much larger than the separation of the two charges forming a dipole ($r \gg 2a$), the fields due to q and $-q$ nearly cancel out. The electric field due to a dipole therefore falls off, at large distance, faster than like $1/r^2$ (the dependence on r of the field due to a single charge q). These qualitative ideas are borne out by the explicit calculation as follows:

1.11.1 The field of an electric dipole

The electric field of the pair of charges ($-q$ and q) at any point in space can be found out from Coulomb's law and the superposition principle. The results are simple for the following two cases: (i) when the point is on the dipole axis, and (ii) when it is in the *equatorial plane* of the dipole, i.e., on a plane perpendicular to the dipole axis through its centre. The electric field at any general point P is obtained by adding the electric fields \mathbf{E}_{-q} due to the charge $-q$ and \mathbf{E}_{+q} due to the charge q , by the parallelogram law of vectors.

(i) For points on the axis

Let the point P be at distance r from the centre of the dipole on the side of the charge q , as shown in Fig. 1.20(a). Then

$$\mathbf{E}_{-q} = -\frac{q}{4\pi\epsilon_0(r+a)^2} \hat{\mathbf{p}} \quad [1.13(a)]$$

where $\hat{\mathbf{p}}$ is the unit vector along the dipole axis (from $-q$ to q). Also

$$\mathbf{E}_{+q} = \frac{q}{4\pi\epsilon_0(r-a)^2} \hat{\mathbf{p}} \quad [1.13(b)]$$

Physics

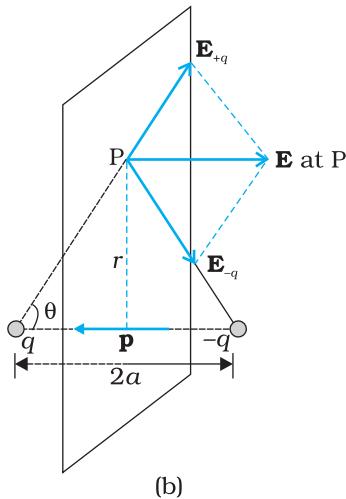
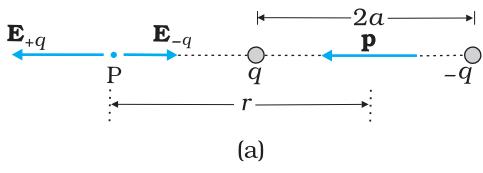


FIGURE 1.20 Electric field of a dipole
at (a) a point on the axis, (b) a point on the equatorial plane of the dipole.
 \mathbf{p} is the dipole moment vector of magnitude $p = q \times 2a$ and directed from $-q$ to q .

The total field at P is

$$\begin{aligned}\mathbf{E} &= \mathbf{E}_{+q} + \mathbf{E}_{-q} = \frac{q}{4\pi\epsilon_0} \left[\frac{1}{(r-a)^2} - \frac{1}{(r+a)^2} \right] \hat{\mathbf{p}} \\ &= \frac{q}{4\pi\epsilon_0} \frac{4ar}{(r^2-a^2)^2} \hat{\mathbf{p}}\end{aligned}\quad (1.14)$$

For $r \gg a$

$$\mathbf{E} = \frac{4qa}{4\pi\epsilon_0 r^3} \hat{\mathbf{p}} \quad (r \gg a) \quad (1.15)$$

(ii) For points on the equatorial plane

The magnitudes of the electric fields due to the two charges $+q$ and $-q$ are given by

$$E_{+q} = \frac{q}{4\pi\epsilon_0} \frac{1}{r^2+a^2} \quad [1.16(a)]$$

$$E_{-q} = \frac{q}{4\pi\epsilon_0} \frac{1}{r^2+a^2} \quad [1.16(b)]$$

and are equal.

The directions of \mathbf{E}_{+q} and \mathbf{E}_{-q} are as shown in Fig. 1.20(b). Clearly, the components normal to the dipole axis cancel away. The components along the dipole axis add up. The total electric field is opposite to $\hat{\mathbf{p}}$. We have

$$\begin{aligned}\mathbf{E} &= -(E_{+q} + E_{-q}) \cos\theta \hat{\mathbf{p}} \\ &= -\frac{2qa}{4\pi\epsilon_0 (r^2+a^2)^{3/2}} \hat{\mathbf{p}}\end{aligned}\quad (1.17)$$

At large distances ($r \gg a$), this reduces to

$$\mathbf{E} = -\frac{2qa}{4\pi\epsilon_0 r^3} \hat{\mathbf{p}} \quad (r \gg a) \quad (1.18)$$

From Eqs. (1.15) and (1.18), it is clear that the dipole field at large distances does not involve q and a separately; it depends on the product qa . This suggests the definition of dipole moment. The *dipole moment vector* \mathbf{p} of an electric dipole is defined by

$$\mathbf{p} = q \times 2a \hat{\mathbf{p}} \quad (1.19)$$

that is, it is a vector whose magnitude is charge q times the separation $2a$ (between the pair of charges $q, -q$) and the direction is along the line from $-q$ to q . In terms of \mathbf{p} , the electric field of a dipole at large distances takes simple forms:

At a point on the dipole axis

$$\mathbf{E} = \frac{2\mathbf{p}}{4\pi\epsilon_0 r^3} \quad (r \gg a) \quad (1.20)$$

At a point on the equatorial plane

$$\mathbf{E} = -\frac{\mathbf{p}}{4\pi\epsilon_0 r^3} \quad (r \gg a) \quad (1.21)$$

Notice the important point that the dipole field at large distances falls off not as $1/r^2$ but as $1/r^3$. Further, the magnitude and the direction of the dipole field depends not only on the distance r but also on the angle between the position vector \mathbf{r} and the dipole moment \mathbf{p} .

We can think of the limit when the dipole size $2a$ approaches zero, the charge q approaches infinity in such a way that the product $p = q \times 2a$ is finite. Such a dipole is referred to as a *point dipole*. For a point dipole, Eqs. (1.20) and (1.21) are exact, true for any r .

1.11.2 Physical significance of dipoles

In most molecules, the centres of positive charges and of negative charges* lie at the same place. Therefore, their dipole moment is zero. CO_2 and CH_4 are of this type of molecules. However, they develop a dipole moment when an electric field is applied. But in some molecules, the centres of negative charges and of positive charges do not coincide. Therefore they have a permanent electric dipole moment, even in the absence of an electric field. Such molecules are called polar molecules. Water molecules, H_2O , is an example of this type. Various materials give rise to interesting properties and important applications in the presence or absence of electric field.

Example 1.10 Two charges $\pm 10 \mu\text{C}$ are placed 5.0 mm apart. Determine the electric field at (a) a point P on the axis of the dipole 15 cm away from its centre O on the side of the positive charge, as shown in Fig. 1.21(a), and (b) a point Q, 15 cm away from O on a line passing through O and normal to the axis of the dipole, as shown in Fig. 1.21(b).

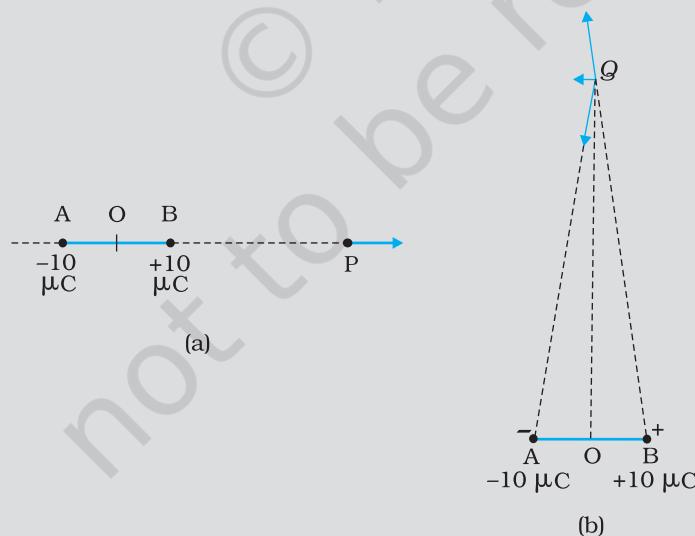


FIGURE 1.21

EXAMPLE 1.10

* Centre of a collection of positive point charges is defined much the same way

as the centre of mass: $\mathbf{r}_{\text{cm}} = \frac{\sum_i q_i \mathbf{r}_i}{\sum_i q_i}$.

Solution (a) Field at P due to charge $+10 \mu\text{C}$

$$= \frac{10^{-5} \text{ C}}{4\pi(8.854 \times 10^{-12} \text{ C}^2 \text{ N}^{-1} \text{ m}^{-2})} \times \frac{1}{(15 - 0.25)^2 \times 10^{-4} \text{ m}^2}$$

$$= 4.13 \times 10^6 \text{ N C}^{-1} \text{ along BP}$$

Field at P due to charge $-10 \mu\text{C}$

$$= \frac{10^{-5} \text{ C}}{4\pi(8.854 \times 10^{-12} \text{ C}^2 \text{ N}^{-1} \text{ m}^{-2})} \times \frac{1}{(15 + 0.25)^2 \times 10^{-4} \text{ m}^2}$$

$$= 3.86 \times 10^6 \text{ N C}^{-1} \text{ along PA}$$

The resultant electric field at P due to the two charges at A and B is
 $= 2.7 \times 10^5 \text{ N C}^{-1}$ along BP.

In this example, the ratio OP/OB is quite large ($= 60$). Thus, we can expect to get approximately the same result as above by directly using the formula for electric field at a far-away point on the axis of a dipole. For a dipole consisting of charges $\pm q$, $2a$ distance apart, the electric field at a distance r from the centre on the axis of the dipole has a magnitude

$$E = \frac{2p}{4\pi\epsilon_0 r^3} \quad (r/a \gg 1)$$

where $p = 2a q$ is the magnitude of the dipole moment.

The direction of electric field on the dipole axis is always along the direction of the dipole moment vector (i.e., from $-q$ to q). Here, $p = 10^{-5} \text{ C} \times 5 \times 10^{-3} \text{ m} = 5 \times 10^{-8} \text{ C m}$

Therefore,

$$E = \frac{2 \times 5 \times 10^{-8} \text{ C m}}{4\pi(8.854 \times 10^{-12} \text{ C}^2 \text{ N}^{-1} \text{ m}^{-2})} \times \frac{1}{(15)^3 \times 10^{-6} \text{ m}^3} = 2.6 \times 10^5 \text{ N C}^{-1}$$

along the dipole moment direction AB, which is close to the result obtained earlier.

(b) Field at Q due to charge $+10 \mu\text{C}$ at B

$$= \frac{10^{-5} \text{ C}}{4\pi(8.854 \times 10^{-12} \text{ C}^2 \text{ N}^{-1} \text{ m}^{-2})} \times \frac{1}{[15^2 + (0.25)^2] \times 10^{-4} \text{ m}^2}$$

$$= 3.99 \times 10^6 \text{ N C}^{-1} \text{ along BQ}$$

Field at Q due to charge $-10 \mu\text{C}$ at A

$$= \frac{10^{-5} \text{ C}}{4\pi(8.854 \times 10^{-12} \text{ C}^2 \text{ N}^{-1} \text{ m}^{-2})} \times \frac{1}{[15^2 + (0.25)^2] \times 10^{-4} \text{ m}^2}$$

$$= 3.99 \times 10^6 \text{ N C}^{-1} \text{ along QA.}$$

Clearly, the components of these two forces with equal magnitudes cancel along the direction OQ but add up along the direction parallel to BA. Therefore, the resultant electric field at Q due to the two charges at A and B is

$$= 2 \times \frac{0.25}{\sqrt{15^2 + (0.25)^2}} \times 3.99 \times 10^6 \text{ N C}^{-1} \text{ along BA}$$

$$= 1.33 \times 10^5 \text{ N C}^{-1} \text{ along BA.}$$

As in (a), we can expect to get approximately the same result by directly using the formula for dipole field at a point on the normal to the axis of the dipole:

$$\begin{aligned}
 E &= \frac{p}{4\pi\epsilon_0 r^3} & (r/a \gg 1) \\
 &= \frac{5 \times 10^{-8} \text{ Cm}}{4\pi(8.854 \times 10^{-12} \text{ C}^2 \text{ N}^{-1} \text{ m}^{-2})} \times \frac{1}{(15)^3 \times 10^{-6} \text{ m}^3} \\
 &= 1.33 \times 10^5 \text{ N C}^{-1}.
 \end{aligned}$$

The direction of electric field in this case is opposite to the direction of the dipole moment vector. Again the result agrees with that obtained before.

EXAMPLE 1.10

1.12 DIPOLE IN A UNIFORM EXTERNAL FIELD

Consider a permanent dipole of dipole moment \mathbf{p} in a uniform external field \mathbf{E} , as shown in Fig. 1.22. (By permanent dipole, we mean that \mathbf{p} exists irrespective of \mathbf{E} ; it has not been induced by \mathbf{E} .)

There is a force $q\mathbf{E}$ on q and a force $-q\mathbf{E}$ on $-q$. The net force on the dipole is zero, since \mathbf{E} is uniform. However, the charges are separated, so the forces act at different points, resulting in a torque on the dipole. When the net force is zero, the torque (couple) is independent of the origin. Its magnitude equals the magnitude of each force multiplied by the arm of the couple (perpendicular distance between the two antiparallel forces).

$$\begin{aligned}
 \text{Magnitude of torque} &= qE \times 2a \sin\theta \\
 &= 2qaE \sin\theta
 \end{aligned}$$

Its direction is normal to the plane of the paper, coming out of it.

The magnitude of $\mathbf{p} \times \mathbf{E}$ is also $pE \sin\theta$ and its direction is normal to the paper, coming out of it. Thus,

$$\tau = \mathbf{p} \times \mathbf{E} \quad (1.22)$$

This torque will tend to align the dipole with the field \mathbf{E} . When \mathbf{p} is aligned with \mathbf{E} , the torque is zero.

What happens if the field is not uniform? In that case, the net force will evidently be non-zero. In addition there will, in general, be a torque on the system as before. The general case is involved, so let us consider the simpler situations when \mathbf{p} is parallel to \mathbf{E} or antiparallel to \mathbf{E} . In either case, the net torque is zero, but there is a net force on the dipole if \mathbf{E} is not uniform.

Figure 1.23 is self-explanatory. It is easily seen that when \mathbf{p} is parallel to \mathbf{E} , the dipole has a net force in the direction of increasing field. When \mathbf{p} is antiparallel to \mathbf{E} , the net force on the dipole is in the direction of decreasing field. In general, the force depends on the orientation of \mathbf{p} with respect to \mathbf{E} .

This brings us to a common observation in frictional electricity. A comb run through dry hair attracts pieces of paper. The comb, as we know, acquires charge through friction. But the paper is not charged. What then explains the attractive force? Taking the clue from the preceding

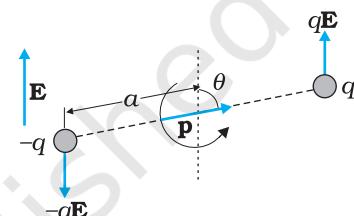
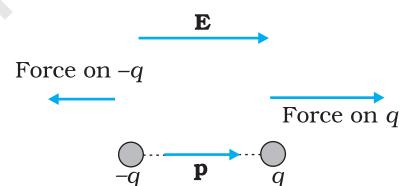
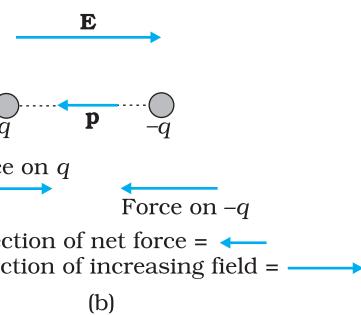


FIGURE 1.22 Dipole in a uniform electric field.



Direction of net force =
Direction of increasing field =

(a)



Direction of net force =
Direction of increasing field =

(b)

FIGURE 1.23 Electric force on a dipole: (a) \mathbf{E} parallel to \mathbf{p} , (b) \mathbf{E} antiparallel to \mathbf{p} .

discussion, the charged comb ‘polarizes’ the piece of paper, i.e., induces a net dipole moment in the direction of field. Further, the electric field due to the comb is not uniform. In this situation, it is easily seen that the paper should move in the direction of the comb!

1.13 CONTINUOUS CHARGE DISTRIBUTION

We have so far dealt with charge configurations involving discrete charges q_1, q_2, \dots, q_n . One reason why we restricted to discrete charges is that the mathematical treatment is simpler and does not involve calculus. For many purposes, however, it is impractical to work in terms of discrete charges and we need to work with continuous charge distributions. For example, on the surface of a charged conductor, it is impractical to specify the charge distribution in terms of the locations of the microscopic charged constituents. It is more feasible to consider an area element ΔS (Fig. 1.24) on the surface of the conductor (which is very small on the macroscopic scale but big enough to include a very large number of electrons) and specify the charge ΔQ on that element. We then define a *surface charge density* σ at the area element by

$$\sigma = \frac{\Delta Q}{\Delta S} \quad (1.23)$$

We can do this at different points on the conductor and thus arrive at a continuous function σ , called the surface charge density. The surface charge density σ so defined ignores the quantisation of charge and the discontinuity in charge distribution at the microscopic level*. σ represents macroscopic surface charge density, which in a sense, is a smoothed out average of the microscopic charge density over an area element ΔS which, as said before, is large microscopically but small macroscopically. The units for σ are C/m^2 .

Similar considerations apply for a line charge distribution and a volume charge distribution. The *linear charge density* λ of a wire is defined by

$$\lambda = \frac{\Delta Q}{\Delta l} \quad (1.24)$$

where Δl is a small line element of wire on the macroscopic scale that, however, includes a large number of microscopic charged constituents, and ΔQ is the charge contained in that line element. The units for λ are C/m . The *volume charge density* (sometimes simply called charge density) is defined in a similar manner:

$$\rho = \frac{\Delta Q}{\Delta V} \quad (1.25)$$

where ΔQ is the charge included in the macroscopically small volume element ΔV that includes a large number of microscopic charged constituents. The units for ρ are C/m^3 .

The notion of continuous charge distribution is similar to that we adopt for continuous mass distribution in mechanics. When we refer to

* At the microscopic level, charge distribution is discontinuous, because they are discrete charges separated by intervening space where there is no charge.

the density of a liquid, we are referring to its macroscopic density. We regard it as a continuous fluid and ignore its discrete molecular constitution.

The field due to a continuous charge distribution can be obtained in much the same way as for a system of discrete charges, Eq. (1.10). Suppose a continuous charge distribution in space has a charge density ρ . Choose any convenient origin O and let the position vector of any point in the charge distribution be \mathbf{r} . The charge density ρ may vary from point to point, i.e., it is a function of \mathbf{r} . Divide the charge distribution into small volume elements of size ΔV . The charge in a volume element ΔV is $\rho \Delta V$.

Now, consider any general point P (inside or outside the distribution) with position vector \mathbf{R} (Fig. 1.24). Electric field due to the charge $\rho \Delta V$ is given by Coulomb's law:

$$\Delta \mathbf{E} = \frac{1}{4\pi\epsilon_0} \frac{\rho \Delta V}{r'^2} \hat{\mathbf{r}}' \quad (1.26)$$

where r' is the distance between the charge element and P, and $\hat{\mathbf{r}}'$ is a unit vector in the direction from the charge element to P. By the superposition principle, the total electric field due to the charge distribution is obtained by summing over electric fields due to different volume elements:

$$\mathbf{E} \approx \frac{1}{4\pi\epsilon_0} \sum_{all \Delta V} \frac{\rho \Delta V}{r'^2} \hat{\mathbf{r}}' \quad (1.27)$$

Note that ρ , r' , $\hat{\mathbf{r}}'$ all can vary from point to point. In a strict mathematical method, we should let $\Delta V \rightarrow 0$ and the sum then becomes an integral; but we omit that discussion here, for simplicity. In short, using Coulomb's law and the superposition principle, electric field can be determined for any charge distribution, discrete or continuous or part discrete and part continuous.

1.14 GAUSS'S LAW

As a simple application of the notion of electric flux, let us consider the total flux through a sphere of radius r , which encloses a point charge q at its centre. Divide the sphere into small area elements, as shown in Fig. 1.25.

The flux through an area element $\Delta \mathbf{S}$ is

$$\Delta\phi = \mathbf{E} \cdot \Delta \mathbf{S} = \frac{q}{4\pi\epsilon_0 r^2} \hat{\mathbf{r}} \cdot \Delta \mathbf{S} \quad (1.28)$$

where we have used Coulomb's law for the electric field due to a single charge q . The unit vector $\hat{\mathbf{r}}$ is along the radius vector from the centre to the area element. Now, since the normal to a sphere at every point is along the radius vector at that point, the area element $\Delta \mathbf{S}$ and $\hat{\mathbf{r}}$ have the same direction. Therefore,

$$\Delta\phi = \frac{q}{4\pi\epsilon_0 r^2} \Delta S \quad (1.29)$$

since the magnitude of a unit vector is 1.

The total flux through the sphere is obtained by adding up flux through all the different area elements:

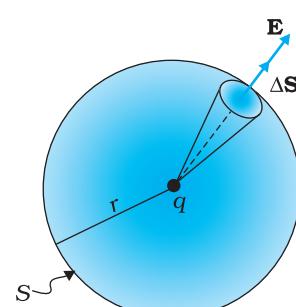


FIGURE 1.25 Flux through a sphere enclosing a point charge q at its centre.

Physics

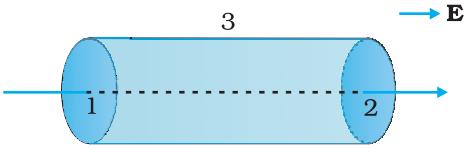


FIGURE 1.26 Calculation of the flux of uniform electric field through the surface of a cylinder.

$$\phi = \sum_{all \Delta S} \frac{q}{4\pi \epsilon_0 r^2} \Delta S$$

Since each area element of the sphere is at the same distance r from the charge,

$$\phi = \frac{q}{4\pi \epsilon_0 r^2} \sum_{all \Delta S} \Delta S = \frac{q}{4\pi \epsilon_0 r^2} S$$

Now S , the total area of the sphere, equals $4\pi r^2$. Thus,

$$\phi = \frac{q}{4\pi \epsilon_0 r^2} \times 4\pi r^2 = \frac{q}{\epsilon_0} \quad (1.30)$$

Equation (1.30) is a simple illustration of a general result of electrostatics called Gauss's law.

We state *Gauss's law* without proof:

Electric flux through a closed surface S

$$= q/\epsilon_0 \quad (1.31)$$

q = total charge enclosed by S .

The law implies that the total electric flux through a closed surface is zero if no charge is enclosed by the surface. We can see that explicitly in the simple situation of Fig. 1.26.

Here the electric field is uniform and we are considering a closed cylindrical surface, with its axis parallel to the uniform field \mathbf{E} . The total flux ϕ through the surface is $\phi = \phi_1 + \phi_2 + \phi_3$, where ϕ_1 and ϕ_2 represent the flux through the surfaces 1 and 2 (of circular cross-section) of the cylinder and ϕ_3 is the flux through the curved cylindrical part of the closed surface. Now the normal to the surface 3 at every point is perpendicular to \mathbf{E} , so by definition of flux, $\phi_3 = 0$. Further, the outward normal to 2 is along \mathbf{E} while the outward normal to 1 is opposite to \mathbf{E} . Therefore,

$$\phi_1 = -ES_1, \quad \phi_2 = +ES_2$$

$$S_1 = S_2 = S$$

where S is the area of circular cross-section. Thus, the total flux is zero, as expected by Gauss's law. Thus, whenever you find that the net electric flux through a closed surface is zero, we conclude that the total charge contained in the closed surface is zero.

The great significance of Gauss's law Eq. (1.31), is that it is true in general, and not only for the simple cases we have considered above. Let us note some important points regarding this law:

- (i) Gauss's law is true for any closed surface, no matter what its shape or size.
- (ii) The term q on the right side of Gauss's law, Eq. (1.31), includes the sum of all charges enclosed by the surface. The charges may be located anywhere inside the surface.
- (iii) In the situation when the surface is so chosen that there are some charges inside and some outside, the electric field [whose flux appears on the left side of Eq. (1.31)] is due to all the charges, both inside and outside S . The term q on the right side of Gauss's law, however, represents only the total charge inside S .

- (iv) The surface that we choose for the application of Gauss's law is called the Gaussian surface. You may choose any Gaussian surface and apply Gauss's law. However, take care not to let the Gaussian surface pass through any discrete charge. This is because electric field due to a system of discrete charges is not well defined at the location of any charge. (As you go close to the charge, the field grows without any bound.) However, the Gaussian surface can pass through a continuous charge distribution.
- (v) Gauss's law is often useful towards a much easier calculation of the electrostatic field *when the system has some symmetry*. This is facilitated by the choice of a suitable Gaussian surface.
- (vi) Finally, Gauss's law is based on the inverse square dependence on distance contained in the Coulomb's law. Any violation of Gauss's law will indicate departure from the inverse square law.

Example 1.11 The electric field components in Fig. 1.27 are $E_x = \alpha x^{1/2}$, $E_y = E_z = 0$, in which $\alpha = 800 \text{ N/C m}^{1/2}$. Calculate (a) the flux through the cube, and (b) the charge within the cube. Assume that $a = 0.1 \text{ m}$.

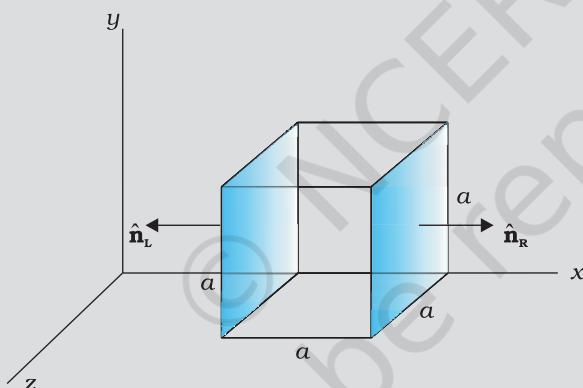


FIGURE 1.27

Solution

- (a) Since the electric field has only an x component, for faces perpendicular to x direction, the angle between \mathbf{E} and $\Delta\mathbf{S}$ is $\pm \pi/2$. Therefore, the flux $\phi = \mathbf{E} \cdot \Delta\mathbf{S}$ is separately zero for each face of the cube except the two shaded ones. Now the magnitude of the electric field at the left face is

$$E_L = \alpha x^{1/2} = \alpha a^{1/2}$$

($x = a$ at the left face).

The magnitude of electric field at the right face is

$$E_R = \alpha x^{1/2} = \alpha (2a)^{1/2}$$

($x = 2a$ at the right face).

The corresponding fluxes are

$$\begin{aligned}\phi_L &= \mathbf{E}_L \cdot \Delta\mathbf{S} = \Delta S \mathbf{E}_L \cdot \hat{\mathbf{n}}_L = E_L \Delta S \cos\theta = -E_L \Delta S, \text{ since } \theta = 180^\circ \\ &= -E_L a^2\end{aligned}$$

$$\begin{aligned}\phi_R &= \mathbf{E}_R \cdot \Delta\mathbf{S} = E_R \Delta S \cos\theta = E_R \Delta S, \text{ since } \theta = 0^\circ \\ &= E_R a^2\end{aligned}$$

Net flux through the cube

EXAMPLE 1.11

EXAMPLE 1.11

$$\begin{aligned}
 &= \phi_R + \phi_L = E_R a^2 - E_L a^2 = a^2 (E_R - E_L) = \alpha a^2 [(2a)^{1/2} - a^{1/2}] \\
 &= \alpha a^{5/2} (\sqrt{2} - 1) \\
 &= 800 (0.1)^{5/2} (\sqrt{2} - 1) \\
 &= 1.05 \text{ N m}^2 \text{ C}^{-1}
 \end{aligned}$$

(b) We can use Gauss's law to find the total charge q inside the cube. We have $\phi = q/\epsilon_0$ or $q = \phi\epsilon_0$. Therefore,

$$q = 1.05 \times 8.854 \times 10^{-12} \text{ C} = 9.27 \times 10^{-12} \text{ C}$$

EXAMPLE 1.12

Example 1.12 An electric field is uniform, and in the positive x direction for positive x , and uniform with the same magnitude but in the negative x direction for negative x . It is given that $\mathbf{E} = 200 \hat{\mathbf{i}}$ N/C for $x > 0$ and $\mathbf{E} = -200 \hat{\mathbf{i}}$ N/C for $x < 0$. A right circular cylinder of length 20 cm and radius 5 cm has its centre at the origin and its axis along the x -axis so that one face is at $x = +10$ cm and the other is at $x = -10$ cm (Fig. 1.28). (a) What is the net outward flux through each flat face? (b) What is the flux through the side of the cylinder? (c) What is the net outward flux through the cylinder? (d) What is the net charge inside the cylinder?

Solution

(a) We can see from the figure that on the left face \mathbf{E} and $\Delta \mathbf{S}$ are parallel. Therefore, the outward flux is

$$\begin{aligned}
 \phi_L &= \mathbf{E} \cdot \Delta \mathbf{S} = -200 \hat{\mathbf{i}} \cdot \Delta \mathbf{S} \\
 &= +200 \Delta S, \text{ since } \hat{\mathbf{i}} \cdot \Delta \mathbf{S} = -\Delta S \\
 &= +200 \times \pi (0.05)^2 = +1.57 \text{ N m}^2 \text{ C}^{-1}
 \end{aligned}$$

On the right face, \mathbf{E} and $\Delta \mathbf{S}$ are parallel and therefore

$$\phi_R = \mathbf{E} \cdot \Delta \mathbf{S} = +1.57 \text{ N m}^2 \text{ C}^{-1}$$

(b) For any point on the side of the cylinder \mathbf{E} is perpendicular to $\Delta \mathbf{S}$ and hence $\mathbf{E} \cdot \Delta \mathbf{S} = 0$. Therefore, the flux out of the side of the cylinder is zero.

(c) Net outward flux through the cylinder

$$\phi = 1.57 + 1.57 + 0 = 3.14 \text{ N m}^2 \text{ C}^{-1}$$

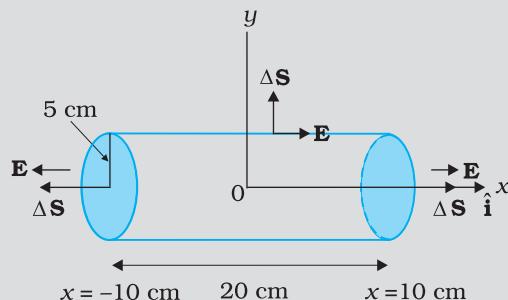


FIGURE 1.28

EXAMPLE 1.12

(d) The net charge within the cylinder can be found by using Gauss's law which gives

$$\begin{aligned}
 q &= \epsilon_0 \phi \\
 &= 3.14 \times 8.854 \times 10^{-12} \text{ C} \\
 &= 2.78 \times 10^{-11} \text{ C}
 \end{aligned}$$

1.15 APPLICATIONS OF GAUSS'S LAW

The electric field due to a general charge distribution is, as seen above, given by Eq. (1.27). In practice, except for some special cases, the summation (or integration) involved in this equation cannot be carried out to give electric field at every point in space. For some symmetric charge configurations, however, it is possible to obtain the electric field in a simple way using the Gauss's law. This is best understood by some examples.

1.15.1 Field due to an infinitely long straight uniformly charged wire

Consider an infinitely long thin straight wire with uniform linear charge density λ . The wire is obviously an axis of symmetry. Suppose we take the radial vector from O to P and rotate it around the wire. The points P, P', P'' so obtained are completely equivalent with respect to the charged wire. This implies that the electric field must have the same magnitude at these points. The direction of electric field at every point must be radial (outward if $\lambda > 0$, inward if $\lambda < 0$). This is clear from Fig. 1.29.

Consider a pair of line elements P_1 and P_2 of the wire, as shown. The electric fields produced by the two elements of the pair when summed give a resultant electric field which is radial (the components normal to the radial vector cancel). This is true for any such pair and hence the total field at any point P is radial. Finally, since the wire is infinite, electric field does not depend on the position of P along the length of the wire. In short, the electric field is everywhere radial in the plane cutting the wire normally, and its magnitude depends only on the radial distance r .

To calculate the field, imagine a cylindrical Gaussian surface, as shown in the Fig. 1.29(b). Since the field is everywhere radial, flux through the two ends of the cylindrical Gaussian surface is zero. At the cylindrical part of the surface, \mathbf{E} is normal to the surface at every point, and its magnitude is constant, since it depends only on r . The surface area of the curved part is $2\pi rl$, where l is the length of the cylinder.

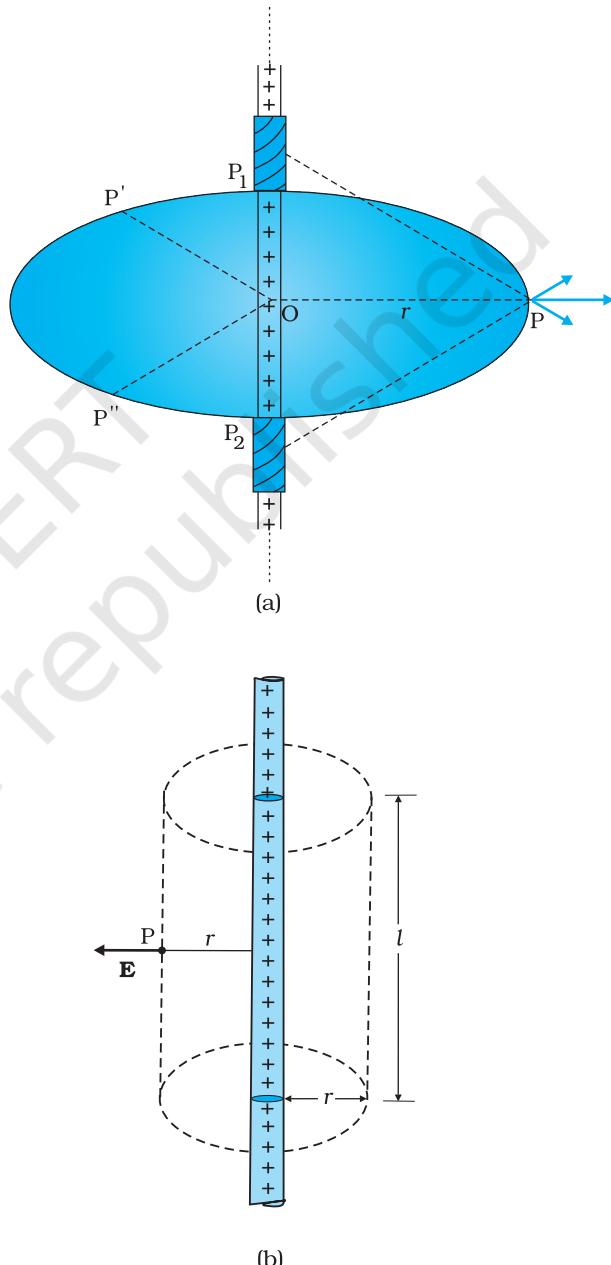


FIGURE 1.29 (a) Electric field due to an infinitely long thin straight wire is radial, (b) The Gaussian surface for a long thin wire of uniform linear charge density.

Flux through the Gaussian surface

$$\begin{aligned} &= \text{flux through the curved cylindrical part of the surface} \\ &= E \times 2\pi rl \end{aligned}$$

The surface includes charge equal to λl . Gauss's law then gives
 $E \times 2\pi rl = \lambda l / \epsilon_0$

$$\text{i.e., } E = \frac{\lambda}{2\pi\epsilon_0 r}$$

Vectorially, \mathbf{E} at any point is given by

$$\mathbf{E} = \frac{\lambda}{2\pi\epsilon_0 r} \hat{\mathbf{n}} \quad (1.32)$$

where $\hat{\mathbf{n}}$ is the radial unit vector in the plane normal to the wire passing through the point. \mathbf{E} is directed outward if λ is positive and inward if λ is negative.

Note that when we write a vector \mathbf{A} as a scalar multiplied by a unit vector, i.e., as $\mathbf{A} = A \hat{\mathbf{a}}$, the scalar A is an algebraic number. It can be negative or positive. The direction of \mathbf{A} will be the same as that of the unit vector $\hat{\mathbf{a}}$ if $A > 0$ and opposite to $\hat{\mathbf{a}}$ if $A < 0$. When we want to restrict to non-negative values, we use the symbol $|\mathbf{A}|$ and call it the modulus of \mathbf{A} . Thus, $|\mathbf{A}| \geq 0$.

Also note that though only the charge enclosed by the surface (λl) was included above, the electric field \mathbf{E} is due to the charge on the entire wire. Further, the assumption that the wire is infinitely long is crucial. Without this assumption, we cannot take \mathbf{E} to be normal to the curved part of the cylindrical Gaussian surface. However, Eq. (1.32) is approximately true for electric field around the central portions of a long wire, where the end effects may be ignored.

1.15.2 Field due to a uniformly charged infinite plane sheet

Let σ be the uniform surface charge density of an infinite plane sheet (Fig. 1.30). We take the x -axis normal to the given plane. By symmetry, the electric field will not depend on y and z coordinates and its direction at every point must be parallel to the x -direction.

We can take the Gaussian surface to be a rectangular parallelepiped of cross sectional area A , as shown. (A cylindrical surface will also do.) As seen from the figure, only the two faces 1 and 2 will contribute to the flux; electric field lines are parallel to the other faces and they, therefore, do not contribute to the total flux.

The unit vector normal to surface 1 is in $-x$ direction while the unit vector normal to surface 2 is in the $+x$ direction. Therefore, flux $\mathbf{E} \cdot \Delta \mathbf{S}$ through both the surfaces are equal and add up. Therefore the net flux through the Gaussian surface is $2 EA$. The charge enclosed by the closed surface is σA . Therefore by Gauss's law,

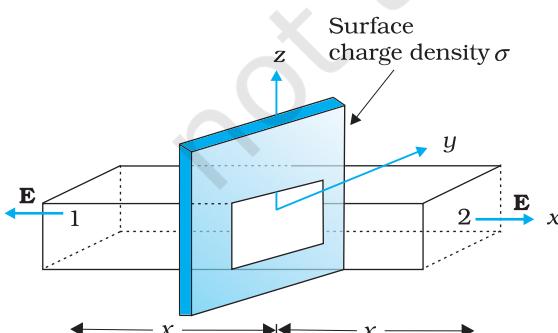


FIGURE 1.30 Gaussian surface for a uniformly charged infinite plane sheet.

$$2EA = \sigma A / \epsilon_0$$

$$\text{or, } E = \sigma / 2\epsilon_0$$

Vectorically,

$$\mathbf{E} = \frac{\sigma}{2\epsilon_0} \hat{\mathbf{n}} \quad (1.33)$$

where $\hat{\mathbf{n}}$ is a unit vector normal to the plane and going away from it.

\mathbf{E} is directed away from the plate if σ is positive and toward the plate if σ is negative. Note that the above application of the Gauss' law has brought out an additional fact: E is independent of x also.

For a finite large planar sheet, Eq. (1.33) is approximately true in the middle regions of the planar sheet, away from the ends.

1.15.3 Field due to a uniformly charged thin spherical shell

Let σ be the uniform surface charge density of a thin spherical shell of radius R (Fig. 1.31). The situation has obvious spherical symmetry. The field at any point P , outside or inside, can depend only on r (the radial distance from the centre of the shell to the point) and must be radial (i.e., along the radius vector).

(i) Field outside the shell: Consider a point P outside the shell with radius vector \mathbf{r} . To calculate \mathbf{E} at P , we take the Gaussian surface to be a sphere of radius r and with centre O , passing through P . All points on this sphere are equivalent relative to the given charged configuration. (That is what we mean by spherical symmetry.) The electric field at each point of the Gaussian surface, therefore, has the same magnitude E and is along the radius vector at each point. Thus, \mathbf{E} and $\Delta\mathbf{S}$ at every point are parallel and the flux through each element is $E \Delta S$. Summing over all ΔS , the flux through the Gaussian surface is $E \times 4 \pi r^2$. The charge enclosed is $\sigma \times 4 \pi R^2$. By Gauss's law

$$E \times 4 \pi r^2 = \frac{\sigma}{\epsilon_0} 4 \pi R^2$$

$$\text{Or, } E = \frac{\sigma R^2}{\epsilon_0 r^2} = \frac{q}{4\pi\epsilon_0 r^2}$$

where $q = 4\pi R^2 \sigma$ is the total charge on the spherical shell.
Vectorially,

$$\mathbf{E} = \frac{q}{4\pi\epsilon_0 r^2} \hat{\mathbf{r}} \quad (1.34)$$

The electric field is directed outward if $q > 0$ and inward if $q < 0$. This, however, is exactly the field produced by a charge q placed at the centre O . Thus for points outside the shell, the field due to a uniformly charged shell is as if the entire charge of the shell is concentrated at its centre.

(ii) Field inside the shell: In Fig. 1.31(b), the point P is inside the shell. The Gaussian surface is again a sphere through P centred at O .

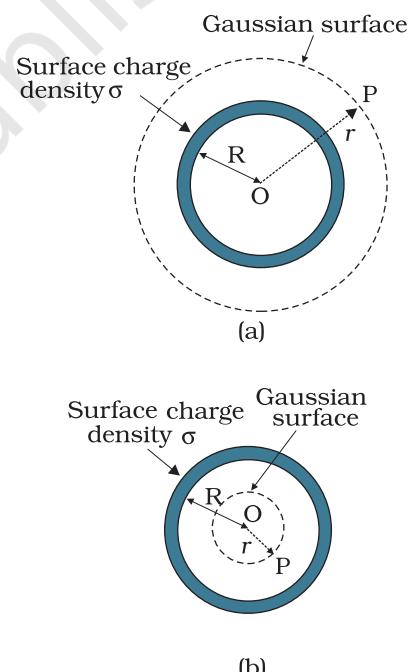


FIGURE 1.31 Gaussian surfaces for a point with (a) $r > R$, (b) $r < R$.

The flux through the Gaussian surface, calculated as before, is $E \times 4 \pi r^2$. However, in this case, the Gaussian surface encloses no charge. Gauss's law then gives

$$E \times 4 \pi r^2 = 0$$

$$\text{i.e., } E = 0 \quad (r < R)$$

(1.35)

that is, the field due to a uniformly charged thin shell is zero at all points inside the shell*. This important result is a direct consequence of Gauss's law which follows from Coulomb's law. The experimental verification of this result confirms the $1/r^2$ dependence in Coulomb's law.

Example 1.13 An early model for an atom considered it to have a positively charged point nucleus of charge Ze , surrounded by a uniform density of negative charge up to a radius R . The atom as a whole is neutral. For this model, what is the electric field at a distance r from the nucleus?

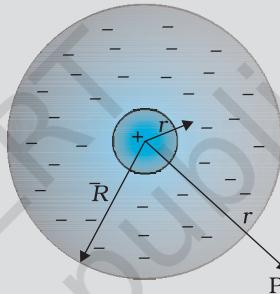


FIGURE 1.32

Solution The charge distribution for this model of the atom is as shown in Fig. 1.32. The total negative charge in the uniform spherical charge distribution of radius R must be $-Ze$, since the atom (nucleus of charge Ze + negative charge) is neutral. This immediately gives us the negative charge density ρ , since we must have

$$\frac{4\pi R^3}{3} \rho = -Ze$$

$$\text{or } \rho = -\frac{3Ze}{4\pi R^3}$$

To find the electric field $\mathbf{E}(\mathbf{r})$ at a point P which is a distance r away from the nucleus, we use Gauss's law. Because of the spherical symmetry of the charge distribution, the magnitude of the electric field $\mathbf{E}(\mathbf{r})$ depends only on the radial distance, no matter what the direction of \mathbf{r} . Its direction is along (or opposite to) the radius vector \mathbf{r} from the origin to the point P . The obvious Gaussian surface is a spherical surface centred at the nucleus. We consider two situations, namely, $r < R$ and $r > R$.

(i) $r < R$: The electric flux ϕ enclosed by the spherical surface is

$$\phi = E(r) \times 4\pi r^2$$

where $E(r)$ is the magnitude of the electric field at r . This is because

EXAMPLE 1.13

* Compare this with a uniform mass shell discussed in Section 8.5 of Class XI Textbook of Physics.

the field at any point on the spherical Gaussian surface has the same direction as the normal to the surface there, and has the same magnitude at all points on the surface.

The charge q enclosed by the Gaussian surface is the positive nuclear charge and the negative charge within the sphere of radius r ,

$$\text{i.e., } q = Z e + \frac{4\pi r^3}{3} \rho$$

Substituting for the charge density ρ obtained earlier, we have

$$q = Z e - Z e \frac{r^3}{R^3}$$

Gauss's law then gives,

$$E(r) = \frac{Z e}{4 \pi \epsilon_0} \left(\frac{1}{r^2} - \frac{r}{R^3} \right); \quad r < R$$

The electric field is directed radially outward.

(ii) $r > R$ In this case, the total charge enclosed by the Gaussian spherical surface is zero since the atom is neutral. Thus, from Gauss's law,

$$E(r) \times 4\pi r^2 = 0 \text{ or } E(r) = 0; \quad r > R$$

At $r = R$, both cases give the same result: $E = 0$.

EXAMPLE 1.13

ON SYMMETRY OPERATIONS

In Physics, we often encounter systems with various symmetries. Consideration of these symmetries helps one arrive at results much faster than otherwise by a straightforward calculation. Consider, for example an infinite uniform sheet of charge (surface charge density σ) along the y - z plane. This system is unchanged if (a) translated parallel to the y - z plane in any direction, (b) rotated about the x -axis through any angle. As the system is unchanged under such symmetry operation, so must its properties be. In particular, in this example, the electric field \mathbf{E} must be unchanged.

Translation symmetry along the y -axis shows that the electric field must be the same at a point $(0, y_1, 0)$ as at $(0, y_2, 0)$. Similarly translational symmetry along the z -axis shows that the electric field at two points $(0, 0, z_1)$ and $(0, 0, z_2)$ must be the same. By using rotation symmetry around the x -axis, we can conclude that \mathbf{E} must be perpendicular to the y - z plane, that is, it must be parallel to the x -direction.

Try to think of a symmetry now which will tell you that the magnitude of the electric field is a constant, independent of the x -coordinate. It thus turns out that the magnitude of the electric field due to a uniformly charged infinite conducting sheet is the same at all points in space. The direction, however, is opposite of each other on either side of the sheet.

Compare this with the effort needed to arrive at this result by a direct calculation using Coulomb's law.

SUMMARY

1. Electric and magnetic forces determine the properties of atoms, molecules and bulk matter.
2. From simple experiments on frictional electricity, one can infer that there are two types of charges in nature; and that like charges repel and unlike charges attract. By convention, the charge on a glass rod rubbed with silk is positive; that on a plastic rod rubbed with fur is then negative.
3. Conductors allow movement of electric charge through them, insulators do not. In metals, the mobile charges are electrons; in electrolytes both positive and negative ions are mobile.
4. Electric charge has three basic properties: quantisation, additivity and conservation.

Quantisation of electric charge means that total charge (q) of a body is always an integral multiple of a basic quantum of charge (e) i.e., $q = n e$, where $n = 0, \pm 1, \pm 2, \pm 3, \dots$. Proton and electron have charges $+e, -e$, respectively. For macroscopic charges for which n is a very large number, quantisation of charge can be ignored.

Additivity of electric charges means that the total charge of a system is the algebraic sum (i.e., the sum taking into account proper signs) of all individual charges in the system.

Conservation of electric charges means that the total charge of an isolated system remains unchanged with time. This means that when bodies are charged through friction, there is a transfer of electric charge from one body to another, but no creation or destruction of charge.

5. *Coulomb's Law:* The mutual electrostatic force between two point charges q_1 and q_2 is proportional to the product $q_1 q_2$ and inversely proportional to the square of the distance r_{21} separating them. Mathematically,

$$\mathbf{F}_{21} = \text{force on } q_2 \text{ due to } q_1 = \frac{k (q_1 q_2)}{r_{21}^2} \hat{\mathbf{r}}_{21}$$

where $\hat{\mathbf{r}}_{21}$ is a unit vector in the direction from q_1 to q_2 and $k = \frac{1}{4\pi\epsilon_0}$ is the constant of proportionality.

In SI units, the unit of charge is coulomb. The experimental value of the constant ϵ_0 is

$$\epsilon_0 = 8.854 \times 10^{-12} \text{ C}^2 \text{ N}^{-1} \text{ m}^{-2}$$

The approximate value of k is

$$k = 9 \times 10^9 \text{ N m}^2 \text{ C}^{-2}$$

6. The ratio of electric force and gravitational force between a proton and an electron is

$$\frac{k e^2}{G m_e m_p} \approx 2.4 \times 10^{39}$$

7. *Superposition Principle:* The principle is based on the property that the forces with which two charges attract or repel each other are not affected by the presence of a third (or more) additional charge(s). For an assembly of charges q_1, q_2, q_3, \dots , the force on any charge, say q_1 , is

Electric Charges and Fields

the vector sum of the force on q_1 due to q_2 , the force on q_1 due to q_3 , and so on. For each pair, the force is given by the Coulomb's law for two charges stated earlier.

8. The electric field \mathbf{E} at a point due to a charge configuration is the force on a small positive test charge q placed at the point divided by the magnitude of the charge. Electric field due to a point charge q has a magnitude $|q|/4\pi\epsilon_0 r^2$; it is radially outwards from q , if q is positive, and radially inwards if q is negative. Like Coulomb force, electric field also satisfies superposition principle.
9. An electric field line is a curve drawn in such a way that the tangent at each point on the curve gives the direction of electric field at that point. The relative closeness of field lines indicates the relative strength of electric field at different points; they crowd near each other in regions of strong electric field and are far apart where the electric field is weak. In regions of constant electric field, the field lines are uniformly spaced parallel straight lines.
10. Some of the important properties of field lines are: (i) Field lines are continuous curves without any breaks. (ii) Two field lines cannot cross each other. (iii) Electrostatic field lines start at positive charges and end at negative charges —they cannot form closed loops.
11. An electric dipole is a pair of equal and opposite charges q and $-q$ separated by some distance $2a$. Its dipole moment vector \mathbf{p} has magnitude $2qa$ and is in the direction of the dipole axis from $-q$ to q .
12. Field of an electric dipole in its equatorial plane (i.e., the plane perpendicular to its axis and passing through its centre) at a distance r from the centre:

$$\mathbf{E} = \frac{-\mathbf{p}}{4\pi\epsilon_0} \frac{1}{(a^2 + r^2)^{3/2}}$$

$$\approx \frac{-\mathbf{p}}{4\pi\epsilon_0 r^3}, \quad \text{for } r \gg a$$

Dipole electric field on the axis at a distance r from the centre:

$$\mathbf{E} = \frac{2\mathbf{p}r}{4\pi\epsilon_0(r^2 - a^2)^2}$$

$$\approx \frac{2\mathbf{p}}{4\pi\epsilon_0 r^3} \quad \text{for } r \gg a$$

The $1/r^3$ dependence of dipole electric fields should be noted in contrast to the $1/r^2$ dependence of electric field due to a point charge.

13. In a uniform electric field \mathbf{E} , a dipole experiences a torque τ given by

$$\tau = \mathbf{p} \times \mathbf{E}$$

but experiences no net force.

14. The flux $\Delta\phi$ of electric field \mathbf{E} through a small area element $\Delta\mathbf{s}$ is given by

$$\Delta\phi = \mathbf{E} \cdot \Delta\mathbf{s}$$

The vector area element $\Delta\mathbf{s}$ is

$$\Delta\mathbf{s} = \Delta S \hat{\mathbf{n}}$$

where ΔS is the magnitude of the area element and $\hat{\mathbf{n}}$ is normal to the area element, which can be considered planar for sufficiently small ΔS .

Physics

For an area element of a closed surface, $\hat{\mathbf{n}}$ is taken to be the direction of *outward* normal, by convention.

15. *Gauss's law:* The flux of electric field through any closed surface S is $1/\epsilon_0$ times the total charge enclosed by S. The law is especially useful in determining electric field \mathbf{E} , when the source distribution has simple symmetry:

(i) *Thin infinitely long straight wire of uniform linear charge density λ*

$$\mathbf{E} = \frac{\lambda}{2\pi\epsilon_0 r} \hat{\mathbf{n}}$$

where r is the perpendicular distance of the point from the wire and $\hat{\mathbf{n}}$ is the radial unit vector in the plane normal to the wire passing through the point.

(ii) *Infinite thin plane sheet of uniform surface charge density σ*

$$\mathbf{E} = \frac{\sigma}{2\epsilon_0} \hat{\mathbf{n}}$$

where $\hat{\mathbf{n}}$ is a unit vector normal to the plane, outward on either side.

(iii) *Thin spherical shell of uniform surface charge density σ*

$$\mathbf{E} = \frac{q}{4\pi\epsilon_0 r^2} \hat{\mathbf{r}} \quad (r \geq R)$$

$$\mathbf{E} = 0 \quad (r < R)$$

where r is the distance of the point from the centre of the shell and R the radius of the shell. q is the total charge of the shell: $q = 4\pi R^2 \sigma$.

The electric field outside the shell is as though the total charge is concentrated at the centre. The same result is true for a solid sphere of uniform volume charge density. The field is zero at all points inside the shell

Physical quantity	Symbol	Dimensions	Unit	Remarks
Vector area element	$\Delta \mathbf{S}$	$[L^2]$	m^2	$\Delta \mathbf{S} = \Delta S \hat{\mathbf{n}}$
Electric field	\mathbf{E}	$[MLT^{-3}A^{-1}]$	$V m^{-1}$	
Electric flux	ϕ	$[ML^3 T^{-3}A^{-1}]$	$V m$	$\Delta\phi = \mathbf{E} \cdot \Delta \mathbf{S}$
Dipole moment	\mathbf{p}	$[LTA]$	$C m$	Vector directed from negative to positive charge
Charge density				
linear	λ	$[L^{-1} TA]$	$C m^{-1}$	Charge/length
surface	σ	$[L^{-2} TA]$	$C m^{-2}$	Charge/area
volume	ρ	$[L^{-3} TA]$	$C m^{-3}$	Charge/volume

POINTS TO PONDER

1. You might wonder why the protons, all carrying positive charges, are compactly residing inside the nucleus. Why do they not fly away? You will learn that there is a third kind of a fundamental force, called the strong force which holds them together. The range of distance where this force is effective is, however, very small $\sim 10^{-14}$ m. This is precisely the size of the nucleus. Also the electrons are not allowed to sit on top of the protons, i.e. inside the nucleus, due to the laws of quantum mechanics. This gives the atoms their structure as they exist in nature.
2. Coulomb force and gravitational force follow the same inverse-square law. But gravitational force has only one sign (always attractive), while Coulomb force can be of both signs (attractive and repulsive), allowing possibility of cancellation of electric forces. This is how gravity, despite being a much weaker force, can be a dominating and more pervasive force in nature.
3. The constant of proportionality k in Coulomb's law is a matter of choice if the unit of charge is to be defined using Coulomb's law. In SI units, however, what is defined is the unit of current (A) via its magnetic effect (Ampere's law) and the unit of charge (coulomb) is simply defined by ($1\text{C} = 1\text{ A s}$). In this case, the value of k is no longer arbitrary; it is approximately $9 \times 10^9 \text{ N m}^2 \text{ C}^{-2}$.
4. The rather large value of k , i.e., the large size of the unit of charge (1C) from the point of view of electric effects arises because (as mentioned in point 3 already) the unit of charge is defined in terms of magnetic forces (forces on current-carrying wires) which are generally much weaker than the electric forces. Thus while 1 ampere is a unit of reasonable size for magnetic effects, $1\text{ C} = 1\text{ A s}$, is too big a unit for electric effects.
5. The additive property of charge is not an 'obvious' property. It is related to the fact that electric charge has no direction associated with it; charge is a scalar.
6. Charge is not only a scalar (or invariant) under rotation; it is also invariant for frames of reference in relative motion. This is not always true for every scalar. For example, kinetic energy is a scalar under rotation, but is not invariant for frames of reference in relative motion.
7. Conservation of total charge of an isolated system is a property independent of the scalar nature of charge noted in point 6. Conservation refers to invariance in time in a given frame of reference. A quantity may be scalar but not conserved (like kinetic energy in an inelastic collision). On the other hand, one can have conserved vector quantity (e.g., angular momentum of an isolated system).
8. Quantisation of electric charge is a basic (unexplained) law of nature; interestingly, there is no analogous law on quantisation of mass.
9. Superposition principle should not be regarded as 'obvious', or equated with the law of addition of vectors. It says two things: force on one charge due to another charge is unaffected by the presence of other charges, and there are no additional three-body, four-body, etc., forces which arise only when there are more than two charges.
10. The electric field due to a discrete charge configuration is not defined at the locations of the discrete charges. For continuous volume charge distribution, it is defined at any point in the distribution. For a surface charge distribution, electric field is discontinuous across the surface.

11. The electric field due to a charge configuration with total charge zero is not zero; but for distances large compared to the size of the configuration, its field falls off faster than $1/r^2$, typical of field due to a single charge. An electric dipole is the simplest example of this fact.

EXERCISES

- 1.1** What is the force between two small charged spheres having charges of $2 \times 10^{-7}\text{C}$ and $3 \times 10^{-7}\text{C}$ placed 30 cm apart in air?
- 1.2** The electrostatic force on a small sphere of charge $0.4\ \mu\text{C}$ due to another small sphere of charge $-0.8\ \mu\text{C}$ in air is 0.2 N. (a) What is the distance between the two spheres? (b) What is the force on the second sphere due to the first?
- 1.3** Check that the ratio $ke^2/G m_e m_p$ is dimensionless. Look up a Table of Physical Constants and determine the value of this ratio. What does the ratio signify?
- 1.4** (a) Explain the meaning of the statement 'electric charge of a body is quantised'.
 (b) Why can one ignore quantisation of electric charge when dealing with macroscopic i.e., large scale charges?
- 1.5** When a glass rod is rubbed with a silk cloth, charges appear on both. A similar phenomenon is observed with many other pairs of bodies. Explain how this observation is consistent with the law of conservation of charge.
- 1.6** Four point charges $q_A = 2\ \mu\text{C}$, $q_B = -5\ \mu\text{C}$, $q_C = 2\ \mu\text{C}$, and $q_D = -5\ \mu\text{C}$ are located at the corners of a square ABCD of side 10 cm. What is the force on a charge of $1\ \mu\text{C}$ placed at the centre of the square?
- 1.7** (a) An electrostatic field line is a continuous curve. That is, a field line cannot have sudden breaks. Why not?
 (b) Explain why two field lines never cross each other at any point?
- 1.8** Two point charges $q_A = 3\ \mu\text{C}$ and $q_B = -3\ \mu\text{C}$ are located 20 cm apart in vacuum.
 (a) What is the electric field at the midpoint O of the line AB joining the two charges?
 (b) If a negative test charge of magnitude $1.5 \times 10^{-9}\ \text{C}$ is placed at this point, what is the force experienced by the test charge?
- 1.9** A system has two charges $q_A = 2.5 \times 10^{-7}\ \text{C}$ and $q_B = -2.5 \times 10^{-7}\ \text{C}$ located at points A: (0, 0, -15 cm) and B: (0, 0, +15 cm), respectively. What are the total charge and electric dipole moment of the system?
- 1.10** An electric dipole with dipole moment $4 \times 10^{-9}\ \text{C m}$ is aligned at 30° with the direction of a uniform electric field of magnitude $5 \times 10^4\ \text{NC}^{-1}$. Calculate the magnitude of the torque acting on the dipole.
- 1.11** A polythene piece rubbed with wool is found to have a negative charge of $3 \times 10^{-7}\ \text{C}$.
 (a) Estimate the number of electrons transferred (from which to which?)
 (b) Is there a transfer of mass from wool to polythene?
- 1.12** (a) Two insulated charged copper spheres A and B have their centres separated by a distance of 50 cm. What is the mutual force of

Electric Charges and Fields

electrostatic repulsion if the charge on each is 6.5×10^{-7} C? The radii of A and B are negligible compared to the distance of separation.

- (b) What is the force of repulsion if each sphere is charged double the above amount, and the distance between them is halved?

1.13 Suppose the spheres A and B in Exercise 1.12 have identical sizes. A third sphere of the same size but uncharged is brought in contact with the first, then brought in contact with the second, and finally removed from both. What is the new force of repulsion between A and B?

1.14 Figure 1.33 shows tracks of three charged particles in a uniform electrostatic field. Give the signs of the three charges. Which particle has the highest charge to mass ratio?

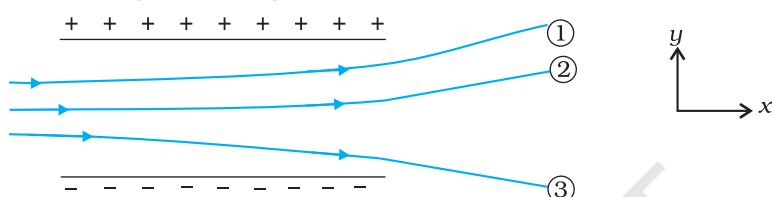


FIGURE 1.33

1.15 Consider a uniform electric field $\mathbf{E} = 3 \times 10^3 \hat{\mathbf{i}}$ N/C. (a) What is the flux of this field through a square of 10 cm on a side whose plane is parallel to the yz plane? (b) What is the flux through the same square if the normal to its plane makes a 60° angle with the x -axis?

1.16 What is the net flux of the uniform electric field of Exercise 1.15 through a cube of side 20 cm oriented so that its faces are parallel to the coordinate planes?

1.17 Careful measurement of the electric field at the surface of a black box indicates that the net outward flux through the surface of the box is 8.0×10^3 Nm²/C. (a) What is the net charge inside the box? (b) If the net outward flux through the surface of the box were zero, could you conclude that there were no charges inside the box? Why or Why not?

1.18 A point charge $+10 \mu\text{C}$ is a distance 5 cm directly above the centre of a square of side 10 cm, as shown in Fig. 1.34. What is the magnitude of the electric flux through the square? (Hint: Think of the square as one face of a cube with edge 10 cm.)

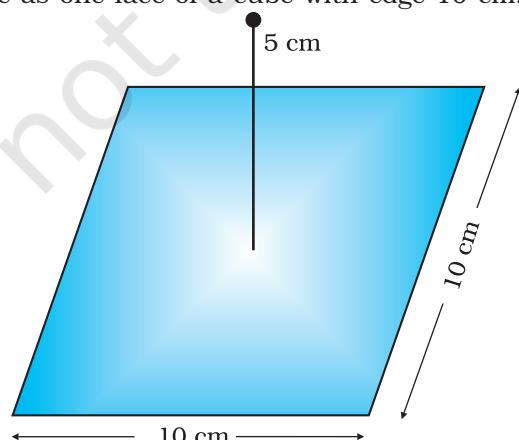


FIGURE 1.34

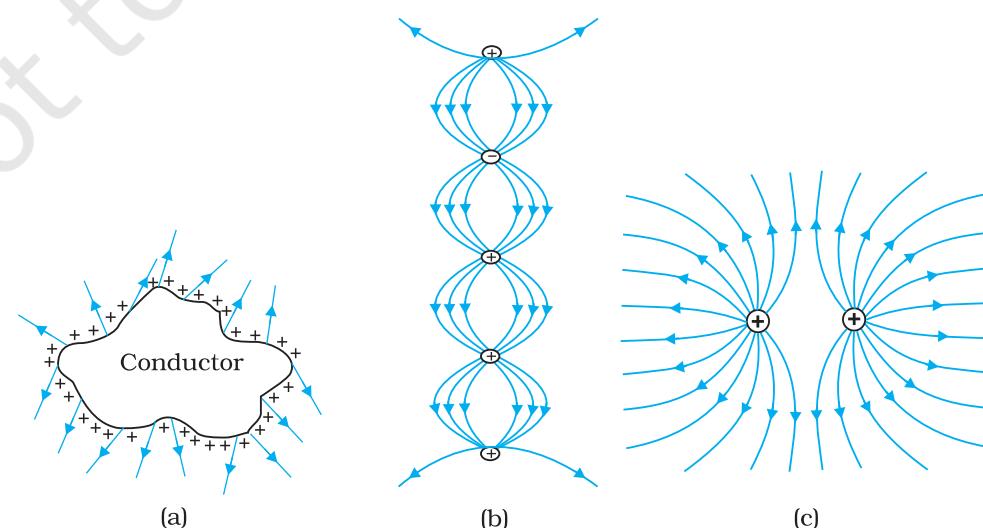


Physics

- 1.19** A point charge of $2.0 \mu\text{C}$ is at the centre of a cubic Gaussian surface 9.0 cm on edge. What is the net electric flux through the surface?
- 1.20** A point charge causes an electric flux of $-1.0 \times 10^3 \text{ Nm}^2/\text{C}$ to pass through a spherical Gaussian surface of 10.0 cm radius centred on the charge. (a) If the radius of the Gaussian surface were doubled, how much flux would pass through the surface? (b) What is the value of the point charge?
- 1.21** A conducting sphere of radius 10 cm has an unknown charge. If the electric field 20 cm from the centre of the sphere is $1.5 \times 10^3 \text{ N/C}$ and points radially inward, what is the net charge on the sphere?
- 1.22** A uniformly charged conducting sphere of 2.4 m diameter has a surface charge density of $80.0 \mu\text{C/m}^2$. (a) Find the charge on the sphere. (b) What is the total electric flux leaving the surface of the sphere?
- 1.23** An infinite line charge produces a field of $9 \times 10^4 \text{ N/C}$ at a distance of 2 cm. Calculate the linear charge density.
- 1.24** Two large, thin metal plates are parallel and close to each other. On their inner faces, the plates have surface charge densities of opposite signs and of magnitude $17.0 \times 10^{-22} \text{ C/m}^2$. What is \mathbf{E} : (a) in the outer region of the first plate, (b) in the outer region of the second plate, and (c) between the plates?

ADDITIONAL EXERCISES

- 1.25** An oil drop of 12 excess electrons is held stationary under a constant electric field of $2.55 \times 10^4 \text{ NC}^{-1}$ in Millikan's oil drop experiment. The density of the oil is 1.26 g cm^{-3} . Estimate the radius of the drop. ($g = 9.81 \text{ m s}^{-2}$; $e = 1.60 \times 10^{-19} \text{ C}$).
- 1.26** Which among the curves shown in Fig. 1.35 cannot possibly represent electrostatic field lines?



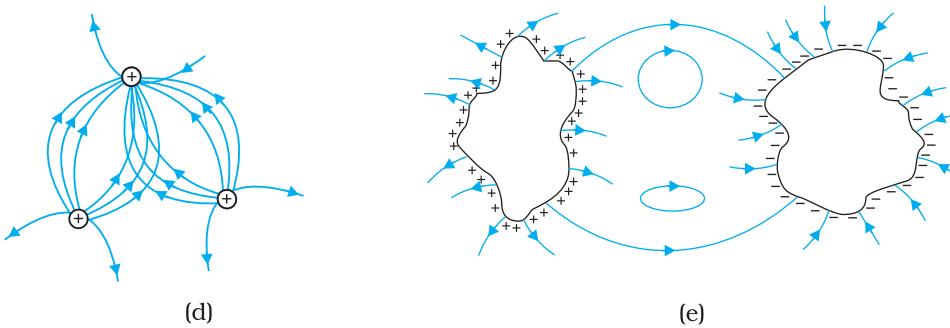


FIGURE 1.35

- 1.27** In a certain region of space, electric field is along the z-direction throughout. The magnitude of electric field is, however, not constant but increases uniformly along the positive z-direction, at the rate of 10^5 NC^{-1} per metre. What are the force and torque experienced by a system having a total dipole moment equal to 10^{-7} Cm in the negative z-direction ?
- 1.28** (a) A conductor A with a cavity as shown in Fig. 1.36(a) is given a charge Q . Show that the entire charge must appear on the outer surface of the conductor. (b) Another conductor B with charge q is inserted into the cavity keeping B insulated from A. Show that the total charge on the outside surface of A is $Q + q$ [Fig. 1.36(b)]. (c) A sensitive instrument is to be shielded from the strong electrostatic fields in its environment. Suggest a possible way.

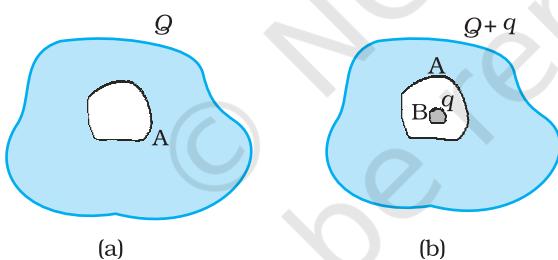


FIGURE 1.36

- 1.29** A hollow charged conductor has a tiny hole cut into its surface. Show that the electric field in the hole is $(\sigma/2\epsilon_0) \hat{\mathbf{n}}$, where $\hat{\mathbf{n}}$ is the unit vector in the outward normal direction, and σ is the surface charge density near the hole.
- 1.30** Obtain the formula for the electric field due to a long thin wire of uniform linear charge density λ without using Gauss's law. [Hint: Use Coulomb's law directly and evaluate the necessary integral.]
- 1.31** It is now believed that protons and neutrons (which constitute nuclei of ordinary matter) are themselves built out of more elementary units called quarks. A proton and a neutron consist of three quarks each. Two types of quarks, the so called 'up' quark (denoted by u) of charge $+(2/3) e$, and the 'down' quark (denoted by d) of charge $(-1/3) e$, together with electrons build up ordinary matter. (Quarks of other types have also been found which give rise to different unusual varieties of matter.) Suggest a possible quark composition of a proton and neutron.



Physics

- 1.32** (a) Consider an arbitrary electrostatic field configuration. A small test charge is placed at a null point (i.e., where $\mathbf{E} = 0$) of the configuration. Show that the equilibrium of the test charge is necessarily unstable.
(b) Verify this result for the simple configuration of two charges of the same magnitude and sign placed a certain distance apart.
- 1.33** A particle of mass m and charge $(-q)$ enters the region between the two charged plates initially moving along x -axis with speed v_x (like particle 1 in Fig. 1.33). The length of plate is L and an uniform electric field E is maintained between the plates. Show that the vertical deflection of the particle at the far edge of the plate is $qEL^2/(2m v_x^2)$.
Compare this motion with motion of a projectile in gravitational field discussed in Section 4.10 of Class XI Textbook of Physics.
- 1.34** Suppose that the particle in Exercise in 1.33 is an electron projected with velocity $v_x = 2.0 \times 10^6 \text{ m s}^{-1}$. If E between the plates separated by 0.5 cm is $9.1 \times 10^2 \text{ N/C}$, where will the electron strike the upper plate? ($|e|=1.6 \times 10^{-19} \text{ C}$, $m_e = 9.1 \times 10^{-31} \text{ kg.}$)

Chapter Two

ELECTROSTATIC POTENTIAL AND CAPACITANCE



2.1 INTRODUCTION

In Chapters 6 and 8 (Class XI), the notion of potential energy was introduced. When an external force does work in taking a body from a point to another against a force like spring force or gravitational force, that work gets stored as potential energy of the body. When the external force is removed, the body moves, gaining kinetic energy and losing an equal amount of potential energy. The sum of kinetic and potential energies is thus conserved. Forces of this kind are called conservative forces. Spring force and gravitational force are examples of conservative forces.

Coulomb force between two (stationary) charges is also a conservative force. This is not surprising, since both have inverse-square dependence on distance and differ mainly in the proportionality constants – the masses in the gravitational law are replaced by charges in Coulomb's law. Thus, like the potential energy of a mass in a gravitational field, we can define electrostatic potential energy of a charge in an electrostatic field.

Consider an electrostatic field \mathbf{E} due to some charge configuration. First, for simplicity, consider the field \mathbf{E} due to a charge Q placed at the origin. Now, imagine that we bring a test charge q from a point R to a point P against the repulsive force on it due to the charge Q . With reference

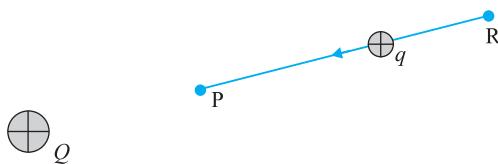


FIGURE 2.1 A test charge $q (> 0)$ is moved from the point R to the point P against the repulsive force on it by the charge $Q (> 0)$ placed at the origin.

to Fig. 2.1, this will happen if Q and q are both positive or both negative. For definiteness, let us take $Q, q > 0$.

Two remarks may be made here. First, we assume that the test charge q is so small that it does not disturb the original configuration, namely the charge Q at the origin (or else, we keep Q fixed at the origin by some unspecified force). Second, in bringing the charge q from R to P, we apply an external force \mathbf{F}_{ext} just enough to counter the repulsive electric force \mathbf{F}_E (i.e., $\mathbf{F}_{\text{ext}} = -\mathbf{F}_E$). This means there is no net force on or acceleration of the charge q when it is brought from R to P, i.e., it is brought with infinitesimally slow constant speed. In

this situation, work done by the external force is the negative of the work done by the electric force, and gets fully stored in the form of potential energy of the charge q . If the external force is removed on reaching P, the electric force will take the charge away from Q – the stored energy (potential energy) at P is used to provide kinetic energy to the charge q in such a way that the sum of the kinetic and potential energies is conserved.

Thus, work done by external forces in moving a charge q from R to P is

$$\begin{aligned} W_{RP} &= \int_R^P \mathbf{F}_{\text{ext}} \cdot d\mathbf{r} \\ &= - \int_R^P \mathbf{F}_E \cdot d\mathbf{r} \end{aligned} \quad (2.1)$$

This work done is against electrostatic repulsive force and gets stored as potential energy.

At every point in electric field, a particle with charge q possesses a certain electrostatic potential energy, this work done increases its potential energy by an amount equal to potential energy difference between points R and P.

Thus, potential energy difference

$$\Delta U = U_P - U_R = W_{RP} \quad (2.2)$$

(Note here that this displacement is in an opposite sense to the electric force and hence work done by electric field is negative, i.e., $-W_{RP}$.)

Therefore, we can define electric potential energy difference between two points as the work required to be done by an external force in moving (without accelerating) charge q from one point to another for electric field of any arbitrary charge configuration.

Two important comments may be made at this stage:

- (i) The right side of Eq. (2.2) depends only on the initial and final positions of the charge. It means that the work done by an electrostatic field in moving a charge from one point to another depends only on the initial and the final points and is independent of the path taken to go from one point to the other. This is the fundamental characteristic of a conservative force. The concept of the potential energy would not be meaningful if the work depended on the path. The path-independence of work done by an electrostatic field can be proved using the Coulomb's law. We omit this proof here.

- (ii) Equation (2.2) defines *potential energy difference* in terms of the physically meaningful quantity *work*. Clearly, potential energy so defined is undetermined to within an additive constant. What this means is that the actual value of potential energy is not physically significant; it is only the difference of potential energy that is significant. We can always add an arbitrary constant α to potential energy at every point, since this will not change the potential energy difference:

$$(U_P + \alpha) - (U_R + \alpha) = U_P - U_R$$

Put it differently, there is a freedom in choosing the point where potential energy is zero. A convenient choice is to have electrostatic potential energy zero at infinity. With this choice, if we take the point R at infinity, we get from Eq. (2.2)

$$W_{\infty P} = U_P - U_\infty = U_P \quad (2.3)$$

Since the point P is arbitrary, Eq. (2.3) provides us with a definition of potential energy of a charge q at any point. *Potential energy of charge q at a point* (in the presence of field due to any charge configuration) is the work done by the external force (equal and opposite to the electric force) in bringing the charge q from infinity to that point.

2.2 ELECTROSTATIC POTENTIAL

Consider any general static charge configuration. We define potential energy of a test charge q in terms of the work done on the charge q . This work is obviously proportional to q , since the force at any point is $q\mathbf{E}$, where \mathbf{E} is the electric field at that point due to the given charge configuration. It is, therefore, convenient to divide the work by the amount of charge q , so that the resulting quantity is independent of q . In other words, work done per unit test charge is characteristic of the electric field associated with the charge configuration. This leads to the idea of electrostatic potential V due to a given charge configuration. From Eq. (2.1), we get:

Work done by external force in bringing a unit positive charge from point R to P

$$= V_P - V_R \left(= \frac{U_P - U_R}{q} \right) \quad (2.4)$$

where V_P and V_R are the electrostatic potentials at P and R, respectively. Note, as before, that it is not the actual value of potential but the potential difference that is physically significant. If, as before, we choose the potential to be zero at infinity, Eq. (2.4) implies:

Work done by an external force in bringing a unit positive charge from infinity to a point = electrostatic potential (V) at that point.



Count Alessandro Volta

(1745 – 1827) Italian physicist, professor at Pavia. Volta established that the *animal electricity* observed by Luigi Galvani, 1737–1798, in experiments with frog muscle tissue placed in contact with dissimilar metals, was not due to any exceptional property of animal tissues but was also generated whenever any wet body was sandwiched between dissimilar metals. This led him to develop the first *voltaic pile*, or battery, consisting of a large stack of moist disks of cardboard (electrolyte) sandwiched between disks of metal (electrodes).

COUNT ALESSANDRO VOLTA (1745 – 1827)

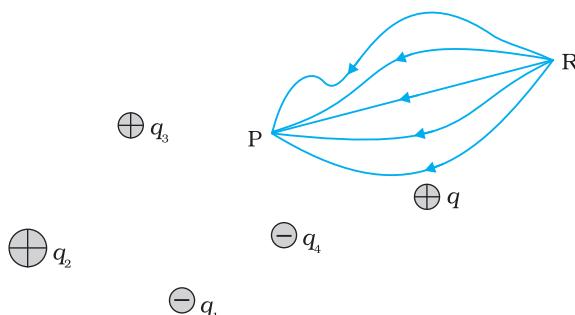


FIGURE 2.2 Work done on a test charge q by the electrostatic field due to any given charge configuration is independent of the path, and depends only on its initial and final positions.

In other words, the electrostatic potential (V) at any point in a region with electrostatic field is the work done in bringing a unit positive charge (without acceleration) from infinity to that point.

The qualifying remarks made earlier regarding potential energy also apply to the definition of potential. To obtain the work done per unit test charge, we should take an infinitesimal test charge δq , obtain the work done δW in bringing it from infinity to the point and determine the ratio $\delta W/\delta q$. Also, the external force at every point of the path is to be equal and opposite to the electrostatic force on the test charge at that point.

2.3 POTENTIAL DUE TO A POINT CHARGE

Consider a point charge Q at the origin (Fig. 2.3). For definiteness, take Q to be positive. We wish to determine the potential at any point P with position vector \mathbf{r} from the origin. For that we must calculate the work done in bringing a unit positive test charge from infinity to the point P . For $Q > 0$, the work done against the repulsive force on the test charge is positive. Since work done is independent of the path, we choose a convenient path – along the radial direction from infinity to the point P .

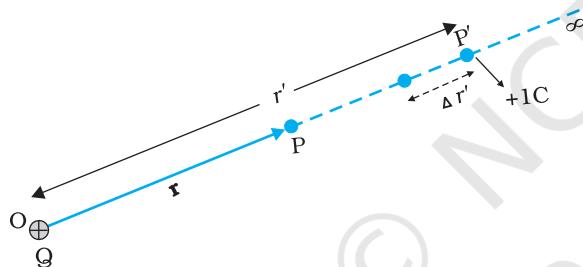


FIGURE 2.3 Work done in bringing a unit positive test charge from infinity to the point P , against the repulsive force of charge Q ($Q > 0$), is the potential at P due to the charge Q .

At some intermediate point P' on the path, the electrostatic force on a unit positive charge is

$$\frac{Q \times 1}{4\pi\epsilon_0 r'^2} \hat{\mathbf{r}}' \quad (2.5)$$

where $\hat{\mathbf{r}}'$ is the unit vector along OP' . Work done against this force from \mathbf{r}' to $\mathbf{r}' + \Delta\mathbf{r}'$ is

$$\Delta W = -\frac{Q}{4\pi\epsilon_0 r'^2} \Delta r' \quad (2.6)$$

The negative sign appears because for $\Delta r' < 0$, ΔW is positive. Total work done (W) by the external force is obtained by integrating Eq. (2.6) from $r' = \infty$ to $r' = r$,

$$W = -\int_{\infty}^r \frac{Q}{4\pi\epsilon_0 r'^2} dr' = \frac{Q}{4\pi\epsilon_0 r'} \Big|_{\infty}^r = \frac{Q}{4\pi\epsilon_0 r} \quad (2.7)$$

This, by definition is the potential at P due to the charge Q

$$V(r) = \frac{Q}{4\pi\epsilon_0 r} \quad (2.8)$$

Equation (2.8) is true for any sign of the charge Q , though we considered $Q > 0$ in its derivation. For $Q < 0$, $V < 0$, i.e., work done (by the external force) per unit positive test charge in bringing it from infinity to the point is negative. This is equivalent to saying that work done by the electrostatic force in bringing the unit positive charge from infinity to the point P is positive. [This is as it should be, since for $Q < 0$, the force on a unit positive test charge is attractive, so that the electrostatic force and the displacement (from infinity to P) are in the same direction.] Finally, we note that Eq. (2.8) is consistent with the choice that potential at infinity be zero.

Figure (2.4) shows how the electrostatic potential ($\propto 1/r$) and the electrostatic field ($\propto 1/r^2$) varies with r .

Example 2.1

- Calculate the potential at a point P due to a charge of 4×10^{-7} C located 9 cm away.
- Hence obtain the work done in bringing a charge of 2×10^{-9} C from infinity to the point P. Does the answer depend on the path along which the charge is brought?

Solution

$$(a) V = \frac{1}{4\pi\epsilon_0} \frac{Q}{r} = 9 \times 10^9 \text{ Nm}^2 \text{ C}^{-2} \times \frac{4 \times 10^{-7} \text{ C}}{0.09 \text{ m}} \\ = 4 \times 10^4 \text{ V}$$

$$(b) W = qV = 2 \times 10^{-9} \text{ C} \times 4 \times 10^4 \text{ V} \\ = 8 \times 10^{-5} \text{ J}$$

No, work done will be path independent. Any arbitrary infinitesimal path can be resolved into two perpendicular displacements: One along \mathbf{r} and another perpendicular to \mathbf{r} . The work done corresponding to the later will be zero.

EXAMPLE 2.1

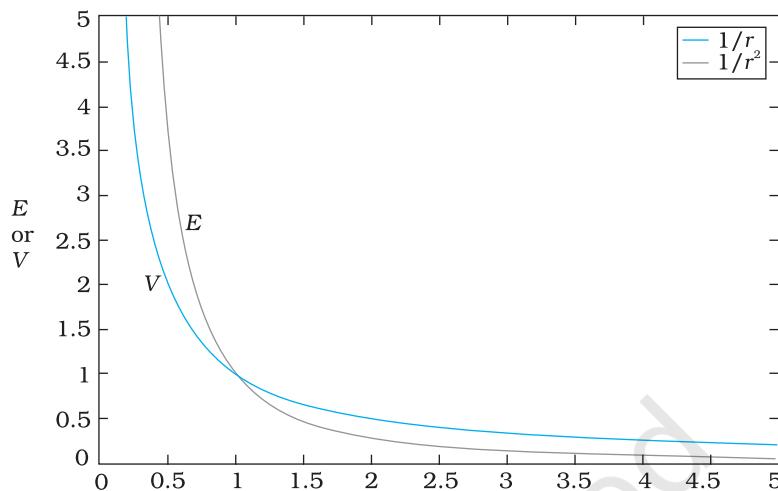


FIGURE 2.4 Variation of potential V with r [in units of $(Q/4\pi\epsilon_0) \text{ m}^{-1}$] (blue curve) and field with r [in units of $(Q/4\pi\epsilon_0) \text{ m}^{-2}$] (black curve) for a point charge Q .

2.4 POTENTIAL DUE TO AN ELECTRIC DIPOLE

As we learnt in the last chapter, an electric dipole consists of two charges q and $-q$ separated by a (small) distance $2a$. Its total charge is zero. It is characterised by a dipole moment vector \mathbf{p} whose magnitude is $q \times 2a$ and which points in the direction from $-q$ to q (Fig. 2.5). We also saw that the electric field of a dipole at a point with position vector \mathbf{r} depends not just on the magnitude r , but also on the angle between \mathbf{r} and \mathbf{p} . Further,

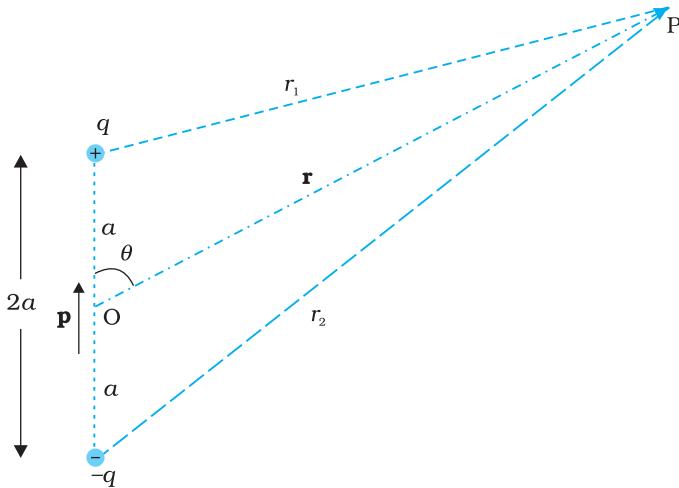


FIGURE 2.5 Quantities involved in the calculation of potential due to a dipole.

Now, by geometry,

$$r_1^2 = r^2 + a^2 - 2ar \cos\theta$$

$$r_2^2 = r^2 + a^2 + 2ar \cos\theta \quad (2.10)$$

We take r much greater than a ($r \gg a$) and retain terms only upto the first order in a/r

$$\approx r^2 \left(1 - \frac{2a \cos \theta}{r} \right) \quad (2.11)$$

Similarly,

$$r_2^2 \approx r^2 \left(1 + \frac{2a \cos \theta}{r} \right) \quad (2.12)$$

Using the Binomial theorem and retaining terms upto the first order in a/r ; we obtain,

$$\frac{1}{r_1} \approx \frac{1}{r} \left(1 - \frac{2a \cos \theta}{r} \right)^{-1/2} \approx \frac{1}{r} \left(1 + \frac{a}{r} \cos \theta \right) \quad [2.13(a)]$$

$$\frac{1}{r_2} \approx \frac{1}{r} \left(1 + \frac{2a \cos \theta}{r} \right)^{-1/2} \approx \frac{1}{r} \left(1 - \frac{a}{r} \cos \theta \right) \quad [2.13(b)]$$

Using Eqs. (2.9) and (2.13) and $p = 2qa$, we get

$$V = \frac{q}{4\pi\epsilon_0} \frac{2a \cos \theta}{r^2} = \frac{p \cos \theta}{4\pi\epsilon_0 r^2} \quad (2.14)$$

Now, $p \cos \theta = \mathbf{p} \cdot \hat{\mathbf{r}}$

the field falls off, at large distance, not as $1/r^2$ (typical of field due to a single charge) but as $1/r^3$. We, now, determine the electric potential due to a dipole and contrast it with the potential due to a single charge.

As before, we take the origin at the centre of the dipole. Now we know that the electric field obeys the superposition principle. Since potential is related to the work done by the field, electrostatic potential also follows the superposition principle. Thus, the potential due to the dipole is the sum of potentials due to the charges q and $-q$

$$V = \frac{1}{4\pi\epsilon_0} \left(\frac{q}{r_1} - \frac{q}{r_2} \right) \quad (2.9)$$

where r_1 and r_2 are the distances of the point P from q and $-q$, respectively.

where $\hat{\mathbf{r}}$ is the unit vector along the position vector \mathbf{OP} .

The electric potential of a dipole is then given by

$$V = \frac{1}{4\pi\epsilon_0} \frac{\mathbf{p} \cdot \hat{\mathbf{r}}}{r^2}; \quad (r \gg a) \quad (2.15)$$

Equation (2.15) is, as indicated, approximately true only for distances large compared to the size of the dipole, so that higher order terms in a/r are negligible. For a point dipole \mathbf{p} at the origin, Eq. (2.15) is, however, exact.

From Eq. (2.15), potential on the dipole axis ($\theta = 0, \pi$) is given by

$$V = \pm \frac{1}{4\pi\epsilon_0} \frac{p}{r^2} \quad (2.16)$$

(Positive sign for $\theta = 0$, negative sign for $\theta = \pi$.) The potential in the equatorial plane ($\theta = \pi/2$) is zero.

The important contrasting features of electric potential of a dipole from that due to a single charge are clear from Eqs. (2.8) and (2.15):

- (i) The potential due to a dipole depends not just on r but also on the angle between the position vector \mathbf{r} and the dipole moment vector \mathbf{p} . (It is, however, axially symmetric about \mathbf{p} . That is, if you rotate the position vector \mathbf{r} about \mathbf{p} , keeping θ fixed, the points corresponding to P on the cone so generated will have the same potential as at P.)
- (ii) The electric dipole potential falls off, at large distance, as $1/r^2$, not as $1/r$, characteristic of the potential due to a single charge. (You can refer to the Fig. 2.5 for graphs of $1/r^2$ versus r and $1/r$ versus r , drawn there in another context.)

2.5 POTENTIAL DUE TO A SYSTEM OF CHARGES

Consider a system of charges q_1, q_2, \dots, q_n with position vectors $\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n$ relative to some origin (Fig. 2.6). The potential V_1 at P due to the charge q_1 is

$$V_1 = \frac{1}{4\pi\epsilon_0} \frac{q_1}{r_{1P}}$$

where r_{1P} is the distance between q_1 and P.

Similarly, the potential V_2 at P due to q_2 and V_3 due to q_3 are given by

$$V_2 = \frac{1}{4\pi\epsilon_0} \frac{q_2}{r_{2P}}, \quad V_3 = \frac{1}{4\pi\epsilon_0} \frac{q_3}{r_{3P}}$$

where r_{2P} and r_{3P} are the distances of P from charges q_2 and q_3 , respectively; and so on for the potential due to other charges. By the superposition principle, the potential V at P due to the total charge configuration is the algebraic sum of the potentials due to the individual charges

$$V = V_1 + V_2 + \dots + V_n \quad (2.17)$$

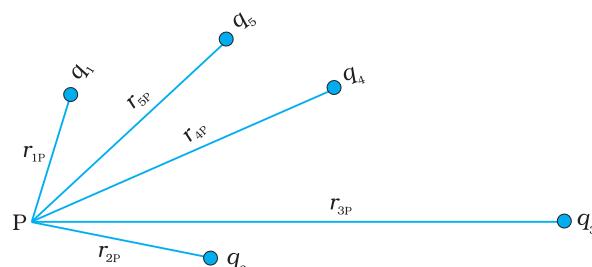


FIGURE 2.6 Potential at a point due to a system of charges is the sum of potentials due to individual charges.

$$= \frac{1}{4\pi\epsilon_0} \left(\frac{q_1}{r_{1P}} + \frac{q_2}{r_{2P}} + \dots + \frac{q_n}{r_{nP}} \right) \quad (2.18)$$

If we have a continuous charge distribution characterised by a charge density $\rho(\mathbf{r})$, we divide it, as before, into small volume elements each of size Δv and carrying a charge $\rho\Delta v$. We then determine the potential due to each volume element and sum (strictly speaking, integrate) over all such contributions, and thus determine the potential due to the entire distribution.

We have seen in Chapter 1 that for a uniformly charged spherical shell, the electric field outside the shell is as if the entire charge is concentrated at the centre. Thus, the potential outside the shell is given by

$$V = \frac{1}{4\pi\epsilon_0} \frac{q}{r} \quad (r \geq R) \quad [2.19(a)]$$

where q is the total charge on the shell and R its radius. The electric field inside the shell is zero. This implies (Section 2.6) that potential is constant inside the shell (as no work is done in moving a charge inside the shell), and, therefore, equals its value at the surface, which is

$$V = \frac{1}{4\pi\epsilon_0} \frac{q}{R} \quad [2.19(b)]$$

Example 2.2 Two charges 3×10^{-8} C and -2×10^{-8} C are located 15 cm apart. At what point on the line joining the two charges is the electric potential zero? Take the potential at infinity to be zero.

Solution Let us take the origin O at the location of the positive charge. The line joining the two charges is taken to be the x -axis; the negative charge is taken to be on the right side of the origin (Fig. 2.7).



FIGURE 2.7

Let P be the required point on the x -axis where the potential is zero. If x is the x -coordinate of P, obviously x must be positive. (There is no possibility of potentials due to the two charges adding up to zero for $x < 0$.) If x lies between O and A, we have

$$\frac{1}{4\pi\epsilon_0} \left[\frac{3 \times 10^{-8}}{x \times 10^{-2}} - \frac{2 \times 10^{-8}}{(15-x) \times 10^{-2}} \right] = 0$$

where x is in cm. That is,

$$\frac{3}{x} - \frac{2}{15-x} = 0$$

which gives $x = 9$ cm.

If x lies on the extended line OA, the required condition is

$$\frac{3}{x} - \frac{2}{x-15} = 0$$

which gives

$$x = 45 \text{ cm}$$

Thus, electric potential is zero at 9 cm and 45 cm away from the positive charge on the side of the negative charge. Note that the formula for potential used in the calculation required choosing potential to be zero at infinity.

Example 2.3 Figures 2.8 (a) and (b) show the field lines of a positive and negative point charge respectively.

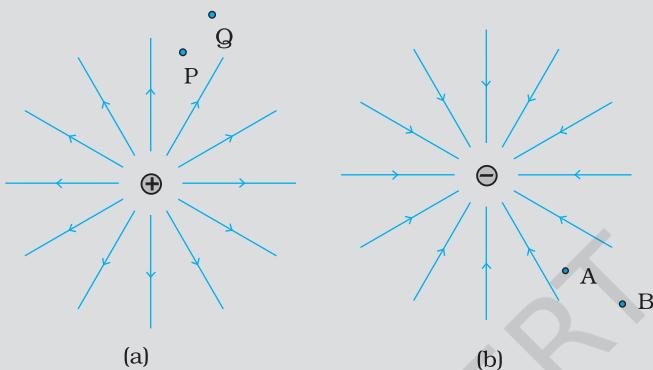


FIGURE 2.8

- (a) Give the signs of the potential difference $V_p - V_q$; $V_b - V_a$.
- (b) Give the sign of the potential energy difference of a small negative charge between the points Q and P; A and B.
- (c) Give the sign of the work done by the field in moving a small positive charge from Q to P.
- (d) Give the sign of the work done by the external agency in moving a small negative charge from B to A.
- (e) Does the kinetic energy of a small negative charge increase or decrease in going from B to A?

Solution

- (a) As $V \propto \frac{1}{r}$, $V_p > V_q$. Thus, $(V_p - V_q)$ is positive. Also V_b is less negative than V_a . Thus, $V_b > V_a$ or $(V_b - V_a)$ is positive.
- (b) A small negative charge will be attracted towards positive charge. The negative charge moves from higher potential energy to lower potential energy. Therefore the sign of potential energy difference of a small negative charge between Q and P is positive. Similarly, $(P.E.)_a > (P.E.)_b$ and hence sign of potential energy differences is positive.
- (c) In moving a small positive charge from Q to P, work has to be done by an external agency against the electric field. Therefore, work done by the field is negative.
- (d) In moving a small negative charge from B to A work has to be done by the external agency. It is positive.
- (e) Due to force of repulsion on the negative charge, velocity decreases and hence the kinetic energy decreases in going from B to A.

EXAMPLE 2.2



Electric potential, equipotential surfaces:
<http://video.mit.edu/watch/4-electrostatic-potential-electric-energy-ev-conservative-field-equipotential-surfaces-12584/>

EXAMPLE 2.3

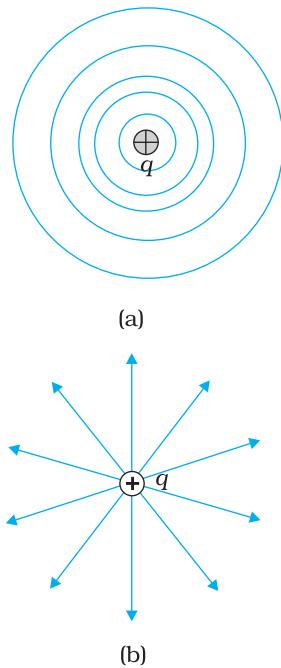


FIGURE 2.9 For a single charge q (a) equipotential surfaces are spherical surfaces centred at the charge, and (b) electric field lines are radial, starting from the charge if $q > 0$.

2.6 EQUIPOTENTIAL SURFACES

An equipotential surface is a surface with a constant value of potential at all points on the surface. For a single charge q , the potential is given by Eq. (2.8):

$$V = \frac{1}{4\pi\epsilon_0} \frac{q}{r}$$

This shows that V is a constant if r is constant. Thus, equipotential surfaces of a single point charge are concentric spherical surfaces centred at the charge.

Now the electric field lines for a single charge q are radial lines starting from or ending at the charge, depending on whether q is positive or negative. Clearly, the electric field at every point is normal to the equipotential surface passing through that point. This is true in general: *for any charge configuration, equipotential surface through a point is normal to the electric field at that point.* The proof of this statement is simple.

If the field were not normal to the equipotential surface, it would have non-zero component along the surface. To move a unit test charge against the direction of the component of the field, work would have to be done. But this is in contradiction to the definition of an equipotential surface: there is no potential difference between any two points on the surface and no work is required to move a test charge on the surface. The electric field must, therefore, be normal to the equipotential surface at every point. Equipotential surfaces offer an alternative visual picture in addition to the picture of electric field lines around a charge configuration.

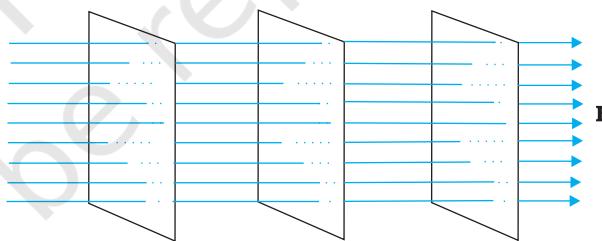


FIGURE 2.10 Equipotential surfaces for a uniform electric field.

For a uniform electric field \mathbf{E} , say, along the x -axis, the equipotential surfaces are planes normal to the x -axis, i.e., planes parallel to the y - z plane (Fig. 2.10). Equipotential surfaces for (a) a dipole and (b) two identical positive charges are shown in Fig. 2.11.

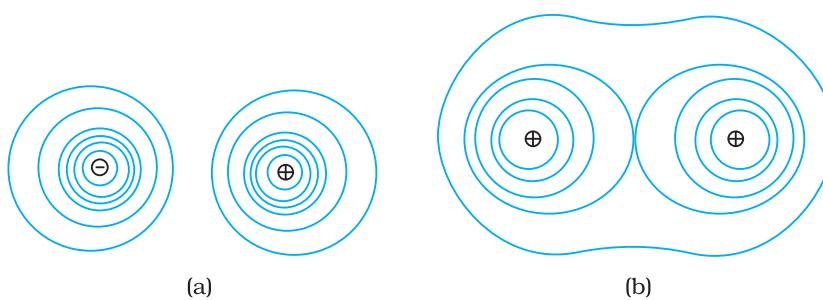


FIGURE 2.11 Some equipotential surfaces for (a) a dipole, (b) two identical positive charges.

2.6.1 Relation between field and potential

Consider two closely spaced equipotential surfaces A and B (Fig. 2.12) with potential values V and $V + \delta V$, where δV is the change in V in the direction of the electric field \mathbf{E} . Let P be a point on the surface B. δl is the perpendicular distance of the surface A from P. Imagine that a unit positive charge is moved along this perpendicular from the surface B to surface A against the electric field. The work done in this process is $|\mathbf{E}| \delta l$.

This work equals the potential difference $V_A - V_B$.

Thus,

$$|\mathbf{E}| \delta l = V - (V + \delta V) = -\delta V$$

$$\text{i.e., } |\mathbf{E}| = -\frac{\delta V}{\delta l} \quad (2.20)$$

Since δV is negative, $\delta V = -|\delta V|$. we can rewrite Eq (2.20) as

$$|\mathbf{E}| = -\frac{\delta V}{\delta l} = +\frac{|\delta V|}{\delta l} \quad (2.21)$$

We thus arrive at two important conclusions concerning the relation between electric field and potential:

- (i) *Electric field is in the direction in which the potential decreases steepest.*
- (ii) *Its magnitude is given by the change in the magnitude of potential per unit displacement normal to the equipotential surface at the point.*

2.7 POTENTIAL ENERGY OF A SYSTEM OF CHARGES

Consider first the simple case of two charges q_1 and q_2 with position vector \mathbf{r}_1 and \mathbf{r}_2 relative to some origin. Let us calculate the work done (externally) in building up this configuration. This means that we consider the charges q_1 and q_2 initially at infinity and determine the work done by an external agency to bring the charges to the given locations. Suppose, first the charge q_1 is brought from infinity to the point \mathbf{r}_1 . There is no external field against which work needs to be done, so work done in bringing q_1 from infinity to \mathbf{r}_1 is zero. This charge produces a potential in space given by

$$V_1 = \frac{1}{4\pi\epsilon_0} \frac{q_1}{r_{1P}}$$

where r_{1P} is the distance of a point P in space from the location of q_1 . From the definition of potential, work done in bringing charge q_2 from infinity to the point \mathbf{r}_2 is q_2 times the potential at \mathbf{r}_2 due to q_1 :

$$\text{work done on } q_2 = \frac{1}{4\pi\epsilon_0} \frac{q_1 q_2}{r_{12}}$$

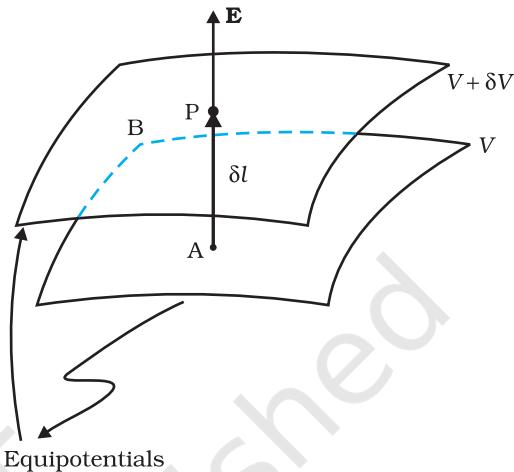


FIGURE 2.12 From the potential to the field.

Physics

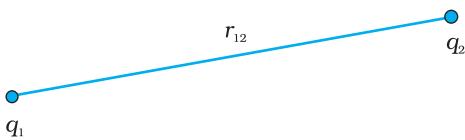


FIGURE 2.13 Potential energy of a system of charges q_1 and q_2 is directly proportional to the product of charges and inversely to the distance between them.

where r_{12} is the distance between points 1 and 2.

Since electrostatic force is conservative, this work gets stored in the form of potential energy of the system. Thus, the potential energy of a system of two charges q_1 and q_2 is

$$U = \frac{1}{4\pi\epsilon_0} \frac{q_1 q_2}{r_{12}} \quad (2.22)$$

Obviously, if q_2 was brought first to its present location and q_1 brought later, the potential energy U would be the same.

More generally, the potential energy expression, Eq. (2.22), is unaltered whatever way the charges are brought to the specified locations, because of path-independence of work for electrostatic force.

Equation (2.22) is true for any sign of q_1 and q_2 . If $q_1 q_2 > 0$, potential energy is positive. This is as expected, since for like charges ($q_1 q_2 > 0$), electrostatic force is repulsive and a positive amount of work is needed to be done against this force to bring the charges from infinity to a finite distance apart. For unlike charges ($q_1 q_2 < 0$), the electrostatic force is attractive. In that case, a positive amount of work is needed against this force to take the charges from the given location to infinity. In other words, a negative amount of work is needed for the reverse path (from infinity to the present locations), so the potential energy is negative.

Equation (2.22) is easily generalised for a system of any number of point charges. Let us calculate the potential energy of a system of three charges q_1 , q_2 and q_3 located at \mathbf{r}_1 , \mathbf{r}_2 , \mathbf{r}_3 , respectively. To bring q_1 first from infinity to \mathbf{r}_1 , no work is required. Next we bring q_2 from infinity to \mathbf{r}_2 . As before, work done in this step is

$$q_2 V_1(\mathbf{r}_2) = \frac{1}{4\pi\epsilon_0} \frac{q_1 q_2}{r_{12}} \quad (2.23)$$

The charges q_1 and q_2 produce a potential, which at any point P is given by

$$V_{1,2} = \frac{1}{4\pi\epsilon_0} \left(\frac{q_1}{r_{1P}} + \frac{q_2}{r_{2P}} \right) \quad (2.24)$$

Work done next in bringing q_3 from infinity to the point \mathbf{r}_3 is q_3 times

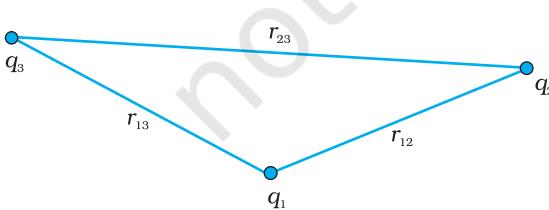


FIGURE 2.14 Potential energy of a system of three charges is given by Eq. (2.26), with the notation given in the figure.

$$q_3 V_{1,2}(\mathbf{r}_3) = \frac{1}{4\pi\epsilon_0} \left(\frac{q_1 q_3}{r_{13}} + \frac{q_2 q_3}{r_{23}} \right) \quad (2.25)$$

The total work done in assembling the charges at the given locations is obtained by adding the work done in different steps [Eq. (2.23) and Eq. (2.25)],

$$U = \frac{1}{4\pi\epsilon_0} \left(\frac{q_1 q_2}{r_{12}} + \frac{q_1 q_3}{r_{13}} + \frac{q_2 q_3}{r_{23}} \right) \quad (2.26)$$

Again, because of the conservative nature of the electrostatic force (or equivalently, the path independence of work done), the final expression for U , Eq. (2.26), is independent of the manner in which the configuration is assembled. *The potential energy*

is characteristic of the present state of configuration, and not the way the state is achieved.

Example 2.4 Four charges are arranged at the corners of a square ABCD of side d , as shown in Fig. 2.15.(a) Find the work required to put together this arrangement. (b) A charge q_0 is brought to the centre E of the square, the four charges being held fixed at its corners. How much extra work is needed to do this?

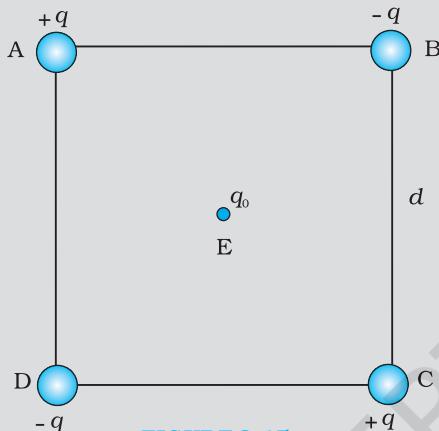


FIGURE 2.15

Solution

(a) Since the work done depends on the final arrangement of the charges, and not on how they are put together, we calculate work needed for one way of putting the charges at A, B, C and D. Suppose, first the charge $+q$ is brought to A, and then the charges $-q$, $+q$, and $-q$ are brought to B, C and D, respectively. The total work needed can be calculated in steps:

- Work needed to bring charge $+q$ to A when no charge is present elsewhere: this is zero.
- Work needed to bring $-q$ to B when $+q$ is at A. This is given by (charge at B) \times (electrostatic potential at B due to charge $+q$ at A)

$$= -q \times \left(\frac{q}{4\pi\epsilon_0 d} \right) = -\frac{q^2}{4\pi\epsilon_0 d}$$

- Work needed to bring charge $+q$ to C when $+q$ is at A and $-q$ is at B. This is given by (charge at C) \times (potential at C due to charges at A and B)

$$\begin{aligned} &= +q \left(\frac{+q}{4\pi\epsilon_0 d\sqrt{2}} + \frac{-q}{4\pi\epsilon_0 d} \right) \\ &= \frac{-q^2}{4\pi\epsilon_0 d} \left(1 - \frac{1}{\sqrt{2}} \right) \end{aligned}$$

- Work needed to bring $-q$ to D when $+q$ at A, $-q$ at B, and $+q$ at C. This is given by (charge at D) \times (potential at D due to charges at A, B and C)

$$\begin{aligned} &= -q \left(\frac{+q}{4\pi\epsilon_0 d} + \frac{-q}{4\pi\epsilon_0 d\sqrt{2}} + \frac{q}{4\pi\epsilon_0 d} \right) \\ &= \frac{-q^2}{4\pi\epsilon_0 d} \left(2 - \frac{1}{\sqrt{2}} \right) \end{aligned}$$

EXAMPLE 2.4

Add the work done in steps (i), (ii), (iii) and (iv). The total work required is

$$\begin{aligned} &= \frac{-q^2}{4\pi\epsilon_0 d} \left\{ (0) + (1) + \left(1 - \frac{1}{\sqrt{2}} \right) + \left(2 - \frac{1}{\sqrt{2}} \right) \right\} \\ &= \frac{-q^2}{4\pi\epsilon_0 d} (4 - \sqrt{2}) \end{aligned}$$

The work done depends only on the arrangement of the charges, and not how they are assembled. By definition, this is the total electrostatic energy of the charges.

(Students may try calculating same work/energy by taking charges in any other order they desire and convince themselves that the energy will remain the same.)

(b) The extra work necessary to bring a charge q_0 to the point E when the four charges are at A, B, C and D is $q_0 \times$ (electrostatic potential at E due to the charges at A, B, C and D). The electrostatic potential at E is clearly zero since potential due to A and C is cancelled by that due to B and D. Hence no work is required to bring any charge to point E.

2.8 POTENTIAL ENERGY IN AN EXTERNAL FIELD

2.8.1 Potential energy of a single charge

In Section 2.7, the source of the electric field was specified – the charges and their locations - and the potential energy of the system of those charges was determined. In this section, we ask a related but a distinct question. What is the potential energy of a charge q in a given field? This question was, in fact, the starting point that led us to the notion of the electrostatic potential (Sections 2.1 and 2.2). But here we address this question again to clarify in what way it is different from the discussion in Section 2.7.

The main difference is that we are now concerned with the potential energy of a charge (or charges) in an *external field*. The external field **E** is *not* produced by the given charge(s) whose potential energy we wish to calculate. **E** is produced by sources external to the given charge(s). The external sources may be known, but often they are unknown or unspecified; what is specified is the electric field **E** or the electrostatic potential V due to the external sources. We assume that the charge q does not significantly affect the sources producing the external field. This is true if q is very small, or the external sources are held fixed by other unspecified forces. Even if q is finite, its influence on the external sources may still be ignored in the situation when very strong sources far away at infinity produce a finite field **E** in the region of interest. Note again that we are interested in determining the potential energy of a given charge q (and later, a system of charges) in the external field; we are not interested in the potential energy of the sources producing the external electric field.

The external electric field **E** and the corresponding external potential V may vary from point to point. By definition, V at a point P is the work done in bringing a unit positive charge from infinity to the point P.

(We continue to take potential at infinity to be zero.) Thus, work done in bringing a charge q from infinity to the point P in the external field is qV . This work is stored in the form of potential energy of q . If the point P has position vector \mathbf{r} relative to some origin, we can write:

Potential energy of q at \mathbf{r} in an external field

$$= qV(\mathbf{r}) \quad (2.27)$$

where $V(\mathbf{r})$ is the external potential at the point \mathbf{r} .

Thus, if an electron with charge $q = e = 1.6 \times 10^{-19} \text{ C}$ is accelerated by a potential difference of $\Delta V = 1 \text{ volt}$, it would gain energy of $q\Delta V = 1.6 \times 10^{-19} \text{ J}$. This unit of energy is defined as 1 *electron volt* or 1eV, i.e., $1 \text{ eV} = 1.6 \times 10^{-19} \text{ J}$. The units based on eV are most commonly used in atomic, nuclear and particle physics, ($1 \text{ keV} = 10^3 \text{ eV} = 1.6 \times 10^{-16} \text{ J}$, $1 \text{ MeV} = 10^6 \text{ eV} = 1.6 \times 10^{-13} \text{ J}$, $1 \text{ GeV} = 10^9 \text{ eV} = 1.6 \times 10^{-10} \text{ J}$ and $1 \text{ TeV} = 10^{12} \text{ eV} = 1.6 \times 10^{-7} \text{ J}$). [This has already been defined on Page 117, XI Physics Part I, Table 6.1.]

2.8.2 Potential energy of a system of two charges in an external field

Next, we ask: what is the potential energy of a system of two charges q_1 and q_2 located at \mathbf{r}_1 and \mathbf{r}_2 , respectively, in an external field? First, we calculate the work done in bringing the charge q_1 from infinity to \mathbf{r}_1 . Work done in this step is $q_1 V(\mathbf{r}_1)$, using Eq. (2.27). Next, we consider the work done in bringing q_2 to \mathbf{r}_2 . In this step, work is done not only against the external field \mathbf{E} but also against the field due to q_1 .

Work done on q_2 against the external field

$$= q_2 V(\mathbf{r}_2)$$

Work done on q_2 against the field due to q_1

$$= \frac{q_1 q_2}{4\pi\epsilon_0 r_{12}}$$

where r_{12} is the distance between q_1 and q_2 . We have made use of Eqs. (2.27) and (2.22). By the superposition principle for fields, we add up the work done on q_2 against the two fields (\mathbf{E} and that due to q_1):

Work done in bringing q_2 to \mathbf{r}_2

$$= q_2 V(\mathbf{r}_2) + \frac{q_1 q_2}{4\pi\epsilon_0 r_{12}} \quad (2.28)$$

Thus,

Potential energy of the system

= the total work done in assembling the configuration

$$= q_1 V(\mathbf{r}_1) + q_2 V(\mathbf{r}_2) + \frac{q_1 q_2}{4\pi\epsilon_0 r_{12}} \quad (2.29)$$

Example 2.5

- (a) Determine the electrostatic potential energy of a system consisting of two charges $7 \mu\text{C}$ and $-2 \mu\text{C}$ (and with no external field) placed at $(-9 \text{ cm}, 0, 0)$ and $(9 \text{ cm}, 0, 0)$ respectively.
- (b) How much work is required to separate the two charges infinitely away from each other?

EXAMPLE 2.5

EXAMPLE 2.5

- (c) Suppose that the same system of charges is now placed in an external electric field $E = A(1/r^2)$; $A = 9 \times 10^5 \text{ C m}^{-2}$. What would the electrostatic energy of the configuration be?

Solution

$$(a) U = \frac{1}{4\pi\epsilon_0} \frac{q_1 q_2}{r} = 9 \times 10^9 \times \frac{7 \times (-2) \times 10^{-12}}{0.18} = -0.7 \text{ J.}$$

$$(b) W = U_2 - U_1 = 0 - U = 0 - (-0.7) = 0.7 \text{ J.}$$

(c) The mutual interaction energy of the two charges remains unchanged. In addition, there is the energy of interaction of the two charges with the external electric field. We find,

$$q_1 V(\mathbf{r}_1) + q_2 V(\mathbf{r}_2) = A \frac{7 \mu\text{C}}{0.09\text{m}} + A \frac{-2 \mu\text{C}}{0.09\text{m}}$$

and the net electrostatic energy is

$$\begin{aligned} q_1 V(\mathbf{r}_1) + q_2 V(\mathbf{r}_2) + \frac{q_1 q_2}{4\pi\epsilon_0 r_{12}} &= A \frac{7 \mu\text{C}}{0.09\text{m}} + A \frac{-2 \mu\text{C}}{0.09\text{m}} - 0.7 \text{ J} \\ &= 70 - 20 - 0.7 = 49.3 \text{ J} \end{aligned}$$

2.8.3 Potential energy of a dipole in an external field

Consider a dipole with charges $q_1 = +q$ and $q_2 = -q$ placed in a uniform electric field \mathbf{E} , as shown in Fig. 2.16.

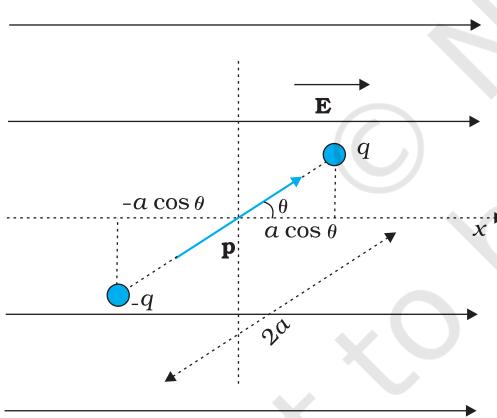


FIGURE 2.16 Potential energy of a dipole in a uniform external field.

As seen in the last chapter, in a uniform electric field, the dipole experiences no net force; but experiences a torque τ given by

$$\tau = \mathbf{p} \times \mathbf{E} \quad (2.30)$$

which will tend to rotate it (unless \mathbf{p} is parallel or antiparallel to \mathbf{E}). Suppose an external torque τ_{ext} is applied in such a manner that it just neutralises this torque and rotates it in the plane of paper from angle θ_0 to angle θ_1 at an infinitesimal angular speed and *without angular acceleration*. The amount of work done by the external torque will be given by

$$\begin{aligned} W &= \int_{\theta_0}^{\theta_1} \tau_{\text{ext}}(\theta) d\theta = \int_{\theta_0}^{\theta_1} pE \sin \theta d\theta \\ &= pE (\cos \theta_0 - \cos \theta_1) \end{aligned} \quad (2.31)$$

This work is stored as the potential energy of the system. We can then associate potential energy $U(\theta)$ with an inclination θ of the dipole. Similar to other potential energies, there is a freedom in choosing the angle where the potential energy U is taken to be zero. A natural choice is to take $\theta_0 = \pi/2$. (An explanation for it is provided towards the end of discussion.) We can then write,

$$U(\theta) = pE \left(\cos \frac{\pi}{2} - \cos \theta \right) = -pE \cos \theta = -\mathbf{p} \cdot \mathbf{E} \quad (2.32)$$

Electrostatic Potential and Capacitance

This expression can alternately be understood also from Eq. (2.29). We apply Eq. (2.29) to the present system of two charges $+q$ and $-q$. The potential energy expression then reads

$$U'(\theta) = q[V(\mathbf{r}_1) - V(\mathbf{r}_2)] - \frac{q^2}{4\pi\epsilon_0 \times 2a} \quad (2.33)$$

Here, \mathbf{r}_1 and \mathbf{r}_2 denote the position vectors of $+q$ and $-q$. Now, the potential difference between positions \mathbf{r}_1 and \mathbf{r}_2 equals the work done in bringing a unit positive charge against field from \mathbf{r}_2 to \mathbf{r}_1 . The displacement parallel to the force is $2a \cos\theta$. Thus, $[V(\mathbf{r}_1) - V(\mathbf{r}_2)] = -E \times 2a \cos\theta$. We thus obtain,

$$U'(\theta) = -pE \cos\theta - \frac{q^2}{4\pi\epsilon_0 \times 2a} = -\mathbf{p} \cdot \mathbf{E} - \frac{q^2}{4\pi\epsilon_0 \times 2a} \quad (2.34)$$

We note that $U'(\theta)$ differs from $U(\theta)$ by a quantity which is just a constant for a given dipole. Since a constant is insignificant for potential energy, we can drop the second term in Eq. (2.34) and it then reduces to Eq. (2.32).

We can now understand why we took $\theta_0=\pi/2$. In this case, the work done against the *external* field \mathbf{E} in bringing $+q$ and $-q$ are equal and opposite and cancel out, i.e., $q[V(\mathbf{r}_1) - V(\mathbf{r}_2)] = 0$.

Example 2.6 A molecule of a substance has a permanent electric dipole moment of magnitude 10^{-29} C m. A mole of this substance is polarised (at low temperature) by applying a strong electrostatic field of magnitude 10^6 V m $^{-1}$. The direction of the field is suddenly changed by an angle of 60° . Estimate the heat released by the substance in aligning its dipoles along the new direction of the field. For simplicity, assume 100% polarisation of the sample.

Solution Here, dipole moment of each molecules = 10^{-29} C m
As 1 mole of the substance contains 6×10^{23} molecules,
total dipole moment of all the molecules, $p = 6 \times 10^{23} \times 10^{-29}$ C m
 $= 6 \times 10^{-6}$ C m

Initial potential energy, $U_i = -pE \cos\theta = -6 \times 10^{-6} \times 10^6 \cos 0^\circ = -6$ J
Final potential energy (when $\theta = 60^\circ$), $U_f = -6 \times 10^{-6} \times 10^6 \cos 60^\circ = -3$ J
Change in potential energy = -3 J - (-6) J = 3 J

So, there is loss in potential energy. This must be the energy released by the substance in the form of heat in aligning its dipoles.

EXAMPLE 2.6

2.9 ELECTROSTATICS OF CONDUCTORS

Conductors and insulators were described briefly in Chapter 1. Conductors contain mobile charge carriers. In metallic conductors, these charge carriers are electrons. In a metal, the outer (valence) electrons part away from their atoms and are free to move. These electrons are free within the metal but not free to leave the metal. The free electrons form a kind of ‘gas’; they collide with each other and with the ions, and move randomly in different directions. In an external electric field, they drift against the direction of the field. The positive ions made up of the nuclei and the bound electrons remain held in their fixed positions. In electrolytic conductors, the charge carriers are both positive and negative ions; but

the situation in this case is more involved – the movement of the charge carriers is affected both by the external electric field as also by the so-called chemical forces (see Chapter 3). We shall restrict our discussion to metallic solid conductors. Let us note important results regarding electrostatics of conductors.

1. Inside a conductor, electrostatic field is zero

Consider a conductor, neutral or charged. There may also be an external electrostatic field. In the static situation, when there is no current inside or on the surface of the conductor, the electric field is zero everywhere inside the conductor. This fact can be taken as the defining property of a conductor. A conductor has free electrons. As long as electric field is not zero, the free charge carriers would experience force and drift. In the static situation, the free charges have so distributed themselves that the electric field is zero everywhere inside. *Electrostatic field is zero inside a conductor.*

2. At the surface of a charged conductor, electrostatic field must be normal to the surface at every point

If \mathbf{E} were not normal to the surface, it would have some non-zero component along the surface. Free charges on the surface of the conductor would then experience force and move. In the static situation, therefore, \mathbf{E} should have no tangential component. Thus *electrostatic field at the surface of a charged conductor must be normal to the surface at every point.* (For a conductor without any surface charge density, field is zero even at the surface.) See result 5.

3. The interior of a conductor can have no excess charge in the static situation

A neutral conductor has equal amounts of positive and negative charges in every small volume or surface element. When the conductor is charged, the excess charge can reside only on the surface in the static situation. This follows from the Gauss's law. Consider any arbitrary volume element v inside a conductor. On the closed surface S bounding the volume element v , electrostatic field is zero. Thus the total electric flux through S is zero. Hence, by Gauss's law, there is no net charge enclosed by S . But the surface S can be made as small as you like, i.e., the volume v can be made vanishingly small. This means *there is no net charge at any point inside the conductor, and any excess charge must reside at the surface.*

4. Electrostatic potential is constant throughout the volume of the conductor and has the same value (as inside) on its surface

This follows from results 1 and 2 above. Since $\mathbf{E} = 0$ inside the conductor and has no tangential component on the surface, no work is done in moving a small test charge within the conductor and on its surface. That is, there is no potential difference between any two points inside or on the surface of the conductor. Hence, the result. If the conductor is charged,

electric field normal to the surface exists; this means potential will be different for the surface and a point just outside the surface.

In a system of conductors of arbitrary size, shape and charge configuration, each conductor is characterised by a constant value of potential, but this constant may differ from one conductor to the other.

5. Electric field at the surface of a charged conductor

$$\mathbf{E} = \frac{\sigma}{\epsilon_0} \hat{\mathbf{n}} \quad (2.35)$$

where σ is the surface charge density and $\hat{\mathbf{n}}$ is a unit vector normal to the surface in the outward direction.

To derive the result, choose a pill box (a short cylinder) as the Gaussian surface about any point P on the surface, as shown in Fig. 2.17. The pill box is partly inside and partly outside the surface of the conductor. It has a small area of cross section δS and negligible height.

Just inside the surface, the electrostatic field is zero; just outside, the field is normal to the surface with magnitude E . Thus, the contribution to the total flux through the pill box comes only from the outside (circular) cross-section of the pill box. This equals $\pm E\delta S$ (positive for $\sigma > 0$, negative for $\sigma < 0$), since over the small area δS , \mathbf{E} may be considered constant and \mathbf{E} and δS are parallel or antiparallel. The charge enclosed by the pill box is $\sigma\delta S$.

By Gauss's law

$$E\delta S = \frac{|\sigma|\delta S}{\epsilon_0}$$

$$E = \frac{|\sigma|}{\epsilon_0} \quad (2.36)$$

Including the fact that electric field is normal to the surface, we get the vector relation, Eq. (2.35), which is true for both signs of σ . For $\sigma > 0$, electric field is normal to the surface outward; for $\sigma < 0$, electric field is normal to the surface inward.

6. Electrostatic shielding

Consider a conductor with a cavity, with no charges inside the cavity. A remarkable result is that the electric field inside the cavity is zero, whatever be the size and shape of the cavity and whatever be the charge on the conductor and the external fields in which it might be placed. We have proved a simple case of this result already: the electric field inside a charged spherical shell is zero. The proof of the result for the shell makes use of the spherical symmetry of the shell (see Chapter 1). But the vanishing of electric field in the (charge-free) cavity of a conductor is, as mentioned above, a very general result. A related result is that even if the conductor

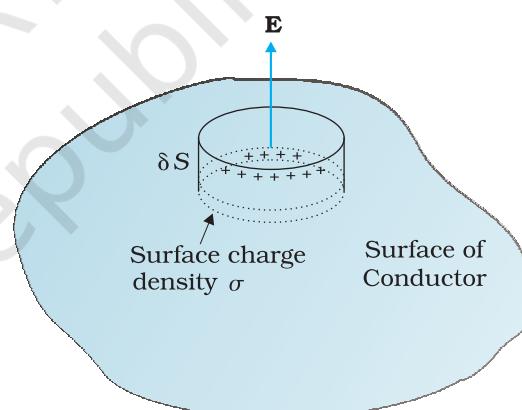


FIGURE 2.17 The Gaussian surface (a pill box) chosen to derive Eq. (2.35) for electric field at the surface of a charged conductor.

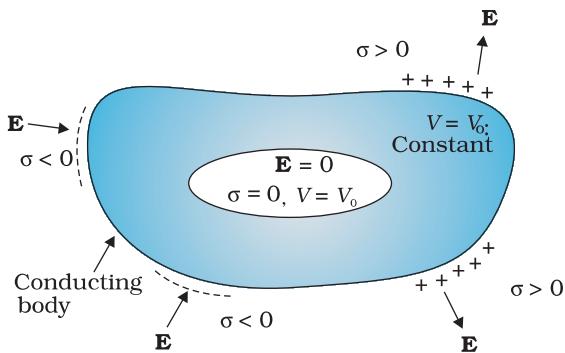


FIGURE 2.18 The electric field inside a cavity of any conductor is zero. All charges reside only on the outer surface of a conductor with cavity. (There are no charges placed in the cavity.)

is charged or charges are induced on a neutral conductor by an external field, all charges reside only on the outer surface of a conductor with cavity.

The proofs of the results noted in Fig. 2.18 are omitted here, but we note their important implication. Whatever be the charge and field configuration outside, any cavity in a conductor remains shielded from outside electric influence: *the field inside the cavity is always zero*. This is known as *electrostatic shielding*. The effect can be made use of in protecting sensitive instruments from outside electrical influence. Figure 2.19 gives a summary of the important electrostatic properties of a conductor.

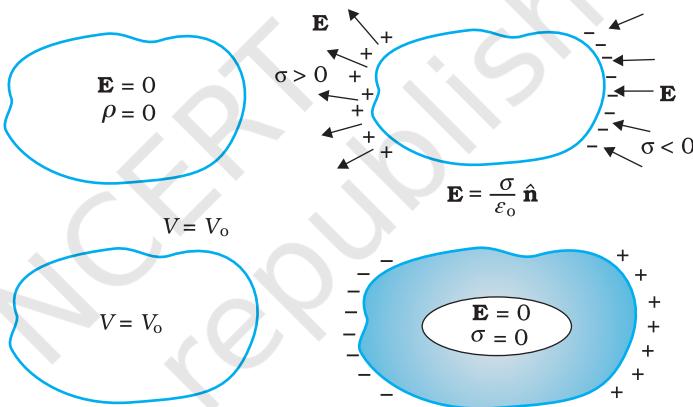


FIGURE 2.19 Some important electrostatic properties of a conductor.

Example 2.7

- A comb run through one's dry hair attracts small bits of paper. Why?
What happens if the hair is wet or if it is a rainy day? (Remember, a paper does not conduct electricity.)
- Ordinary rubber is an insulator. But special rubber tyres of aircraft are made slightly conducting. Why is this necessary?
- Vehicles carrying inflammable materials usually have metallic ropes touching the ground during motion. Why?
- A bird perches on a bare high power line, and nothing happens to the bird. A man standing on the ground touches the same line and gets a fatal shock. Why?

Solution

- This is because the comb gets charged by friction. The molecules in the paper get polarised by the charged comb, resulting in a net force of attraction. If the hair is wet, or if it is rainy day, friction between hair and the comb reduces. The comb does not get charged and thus it will not attract small bits of paper.

- (b) To enable them to conduct charge (produced by friction) to the ground; as too much of static electricity accumulated may result in spark and result in fire.
- (c) Reason similar to (b).
- (d) Current passes only when there is difference in potential.

EXAMPLE 2.7

2.10 DIELECTRICS AND POLARISATION

Dielectrics are non-conducting substances. In contrast to conductors, they have no (or negligible number of) charge carriers. Recall from Section

2.9 what happens when a conductor is placed in an external electric field. The free charge carriers move and charge distribution in the conductor adjusts itself in such a way that the electric field due to induced charges opposes the external field within the conductor. This happens until, in the static situation, the two fields cancel each other and the net electrostatic field in the conductor is zero. In a dielectric, this free movement of charges is not possible. It turns out that the external field induces dipole moment by stretching or re-orienting molecules of the dielectric. The collective effect of all the molecular dipole moments is net charges on the surface of the dielectric which produce a field that opposes the external field. Unlike in a conductor, however, the opposing field so induced does not exactly cancel the external field. It only reduces it. The extent of the effect depends on the nature of the dielectric. To understand the effect, we need to look at the charge distribution of a dielectric at the molecular level.

The molecules of a substance may be polar or non-polar. In a non-polar molecule, the centres of positive and negative charges coincide. The molecule then has no permanent (or intrinsic) dipole moment. Examples of non-polar molecules are oxygen (O_2) and hydrogen (H_2) molecules which, because of their symmetry, have no dipole moment. On the other hand, a polar molecule is one in which the centres of positive and negative charges are separated (even when there is no external field). Such molecules have a permanent dipole moment. An ionic molecule such as HCl or a molecule of water (H_2O) are examples of polar molecules.

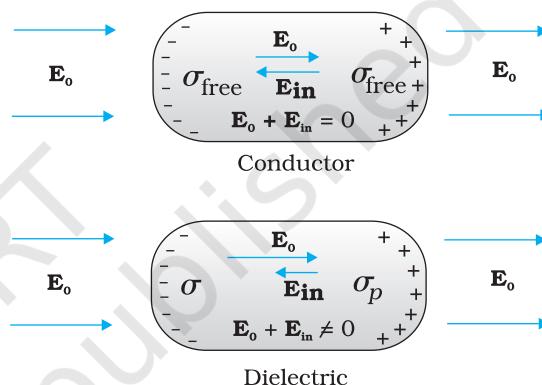


FIGURE 2.20 Difference in behaviour of a conductor and a dielectric in an external electric field.

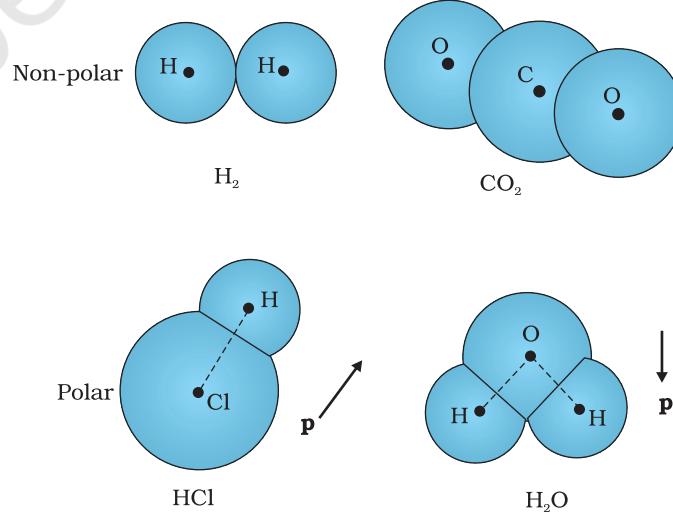


FIGURE 2.21 Some examples of polar and non-polar molecules.

Physics

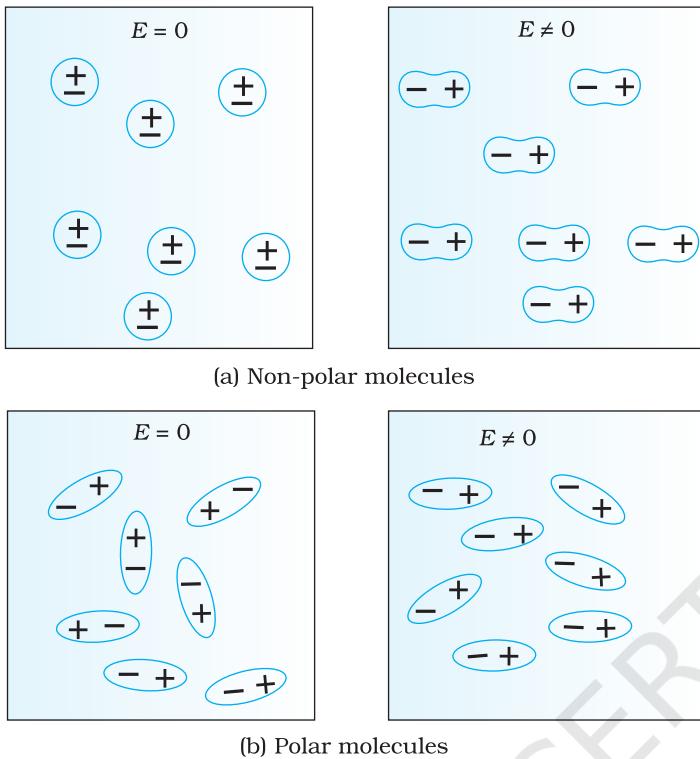


FIGURE 2.22 A dielectric develops a net dipole moment in an external electric field. (a) Non-polar molecules, (b) Polar molecules.

an external field is applied, the individual dipole moments tend to align with the field. When summed over all the molecules, there is then a net dipole moment in the direction of the external field, i.e., the dielectric is polarised. The extent of polarisation depends on the relative strength of two mutually opposite factors: the dipole potential energy in the external field tending to align the dipoles with the field and thermal energy tending to disrupt the alignment. There may be, in addition, the ‘induced dipole moment’ effect as for non-polar molecules, but generally the alignment effect is more important for polar molecules.

Thus in either case, whether polar or non-polar, a dielectric develops a net dipole moment in the presence of an external field. The dipole moment per unit volume is called *polarisation* and is denoted by \mathbf{P} . For linear isotropic dielectrics,

$$\mathbf{P} = \chi_e \mathbf{E} \quad (2.37)$$

where χ_e is a constant characteristic of the dielectric and is known as the *electric susceptibility* of the dielectric medium.

It is possible to relate χ_e to the molecular properties of the substance, but we shall not pursue that here.

The question is: how does the polarised dielectric modify the original external field inside it? Let us consider, for simplicity, a rectangular dielectric slab placed in a uniform external field \mathbf{E}_0 parallel to two of its faces. The field causes a uniform polarisation \mathbf{P} of the dielectric. Thus

In an external electric field, the positive and negative charges of a non-polar molecule are displaced in opposite directions. The displacement stops when the external force on the constituent charges of the molecule is balanced by the restoring force (due to internal fields in the molecule). The non-polar molecule thus develops an induced dipole moment. The dielectric is said to be polarised by the external field. We consider only the simple situation when the induced dipole moment is in the direction of the field and is proportional to the field strength. (Substances for which this assumption is true are called *linear isotropic dielectrics*.) The induced dipole moments of different molecules add up giving a net dipole moment of the dielectric in the presence of the external field.

A dielectric with polar molecules also develops a net dipole moment in an external field, but for a different reason. In the absence of any external field, the different permanent dipoles are oriented randomly due to thermal agitation; so the total dipole moment is zero. When

Electrostatic Potential and Capacitance

every volume element Δv of the slab has a dipole moment $\mathbf{P}\Delta v$ in the direction of the field. The volume element Δv is macroscopically small but contains a very large number of molecular dipoles. Anywhere inside the dielectric, the volume element Δv has no net charge (though it has net dipole moment). This is, because, the positive charge of one dipole sits close to the negative charge of the adjacent dipole. However, at the surfaces of the dielectric normal to the electric field, there is evidently a net charge density. As seen in Fig. 2.23, the positive ends of the dipoles remain unneutralised at the right surface and the negative ends at the left surface. The unbalanced charges are the induced charges due to the external field.

Thus the polarised dielectric is equivalent to two charged surfaces with induced surface charge densities, say σ_p and $-\sigma_p$. Clearly, the field produced by these surface charges opposes the external field. The total field in the dielectric is, thereby, reduced from the case when no dielectric is present. We should note that the surface charge density $\pm\sigma_p$ arises from bound (not free charges) in the dielectric.

2.11 CAPACITORS AND CAPACITANCE

A capacitor is a system of two conductors separated by an insulator (Fig. 2.24). The conductors have charges, say Q_1 and Q_2 , and potentials V_1 and V_2 . Usually, in practice, the two conductors have charges Q and $-Q$, with potential difference $V = V_1 - V_2$ between them. We shall consider only this kind of charge configuration of the capacitor. (Even a single conductor can be used as a capacitor by assuming the other at infinity.) The conductors may be so charged by connecting them to the two terminals of a battery. Q is called the charge of the capacitor, though this, in fact, is the charge on one of the conductors – the total charge of the capacitor is zero.

The electric field in the region between the conductors is proportional to the charge Q . That is, if the charge on the capacitor is, say doubled, the electric field will also be doubled at every point. (This follows from the direct proportionality between field and charge implied by Coulomb's law and the superposition principle.) Now, potential difference V is the work done per unit positive charge in taking a small test charge from the conductor 2 to 1 against the field. Consequently, V is also proportional to Q , and the ratio Q/V is a constant:

$$C = \frac{Q}{V} \quad (2.38)$$

The constant C is called the *capacitance* of the capacitor. C is independent of Q or V , as stated above. The capacitance C depends only on the

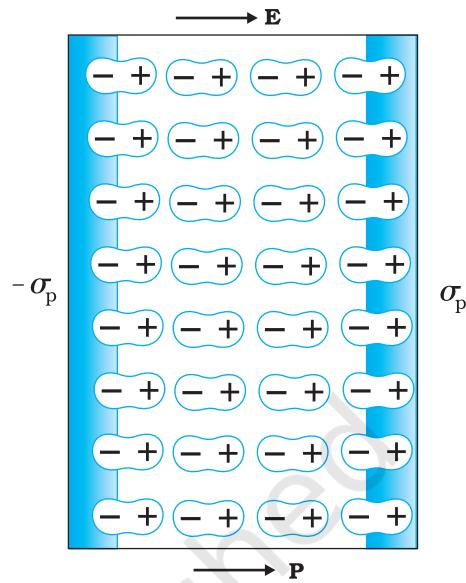


FIGURE 2.23 A uniformly polarised dielectric amounts to induced surface charge density, but no volume charge density.

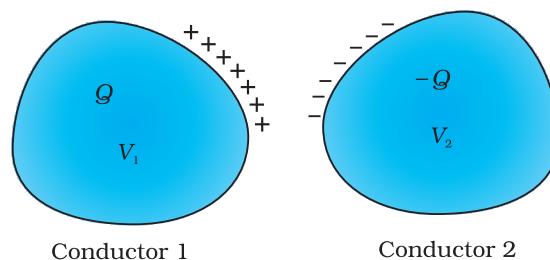


FIGURE 2.24 A system of two conductors separated by an insulator forms a capacitor.

geometrical configuration (shape, size, separation) of the system of two conductors. [As we shall see later, it also depends on the nature of the insulator (dielectric) separating the two conductors.] The SI unit of capacitance is 1 farad ($=1 \text{ coulomb volt}^{-1}$) or $1 \text{ F} = 1 \text{ C V}^{-1}$. A capacitor with fixed capacitance is symbolically shown as $\begin{array}{c} \text{--} \\ \text{|} \end{array}$, while the one with variable capacitance is shown as $\begin{array}{c} \text{+} \\ \text{\textperp} \end{array}$.

Equation (2.38) shows that for large C , V is small for a given Q . This means a capacitor with large capacitance can hold large amount of charge Q at a relatively small V . This is of practical importance. High potential difference implies strong electric field around the conductors. A strong electric field can ionise the surrounding air and accelerate the charges so produced to the oppositely charged plates, thereby neutralising the charge on the capacitor plates, at least partly. In other words, the charge of the capacitor leaks away due to the reduction in insulating power of the intervening medium.

The maximum electric field that a dielectric medium can withstand without break-down (of its insulating property) is called its *dielectric strength*; for air it is about $3 \times 10^6 \text{ Vm}^{-1}$. For a separation between conductors of the order of 1 cm or so, this field corresponds to a potential difference of $3 \times 10^4 \text{ V}$ between the conductors. Thus, for a capacitor to store a large amount of charge without leaking, its capacitance should be high enough so that the potential difference and hence the electric field do not exceed the break-down limits. Put differently, there is a limit to the amount of charge that can be stored on a given capacitor without significant leaking. In practice, a farad is a very big unit; the most common units are its sub-multiples $1 \mu\text{F} = 10^{-6} \text{ F}$, $1 \text{ nF} = 10^{-9} \text{ F}$, $1 \text{ pF} = 10^{-12} \text{ F}$, etc. Besides its use in storing charge, a capacitor is a key element of most ac circuits with important functions, as described in Chapter 7.

2.12 THE PARALLEL PLATE CAPACITOR

A parallel plate capacitor consists of two large plane parallel conducting plates separated by a small distance (Fig. 2.25). We first take the intervening medium between the plates to be vacuum. The effect of a dielectric medium between the plates is discussed in the next section. Let A be the area of each plate and d the separation between them. The two plates have charges Q and $-Q$. Since d is much smaller than the linear dimension of the plates ($d^2 \ll A$), we can use the result on electric field by an infinite plane sheet of uniform surface charge density (Section 1.15). Plate 1 has surface charge density $\sigma = Q/A$ and plate 2 has a surface charge density $-\sigma$. Using Eq. (1.33), the electric field in different regions is:

Outer region I (region above the plate 1),

$$E = \frac{\sigma}{2\varepsilon_0} - \frac{\sigma}{2\varepsilon_0} = 0 \quad (2.39)$$

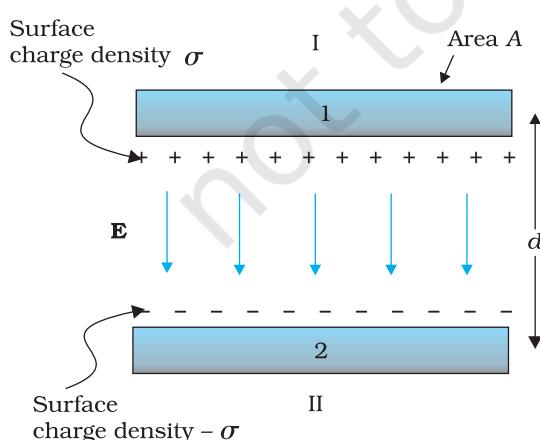


FIGURE 2.25 The parallel plate capacitor.

Electrostatic Potential and Capacitance

Outer region II (region below the plate 2),

$$E = \frac{\sigma}{2\epsilon_0} - \frac{\sigma}{2\epsilon_0} = 0 \quad (2.40)$$

In the inner region between the plates 1 and 2, the electric fields due to the two charged plates add up, giving

$$E = \frac{\sigma}{2\epsilon_0} + \frac{\sigma}{2\epsilon_0} = \frac{\sigma}{\epsilon_0} = \frac{Q}{\epsilon_0 A} \quad (2.41)$$

The direction of electric field is from the positive to the negative plate.

Thus, the electric field is localised between the two plates and is uniform throughout. For plates with finite area, this will not be true near the outer boundaries of the plates. The field lines bend outward at the edges – an effect called ‘fringing of the field’. By the same token, σ will not be strictly uniform on the entire plate. [E and σ are related by Eq. (2.35).] However, for $d^2 \ll A$, these effects can be ignored in the regions sufficiently far from the edges, and the field there is given by Eq. (2.41). Now for uniform electric field, potential difference is simply the electric field times the distance between the plates, that is,

$$V = Ed = \frac{1}{\epsilon_0} \frac{Qd}{A} \quad (2.42)$$

The capacitance C of the parallel plate capacitor is then

$$C = \frac{Q}{V} = \frac{\epsilon_0 A}{d} \quad (2.43)$$

which, as expected, depends only on the geometry of the system. For typical values like $A = 1 \text{ m}^2$, $d = 1 \text{ mm}$, we get

$$C = \frac{8.85 \times 10^{-12} \text{ C}^2 \text{ N}^{-1} \text{ m}^{-2} \times 1 \text{ m}^2}{10^{-3} \text{ m}} = 8.85 \times 10^{-9} \text{ F} \quad (2.44)$$

(You can check that if $1\text{F} = 1\text{C V}^{-1} = 1\text{C (NC}^{-1}\text{m})^{-1} = 1 \text{ C}^2 \text{ N}^{-1} \text{ m}^{-1}$.) This shows that 1F is too big a unit in practice, as remarked earlier. Another way of seeing the ‘bigness’ of 1F is to calculate the area of the plates needed to have $C = 1\text{F}$ for a separation of, say 1 cm:

$$A = \frac{Cd}{\epsilon_0} = \frac{1\text{F} \times 10^{-2} \text{ m}}{8.85 \times 10^{-12} \text{ C}^2 \text{ N}^{-1} \text{ m}^{-2}} = 10^9 \text{ m}^2 \quad (2.45)$$

which is a plate about 30 km in length and breadth!

2.13 EFFECT OF DIELECTRIC ON CAPACITANCE

With the understanding of the behavior of dielectrics in an external field developed in Section 2.10, let us see how the capacitance of a parallel plate capacitor is modified when a dielectric is present. As before, we have two large plates, each of area A , separated by a distance d . The charge on the plates is $\pm Q$, corresponding to the charge density $\pm\sigma$ (with $\sigma = Q/A$). When there is vacuum between the plates,

$$E_0 = \frac{\sigma}{\epsilon_0}$$

PHYSICS

Factors affecting capacitance, capacitors in action
Interactive Java tutorial
<http://micro.magnet.fsu.edu/electromag/java/capacitance/>

Physics

and the potential difference V_0 is

$$V_0 = E_0 d$$

The capacitance C_0 in this case is

$$C_0 = \frac{Q}{V_0} = \epsilon_0 \frac{A}{d} \quad (2.46)$$

Consider next a dielectric inserted between the plates fully occupying the intervening region. The dielectric is polarised by the field and, as explained in Section 2.10, the effect is equivalent to two charged sheets (at the surfaces of the dielectric normal to the field) with surface charge densities σ_p and $-\sigma_p$. The electric field in the dielectric then corresponds to the case when the net surface charge density on the plates is $\pm(\sigma - \sigma_p)$. That is,

$$E = \frac{\sigma - \sigma_p}{\epsilon_0} \quad (2.47)$$

so that the potential difference across the plates is

$$V = Ed = \frac{\sigma - \sigma_p}{\epsilon_0} d \quad (2.48)$$

For linear dielectrics, we expect σ_p to be proportional to E_0 , i.e., to σ . Thus, $(\sigma - \sigma_p)$ is proportional to σ and we can write

$$\sigma - \sigma_p = \frac{\sigma}{K} \quad (2.49)$$

where K is a constant characteristic of the dielectric. Clearly, $K > 1$. We then have

$$V = \frac{\sigma d}{\epsilon_0 K} = \frac{Qd}{A \epsilon_0 K} \quad (2.50)$$

The capacitance C , with dielectric between the plates, is then

$$C = \frac{Q}{V} = \frac{\epsilon_0 K A}{d} \quad (2.51)$$

The product $\epsilon_0 K$ is called the *permittivity* of the medium and is denoted by ϵ

$$\epsilon = \epsilon_0 K \quad (2.52)$$

For vacuum $K = 1$ and $\epsilon = \epsilon_0$; ϵ_0 is called the *permittivity of the vacuum*. The dimensionless ratio

$$K = \frac{\epsilon}{\epsilon_0} \quad (2.53)$$

is called the *dielectric constant* of the substance. As remarked before, from Eq. (2.49), it is clear that K is greater than 1. From Eqs. (2.46) and (2.51)

$$K = \frac{C}{C_0} \quad (2.54)$$

Thus, the dielectric constant of a substance is the factor (>1) by which the capacitance increases from its vacuum value, when the dielectric is inserted fully between the plates of a capacitor. Though we arrived at

Electrostatic Potential and Capacitance

Eq. (2.54) for the case of a parallel plate capacitor, it holds good for any type of capacitor and can, in fact, be viewed in general as a definition of the dielectric constant of a substance.

ELECTRIC DISPLACEMENT

We have introduced the notion of dielectric constant and arrived at Eq. (2.54), without giving the explicit relation between the induced charge density σ_p and the polarisation \mathbf{P} .

We take without proof the result that

$$\sigma_p = \mathbf{P} \cdot \hat{\mathbf{n}}$$

where $\hat{\mathbf{n}}$ is a unit vector along the outward normal to the surface. Above equation is general, true for any shape of the dielectric. For the slab in Fig. 2.23, \mathbf{P} is along $\hat{\mathbf{n}}$ at the right surface and opposite to $\hat{\mathbf{n}}$ at the left surface. Thus at the right surface, induced charge density is positive and at the left surface, it is negative, as guessed already in our qualitative discussion before. Putting the equation for electric field in vector form

$$\mathbf{E} \cdot \hat{\mathbf{n}} = \frac{\sigma - \mathbf{P} \cdot \hat{\mathbf{n}}}{\epsilon_0}$$

$$\text{or } (\epsilon_0 \mathbf{E} + \mathbf{P}) \cdot \hat{\mathbf{n}} = \sigma$$

The quantity $\epsilon_0 \mathbf{E} + \mathbf{P}$ is called the *electric displacement* and is denoted by \mathbf{D} . It is a vector quantity. Thus,

$$\mathbf{D} = \epsilon_0 \mathbf{E} + \mathbf{P}, \mathbf{D} \cdot \hat{\mathbf{n}} = \sigma$$

The significance of \mathbf{D} is this : in vacuum, \mathbf{E} is related to the free charge density σ . When a dielectric medium is present, the corresponding role is taken up by \mathbf{D} . For a dielectric medium, it is \mathbf{D} not \mathbf{E} that is directly related to free charge density σ , as seen in above equation. Since \mathbf{P} is in the same direction as \mathbf{E} , all the three vectors \mathbf{P} , \mathbf{E} and \mathbf{D} are parallel.

The ratio of the magnitudes of \mathbf{D} and \mathbf{E} is

$$\frac{D}{E} = \frac{\sigma \epsilon_0}{\sigma - \sigma_p} = \epsilon_0 K$$

Thus,

$$\mathbf{D} = \epsilon_0 K \mathbf{E}$$

$$\text{and } \mathbf{P} = \mathbf{D} - \epsilon_0 \mathbf{E} = \epsilon_0 (K - 1) \mathbf{E}$$

This gives for the electric susceptibility χ_e defined in Eq. (2.37)

$$\chi_e = \epsilon_0 (K - 1)$$

Example 2.8 A slab of material of dielectric constant K has the same area as the plates of a parallel-plate capacitor but has a thickness $(3/4)d$, where d is the separation of the plates. How is the capacitance changed when the slab is inserted between the plates?

Solution Let $E_0 = V_0/d$ be the electric field between the plates when there is no dielectric and the potential difference is V_0 . If the dielectric is now inserted, the electric field in the dielectric will be $E = E_0/K$. The potential difference will then be

EXAMPLE 2.8

EXAMPLE 2.8

$$V = E_0 \left(\frac{1}{4}d \right) + \frac{E_0}{K} \left(\frac{3}{4}d \right)$$

$$= E_0 d \left(\frac{1}{4} + \frac{3}{4K} \right) = V_0 \frac{K+3}{4K}$$

The potential difference decreases by the factor $(K+3)/K$ while the free charge Q_0 on the plates remains unchanged. The capacitance thus increases

$$C = \frac{Q_0}{V} = \frac{4K}{K+3} \frac{Q_0}{V_0} = \frac{4K}{K+3} C_0$$

2.14 COMBINATION OF CAPACITORS

We can combine several capacitors of capacitance C_1, C_2, \dots, C_n to obtain a system with some effective capacitance C . The effective capacitance depends on the way the individual capacitors are combined. Two simple possibilities are discussed below.

2.14.1 Capacitors in series

Figure 2.26 shows capacitors C_1 and C_2 combined in series.

The left plate of C_1 and the right plate of C_2 are connected to two terminals of a battery and have charges Q and $-Q$, respectively. It then follows that the right plate of C_1 has charge $-Q$ and the left plate of C_2 has charge Q . If this was not so, the net charge on each capacitor would not be zero. This would result in an electric field in the conductor connecting C_1 and C_2 . Charge would flow until the net charge on both C_1 and C_2 is zero and there is no electric field in the conductor connecting C_1 and C_2 . Thus, in the series combination, charges on the two plates ($\pm Q$) are the same on each capacitor. The total potential drop V across the combination is the sum of the potential drops V_1 and V_2 across C_1 and C_2 , respectively.

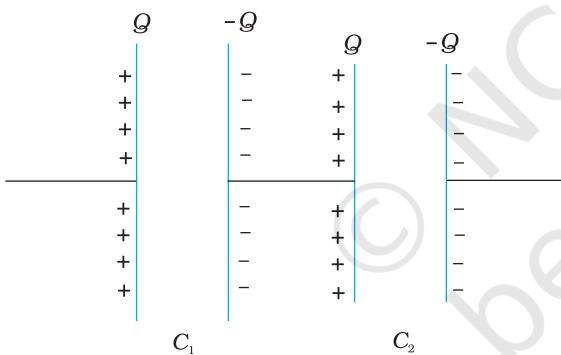


FIGURE 2.26 Combination of two capacitors in series.

$$V = V_1 + V_2 = \frac{Q}{C_1} + \frac{Q}{C_2} \quad (2.55)$$

$$\text{i.e., } \frac{V}{Q} = \frac{1}{C_1} + \frac{1}{C_2}, \quad (2.56)$$

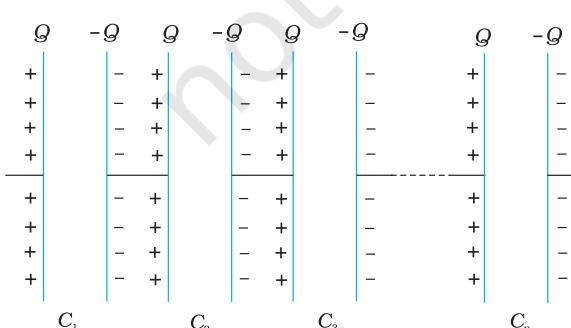
Now we can regard the combination as an effective capacitor with charge Q and potential difference V . The *effective capacitance* of the combination is

$$C = \frac{Q}{V} \quad (2.57)$$

We compare Eq. (2.57) with Eq. (2.56), and obtain

$$\frac{1}{C} = \frac{1}{C_1} + \frac{1}{C_2} \quad (2.58)$$

FIGURE 2.27 Combination of n capacitors in series.



Electrostatic Potential and Capacitance

The proof clearly goes through for any number of capacitors arranged in a similar way. Equation (2.55), for n capacitors arranged in series, generalises to

$$V = V_1 + V_2 + \dots + V_n = \frac{Q}{C_1} + \frac{Q}{C_2} + \dots + \frac{Q}{C_n} \quad (2.59)$$

Following the same steps as for the case of two capacitors, we get the general formula for effective capacitance of a series combination of n capacitors:

$$\frac{1}{C} = \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3} + \dots + \frac{1}{C_n} \quad (2.60)$$

2.14.2 Capacitors in parallel

Figure 2.28 (a) shows two capacitors arranged in parallel. In this case, the same potential difference is applied across both the capacitors. But the plate charges ($\pm Q_1$) on capacitor 1 and the plate charges ($\pm Q_2$) on the capacitor 2 are not necessarily the same:

$$Q_1 = C_1 V, Q_2 = C_2 V \quad (2.61)$$

The equivalent capacitor is one with charge

$$Q = Q_1 + Q_2 \quad (2.62)$$

and potential difference V .

$$Q = CV = C_1 V + C_2 V \quad (2.63)$$

The effective capacitance C is, from Eq. (2.63),

$$C = C_1 + C_2 \quad (2.64)$$

The general formula for effective capacitance C for parallel combination of n capacitors [Fig. 2.28 (b)] follows similarly,

$$Q = Q_1 + Q_2 + \dots + Q_n \quad (2.65)$$

$$\text{i.e., } CV = C_1 V + C_2 V + \dots + C_n V \quad (2.66)$$

which gives

$$C = C_1 + C_2 + \dots + C_n \quad (2.67)$$

Example 2.9 A network of four $10 \mu\text{F}$ capacitors is connected to a 500 V supply, as shown in Fig. 2.29. Determine (a) the equivalent capacitance of the network and (b) the charge on each capacitor. (Note, the *charge on a capacitor* is the charge on the plate with higher potential, equal and opposite to the charge on the plate with lower potential.)

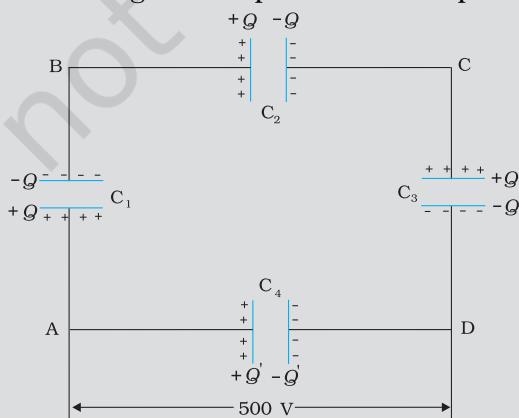


FIGURE 2.29

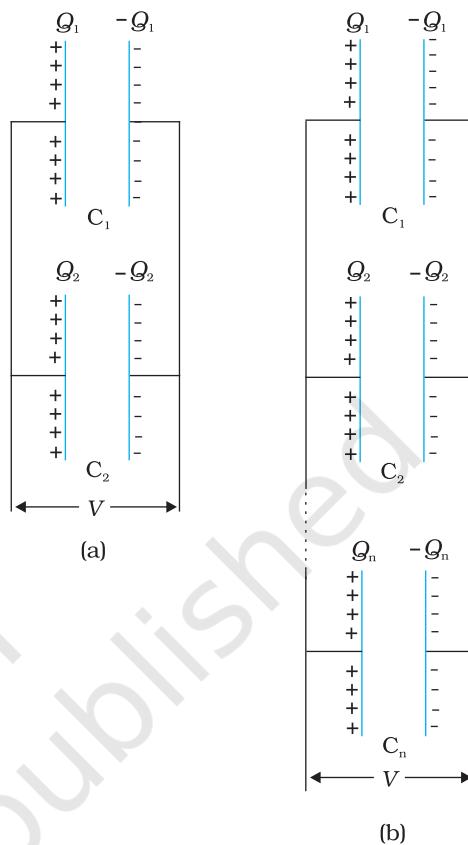


FIGURE 2.28 Parallel combination of (a) two capacitors, (b) n capacitors.

EXAMPLE 2.9

EXAMPLE 2.9

Solution

(a) In the given network, C_1 , C_2 and C_3 are connected in series. The effective capacitance C' of these three capacitors is given by

$$\frac{1}{C'} = \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3}$$

For $C_1 = C_2 = C_3 = 10 \mu\text{F}$, $C' = (10/3) \mu\text{F}$. The network has C' and C_4 connected in parallel. Thus, the equivalent capacitance C of the network is

$$C = C' + C_4 = \left(\frac{10}{3} + 10 \right) \mu\text{F} = 13.3 \mu\text{F}$$

(b) Clearly, from the figure, the charge on each of the capacitors, C_1 , C_2 and C_3 is the same, say Q . Let the charge on C_4 be Q' . Now, since the potential difference across AB is Q/C_1 , across BC is Q/C_2 , across CD is Q/C_3 , we have

$$\frac{Q}{C_1} + \frac{Q}{C_2} + \frac{Q}{C_3} = 500 \text{ V}$$

Also, $Q'/C_4 = 500 \text{ V}$.

This gives for the given value of the capacitances,

$$Q = 500 \text{ V} \times \frac{10}{3} \mu\text{F} = 1.7 \times 10^{-3} \text{ C} \text{ and}$$

$$Q' = 500 \text{ V} \times 10 \mu\text{F} = 5.0 \times 10^{-3} \text{ C}$$

2.15 ENERGY STORED IN A CAPACITOR

A capacitor, as we have seen above, is a system of two conductors with charge Q and $-Q$. To determine the energy stored in this configuration, consider initially two uncharged conductors 1 and 2. Imagine next a process of transferring charge from conductor 2 to conductor 1 bit by bit, so that at the end, conductor 1 gets charge Q . By charge conservation, conductor 2 has charge $-Q$ at the end (Fig 2.30).

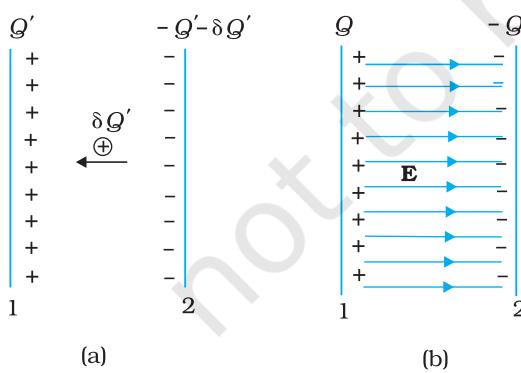


FIGURE 2.30 (a) Work done in a small step of building charge on conductor 1 from Q' to $Q' + \delta Q'$. (b) Total work done in charging the capacitor may be viewed as stored in the energy of electric field between the plates.

In transferring positive charge from conductor 2 to conductor 1, work will be done externally, since at any stage conductor 1 is at a higher potential than conductor 2. To calculate the total work done, we first calculate the work done in a small step involving transfer of an infinitesimal (i.e., vanishingly small) amount of charge. Consider the intermediate situation when the conductors 1 and 2 have charges Q' and $-Q'$ respectively. At this stage, the potential difference V' between conductors 1 to 2 is Q'/C , where C is the capacitance of the system. Next imagine that a small charge $\delta Q'$ is transferred from conductor 2 to 1. Work done in this step ($\delta W'$), resulting in charge Q' on conductor 1 increasing to $Q' + \delta Q'$, is given by

$$\delta W = V' \delta Q' = \frac{Q'}{C} \delta Q' \quad (2.68)$$

Electrostatic Potential and Capacitance

Since $\delta Q'$ can be made as small as we like, Eq. (2.68) can be written as

$$\delta W = \frac{1}{2C}[(Q' + \delta Q')^2 - Q'^2] \quad (2.69)$$

Equations (2.68) and (2.69) are identical because the term of second order in $\delta Q'$, i.e., $\delta Q'^2/2C$, is negligible, since $\delta Q'$ is arbitrarily small. The total work done (W) is the sum of the small work (δW) over the very large number of steps involved in building the charge Q' from zero to Q .

$$W = \sum_{\text{sum over all steps}} \delta W = \sum_{\text{sum over all steps}} \frac{1}{2C}[(Q' + \delta Q')^2 - Q'^2] \quad (2.70)$$

$$= \frac{1}{2C}[\{\delta Q'^2 - 0\} + \{(2\delta Q')^2 - \delta Q'^2\} + \{(3\delta Q')^2 - (2\delta Q')^2\} + \dots + \{Q^2 - (Q - \delta Q)^2\}] \quad (2.71)$$

$$= \frac{1}{2C}[Q^2 - 0] = \frac{Q^2}{2C} \quad (2.72)$$

The same result can be obtained directly from Eq. (2.68) by integration

$$W = \int_0^Q \frac{Q'}{C} \delta Q' = \frac{1}{C} \frac{Q'^2}{2} \Big|_0^Q = \frac{Q^2}{2C}$$

This is not surprising since integration is nothing but summation of a large number of small terms.

We can write the final result, Eq. (2.72) in different ways

$$W = \frac{Q^2}{2C} = \frac{1}{2} CV^2 = \frac{1}{2} QV \quad (2.73)$$

Since electrostatic force is conservative, this work is stored in the form of potential energy of the system. For the same reason, the final result for potential energy [Eq. (2.73)] is independent of the manner in which the charge configuration of the capacitor is built up. When the capacitor discharges, this stored-up energy is released. It is possible to view the potential energy of the capacitor as 'stored' in the electric field between the plates. To see this, consider for simplicity, a parallel plate capacitor [of area A (of each plate) and separation d between the plates].

Energy stored in the capacitor

$$= \frac{1}{2} \frac{Q^2}{C} = \frac{(A\sigma)^2}{2} \times \frac{d}{\epsilon_0 A} \quad (2.74)$$

The surface charge density σ is related to the electric field E between the plates,

$$E = \frac{\sigma}{\epsilon_0} \quad (2.75)$$

From Eqs. (2.74) and (2.75), we get

Energy stored in the capacitor

$$U = (1/2)\epsilon_0 E^2 \times Ad \quad (2.76)$$

Note that Ad is the volume of the region between the plates (where electric field alone exists). If we define *energy density as energy stored per unit volume of space*, Eq (2.76) shows that

Energy density of electric field,

$$u = (1/2)\epsilon_0 E^2 \quad (2.77)$$

Though we derived Eq. (2.77) for the case of a parallel plate capacitor, the result on energy density of an electric field is, in fact, very general and holds true for electric field due to any configuration of charges.

Example 2.10 (a) A 900 pF capacitor is charged by 100 V battery [Fig. 2.31(a)]. How much electrostatic energy is stored by the capacitor? (b) The capacitor is disconnected from the battery and connected to another 900 pF capacitor [Fig. 2.31(b)]. What is the electrostatic energy stored by the system?

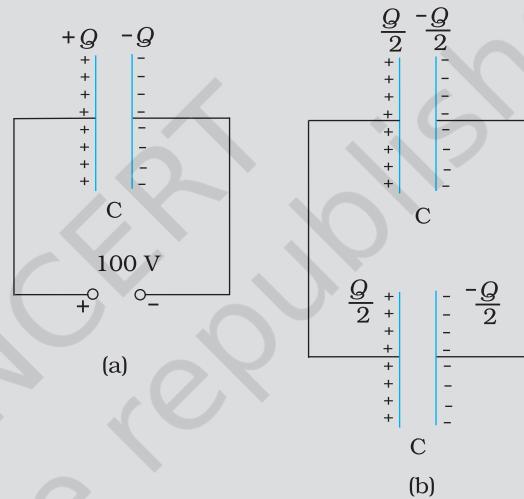


FIGURE 2.31

Solution

(a) The charge on the capacitor is

$$Q = CV = 900 \times 10^{-12} \text{ F} \times 100 \text{ V} = 9 \times 10^{-8} \text{ C}$$

The energy stored by the capacitor is

$$= (1/2) CV^2 = (1/2) QV$$

$$= (1/2) \times 9 \times 10^{-8} \text{ C} \times 100 \text{ V} = 4.5 \times 10^{-6} \text{ J}$$

(b) In the steady situation, the two capacitors have their positive plates at the same potential, and their negative plates at the same potential. Let the common potential difference be V' . The charge on each capacitor is then $Q' = CV'$. By charge conservation, $Q' = Q/2$. This implies $V' = V/2$. The total energy of the system is

$$= 2 \times \frac{1}{2} Q' V' = \frac{1}{4} QV = 2.25 \times 10^{-6} \text{ J}$$

Thus in going from (a) to (b), though no charge is lost; the final energy is only half the initial energy. *Where has the remaining energy gone?*

There is a transient period before the system settles to the situation (b). During this period, a transient current flows from the first capacitor to the second. Energy is lost during this time in the form of heat and electromagnetic radiation.

2.16 VAN DE GRAAFF GENERATOR

This is a machine that can build up high voltages of the order of a few million volts. The resulting large electric fields are used to accelerate charged particles (electrons, protons, ions) to high energies needed for experiments to probe the small scale structure of matter. The principle underlying the machine is as follows.

Suppose we have a large spherical conducting shell of radius R , on which we place a charge Q . This charge spreads itself uniformly all over the sphere. As we have seen in Section 1.14, the field outside the sphere is just that of a point charge Q at the centre; while the field inside the sphere vanishes. So the potential outside is that of a point charge; and inside it is constant, namely the value at the radius R . We thus have:

Potential inside conducting spherical shell of radius R carrying charge Q
= constant

$$= \frac{1}{4\pi\epsilon_0} \frac{Q}{R} \quad (2.78)$$

Now, as shown in Fig. 2.32, let us suppose that in some way we introduce a small sphere of radius r , carrying some charge q , into the large one, and place it at the centre. The potential due to this new charge clearly has the following values at the radii indicated:

Potential due to small sphere of radius r carrying charge q

$$\begin{aligned} &= \frac{1}{4\pi\epsilon_0} \frac{q}{r} \text{ at surface of small sphere} \\ &= \frac{1}{4\pi\epsilon_0} \frac{q}{R} \text{ at large shell of radius } R. \end{aligned} \quad (2.79)$$

Taking both charges q and Q into account we have for the total potential V and the potential difference the values

$$\begin{aligned} V(R) &= \frac{1}{4\pi\epsilon_0} \left(\frac{Q}{R} + \frac{q}{R} \right) \\ V(r) &= \frac{1}{4\pi\epsilon_0} \left(\frac{Q}{R} + \frac{q}{r} \right) \\ V(r) - V(R) &= \frac{q}{4\pi\epsilon_0} \left(\frac{1}{r} - \frac{1}{R} \right) \end{aligned} \quad (2.80)$$

Assume now that q is positive. We see that, independent of the amount of charge Q that may have accumulated on the larger sphere and even if it is positive, the inner sphere is always at a higher potential: the difference $V(r) - V(R)$ is positive. The potential due to Q is constant upto radius R and so cancels out in the difference!

This means that if we now connect the smaller and larger sphere by a wire, the charge q on the former



Van de Graaff generator, principle and demonstration:
<http://www.physics.gla.ac.uk/~kskeldon/PubSci/exhibits/E10/>

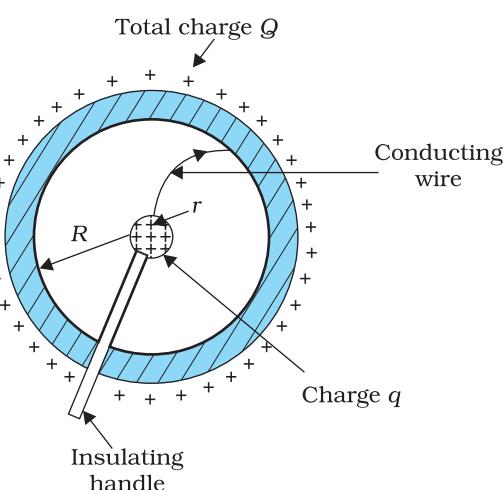


FIGURE 2.32 Illustrating the principle of the electrostatic generator.

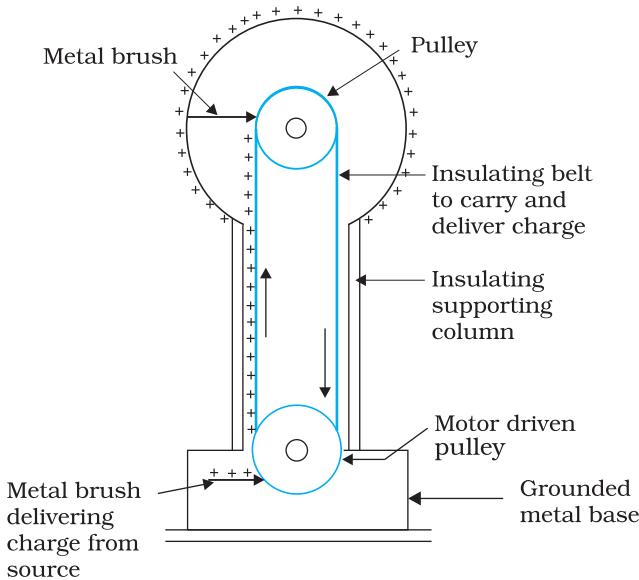


FIGURE 2.33 Principle of construction of Van de Graaff generator.

belt insulating material, like rubber or silk, is wound around two pulleys – one at ground level, one at the centre of the shell. This belt is kept continuously moving by a motor driving the lower pulley. It continuously carries positive charge, sprayed on to it by a brush at ground level, to the top. There it transfers its positive charge to another conducting brush connected to the large shell. Thus positive charge is transferred to the shell, where it spreads out uniformly on the outer surface. In this way, voltage differences of as much as 6 or 8 million volts (with respect to ground) can be built up.

will immediately flow onto the matter, even though the charge Q may be quite large. The natural tendency is for positive charge to move from higher to lower potential. Thus, provided we are somehow able to introduce the small charged sphere into the larger one, we can in this way keep piling up larger and larger amount of charge on the latter. The potential (Eq. 2.78) at the outer sphere would also keep rising, at least until we reach the breakdown field of air.

This is the principle of the van de Graaff generator. It is a machine capable of building up potential difference of a few million volts, and fields close to the breakdown field of air which is about 3×10^6 V/m. A schematic diagram of the van de Graaff generator is given in Fig. 2.33. A large spherical conducting shell (of few metres radius) is supported at a height several meters above the ground on an insulating column. A long narrow endless

SUMMARY

- Electrostatic force is a conservative force. Work done by an external force (equal and opposite to the electrostatic force) in bringing a charge q from a point R to a point P is $V_P - V_R$, which is the difference in potential energy of charge q between the final and initial points.
- Potential at a point is the work done per unit charge (by an external agency) in bringing a charge from infinity to that point. Potential at a point is arbitrary to within an additive constant, since it is the potential difference between two points which is physically significant. If potential at infinity is chosen to be zero; potential at a point with position vector \mathbf{r} due to a point charge Q placed at the origin is given by
$$V(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} \frac{Q}{r}$$
- The electrostatic potential at a point with position vector \mathbf{r} due to a point dipole of dipole moment \mathbf{p} placed at the origin is
$$V(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} \frac{\mathbf{p} \cdot \hat{\mathbf{r}}}{r^2}$$

Electrostatic Potential and Capacitance

The result is true also for a dipole (with charges $-q$ and q separated by $2a$) for $r \gg a$.

- For a charge configuration q_1, q_2, \dots, q_n with position vectors $\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n$, the potential at a point P is given by the superposition principle

$$V = \frac{1}{4\pi\epsilon_0} \left(\frac{q_1}{r_{1P}} + \frac{q_2}{r_{2P}} + \dots + \frac{q_n}{r_{nP}} \right)$$

where r_{1P} is the distance between q_1 and P, as and so on.

- An equipotential surface is a surface over which potential has a constant value. For a point charge, concentric spheres centered at a location of the charge are equipotential surfaces. The electric field \mathbf{E} at a point is perpendicular to the equipotential surface through the point. \mathbf{E} is in the direction of the steepest decrease of potential.
- Potential energy stored in a system of charges is the work done (by an external agency) in assembling the charges at their locations. Potential energy of two charges q_1, q_2 at $\mathbf{r}_1, \mathbf{r}_2$ is given by

$$U = \frac{1}{4\pi\epsilon_0} \frac{q_1 q_2}{r_{12}}$$

where r_{12} is distance between q_1 and q_2 .

- The potential energy of a charge q in an external potential $V(\mathbf{r})$ is $qV(\mathbf{r})$. The potential energy of a dipole moment \mathbf{p} in a uniform electric field \mathbf{E} is $-\mathbf{p} \cdot \mathbf{E}$.
- Electrostatics field \mathbf{E} is zero in the interior of a conductor; just outside the surface of a charged conductor, \mathbf{E} is normal to the surface given by

$\mathbf{E} = \frac{\sigma}{\epsilon_0} \hat{\mathbf{n}}$ where $\hat{\mathbf{n}}$ is the unit vector along the outward normal to the surface and σ is the surface charge density. Charges in a conductor can reside only at its surface. Potential is constant within and on the surface of a conductor. In a cavity within a conductor (with no charges), the electric field is zero.

- A capacitor is a system of two conductors separated by an insulator. Its capacitance is defined by $C = Q/V$, where Q and $-Q$ are the charges on the two conductors and V is the potential difference between them. C is determined purely geometrically, by the shapes, sizes and relative positions of the two conductors. The unit of capacitance is farad: $1 \text{ F} = 1 \text{ C V}^{-1}$. For a parallel plate capacitor (with vacuum between the plates),

$$C = \epsilon_0 \frac{A}{d}$$

where A is the area of each plate and d the separation between them.

- If the medium between the plates of a capacitor is filled with an insulating substance (dielectric), the electric field due to the charged plates induces a net dipole moment in the dielectric. This effect, called polarisation, gives rise to a field in the opposite direction. The net electric field inside the dielectric and hence the potential difference between the plates is thus reduced. Consequently, the capacitance C increases from its value C_0 when there is no medium (vacuum),

$$C = KC_0$$

where K is the dielectric constant of the insulating substance.

Physics

11. For capacitors in the series combination, the total capacitance C is given by

$$\frac{1}{C} = \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3} + \dots$$

In the parallel combination, the total capacitance C is:

$$C = C_1 + C_2 + C_3 + \dots$$

where C_1, C_2, C_3, \dots are individual capacitances.

12. The energy U stored in a capacitor of capacitance C , with charge Q and voltage V is

$$U = \frac{1}{2} QV = \frac{1}{2} CV^2 = \frac{1}{2} \frac{Q^2}{C}$$

The electric energy density (energy per unit volume) in a region with electric field is $(1/2)\epsilon_0 E^2$.

13. A Van de Graaff generator consists of a large spherical conducting shell (a few metre in diameter). By means of a moving belt and suitable brushes, charge is continuously transferred to the shell and potential difference of the order of several million volts is built up, which can be used for accelerating charged particles.

Physical quantity	Symbol	Dimensions	Unit	Remark
Potential	ϕ or V	$[M^1 L^2 T^{-3} A^{-1}]$	V	Potential difference is physically significant
Capacitance	C	$[M^{-1} L^{-2} T^{-4} A^2]$	F	
Polarisation	\mathbf{P}	$[L^{-2} AT]$	$C m^{-2}$	Dipole moment per unit volume
Dielectric constant	K	[Dimensionless]		

POINTS TO PONDER

- Electrostatics deals with forces between charges at rest. But if there is a force on a charge, how can it be at rest? Thus, when we are talking of electrostatic force between charges, it should be understood that each charge is being kept at rest by some unspecified force that opposes the net Coulomb force on the charge.
- A capacitor is so configured that it confines the electric field lines within a small region of space. Thus, even though field may have considerable strength, the potential difference between the two conductors of a capacitor is small.
- Electric field is discontinuous across the surface of a spherical charged shell. It is zero inside and $\frac{\sigma}{\epsilon_0} \hat{\mathbf{n}}$ outside. Electric potential is, however continuous across the surface, equal to $q/4\pi\epsilon_0 R$ at the surface.
- The torque $\mathbf{p} \times \mathbf{E}$ on a dipole causes it to oscillate about \mathbf{E} . Only if there is a dissipative mechanism, the oscillations are damped and the dipole eventually aligns with \mathbf{E} .

5. Potential due to a charge q at its own location is not defined – it is infinite.
6. In the expression $qV(\mathbf{r})$ for potential energy of a charge q , $V(\mathbf{r})$ is the potential due to external charges and not the potential due to q . As seen in point 5, this expression will be ill-defined if $V(\mathbf{r})$ includes potential due to a charge q itself.
7. A cavity inside a conductor is shielded from outside electrical influences. It is worth noting that electrostatic shielding does not work the other way round; that is, if you put charges inside the cavity, the exterior of the conductor is not shielded from the fields by the inside charges.

EXERCISES

- 2.1** Two charges $5 \times 10^{-8} \text{ C}$ and $-3 \times 10^{-8} \text{ C}$ are located 16 cm apart. At what point(s) on the line joining the two charges is the electric potential zero? Take the potential at infinity to be zero.
- 2.2** A regular hexagon of side 10 cm has a charge $5 \mu\text{C}$ at each of its vertices. Calculate the potential at the centre of the hexagon.
- 2.3** Two charges $2 \mu\text{C}$ and $-2 \mu\text{C}$ are placed at points A and B 6 cm apart.
 - (a) Identify an equipotential surface of the system.
 - (b) What is the direction of the electric field at every point on this surface?
- 2.4** A spherical conductor of radius 12 cm has a charge of $1.6 \times 10^{-7} \text{ C}$ distributed uniformly on its surface. What is the electric field
 - (a) inside the sphere
 - (b) just outside the sphere
 - (c) at a point 18 cm from the centre of the sphere?
- 2.5** A parallel plate capacitor with air between the plates has a capacitance of 8 pF ($1\text{pF} = 10^{-12} \text{ F}$). What will be the capacitance if the distance between the plates is reduced by half, and the space between them is filled with a substance of dielectric constant 6?
- 2.6** Three capacitors each of capacitance 9 pF are connected in series.
 - (a) What is the total capacitance of the combination?
 - (b) What is the potential difference across each capacitor if the combination is connected to a 120 V supply?
- 2.7** Three capacitors of capacitances 2 pF , 3 pF and 4 pF are connected in parallel.
 - (a) What is the total capacitance of the combination?
 - (b) Determine the charge on each capacitor if the combination is connected to a 100 V supply.
- 2.8** In a parallel plate capacitor with air between the plates, each plate has an area of $6 \times 10^{-3} \text{ m}^2$ and the distance between the plates is 3 mm. Calculate the capacitance of the capacitor. If this capacitor is connected to a 100 V supply, what is the charge on each plate of the capacitor?

Physics

- 2.9** Explain what would happen if in the capacitor given in Exercise 2.8, a 3 mm thick mica sheet (of dielectric constant = 6) were inserted between the plates,
(a) while the voltage supply remained connected.
(b) after the supply was disconnected.
- 2.10** A 12pF capacitor is connected to a 50V battery. How much electrostatic energy is stored in the capacitor?
- 2.11** A 600pF capacitor is charged by a 200V supply. It is then disconnected from the supply and is connected to another uncharged 600 pF capacitor. How much electrostatic energy is lost in the process?

ADDITIONAL EXERCISES

- 2.12** A charge of 8 mC is located at the origin. Calculate the work done in taking a small charge of -2×10^{-9} C from a point P (0, 0, 3 cm) to a point Q (0, 4 cm, 0), via a point R (0, 6 cm, 9 cm).
- 2.13** A cube of side b has a charge q at each of its vertices. Determine the potential and electric field due to this charge array at the centre of the cube.
- 2.14** Two tiny spheres carrying charges 1.5 μC and 2.5 μC are located 30 cm apart. Find the potential and electric field:
(a) at the mid-point of the line joining the two charges, and
(b) at a point 10 cm from this midpoint in a plane normal to the line and passing through the mid-point.
- 2.15** A spherical conducting shell of inner radius r_1 and outer radius r_2 has a charge Q .
(a) A charge q is placed at the centre of the shell. What is the surface charge density on the inner and outer surfaces of the shell?
(b) Is the electric field inside a cavity (with no charge) zero, even if the shell is not spherical, but has any irregular shape? Explain.
- 2.16** (a) Show that the normal component of electrostatic field has a discontinuity from one side of a charged surface to another given by
- $$(\mathbf{E}_2 - \mathbf{E}_1) \cdot \hat{\mathbf{n}} = \frac{\sigma}{\epsilon_0}$$
- where $\hat{\mathbf{n}}$ is a unit vector normal to the surface at a point and σ is the surface charge density at that point. (The direction of $\hat{\mathbf{n}}$ is from side 1 to side 2.) Hence show that just outside a conductor, the electric field is $\sigma \hat{\mathbf{n}} / \epsilon_0$.
- (b) Show that the tangential component of electrostatic field is continuous from one side of a charged surface to another. [Hint: For (a), use Gauss's law. For, (b) use the fact that work done by electrostatic field on a closed loop is zero.]
- 2.17** A long charged cylinder of linear charged density λ is surrounded by a hollow co-axial conducting cylinder. What is the electric field in the space between the two cylinders?
- 2.18** In a hydrogen atom, the electron and proton are bound at a distance of about 0.53 Å:

Electrostatic Potential and Capacitance

- (a) Estimate the potential energy of the system in eV, taking the zero of the potential energy at infinite separation of the electron from proton.
- (b) What is the minimum work required to free the electron, given that its kinetic energy in the orbit is half the magnitude of potential energy obtained in (a)?
- (c) What are the answers to (a) and (b) above if the zero of potential energy is taken at 1.06 Å separation?
- 2.19** If one of the two electrons of a H_2 molecule is removed, we get a hydrogen molecular ion H_2^+ . In the ground state of an H_2^+ , the two protons are separated by roughly 1.5 Å, and the electron is roughly 1 Å from each proton. Determine the potential energy of the system. Specify your choice of the zero of potential energy.
- 2.20** Two charged conducting spheres of radii a and b are connected to each other by a wire. What is the ratio of electric fields at the surfaces of the two spheres? Use the result obtained to explain why charge density on the sharp and pointed ends of a conductor is higher than on its flatter portions.
- 2.21** Two charges $-q$ and $+q$ are located at points $(0, 0, -a)$ and $(0, 0, a)$, respectively.
- (a) What is the electrostatic potential at the points $(0, 0, z)$ and $(x, y, 0)$?
- (b) Obtain the dependence of potential on the distance r of a point from the origin when $r/a \gg 1$.
- (c) How much work is done in moving a small test charge from the point $(5, 0, 0)$ to $(-7, 0, 0)$ along the x -axis? Does the answer change if the path of the test charge between the same points is not along the x -axis?
- 2.22** Figure 2.34 shows a charge array known as an *electric quadrupole*. For a point on the axis of the quadrupole, obtain the dependence of potential on r for $r/a \gg 1$, and contrast your results with that due to an electric dipole, and an electric monopole (i.e., a single charge).

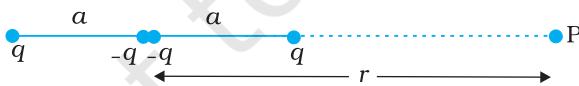


FIGURE 2.34

- 2.23** An electrical technician requires a capacitance of $2 \mu F$ in a circuit across a potential difference of 1 kV. A large number of $1 \mu F$ capacitors are available to him each of which can withstand a potential difference of not more than 400 V. Suggest a possible arrangement that requires the minimum number of capacitors.
- 2.24** What is the area of the plates of a $2 F$ parallel plate capacitor, given that the separation between the plates is 0.5 cm? [You will realise from your answer why ordinary capacitors are in the range of μF or less. However, electrolytic capacitors do have a much larger capacitance ($0.1 F$) because of very minute separation between the conductors.]

Physics

- 2.25** Obtain the equivalent capacitance of the network in Fig. 2.35. For a 300 V supply, determine the charge and voltage across each capacitor.

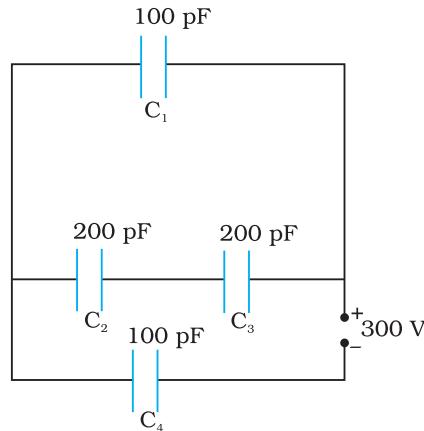


FIGURE 2.35

- 2.26** The plates of a parallel plate capacitor have an area of 90 cm^2 each and are separated by 2.5 mm. The capacitor is charged by connecting it to a 400 V supply.
- How much electrostatic energy is stored by the capacitor?
 - View this energy as stored in the electrostatic field between the plates, and obtain the energy per unit volume u . Hence arrive at a relation between u and the magnitude of electric field E between the plates.
- 2.27** A $4 \mu\text{F}$ capacitor is charged by a 200 V supply. It is then disconnected from the supply, and is connected to another uncharged $2 \mu\text{F}$ capacitor. How much electrostatic energy of the first capacitor is lost in the form of heat and electromagnetic radiation?
- 2.28** Show that the force on each plate of a parallel plate capacitor has a magnitude equal to $(\frac{1}{2}) QE$, where Q is the charge on the capacitor, and E is the magnitude of electric field between the plates. Explain the origin of the factor $\frac{1}{2}$.
- 2.29** A spherical capacitor consists of two concentric spherical conductors, held in position by suitable insulating supports (Fig. 2.36). Show

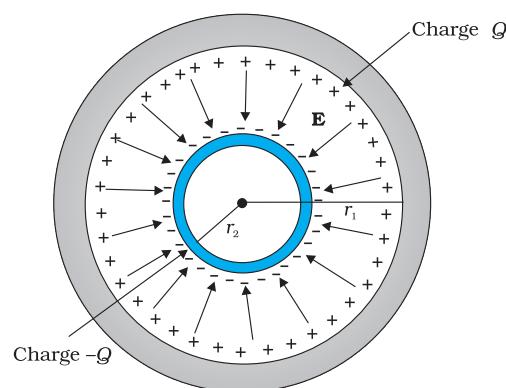


FIGURE 2.36

Electrostatic Potential and Capacitance

that the capacitance of a spherical capacitor is given by

$$C = \frac{4\pi\epsilon_0 r_1 r_2}{r_1 - r_2}$$

where r_1 and r_2 are the radii of outer and inner spheres, respectively.

- 2.30** A spherical capacitor has an inner sphere of radius 12 cm and an outer sphere of radius 13 cm. The outer sphere is earthed and the inner sphere is given a charge of $2.5 \mu\text{C}$. The space between the concentric spheres is filled with a liquid of dielectric constant 32.
- Determine the capacitance of the capacitor.
 - What is the potential of the inner sphere?
 - Compare the capacitance of this capacitor with that of an isolated sphere of radius 12 cm. Explain why the latter is much smaller.
- 2.31** Answer carefully:
- Two large conducting spheres carrying charges Q_1 and Q_2 are brought close to each other. Is the magnitude of electrostatic force between them exactly given by $Q_1 Q_2 / 4\pi\epsilon_0 r^2$, where r is the distance between their centres?
 - If Coulomb's law involved $1/r^3$ dependence (instead of $1/r^2$), would Gauss's law be still true?
 - A small test charge is released at rest at a point in an electrostatic field configuration. Will it travel along the field line passing through that point?
 - What is the work done by the field of a nucleus in a complete circular orbit of the electron? What if the orbit is elliptical?
 - We know that electric field is discontinuous across the surface of a charged conductor. Is electric potential also discontinuous there?
 - What meaning would you give to the capacitance of a single conductor?
 - Guess a possible reason why water has a much greater dielectric constant ($= 80$) than say, mica ($= 6$).
- 2.32** A cylindrical capacitor has two co-axial cylinders of length 15 cm and radii 1.5 cm and 1.4 cm. The outer cylinder is earthed and the inner cylinder is given a charge of $3.5 \mu\text{C}$. Determine the capacitance of the system and the potential of the inner cylinder. Neglect end effects (i.e., bending of field lines at the ends).
- 2.33** A parallel plate capacitor is to be designed with a voltage rating 1 kV, using a material of dielectric constant 3 and dielectric strength about 10^7 Vm^{-1} . (Dielectric strength is the maximum electric field a material can tolerate without breakdown, i.e., without starting to conduct electricity through partial ionisation.) For safety, we should like the field never to exceed, say 10% of the dielectric strength. What minimum area of the plates is required to have a capacitance of 50 pF?
- 2.34** Describe schematically the equipotential surfaces corresponding to
- a constant electric field in the z -direction,
 - a field that uniformly increases in magnitude but remains in a constant (say, z) direction,

Physics

- (c) a single positive charge at the origin, and
- (d) a uniform grid consisting of long equally spaced parallel charged wires in a plane.

2.35 In a Van de Graaff type generator a spherical metal shell is to be a 15×10^6 V electrode. The dielectric strength of the gas surrounding the electrode is 5×10^7 Vm^{-1} . What is the minimum radius of the spherical shell required? (You will learn from this exercise why one cannot build an electrostatic generator using a very small shell which requires a small charge to acquire a high potential.)

2.36 A small sphere of radius r_1 and charge q_1 is enclosed by a spherical shell of radius r_2 and charge q_2 . Show that if q_1 is positive, charge will necessarily flow from the sphere to the shell (when the two are connected by a wire) no matter what the charge q_2 on the shell is.

2.37 Answer the following:

(a) The top of the atmosphere is at about 400 kV with respect to the surface of the earth, corresponding to an electric field that decreases with altitude. Near the surface of the earth, the field is about 100 Vm^{-1} . Why then do we not get an electric shock as we step out of our house into the open? (Assume the house to be a steel cage so there is no field inside!)

(b) A man fixes outside his house one evening a two metre high insulating slab carrying on its top a large aluminium sheet of area 1m^2 . Will he get an electric shock if he touches the metal sheet next morning?

(c) The discharging current in the atmosphere due to the small conductivity of air is known to be 1800 A on an average over the globe. Why then does the atmosphere not discharge itself completely in due course and become electrically neutral? In other words, what keeps the atmosphere charged?

(d) What are the forms of energy into which the electrical energy of the atmosphere is dissipated during a lightning?
(Hint: The earth has an electric field of about 100 Vm^{-1} at its surface in the downward direction, corresponding to a surface charge density = $-10^{-9} \text{ C m}^{-2}$. Due to the slight conductivity of the atmosphere up to about 50 km (beyond which it is good conductor), about + 1800 C is pumped every second into the earth as a whole. The earth, however, does not get discharged since thunderstorms and lightning occurring continually all over the globe pump an equal amount of negative charge on the earth.)

Chapter Three

CURRENT

ELECTRICITY



3.1 INTRODUCTION

In Chapter 1, all charges whether free or bound, were considered to be at rest. Charges in motion constitute an electric current. Such currents occur naturally in many situations. Lightning is one such phenomenon in which charges flow from the clouds to the earth through the atmosphere, sometimes with disastrous results. The flow of charges in lightning is not steady, but in our everyday life we see many devices where charges flow in a steady manner, like water flowing smoothly in a river. A torch and a cell-driven clock are examples of such devices. In the present chapter, we shall study some of the basic laws concerning steady electric currents.

3.2 ELECTRIC CURRENT

Imagine a small area held normal to the direction of flow of charges. Both the positive and the negative charges may flow forward and backward across the area. In a given time interval t , let q_+ be the net amount (*i.e.*, forward *minus* backward) of positive charge that flows in the forward direction across the area. Similarly, let q_- be the net amount of negative charge flowing across the area in the forward direction. The net amount of charge flowing across the area in the forward direction in the time interval t , then, is $q = q_+ - q_-$. This is proportional to t for steady current

and the quotient

$$I = \frac{q}{t} \quad (3.1)$$

is defined to be the *current* across the area in the forward direction. (If it turns out to be a negative number, it implies a current in the backward direction.)

Currents are not always steady and hence more generally, we define the current as follows. Let ΔQ be the net charge flowing across a cross-section of a conductor during the time interval Δt [i.e., between times t and $(t + \Delta t)$]. Then, the current at time t across the cross-section of the conductor is defined as the value of the ratio of ΔQ to Δt in the limit of Δt tending to zero,

$$I(t) = \lim_{\Delta t \rightarrow 0} \frac{\Delta Q}{\Delta t} \quad (3.2)$$

In SI units, the unit of current is ampere. An ampere is defined through magnetic effects of currents that we will study in the following chapter. An ampere is typically the order of magnitude of currents in domestic appliances. An average lightning carries currents of the order of tens of thousands of amperes and at the other extreme, currents in our nerves are in microamperes.

3.3 ELECTRIC CURRENTS IN CONDUCTORS

An electric charge will experience a force if an electric field is applied. If it is free to move, it will thus move contributing to a current. In nature, free charged particles do exist like in upper strata of atmosphere called the *ionosphere*. However, in atoms and molecules, the negatively charged electrons and the positively charged nuclei are bound to each other and are thus not free to move. Bulk matter is made up of many molecules, a gram of water, for example, contains approximately 10^{22} molecules. These molecules are so closely packed that the electrons are no longer attached to individual nuclei. In some materials, the electrons will still be bound, i.e., they will not accelerate even if an electric field is applied. In other materials, notably metals, some of the electrons are practically free to move within the bulk material. These materials, generally called conductors, develop electric currents in them when an electric field is applied.

If we consider solid conductors, then of course the atoms are tightly bound to each other so that the current is carried by the negatively charged electrons. There are, however, other types of conductors like electrolytic solutions where positive and negative charges both can move. In our discussions, we will focus only on solid conductors so that the current is carried by the negatively charged electrons in the background of fixed positive ions.

Consider first the case when no electric field is present. The electrons will be moving due to thermal motion during which they collide with the fixed ions. An electron colliding with an ion emerges with the same speed as before the collision. However, the direction of its velocity after the collision is completely random. At a given time, there is no preferential direction for the velocities of the electrons. Thus on the average, the

number of electrons travelling in any direction will be equal to the number of electrons travelling in the opposite direction. So, there will be no net electric current.

Let us now see what happens to such a piece of conductor if an electric field is applied. To focus our thoughts, imagine the conductor in the shape of a cylinder of radius R (Fig. 3.1). Suppose we now take two thin circular discs of a dielectric of the same radius and put positive charge $+Q$ distributed over one disc and similarly $-Q$ at the other disc. We attach the two discs on the two flat surfaces of the cylinder. An electric field will be created and is directed from the positive towards the negative charge. The electrons will be accelerated due to this field towards $+Q$. They will thus move to neutralise the charges. The electrons, as long as they are moving, will constitute an electric current. Hence in the situation considered, there will be a current for a very short while and no current thereafter.

We can also imagine a mechanism where the ends of the cylinder are supplied with fresh charges to make up for any charges neutralised by electrons moving inside the conductor. In that case, there will be a steady electric field in the body of the conductor. This will result in a continuous current rather than a current for a short period of time. Mechanisms, which maintain a steady electric field are cells or batteries that we shall study later in this chapter. In the next sections, we shall study the steady current that results from a steady electric field in conductors.

3.4 OHM'S LAW

A basic law regarding flow of currents was discovered by G.S. Ohm in 1828, long before the physical mechanism responsible for flow of currents was discovered. Imagine a conductor through which a current I is flowing and let V be the potential difference between the ends of the conductor. Then Ohm's law states that

$$V \propto I$$

$$\text{or, } V = RI \quad (3.3)$$

where the constant of proportionality R is called the *resistance* of the conductor. The SI units of resistance is *ohm*, and is denoted by the symbol Ω . The resistance R not only depends on the material of the conductor but also on the dimensions of the conductor. The dependence of R on the dimensions of the conductor can easily be determined as follows.

Consider a conductor satisfying Eq. (3.3) to be in the form of a slab of length l and cross sectional area A [Fig. 3.2(a)]. Imagine placing two such identical slabs side by side [Fig. 3.2(b)], so that the length of the combination is $2l$. The current flowing through the combination is the same as that flowing through either of the slabs. If V is the potential difference across the ends of the first slab, then V is also the potential difference across the ends of the second slab since the second slab is

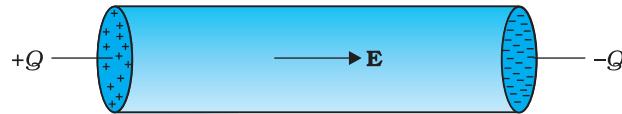


FIGURE 3.1 Charges $+Q$ and $-Q$ put at the ends of a metallic cylinder. The electrons will drift because of the electric field created to neutralise the charges. The current thus will stop after a while unless the charges $+Q$ and $-Q$ are continuously replenished.

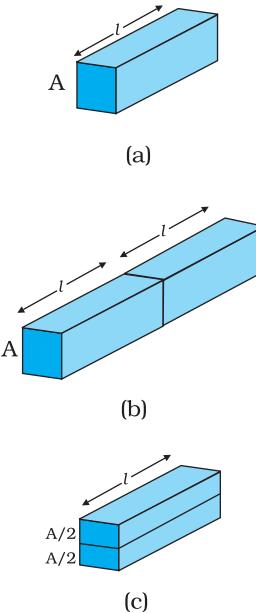


FIGURE 3.2
Illustrating the relation $R = \rho l / A$ for a rectangular slab of length l and area of cross-section A .

Physics

GEORG SIMON OHM (1787-1854)



Georg Simon Ohm (1787-1854) German physicist, professor at Munich. Ohm was led to his law by an analogy between the conduction of heat: the electric field is analogous to the temperature gradient, and the electric current is analogous to the heat flow.

identical to the first and the same current I flows through both. The potential difference across the ends of the combination is clearly sum of the potential difference across the two individual slabs and hence equals $2V$. The current through the combination is I and the resistance of the combination R_C is [from Eq. (3.3)],

$$R_C = \frac{2V}{I} = 2R \quad (3.4)$$

since $V/I = R$, the resistance of either of the slabs. Thus, doubling the length of a conductor doubles the resistance. In general, then resistance is proportional to length,

$$R \propto l \quad (3.5)$$

Next, imagine dividing the slab into two by cutting it lengthwise so that the slab can be considered as a combination of two identical slabs of length l , but each having a cross sectional area of $A/2$ [Fig. 3.2(c)].

For a given voltage V across the slab, if I is the current through the entire slab, then clearly the current flowing through each of the two half-slabs is $I/2$. Since the potential difference across the ends of the half-slabs is V , i.e., the same as across the full slab, the resistance of each of the half-slabs R_1 is

$$R_1 = \frac{V}{(I/2)} = 2 \frac{V}{I} = 2R. \quad (3.6)$$

Thus, halving the area of the cross-section of a conductor doubles the resistance. In general, then the resistance R is inversely proportional to the cross-sectional area,

$$R \propto \frac{1}{A} \quad (3.7)$$

Combining Eqs. (3.5) and (3.7), we have

$$R \propto \frac{l}{A} \quad (3.8)$$

and hence for a given conductor

$$R = \rho \frac{l}{A} \quad (3.9)$$

where the constant of proportionality ρ depends on the material of the conductor but not on its dimensions. ρ is called *resistivity*.

Using the last equation, Ohm's law reads

$$V = I \times R = \frac{I\rho}{A} \quad (3.10)$$

Current per unit area (taken normal to the current), I/A , is called *current density* and is denoted by j . The SI units of the current density are A/m^2 . Further, if E is the magnitude of uniform electric field in the conductor whose length is l , then the potential difference V across its ends is El . Using these, the last equation reads

$$E l = j \rho l$$

$$\text{or, } E = j \rho \quad (3.11)$$

The above relation for *magnitudes* E and j can indeed be cast in a *vector* form. The current density, (which we have defined as the current through unit area *normal* to the current) is also directed along \mathbf{E} , and is also a vector \mathbf{j} ($\equiv j \mathbf{E}/E$). Thus, the last equation can be written as,

$$\mathbf{E} = \mathbf{j} \rho \quad (3.12)$$

$$\text{or, } \mathbf{j} = \sigma \mathbf{E} \quad (3.13)$$

where $\sigma \equiv 1/\rho$ is called the *conductivity*. Ohm's law is often stated in an equivalent form, Eq. (3.13) in addition to Eq.(3.3). In the next section, we will try to understand the origin of the Ohm's law as arising from the characteristics of the drift of electrons.

3.5 DRIFT OF ELECTRONS AND THE ORIGIN OF RESISTIVITY

As remarked before, an electron will suffer collisions with the heavy fixed ions, but after collision, it will emerge with the same speed but in random directions. If we consider all the electrons, their average velocity will be zero since their directions are random. Thus, if there are N electrons and the velocity of the i^{th} electron ($i = 1, 2, 3, \dots, N$) at a given time is \mathbf{v}_i , then

$$\frac{1}{N} \sum_{i=1}^N \mathbf{v}_i = 0 \quad (3.14)$$

Consider now the situation when an electric field is present. Electrons will be accelerated due to this field by

$$\mathbf{a} = \frac{-e \mathbf{E}}{m} \quad (3.15)$$

where $-e$ is the charge and m is the mass of an electron. Consider again the i^{th} electron at a given time t . This electron would have had its last collision some time before t , and let t_i be the time elapsed after its last collision. If \mathbf{v}_i was its velocity immediately after the last collision, then its velocity \mathbf{V}_i at time t is

$$\mathbf{V}_i = \mathbf{v}_i - \frac{e \mathbf{E}}{m} t_i \quad (3.16)$$

since starting with its last collision it was accelerated (Fig. 3.3) with an acceleration given by Eq. (3.15) for a time interval t_i . The average velocity of the electrons at time t is the average of all the \mathbf{V}_i 's. The average of \mathbf{v}_i 's is zero [Eq. (3.14)] since immediately after any collision, the direction of the velocity of an electron is completely random. The collisions of the electrons do not occur at regular intervals but at random times. Let us denote by τ , the average time between successive collisions. Then at a given time, some of the electrons would have spent

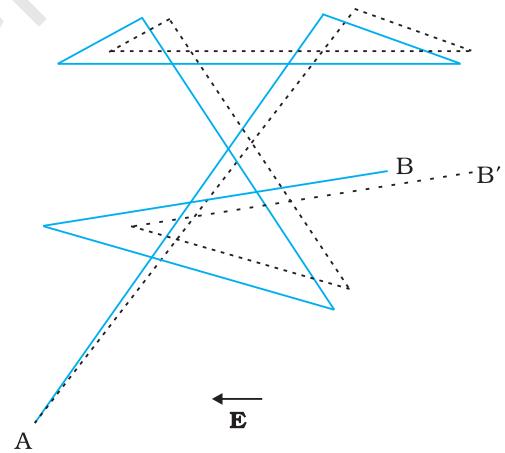


FIGURE 3.3 A schematic picture of an electron moving from a point A to another point B through repeated collisions, and straight line travel between collisions (full lines). If an electric field is applied as shown, the electron ends up at point B' (dotted lines). A slight drift in a direction opposite the electric field is visible.

time more than τ and some less than τ . In other words, the time t_i in Eq. (3.16) will be less than τ for some and more than τ for others as we go through the values of $i = 1, 2, \dots, N$. The average value of t_i then is τ (known as *relaxation time*). Thus, averaging Eq. (3.16) over the N -electrons at any given time t gives us for the average velocity \mathbf{v}_d

$$\begin{aligned}\mathbf{v}_d &\equiv (\mathbf{v}_i)_{\text{average}} = (\mathbf{v}_i)_{\text{average}} - \frac{e\mathbf{E}}{m} (t_i)_{\text{average}} \\ &= 0 - \frac{e\mathbf{E}}{m} \tau = -\frac{e\mathbf{E}}{m} \tau\end{aligned}\quad (3.17)$$

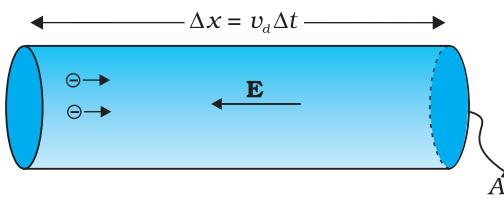


FIGURE 3.4 Current in a metallic conductor. The magnitude of current density in a metal is the magnitude of charge contained in a cylinder of unit area and length v_d .

This last result is surprising. It tells us that the electrons move with an average velocity which is independent of time, although electrons are accelerated. This is the phenomenon of drift and the velocity \mathbf{v}_d in Eq. (3.17) is called the *drift velocity*.

Because of the drift, there will be net transport of charges across any area perpendicular to \mathbf{E} . Consider a planar area A , located inside the conductor such that the normal to the area is parallel to \mathbf{E} (Fig. 3.4). Then because of the drift, in an infinitesimal amount of time Δt , all electrons to the left of the area at distances up to $|\mathbf{v}_d| \Delta t$ would have crossed the area. If n is the number of free electrons per unit volume in the metal, then there are $n \Delta t |\mathbf{v}_d| A$ such electrons.

Since each electron carries a charge $-e$, the total charge transported across this area A to the right in time Δt is $-ne A |\mathbf{v}_d| \Delta t$. \mathbf{E} is directed towards the left and hence the total charge transported along \mathbf{E} across the area is negative of this. The amount of charge crossing the area A in time Δt is by definition [Eq. (3.2)] $I \Delta t$, where I is the magnitude of the current. Hence,

$$I \Delta t = +ne A |\mathbf{v}_d| \Delta t \quad (3.18)$$

Substituting the value of $|\mathbf{v}_d|$ from Eq. (3.17)

$$I \Delta t = \frac{e^2 A}{m} \tau n \Delta t |\mathbf{E}| \quad (3.19)$$

By definition I is related to the magnitude $|\mathbf{j}|$ of the current density by

$$I = |\mathbf{j}| A \quad (3.20)$$

Hence, from Eqs.(3.19) and (3.20),

$$|\mathbf{j}| = \frac{ne^2}{m} \tau |\mathbf{E}| \quad (3.21)$$

The vector \mathbf{j} is parallel to \mathbf{E} and hence we can write Eq. (3.21) in the vector form

$$\mathbf{j} = \frac{ne^2}{m} \tau \mathbf{E} \quad (3.22)$$

Comparison with Eq. (3.13) shows that Eq. (3.22) is exactly the Ohm's law, if we identify the conductivity σ as

$$\sigma = \frac{ne^2}{m} \tau \quad (3.23)$$

We thus see that a very simple picture of electrical conduction reproduces Ohm's law. We have, of course, made assumptions that τ and n are constants, independent of E . We shall, in the next section, discuss the limitations of Ohm's law.

Example 3.1 (a) Estimate the average drift speed of conduction electrons in a copper wire of cross-sectional area $1.0 \times 10^{-7} \text{ m}^2$ carrying a current of 1.5 A . Assume that each copper atom contributes roughly one conduction electron. The density of copper is $9.0 \times 10^3 \text{ kg/m}^3$, and its atomic mass is 63.5 u . (b) Compare the drift speed obtained above with, (i) thermal speeds of copper atoms at ordinary temperatures, (ii) speed of propagation of electric field along the conductor which causes the drift motion.

Solution

(a) The direction of drift velocity of conduction electrons is opposite to the electric field direction, i.e., electrons drift in the direction of increasing potential. The drift speed v_d is given by Eq. (3.18)

$$v_d = (I/neA)$$

Now, $e = 1.6 \times 10^{-19} \text{ C}$, $A = 1.0 \times 10^{-7} \text{ m}^2$, $I = 1.5 \text{ A}$. The density of conduction electrons, n is equal to the number of atoms per cubic metre (assuming one conduction electron per Cu atom as is reasonable from its valence electron count of one). A cubic metre of copper has a mass of $9.0 \times 10^3 \text{ kg}$. Since 6.0×10^{23} copper atoms have a mass of 63.5 g ,

$$\begin{aligned} n &= \frac{6.0 \times 10^{23}}{63.5} \times 9.0 \times 10^6 \\ &= 8.5 \times 10^{28} \text{ m}^{-3} \end{aligned}$$

which gives,

$$\begin{aligned} v_d &= \frac{1.5}{8.5 \times 10^{28} \times 1.6 \times 10^{-19} \times 1.0 \times 10^{-7}} \\ &= 1.1 \times 10^{-3} \text{ m s}^{-1} = 1.1 \text{ mm s}^{-1} \end{aligned}$$

- (b) (i) At a temperature T , the thermal speed* of a copper atom of mass M is obtained from $\langle (1/2) Mv^2 \rangle = (3/2) k_B T$ and is thus typically of the order of $\sqrt{k_B T/M}$, where k_B is the Boltzmann constant. For copper at 300 K , this is about $2 \times 10^2 \text{ m/s}$. This figure indicates the random vibrational speeds of copper atoms in a conductor. Note that the drift speed of electrons is much smaller, about 10^{-5} times the typical thermal speed at ordinary temperatures.
(ii) An electric field travelling along the conductor has a speed of an electromagnetic wave, namely equal to $3.0 \times 10^8 \text{ m s}^{-1}$ (You will learn about this in Chapter 8). The drift speed is, in comparison, extremely small; smaller by a factor of 10^{-11} .

EXAMPLE 3.1

* See Eq. (13.23) of Chapter 13 from Class XI book.

EXAMPLE 3.2
Example 3.2

- In Example 3.1, the electron drift speed is estimated to be only a few mm s^{-1} for currents in the range of a few amperes? How then is current established almost the instant a circuit is closed?
- The electron drift arises due to the force experienced by electrons in the electric field inside the conductor. But force should cause acceleration. Why then do the electrons acquire a steady average drift speed?
- If the electron drift speed is so small, and the electron's charge is small, how can we still obtain large amounts of current in a conductor?
- When electrons drift in a metal from lower to higher potential, does it mean that all the 'free' electrons of the metal are moving in the same direction?
- Are the paths of electrons straight lines between successive collisions (with the positive ions of the metal) in the (i) absence of electric field, (ii) presence of electric field?

Solution

- Electric field is established throughout the circuit, almost instantly (with the speed of light) causing at every point a *local electron drift*. Establishment of a current does not have to wait for electrons from one end of the conductor travelling to the other end. However, it does take a little while for the current to reach its steady value.
- Each 'free' electron does accelerate, increasing its drift speed until it collides with a positive ion of the metal. It loses its drift speed after collision but starts to accelerate and increases its drift speed again only to suffer a collision again and so on. On the average, therefore, electrons acquire only a drift speed.
- Simple, because the electron number density is enormous, $\sim 10^{29} \text{ m}^{-3}$.
- By no means. The drift velocity is superposed over the large random velocities of electrons.
- In the absence of electric field, the paths are straight lines; in the presence of electric field, the paths are, in general, curved.

3.5.1 Mobility

As we have seen, conductivity arises from mobile charge carriers. In metals, these mobile charge carriers are electrons; in an ionised gas, they are electrons and positive charged ions; in an electrolyte, these can be both positive and negative ions.

An important quantity is the *mobility* μ defined as the magnitude of the drift velocity per unit electric field:

$$\mu = \frac{|\mathbf{v}_d|}{E} \quad (3.24)$$

The SI unit of mobility is m^2/Vs and is 10^4 of the mobility in practical units (cm^2/Vs). Mobility is positive. From Eq. (3.17), we have

$$v_d = \frac{e \tau E}{m}$$

Hence,

$$\mu = \frac{v_d}{E} = \frac{e \tau}{m}$$

where τ is the average collision time for electrons.

3.6 LIMITATIONS OF OHM'S LAW

Although Ohm's law has been found valid over a large class of materials, there do exist materials and devices used in electric circuits where the proportionality of V and I does not hold. The deviations broadly are one or more of the following types:

- (a) V ceases to be proportional to I (Fig. 3.5).
- (b) The relation between V and I depends on the sign of V . In other words, if I is the current for a certain V , then reversing the direction of V keeping its magnitude fixed, does not produce a current of the same magnitude as I in the opposite direction (Fig. 3.6). This happens, for example, in a diode which we will study in Chapter 14.

(3.25)

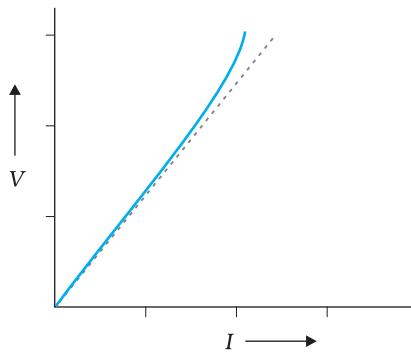


FIGURE 3.5 The dashed line represents the linear Ohm's law. The solid line is the voltage V versus current I for a good conductor.

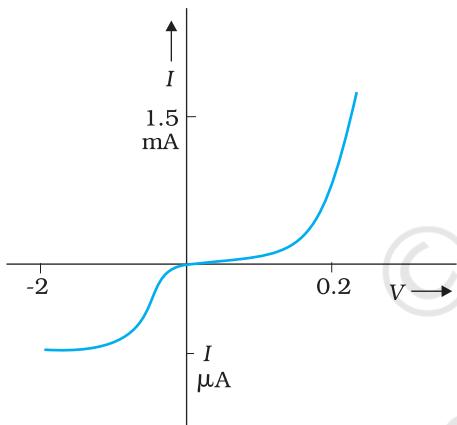


FIGURE 3.6 Characteristic curve of a diode. Note the different scales for negative and positive values of the voltage and current.

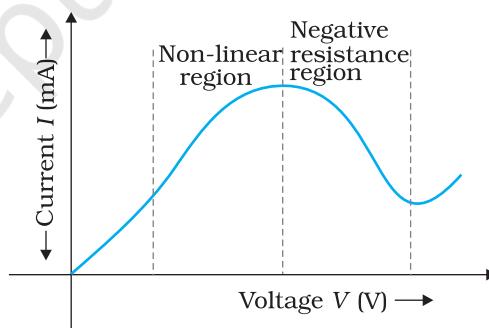


FIGURE 3.7 Variation of current versus voltage for GaAs.

- (c) The relation between V and I is not unique, i.e., there is more than one value of V for the same current I (Fig. 3.7). A material exhibiting such behaviour is GaAs.

Materials and devices not obeying Ohm's law in the form of Eq. (3.3) are actually widely used in electronic circuits. In this and a few subsequent chapters, however, we will study the electrical currents in materials that obey Ohm's law.

3.7 RESISTIVITY OF VARIOUS MATERIALS

The resistivities of various common materials are listed in Table 3.1. The materials are classified as conductors, semiconductors and insulators

Physics

depending on their resistivities, in an increasing order of their values. Metals have low resistivities in the range of $10^{-8} \Omega\text{m}$ to $10^{-6} \Omega\text{m}$. At the other end are insulators like ceramic, rubber and plastics having resistivities 10^{18} times greater than metals or more. In between the two are the semiconductors. These, however, have resistivities characteristically decreasing with a rise in temperature. The resistivities of semiconductors are also affected by presence of small amount of impurities. This last feature is exploited in use of semiconductors for electronic devices.

TABLE 3.1 RESISTIVITIES OF SOME MATERIALS

Material	Resistivity, ρ [$\Omega\text{ m}$] at 0°C	Temperature coefficient of resistivity, α [$^\circ\text{C}$] ⁻¹ $\frac{1}{\rho} \left(\frac{d\rho}{dT} \right)$ at 0°C
Conductors		
Silver	1.6×10^{-8}	0.0041
Copper	1.7×10^{-8}	0.0068
Aluminium	2.7×10^{-8}	0.0043
Tungsten	5.6×10^{-8}	0.0045
Iron	10×10^{-8}	0.0065
Platinum	11×10^{-8}	0.0039
Mercury	98×10^{-8}	0.0009
Nichrome (alloy of Ni, Fe, Cr)	$\sim 100 \times 10^{-8}$	0.0004
Manganin (alloy)	48×10^{-8}	0.002×10^{-3}
Semiconductors		
Carbon (graphite)	3.5×10^{-5}	- 0.0005
Germanium	0.46	- 0.05
Silicon	2300	- 0.07
Insulators		
Pure Water	2.5×10^5	
Glass	$10^{10} - 10^{14}$	
Hard Rubber	$10^{13} - 10^{16}$	
NaCl	$\sim 10^{14}$	
Fused Quartz	$\sim 10^{16}$	

Commercially produced resistors for domestic use or in laboratories are of two major types: *wire bound resistors* and *carbon resistors*. Wire bound resistors are made by winding the wires of an alloy, viz., manganin, constantan, nichrome or similar ones. The choice of these materials is dictated mostly by the fact that their resistivities are relatively insensitive to temperature. These resistances are typically in the range of a fraction of an ohm to a few hundred ohms.

Resistors in the higher range are made mostly from carbon. Carbon resistors are compact, inexpensive and thus find extensive use in electronic circuits. Carbon resistors are small in size and hence their values are given using a colour code.

TABLE 3.2 RESISTOR COLOUR CODES

Colour	Number	Multiplier	Tolerance (%)
Black	0	1	
Brown	1	10^1	
Red	2	10^2	
Orange	3	10^3	
Yellow	4	10^4	
Green	5	10^5	
Blue	6	10^6	
Violet	7	10^7	
Gray	8	10^8	
White	9	10^9	
Gold		10^{-1}	5
Silver		10^{-2}	10
No colour			20

The resistors have a set of co-axial coloured rings on them whose significance are listed in Table 3.2. The first two *bands* from the end indicate the first two significant figures of the resistance in ohms. The third band indicates the decimal multiplier (as listed in Table 3.2). The last band stands for tolerance or possible variation in percentage about the indicated values. Sometimes, this last band is absent and that indicates a tolerance of 20% (Fig. 3.8). For example, if the four colours are orange, blue, yellow and gold, the resistance value is $36 \times 10^4 \Omega$, with a tolerance value of 5%.

3.8 TEMPERATURE DEPENDENCE OF RESISTIVITY

The resistivity of a material is found to be dependent on the temperature. Different materials do not exhibit the same dependence on temperatures. Over a limited range of temperatures, that is not too large, the resistivity of a metallic conductor is approximately given by,

$$\rho_T = \rho_0 [1 + \alpha (T - T_0)] \quad (3.26)$$

where ρ_T is the resistivity at a temperature T and ρ_0 is the same at a reference temperature T_0 . α is called the *temperature co-efficient of resistivity*, and from Eq. (3.26), the dimension of α is $(\text{Temperature})^{-1}$.

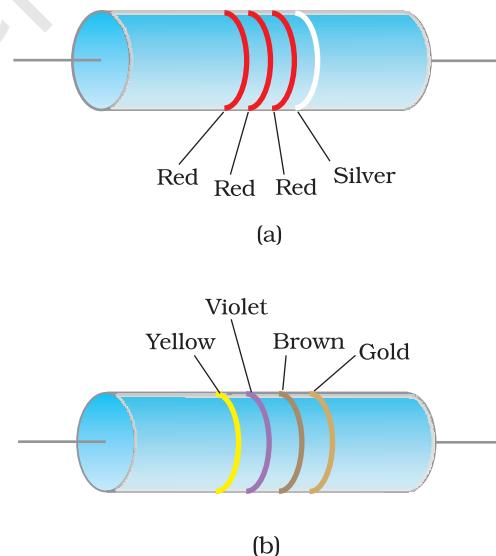


FIGURE 3.8 Colour coded resistors
 (a) $(22 \times 10^2 \Omega) \pm 10\%$,
 (b) $(47 \times 10 \Omega) \pm 5\%$.

Physics

For metals, α is positive and values of α for some metals at $T_0 = 0^\circ\text{C}$ are listed in Table 3.1.

The relation of Eq. (3.26) implies that a graph of ρ_T plotted against T would be a straight line. At temperatures much lower than 0°C , the graph, however, deviates considerably from a straight line (Fig. 3.9).

Equation (3.26) thus, can be used approximately over a limited range of T around any reference temperature T_0 , where the graph can be approximated as a straight line.

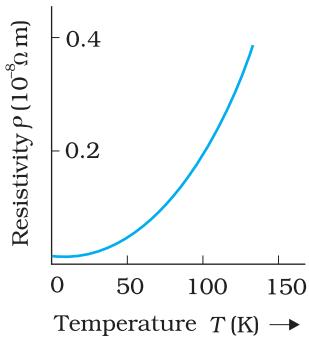


FIGURE 3.9
Resistivity ρ_T of copper as a function of temperature T .

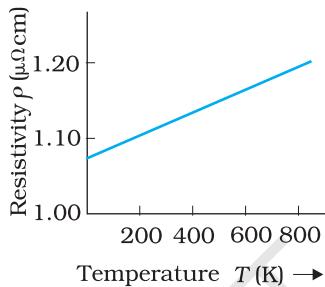


FIGURE 3.10 Resistivity ρ_T of nichrome as a function of absolute temperature T .

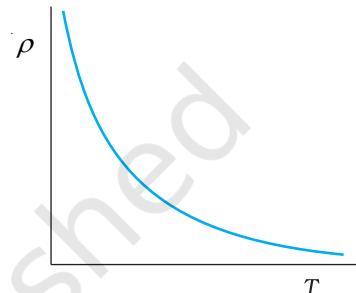


FIGURE 3.11
Temperature dependence of resistivity for a typical semiconductor.

Some materials like Nichrome (which is an alloy of nickel, iron and chromium) exhibit a very weak dependence of resistivity with temperature (Fig. 3.10). Manganin and constantan have similar properties. These materials are thus widely used in wire bound standard resistors since their resistance values would change very little with temperatures.

Unlike metals, the resistivities of semiconductors decrease with increasing temperatures. A typical dependence is shown in Fig. 3.11.

We can qualitatively understand the temperature dependence of resistivity, in the light of our derivation of Eq. (3.23). From this equation, resistivity of a material is given by

$$\rho = \frac{1}{\sigma} = \frac{m}{n e^2 \tau} \quad (3.27)$$

ρ thus depends inversely both on the number n of free electrons per unit volume and on the average time τ between collisions. As we increase temperature, average speed of the electrons, which act as the carriers of current, increases resulting in more frequent collisions. The average time of collisions τ , thus decreases with temperature.

In a metal, n is not dependent on temperature to any appreciable extent and thus the decrease in the value of τ with rise in temperature causes ρ to increase as we have observed.

For insulators and semiconductors, however, n increases with temperature. This increase more than compensates any decrease in τ in Eq.(3.23) so that for such materials, ρ decreases with temperature.

Example 3.3 An electric toaster uses nichrome for its heating element. When a negligibly small current passes through it, its resistance at room temperature ($27.0\text{ }^{\circ}\text{C}$) is found to be $75.3\text{ }\Omega$. When the toaster is connected to a 230 V supply, the current settles, after a few seconds, to a steady value of 2.68 A . What is the steady temperature of the nichrome element? The temperature coefficient of resistance of nichrome averaged over the temperature range involved, is $1.70 \times 10^{-4}\text{ }^{\circ}\text{C}^{-1}$.

Solution When the current through the element is very small, heating effects can be ignored and the temperature T_1 of the element is the same as room temperature. When the toaster is connected to the supply, its initial current will be slightly higher than its steady value of 2.68 A . But due to heating effect of the current, the temperature will rise. This will cause an increase in resistance and a slight decrease in current. In a few seconds, a steady state will be reached when temperature will rise no further, and both the resistance of the element and the current drawn will achieve steady values. The resistance R_2 at the steady temperature T_2 is

$$R_2 = \frac{230\text{ V}}{2.68\text{ A}} = 85.8\text{ }\Omega$$

Using the relation

$$R_2 = R_1 [1 + \alpha (T_2 - T_1)]$$

with $\alpha = 1.70 \times 10^{-4}\text{ }^{\circ}\text{C}^{-1}$, we get

$$T_2 - T_1 = \frac{(85.8 - 75.3)}{(75.3) \times 1.70 \times 10^{-4}} = 820\text{ }^{\circ}\text{C}$$

that is, $T_2 = (820 + 27.0)\text{ }^{\circ}\text{C} = 847\text{ }^{\circ}\text{C}$

Thus, the steady temperature of the heating element (when heating effect due to the current equals heat loss to the surroundings) is $847\text{ }^{\circ}\text{C}$.

EXAMPLE 3.3

Example 3.4 The resistance of the platinum wire of a platinum resistance thermometer at the ice point is $5\text{ }\Omega$ and at steam point is $5.39\text{ }\Omega$. When the thermometer is inserted in a hot bath, the resistance of the platinum wire is $5.795\text{ }\Omega$. Calculate the temperature of the bath.

Solution $R_0 = 5\text{ }\Omega$, $R_{100} = 5.23\text{ }\Omega$ and $R_t = 5.795\text{ }\Omega$

$$\begin{aligned} \text{Now, } t &= \frac{R_t - R_0}{R_{100} - R_0} \times 100, \quad R_t = R_0 (1 + \alpha t) \\ &= \frac{5.795 - 5}{5.23 - 5} \times 100 \\ &= \frac{0.795}{0.23} \times 100 = 345.65\text{ }^{\circ}\text{C} \end{aligned}$$

EXAMPLE 3.4

3.9 ELECTRICAL ENERGY, POWER

Consider a conductor with end points A and B, in which a current I is flowing from A to B. The electric potential at A and B are denoted by $V(A)$

Physics

and $V(B)$ respectively. Since current is flowing from A to B, $V(A) > V(B)$ and the potential difference across AB is $V = V(A) - V(B) > 0$.

In a time interval Δt , an amount of charge $\Delta Q = I \Delta t$ travels from A to B. The potential energy of the charge at A, by definition, was $Q V(A)$ and similarly at B, it is $Q V(B)$. Thus, change in its potential energy ΔU_{pot} is

$$\begin{aligned}\Delta U_{\text{pot}} &= \text{Final potential energy} - \text{Initial potential energy} \\ &= \Delta Q[(V(B) - V(A))] = -\Delta Q V \\ &= -I V \Delta t < 0\end{aligned}\quad (3.28)$$

If charges moved without collisions through the conductor, their kinetic energy would also change so that the total energy is unchanged. Conservation of total energy would then imply that,

$$\Delta K = -\Delta U_{\text{pot}} \quad (3.29)$$

that is,

$$\Delta K = I V \Delta t > 0 \quad (3.30)$$

Thus, in case charges were moving freely through the conductor under the action of electric field, their kinetic energy would increase as they move. We have, however, seen earlier that on the average, charge carriers do not move with acceleration but with a steady drift velocity. This is because of the collisions with ions and atoms during transit. During collisions, the energy gained by the charges thus is shared with the atoms. The atoms vibrate more vigorously, i.e., the conductor heats up. Thus, in an actual conductor, an amount of energy dissipated as heat in the conductor during the time interval Δt is,

$$\Delta W = I V \Delta t \quad (3.31)$$

The energy dissipated per unit time is the power dissipated $P = \Delta W / \Delta t$ and we have,

$$P = I V \quad (3.32)$$

Using Ohm's law $V = IR$, we get

$$P = I^2 R = V^2 / R \quad (3.33)$$

as the power loss ("ohmic loss") in a conductor of resistance R carrying a current I . It is this power which heats up, for example, the coil of an electric bulb to incandescence, radiating out heat and light.

Where does the power come from? As we have reasoned before, we need an external source to keep a steady current through the conductor. It is clearly this source which must supply this power. In the simple circuit shown with a cell (Fig. 3.12), it is the chemical energy of the cell which supplies this power for as long as it can.

The expressions for power, Eqs. (3.32) and (3.33), show the dependence of the power dissipated in a resistor R on the current through it and the voltage across it.

Equation (3.33) has an important application to power transmission. Electrical power is transmitted from power stations to homes and factories, which

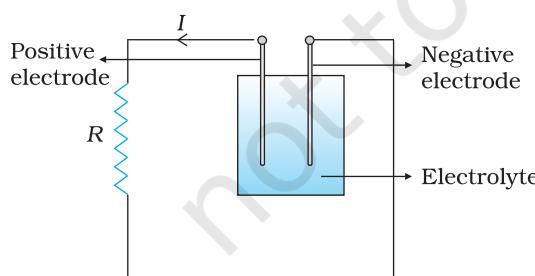


FIGURE 3.12 Heat is produced in the resistor R which is connected across the terminals of a cell. The energy dissipated in the resistor R comes from the chemical energy of the electrolyte.

may be hundreds of miles away, via transmission cables. One obviously wants to minimise the power loss in the transmission cables connecting the power stations to homes and factories. We shall see now how this can be achieved. Consider a device R , to which a power P is to be delivered via transmission cables having a resistance R_c to be dissipated by it finally. If V is the voltage across R and I the current through it, then

$$P = VI \quad (3.34)$$

The connecting wires from the power station to the device has a finite resistance R_c . The power dissipated in the connecting wires, which is wasted is P_c with

$$\begin{aligned} P_c &= I^2 R_c \\ &= \frac{P^2 R_c}{V^2} \end{aligned} \quad (3.35)$$

from Eq. (3.32). Thus, to drive a device of power P , the power wasted in the connecting wires is inversely proportional to V^2 . The transmission cables from power stations are hundreds of miles long and their resistance R_c is considerable. To reduce P_c , these wires carry current at enormous values of V and this is the reason for the high voltage danger signs on transmission lines — a common sight as we move away from populated areas. Using electricity at such voltages is not safe and hence at the other end, a device called a transformer lowers the voltage to a value suitable for use.

3.10 COMBINATION OF RESISTORS – SERIES AND PARALLEL

The current through a single resistor R across which there is a potential difference V is given by Ohm's law $I = V/R$. Resistors are sometimes joined together and there are simple rules for calculation of equivalent resistance of such combination.



FIGURE 3.13 A series combination of two resistor R_1 and R_2 .

Two resistors are said to be in *series* if only one of their end points is joined (Fig. 3.13). If a third resistor is joined with the series combination of the two (Fig. 3.14), then all three are said to be in series. Clearly, we can extend this definition to series combination of any number of resistors.

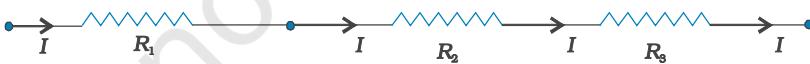


FIGURE 3.14 A series combination of three resistors R_1 , R_2 , R_3 .

Two or more resistors are said to be in *parallel* if one end of all the resistors is joined together and similarly the other ends joined together (Fig. 3.15).

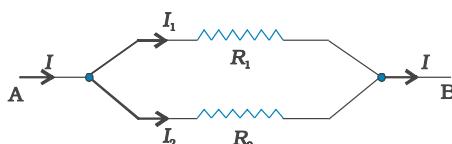


FIGURE 3.15 Two resistors R_1 and R_2 connected in parallel.

Physics

Consider two resistors R_1 and R_2 in series. The charge which leaves R_1 must be entering R_2 . Since current measures the rate of flow of charge, this means that the same current I flows through R_1 and R_2 . By Ohm's law:

Potential difference across $R_1 = V_1 = IR_1$, and

Potential difference across $R_2 = V_2 = IR_2$.

The potential difference V across the combination is $V_1 + V_2$. Hence,

$$V = V_1 + V_2 = I(R_1 + R_2) \quad (3.36)$$

This is as if the combination had an equivalent resistance R_{eq} , which by Ohm's law is

$$R_{eq} = \frac{V}{I} = (R_1 + R_2) \quad (3.37)$$

If we had three resistors connected in series, then similarly

$$V = IR_1 + IR_2 + IR_3 = I(R_1 + R_2 + R_3). \quad (3.38)$$

This obviously can be extended to a series combination of any number n of resistors R_1, R_2, \dots, R_n . The equivalent resistance R_{eq} is

$$R_{eq} = R_1 + R_2 + \dots + R_n \quad (3.39)$$

Consider now the parallel combination of two resistors (Fig. 3.15). The charge that flows in at A from the left flows out partly through R_1 and partly through R_2 . The currents I, I_1, I_2 shown in the figure are the rates of flow of charge at the points indicated. Hence,

$$I = I_1 + I_2 \quad (3.40)$$

The potential difference between A and B is given by the Ohm's law applied to R_1

$$V = I_1 R_1 \quad (3.41)$$

Also, Ohm's law applied to R_2 gives

$$V = I_2 R_2 \quad (3.42)$$

$$\therefore I = I_1 + I_2 = \frac{V}{R_1} + \frac{V}{R_2} = V \left(\frac{1}{R_1} + \frac{1}{R_2} \right) \quad (3.43)$$

If the combination was replaced by an equivalent resistance R_{eq} , we would have, by Ohm's law

$$I = \frac{V}{R_{eq}} \quad (3.44)$$

Hence,

$$\frac{1}{R_{eq}} = \frac{1}{R_1} + \frac{1}{R_2} \quad (3.45)$$

We can easily see how this extends to three resistors in parallel (Fig. 3.16).

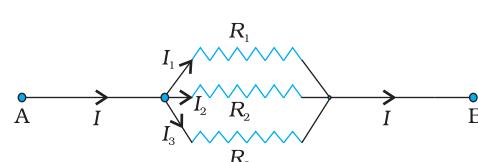


FIGURE 3.16 Parallel combination of three resistors R_1, R_2 and R_3 .

Exactly as before

$$I = I_1 + I_2 + I_3 \quad (3.46)$$

and applying Ohm's law to R_1 , R_2 and R_3 we get,

$$V = I_1 R_1, V = I_2 R_2, V = I_3 R_3 \quad (3.47)$$

So that

$$I = I_1 + I_2 + I_3 = V \left(\frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} \right) \quad (3.48)$$

An equivalent resistance R_{eq} that replaces the combination, would be such that

$$I = \frac{V}{R_{eq}} \quad (3.49)$$

and hence

$$\frac{1}{R_{eq}} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} \quad (3.50)$$

We can reason similarly for any number of resistors in parallel. The equivalent resistance of n resistors $R_1, R_2 \dots, R_n$ is

$$\frac{1}{R_{eq}} = \frac{1}{R_1} + \frac{1}{R_2} + \dots + \frac{1}{R_n} \quad (3.51)$$

These formulae for equivalent resistances can be used to find out currents and voltages in more complicated circuits. Consider for example, the circuit in Fig. (3.17), where there are three resistors R_1 , R_2 and R_3 . R_2 and R_3 are in parallel and hence we can replace them by an equivalent R_{eq}^{23} between point B and C with

$$\frac{1}{R_{eq}^{23}} = \frac{1}{R_2} + \frac{1}{R_3}$$

$$\text{or, } R_{eq}^{23} = \frac{R_2 R_3}{R_2 + R_3} \quad (3.52)$$

The circuit now has R_1 and R_{eq}^{23} in series and hence their combination can be replaced by an equivalent resistance R_{eq}^{123} with

$$R_{eq}^{123} = R_{eq}^{23} + R_1 \quad (3.53)$$

If the voltage between A and C is V , the current I is given by

$$\begin{aligned} I &= \frac{V}{R_{eq}^{123}} = \frac{V}{R_1 + [R_2 R_3 / (R_2 + R_3)]} \\ &= \frac{V(R_2 + R_3)}{R_1 R_2 + R_1 R_3 + R_2 R_3} \end{aligned} \quad (3.54)$$

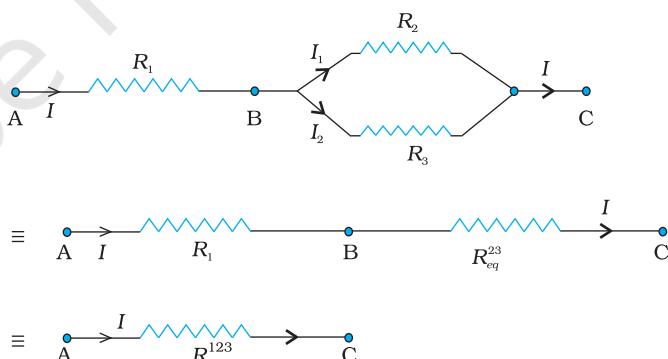


FIGURE 3.17 A combination of three resistors R_1 , R_2 and R_3 . R_2 , R_3 are in parallel with an equivalent resistance R_{eq}^{23} . R_1 and R_{eq}^{23} are in series with an equivalent resistance R_{eq}^{123} .

3.11 CELLS, EMF, INTERNAL RESISTANCE

We have already mentioned that a simple device to maintain a steady current in an electric circuit is the electrolytic cell. Basically a cell has two electrodes, called the positive (P) and the negative (N), as shown in Fig. 3.18. They are immersed in an electrolytic solution. Dipped in the

solution, the electrodes exchange charges with the electrolyte. The positive electrode has a potential difference V_+ ($V_+ > 0$) between itself and the electrolyte solution immediately adjacent to it marked A in the figure. Similarly, the negative electrode develops a negative potential $-V_-$ ($V_- \geq 0$) relative to the electrolyte adjacent to it, marked as B in the figure. When there is no current, the electrolyte has the same potential throughout, so that the potential difference between P and N is $V_+ - (-V_-) = V_+ + V_-$. This difference is called the *electromotive force* (emf) of the cell and is denoted by ε . Thus

$$\varepsilon = V_+ + V_- > 0 \quad (3.55)$$

Note that ε is, actually, a potential difference and *not a force*. The name emf, however, is used because of historical reasons, and was given at a time when the phenomenon was not understood properly.

To understand the significance of ε , consider a resistor R connected across the cell (Fig. 3.18). A current I flows across R from C to D. As explained before, a steady current is maintained because current flows from N to P through the electrolyte. Clearly, across the electrolyte the same current flows through the electrolyte but from N to P, whereas through R , it flows from P to N.

The electrolyte through which a current flows has a finite resistance r , called the *internal resistance*. Consider first the situation when R is infinite so that $I = V/R = 0$, where V is the potential difference between P and N. Now,

$$\begin{aligned} V &= \text{Potential difference between P and A} \\ &\quad + \text{Potential difference between A and B} \\ &\quad + \text{Potential difference between B and N} \\ &= \varepsilon \end{aligned} \quad (3.56)$$

Thus, emf ε is the potential difference between the positive and negative electrodes in an open circuit, i.e., when no current is flowing through the cell.

If however R is finite, I is not zero. In that case the potential difference between P and N is

$$\begin{aligned} V &= V_+ + V_- - Ir \\ &= \varepsilon - Ir \end{aligned} \quad (3.57)$$

Note the negative sign in the expression (Ir) for the potential difference between A and B. This is because the current I flows from B to A in the electrolyte.

In practical calculations, internal resistances of cells in the circuit may be neglected when the current I is such that $\varepsilon \gg Ir$. The actual values of the internal resistances of cells vary from cell to cell. The internal resistance of dry cells, however, is much higher than the common electrolytic cells.

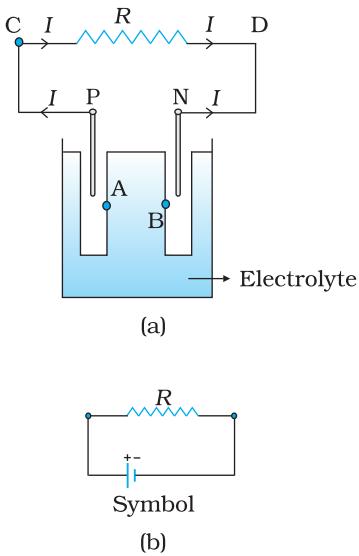


FIGURE 3.18 (a) Sketch of an electrolyte cell with positive terminal P and negative terminal N. The gap between the electrodes is exaggerated for clarity. A and B are points in the electrolyte typically close to P and N. (b) the symbol for a cell, + referring to P and - referring to the N electrode. Electrical connections to the cell are made at P and N.

Current Electricity

We also observe that since V is the potential difference across R , we have from Ohm's law

$$V = I R \quad (3.58)$$

Combining Eqs. (3.57) and (3.58), we get

$$I R = \varepsilon - I r$$

$$\text{Or, } I = \frac{\varepsilon}{R + r} \quad (3.59)$$

The maximum current that can be drawn from a cell is for $R = 0$ and it is $I_{\max} = \varepsilon/r$. However, in most cells the maximum allowed current is much lower than this to prevent permanent damage to the cell.

CHARGES IN CLOUDS

In olden days lightning was considered as an atmospheric flash of supernatural origin. It was believed to be the great weapon of Gods. But today the phenomenon of lightning can be explained scientifically by elementary principles of physics.

Atmospheric electricity arises due to the separation of electric charges. In the ionosphere and magnetosphere strong electric current is generated from the solar-terrestrial interaction. In the lower atmosphere the current is weaker and is maintained by thunderstorm.

There are ice particles in the clouds, which grow, collide, fracture and break apart. The smaller particles acquire positive charge and the larger ones negative charge. These charged particles get separated by updrifts in the clouds and gravity. The upper portion of the cloud becomes positively charged and the middle negatively charged, leading to dipole structure. Sometimes a very weak positive charge is found near the base of the cloud. The ground is positively charged at the time of thunderstorm development. Also cosmic and radioactive radiations ionise air into positive and negative ions and air becomes (weakly) electrically conductive. The separation of charges produce tremendous amount of electrical potential within the cloud as well as between the cloud and ground. This can amount to millions of volts and eventually the electrical resistance in the air breaks down and lightning flash begins and thousands of amperes of current flows. The electric field is of the order of 10^5 V/m . A lightning flash is composed of a series of strokes with an average of about four and the duration of each flash is about 30 seconds. The average peak power per stroke is about 10^{12} watts.

During fair weather also there is charge in the atmosphere. The fair weather electric field arises due to the existence of a surface charge density at ground and an atmospheric conductivity as well as due to the flow of current from the ionosphere to the earth's surface, which is of the order of picoampere / square metre. The surface charge density at ground is negative; the electric field is directed downward. Over land the average electric field is about 120 V/m , which corresponds to a surface charge density of $-1.2 \times 10^{-9} \text{ C/m}^2$. Over the entire earth's surface, the total negative charge amount to about 600 kC . An equal positive charge exists in the atmosphere. This electric field is not noticeable in daily life. The reason why it is not noticed is that virtually everything, including our bodies, is conductor compared to air.

Example 3.5 A network of resistors is connected to a 16 V battery with internal resistance of 1Ω , as shown in Fig. 3.19: (a) Compute the equivalent resistance of the network. (b) Obtain the current in each resistor. (c) Obtain the voltage drops V_{AB} , V_{BC} and V_{CD} .

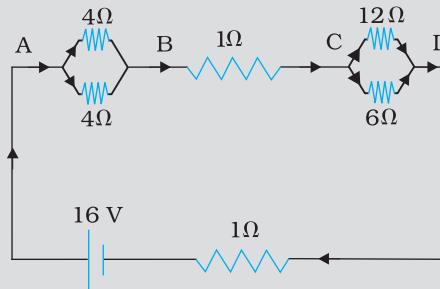


FIGURE 3.19

Solution

(a) The network is a simple series and parallel combination of resistors. First the two 4Ω resistors in parallel are equivalent to a resistor $[(4 \times 4)/(4 + 4)] \Omega = 2\Omega$.

In the same way, the 12Ω and 6Ω resistors in parallel are equivalent to a resistor of $[(12 \times 6)/(12 + 6)] \Omega = 4\Omega$.

The equivalent resistance R of the network is obtained by combining these resistors (2Ω and 4Ω) with 1Ω in series, that is,

$$R = 2\Omega + 4\Omega + 1\Omega = 7\Omega.$$

(b) The total current I in the circuit is

$$I = \frac{\varepsilon}{R+r} = \frac{16V}{(7+1)\Omega} = 2A$$

Consider the resistors between A and B. If I_1 is the current in one of the 4Ω resistors and I_2 the current in the other,

$$I_1 \times 4 = I_2 \times 4$$

that is, $I_1 = I_2$, which is otherwise obvious from the symmetry of the two arms. But $I_1 + I_2 = I = 2A$. Thus,

$$I_1 = I_2 = 1A$$

that is, current in each 4Ω resistor is $1A$. Current in 1Ω resistor between B and C would be $2A$.

Now, consider the resistances between C and D. If I_3 is the current in the 12Ω resistor, and I_4 in the 6Ω resistor,

$$I_3 \times 12 = I_4 \times 6, \text{ i.e., } I_4 = 2I_3$$

$$\text{But, } I_3 + I_4 = I = 2A$$

$$\text{Thus, } I_3 = \left(\frac{2}{3}\right) A, I_4 = \left(\frac{4}{3}\right) A$$

that is, the current in the 12Ω resistor is $(2/3)A$, while the current in the 6Ω resistor is $(4/3)A$.

(c) The voltage drop across AB is

$$V_{AB} = I_1 \times 4 = 1A \times 4\Omega = 4V,$$

This can also be obtained by multiplying the total current between A and B by the equivalent resistance between A and B, that is,

EXAMPLE 3.5

$$V_{AB} = 2 \text{ A} \times 2 \Omega = 4 \text{ V}$$

The voltage drop across BC is

$$V_{BC} = 2 \text{ A} \times 1 \Omega = 2 \text{ V}$$

Finally, the voltage drop across CD is

$$V_{CD} = 12 \Omega \times I_3 = 12 \Omega \times \left(\frac{2}{3}\right) \text{ A} = 8 \text{ V.}$$

This can alternately be obtained by multiplying total current between C and D by the equivalent resistance between C and D, that is,

$$V_{CD} = 2 \text{ A} \times 4 \Omega = 8 \text{ V}$$

Note that the total voltage drop across AD is $4 \text{ V} + 2 \text{ V} + 8 \text{ V} = 14 \text{ V}$. Thus, the terminal voltage of the battery is 14 V, while its emf is 16 V. The loss of the voltage ($= 2 \text{ V}$) is accounted for by the internal resistance 1Ω of the battery [$2 \text{ A} \times 1 \Omega = 2 \text{ V}$].

3.12 CELLS IN SERIES AND IN PARALLEL

Like resistors, cells can be combined together in an electric circuit. And like resistors, one can, for calculating currents and voltages in a circuit, replace a combination of cells by an equivalent cell.

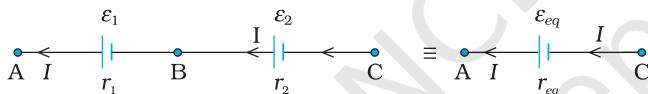


FIGURE 3.20 Two cells of emf's ε_1 and ε_2 in the series. r_1 , r_2 are their internal resistances. For connections across A and C, the combination can be considered as one cell of emf ε_{eq} and an internal resistance r_{eq} .

Consider first two cells in series (Fig. 3.20), where one terminal of the two cells is joined together leaving the other terminal in either cell free. ε_1 , ε_2 are the emf's of the two cells and r_1 , r_2 their internal resistances, respectively.

Let $V(A)$, $V(B)$, $V(C)$ be the potentials at points A, B and C shown in Fig. 3.20. Then $V(A) - V(B)$ is the potential difference between the positive and negative terminals of the first cell. We have already calculated it in Eq. (3.57) and hence,

$$V_{AB} \equiv V(A) - V(B) = \varepsilon_1 - Ir_1 \quad (3.60)$$

Similarly,

$$V_{BC} \equiv V(B) - V(C) = \varepsilon_2 - Ir_2 \quad (3.61)$$

Hence, the potential difference between the terminals A and C of the combination is

$$\begin{aligned} V_{AC} &\equiv V(A) - V(C) = [V(A) - V(B)] + [V(B) - V(C)] \\ &= (\varepsilon_1 + \varepsilon_2) - I(r_1 + r_2) \end{aligned} \quad (3.62)$$

Physics

If we wish to replace the combination by a single cell between A and C of emf ε_{eq} and internal resistance r_{eq} , we would have

$$V_{AC} = \varepsilon_{eq} - Ir_{eq} \quad (3.63)$$

Comparing the last two equations, we get

$$\varepsilon_{eq} = \varepsilon_1 + \varepsilon_2 \quad (3.64)$$

$$\text{and } r_{eq} = r_1 + r_2 \quad (3.65)$$

In Fig. 3.20, we had connected the negative electrode of the first to the positive electrode of the second. If instead we connect the two negatives, Eq. (3.61) would change to $V_{BC} = -\varepsilon_2 - Ir_2$ and we will get

$$\varepsilon_{eq} = \varepsilon_1 - \varepsilon_2 \quad (\varepsilon_1 > \varepsilon_2) \quad (3.66)$$

The rule for series combination clearly can be extended to any number of cells:

- (i) The equivalent emf of a series combination of n cells is just the sum of their individual emf's, and
- (ii) The equivalent internal resistance of a series combination of n cells is just the sum of their internal resistances.

This is so, when the current leaves each cell from the positive electrode. If in the combination, the current leaves any cell from the *negative* electrode, the emf of the cell enters the expression for ε_{eq} with a *negative* sign, as in Eq. (3.66).

Next, consider a parallel combination of the cells (Fig. 3.21). I_1 and I_2 are the currents leaving the positive electrodes of the cells. At the point B_1 , I_1 and I_2 flow in whereas the current I flows out. Since as much charge flows in as out, we have

$$I = I_1 + I_2 \quad (3.67)$$

Let $V(B_1)$ and $V(B_2)$ be the potentials at B_1 and B_2 , respectively. Then, considering the first cell, the potential difference across its terminals is $V(B_1) - V(B_2)$. Hence, from Eq. (3.57)

$$V \equiv V(B_1) - V(B_2) = \varepsilon_1 - I_1 r_1 \quad (3.68)$$

Points B_1 and B_2 are connected exactly similarly to the second cell. Hence considering the second cell, we also have

$$V \equiv V(B_1) - V(B_2) = \varepsilon_2 - I_2 r_2 \quad (3.69)$$

Combining the last three equations

$$\begin{aligned} I &= I_1 + I_2 \\ &= \frac{\varepsilon_1 - V}{r_1} + \frac{\varepsilon_2 - V}{r_2} = \left(\frac{\varepsilon_1}{r_1} + \frac{\varepsilon_2}{r_2} \right) - V \left(\frac{1}{r_1} + \frac{1}{r_2} \right) \end{aligned} \quad (3.70)$$

Hence, V is given by,

$$V = \frac{\varepsilon_1 r_2 + \varepsilon_2 r_1}{r_1 + r_2} - I \frac{r_1 r_2}{r_1 + r_2} \quad (3.71)$$

If we want to replace the combination by a single cell, between B_1 and B_2 , of emf ε_{eq} and internal resistance r_{eq} , we would have

$$V = \varepsilon_{eq} - Ir_{eq} \quad (3.72)$$

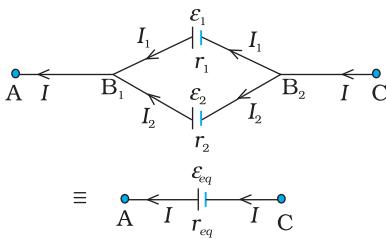


FIGURE 3.21 Two cells in parallel. For connections across A and C, the combination can be replaced by one cell of emf ε_{eq} and internal resistances r_{eq} whose values are given in Eqs. (3.73) and (3.74).

The last two equations should be the same and hence

$$\varepsilon_{eq} = \frac{\varepsilon_1 r_2 + \varepsilon_2 r_1}{r_1 + r_2} \quad (3.73)$$

$$r_{eq} = \frac{r_1 r_2}{r_1 + r_2} \quad (3.74)$$

We can put these equations in a simpler way,

$$\frac{1}{r_{eq}} = \frac{1}{r_1} + \frac{1}{r_2} \quad (3.75)$$

$$\frac{\varepsilon_{eq}}{r_{eq}} = \frac{\varepsilon_1}{r_1} + \frac{\varepsilon_2}{r_2} \quad (3.76)$$

In Fig. (3.21), we had joined the positive terminals together and similarly the two negative ones, so that the currents I_1, I_2 flow out of positive terminals. If the negative terminal of the second is connected to positive terminal of the first, Eqs. (3.75) and (3.76) would still be valid with $\varepsilon_2 \rightarrow -\varepsilon_2$

Equations (3.75) and (3.76) can be extended easily. If there are n cells of emf $\varepsilon_1, \dots, \varepsilon_n$ and of internal resistances r_1, \dots, r_n respectively, connected in parallel, the combination is equivalent to a single cell of emf ε_{eq} and internal resistance r_{eq} , such that

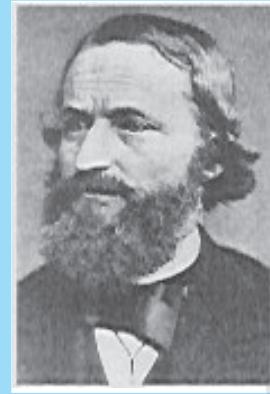
$$\frac{1}{r_{eq}} = \frac{1}{r_1} + \dots + \frac{1}{r_n} \quad (3.77)$$

$$\frac{\varepsilon_{eq}}{r_{eq}} = \frac{\varepsilon_1}{r_1} + \dots + \frac{\varepsilon_n}{r_n} \quad (3.78)$$

3.13 KIRCHHOFF's RULES

Electric circuits generally consist of a number of resistors and cells interconnected sometimes in a complicated way. The formulae we have derived earlier for series and parallel combinations of resistors are not always sufficient to determine all the currents and potential differences in the circuit. Two rules, called *Kirchhoff's rules*, are very useful for analysis of electric circuits.

Given a circuit, we start by labelling currents in each resistor by a symbol, say I , and a directed arrow to indicate that a current I flows along the resistor in the direction indicated. If ultimately I is determined to be positive, the actual current in the resistor is in the direction of the arrow. If I turns out to be negative, the current actually flows in a direction opposite to the arrow. Similarly, for each source (i.e., cell or some other source of electrical power) the positive and negative electrodes are labelled as well as a directed arrow with a symbol for the current flowing through the cell. This will tell us the potential difference, $V = V(P) - V(N) = \varepsilon - Ir$



Gustav Robert Kirchhoff (1824 – 1887) German physicist, professor at Heidelberg and at Berlin. Mainly known for his development of spectroscopy, he also made many important contributions to mathematical physics, among them, his first and second rules for circuits.

GUSTAV ROBERT KIRCHHOFF (1824 – 1887)

[Eq. (3.57) between the positive terminal P and the negative terminal N; I here is the current flowing from N to P through the cell]. If, while labelling the current I through the cell one goes from P to N, then of course

$$V = \varepsilon + Ir \quad (3.79)$$

Having clarified labelling, we now state the rules and the proof:

(a) *Junction rule:* At any junction, the sum of the currents entering the junction is equal to the sum of currents leaving the junction (Fig. 3.22). This applies equally well if instead of a junction of several lines, we consider a point in a line.

The proof of this rule follows from the fact that when currents are steady, there is no accumulation of charges at any junction or at any point in a line. Thus, the total current flowing in, (which is the rate at which charge flows into the junction), must equal the total current flowing out.

(b) *Loop rule:* The algebraic sum of changes in potential around any closed loop involving resistors and cells in the loop is zero (Fig. 3.22).

This rule is also obvious, since electric potential is dependent on the location of the point. Thus starting with any point if we come back to the same point, the total change must be zero. In a closed loop, we do come back to the starting point and hence the rule.

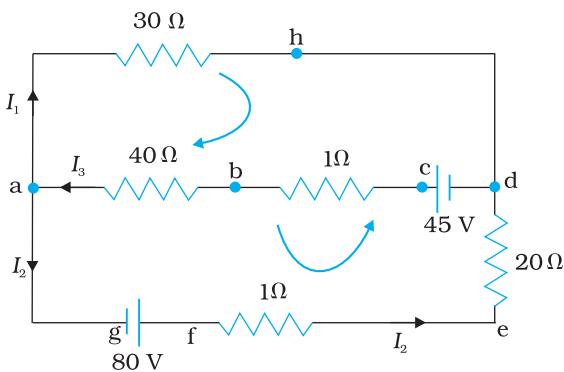
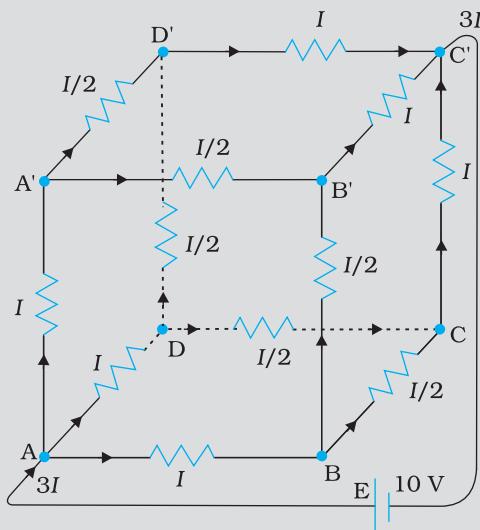


FIGURE 3.22 At junction a the current leaving is $I_1 + I_2$ and current entering is I_3 . The junction rule says $I_3 = I_1 + I_2$. At point h current entering is I_1 . There is only one current leaving h and by junction rule that will also be I_1 . For the loops 'ahdcba' and 'ahdefga', the loop rules give $-30I_1 - 41I_3 + 45 = 0$ and $-30I_1 + 21I_2 - 80 = 0$.

dependent on the location of the point. Thus starting with any point if we come back to the same point, the total change must be zero. In a closed loop, we do come back to the starting point and hence the rule.

Example 3.6 A battery of 10 V and negligible internal resistance is connected across the diagonally opposite corners of a cubical network consisting of 12 resistors each of resistance 1 Ω (Fig. 3.23). Determine the equivalent resistance of the network and the current along each edge of the cube.



EXAMPLE 3.6

Solution The network is not reducible to a simple series and parallel combinations of resistors. There is, however, a clear symmetry in the problem which we can exploit to obtain the equivalent resistance of the network.

The paths AA', AD and AB are obviously symmetrically placed in the network. Thus, the current in each must be the same, say, I . Further, at the corners A', B and D, the incoming current I must split equally into the two outgoing branches. In this manner, the current in all the 12 edges of the cube are easily written down in terms of I , using Kirchhoff's first rule and the symmetry in the problem.

Next take a closed loop, say, ABCC'EA, and apply Kirchhoff's second rule:

$$-IR - (1/2)IR - IR + \varepsilon = 0$$

where R is the resistance of each edge and ε the emf of battery. Thus,

$$\varepsilon = \frac{5}{2}IR$$

The equivalent resistance R_{eq} of the network is

$$R_{eq} = \frac{\varepsilon}{3I} = \frac{5}{6}R$$

For $R = 1\ \Omega$, $R_{eq} = (5/6)\ \Omega$ and for $\varepsilon = 10\ V$, the total current ($= 3I$) in the network is

$$3I = 10\ V/(5/6)\ \Omega = 12\ A, \text{ i.e., } I = 4\ A$$

The current flowing in each edge can now be read off from the Fig. 3.23.

It should be noted that because of the symmetry of the network, the great power of Kirchhoff's rules has not been very apparent in Example 3.6. In a general network, there will be no such simplification due to symmetry, and only by application of Kirchhoff's rules to junctions and closed loops (as many as necessary to solve the unknowns in the network) can we handle the problem. This will be illustrated in Example 3.7.

Example 3.7 Determine the current in each branch of the network shown in Fig. 3.24.

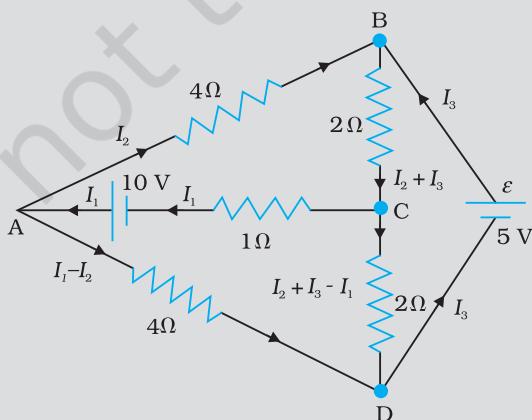


FIGURE 3.24

PHYSICS

Simulation for application of Kirchhoff's rules:
<http://www.phys.hawaii.edu/~teb/optics/java/kirch3/>

EXAMPLE 3.6

EXAMPLE 3.7

EXAMPLE 3.7

Solution Each branch of the network is assigned an unknown current to be determined by the application of Kirchhoff's rules. To reduce the number of unknowns at the outset, the first rule of Kirchhoff is used at every junction to assign the unknown current in each branch. We then have three unknowns I_1 , I_2 and I_3 which can be found by applying the second rule of Kirchhoff to three different closed loops. Kirchhoff's second rule for the closed loop ADCA gives,

$$10 - 4(I_1 - I_2) + 2(I_2 + I_3 - I_1) - I_1 = 0 \quad [3.80(a)]$$

that is, $7I_1 - 6I_2 - 2I_3 = 10$

For the closed loop ABCA, we get

$$10 - 4I_2 - 2(I_2 + I_3) - I_1 = 0 \quad [3.80(b)]$$

that is, $I_1 + 6I_2 + 2I_3 = 10$

For the closed loop BCDEB, we get

$$5 - 2(I_2 + I_3) - 2(I_2 + I_3 - I_1) = 0 \quad [3.80(c)]$$

that is, $2I_1 - 4I_2 - 4I_3 = -5$

Equations (3.80 a, b, c) are three simultaneous equations in three unknowns. These can be solved by the usual method to give

$$I_1 = 2.5\text{ A}, \quad I_2 = \frac{5}{8}\text{ A}, \quad I_3 = 1\frac{7}{8}\text{ A}$$

The currents in the various branches of the network are

$$\begin{aligned} AB : & \frac{5}{8}\text{ A}, \quad CA : 2\frac{1}{2}\text{ A}, \quad DEB : 1\frac{7}{8}\text{ A} \\ AD : & 1\frac{7}{8}\text{ A}, \quad CD : 0\text{ A}, \quad BC : 2\frac{1}{2}\text{ A} \end{aligned}$$

It is easily verified that Kirchhoff's second rule applied to the remaining closed loops does not provide any additional independent equation, that is, the above values of currents satisfy the second rule for every closed loop of the network. For example, the total voltage drop over the closed loop BADEB

$$5\text{ V} + \left(\frac{5}{8} \times 4\right)\text{ V} - \left(\frac{15}{8} \times 4\right)\text{ V}$$

equal to zero, as required by Kirchhoff's second rule.

3.14 WHEATSTONE BRIDGE

As an application of Kirchhoff's rules consider the circuit shown in Fig. 3.25, which is called the *Wheatstone bridge*. The bridge has four resistors R_1 , R_2 , R_3 and R_4 . Across one pair of diagonally opposite points (A and C in the figure) a source is connected. This (i.e., AC) is called the battery arm. Between the other two vertices, B and D, a galvanometer G (which is a device to detect currents) is connected. This line, shown as BD in the figure, is called the galvanometer arm.

For simplicity, we assume that the cell has no internal resistance. In general there will be currents flowing across all the resistors as well as a current I_g through G. Of special interest, is the case of a *balanced* bridge where the resistors are such that $I_g = 0$. We can easily get the balance condition, such that there is no current through G. In this case, the Kirchhoff's junction rule applied to junctions D and B (see the figure)

immediately gives us the relations $I_1 = I_3$ and $I_2 = I_4$. Next, we apply Kirchhoff's loop rule to closed loops ADBA and CBDC. The first loop gives

$$-I_1 R_1 + 0 + I_2 R_2 = 0 \quad (I_g = 0) \quad (3.81)$$

and the second loop gives, upon using $I_3 = I_1$, $I_4 = I_2$

$$I_2 R_4 + 0 - I_1 R_3 = 0 \quad (3.82)$$

From Eq. (3.81), we obtain,

$$\frac{I_1}{I_2} = \frac{R_2}{R_1}$$

whereas from Eq. (3.82), we obtain,

$$\frac{I_1}{I_2} = \frac{R_4}{R_3}$$

Hence, we obtain the condition

$$\frac{R_2}{R_1} = \frac{R_4}{R_3} \quad [3.83(a)]$$

This last equation relating the four resistors is called the *balance condition* for the galvanometer to give zero or null deflection.

The Wheatstone bridge and its balance condition provide a practical method for determination of an unknown resistance. Let us suppose we have an unknown resistance, which we insert in the fourth arm; R_4 is thus not known. Keeping known resistances R_1 and R_2 in the first and second arm of the bridge, we go on varying R_3 till the galvanometer shows a null deflection. The bridge then is balanced, and from the balance condition the value of the unknown resistance R_4 is given by,

$$R_4 = R_3 \frac{R_2}{R_1} \quad [3.83(b)]$$

A practical device using this principle is called the *meter bridge*. It will be discussed in the next section.

Example 3.8 The four arms of a Wheatstone bridge (Fig. 3.26) have the following resistances:

$AB = 100\Omega$, $BC = 10\Omega$, $CD = 5\Omega$, and $DA = 60\Omega$.

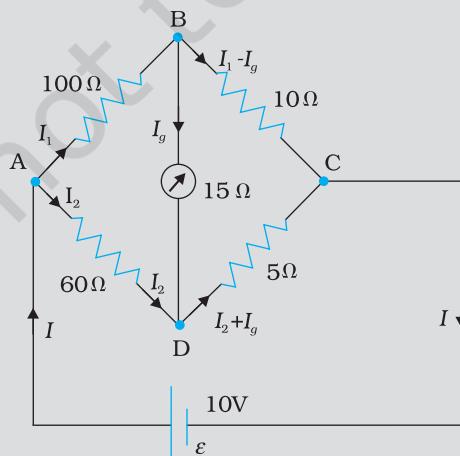


FIGURE 3.26

EXAMPLE 3.8

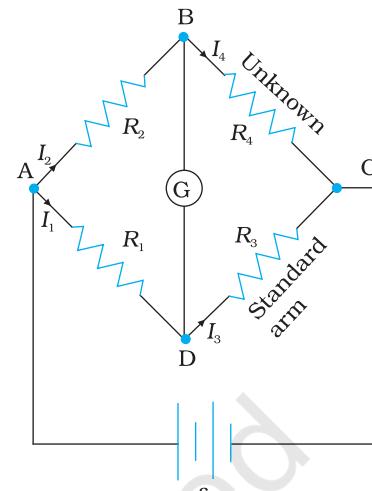


FIGURE 3.25

EXAMPLE 3.8

A galvanometer of 15Ω resistance is connected across BD. Calculate the current through the galvanometer when a potential difference of 10 V is maintained across AC.

Solution Considering the mesh BADB, we have

$$100I_1 + 15I_g - 60I_2 = 0$$

$$\text{or } 20I_1 + 3I_g - 12I_2 = 0 \quad [3.84(a)]$$

Considering the mesh BCDB, we have

$$10(I_1 - I_g) - 15I_g - 5(I_2 + I_g) = 0$$

$$10I_1 - 30I_g - 5I_2 = 0$$

$$2I_1 - 6I_g - I_2 = 0 \quad [3.84(b)]$$

Considering the mesh ADCEA,

$$60I_2 + 5(I_2 + I_g) = 10$$

$$65I_2 + 5I_g = 10$$

$$13I_2 + I_g = 2 \quad [3.84(c)]$$

Multiplying Eq. (3.84b) by 10

$$20I_1 - 60I_g - 10I_2 = 0 \quad [3.84(d)]$$

From Eqs. (3.84d) and (3.84a) we have

$$63I_g - 2I_2 = 0$$

$$I_2 = 31.5I_g \quad [3.84(e)]$$

Substituting the value of I_2 into Eq. [3.84(c)], we get

$$13(31.5I_g) + I_g = 2$$

$$410.5I_g = 2$$

$$I_g = 4.87 \text{ mA.}$$

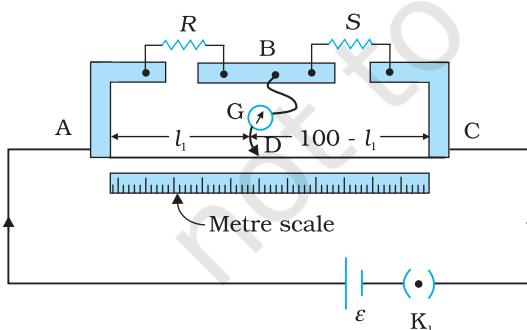


FIGURE 3.27 A meter bridge. Wire AC is 1 m long. R is a resistance to be measured and S is a standard resistance.

3.15 METER BRIDGE

The meter bridge is shown in Fig. 3.27. It consists of a wire of length 1 m and of uniform cross sectional area stretched taut and clamped between two thick metallic strips bent at right angles, as shown. The metallic strip has two gaps across which resistors can be connected. The end points where the wire is clamped are connected to a cell through a key. One end of a galvanometer is connected to the metallic strip midway between the two gaps. The other end of the galvanometer is connected to a 'jockey'. The jockey is essentially a metallic rod whose one end has a knife-edge which can slide over the wire to make electrical connection.

R is an unknown resistance whose value we want to determine. It is connected across one of the gaps. Across the other gap, we connect a

standard known resistance S . The jockey is connected to some point D on the wire, a distance l cm from the end A. The jockey can be moved along the wire. The portion AD of the wire has a resistance $R_{cm}l$, where R_{cm} is the resistance of the wire per unit centimetre. The portion DC of the wire similarly has a resistance $R_{cm}(100-l)$.

The four arms AB, BC, DA and CD [with resistances R , S , $R_{cm}l$ and $R_{cm}(100-l)$] obviously form a Wheatstone bridge with AC as the battery arm and BD the galvanometer arm. If the jockey is moved along the wire, then there will be one position where the galvanometer will show no current. Let the distance of the jockey from the end A at the balance point be $l = l_1$. The four resistances of the bridge at the balance point then are R , S , $R_{cm}l_1$ and $R_{cm}(100-l_1)$. The balance condition, Eq. [3.83(a)] gives

$$\frac{R}{S} = \frac{R_{cm}l_1}{R_{cm}(100-l_1)} = \frac{l_1}{100-l_1} \quad (3.85)$$

Thus, once we have found out l_1 , the unknown resistance R is known in terms of the standard known resistance S by

$$R = S \frac{l_1}{100-l_1} \quad (3.86)$$

By choosing various values of S , we would get various values of l_1 , and calculate R each time. An error in measurement of l_1 would naturally result in an error in R . It can be shown that the percentage error in R can be minimised by adjusting the balance point near the middle of the bridge, i.e., when l_1 is close to 50 cm. (This requires a suitable choice of S .)

Example 3.9 In a metre bridge (Fig. 3.27), the null point is found at a distance of 33.7 cm from A. If now a resistance of 12Ω is connected in parallel with S , the null point occurs at 51.9 cm. Determine the values of R and S .

Solution From the first balance point, we get

$$\frac{R}{S} = \frac{33.7}{66.3} \quad (3.87)$$

After S is connected in parallel with a resistance of 12Ω , the resistance across the gap changes from S to S_{eq} , where

$$S_{eq} = \frac{12S}{S+12}$$

and hence the new balance condition now gives

$$\frac{51.9}{48.1} = \frac{R}{S_{eq}} = \frac{R(S+12)}{12S} \quad (3.88)$$

Substituting the value of R/S from Eq. (3.87), we get

$$\frac{51.9}{48.1} = \frac{S+12}{12} \cdot \frac{33.7}{66.3}$$

which gives $S = 13.5\Omega$. Using the value of R/S above, we get $R = 6.86\Omega$.

3.16 POTENTIOMETER

This is a versatile instrument. It is basically a long piece of uniform wire, sometimes a few meters in length across which a standard cell is connected. In actual design, the wire is sometimes cut in several pieces placed side by side and connected at the ends by thick metal strip. (Fig. 3.28). In the figure, the wires run from A to C. The small vertical portions are the thick metal strips connecting the various sections of the wire.

A current I flows through the wire which can be varied by a variable resistance (rheostat, R) in the circuit. Since the wire is uniform, the potential difference between A and any point at a distance l from A is

$$\varepsilon(l) = \phi l \quad (3.89)$$

where ϕ is the potential drop per unit length.

Figure 3.28 (a) shows an application of the potentiometer to compare the emf of two cells of emf ε_1 and ε_2 . The points marked 1, 2, 3 form a two way key. Consider first a position of the key where 1 and 3 are connected

so that the galvanometer is connected to ε_1 . The jockey is moved along the wire till at a point N_1 , at a distance l_1 from A, there is no deflection in the galvanometer. We can apply Kirchhoff's loop rule to the closed loop AN_1G31A and get,

$$\phi l_1 + 0 - \varepsilon_1 = 0 \quad (3.90)$$

Similarly, if another emf ε_2 is balanced against l_2 (AN_2)

$$\phi l_2 + 0 - \varepsilon_2 = 0 \quad (3.91)$$

From the last two equations

$$\frac{\varepsilon_1}{\varepsilon_2} = \frac{l_1}{l_2} \quad (3.92)$$

This simple mechanism thus allows one to compare the emf's of any two sources. In practice one of the cells is chosen as a standard cell whose emf is known to a high degree of accuracy. The emf of the other cell is then easily calculated from Eq. (3.92).

We can also use a potentiometer to measure internal resistance of a cell [Fig. 3.28 (b)]. For this the cell (emf ε) whose internal resistance (r) is to be determined is connected across a resistance box through a key K_2 , as shown in the figure. With key K_2 open, balance is obtained at length l_1 (AN_1). Then,

$$\varepsilon = \phi l_1 \quad [3.93(a)]$$

When key K_2 is closed, the cell sends a current (I) through the resistance box (R). If V is the terminal potential difference of the cell and balance is obtained at length l_2 (AN_2),

$$V = \phi l_2 \quad [3.93(b)]$$

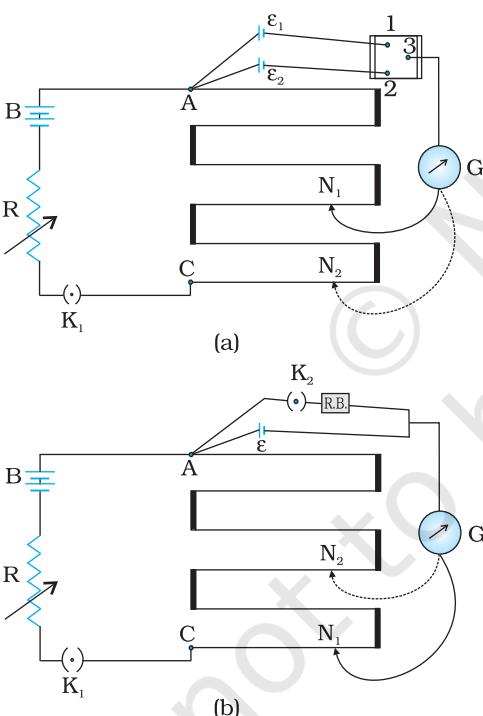


FIGURE 3.28 A potentiometer. G is a galvanometer and R a variable resistance (rheostat). 1, 2, 3 are terminals of a two way key
 (a) circuit for comparing emfs of two cells; (b) circuit for determining internal resistance of a cell.

So, we have $\varepsilon/V = l_1/l_2$

[3.94(a)]

But, $\varepsilon = I(r + R)$ and $V = IR$. This gives

$$\varepsilon/V = (r+R)/R$$

[3.94(b)]

From Eq. [3.94(a)] and [3.94(b)] we have

$$(R+r)/R = l_1/l_2$$

$$r = R \left(\frac{l_1}{l_2} - 1 \right) \quad (3.95)$$

Using Eq. (3.95) we can find the internal resistance of a given cell.

The potentiometer has the advantage that it draws *no current* from the voltage source being measured. As such it is unaffected by the internal resistance of the source.

Example 3.10 A resistance of $R \Omega$ draws current from a potentiometer. The potentiometer has a total resistance $R_0 \Omega$ (Fig. 3.29). A voltage V is supplied to the potentiometer. Derive an expression for the voltage across R when the sliding contact is in the middle of the potentiometer.

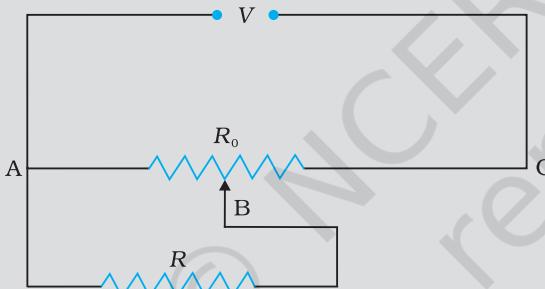


FIGURE 3.29

Solution While the slide is in the middle of the potentiometer only half of its resistance ($R_0/2$) will be between the points A and B. Hence, the total resistance between A and B, say, R_1 , will be given by the following expression:

$$\frac{1}{R_1} = \frac{1}{R} + \frac{1}{(R_0/2)}$$

$$R_1 = \frac{R_0 R}{R_0 + 2R}$$

The total resistance between A and C will be sum of resistance between A and B and B and C, i.e., $R_1 + R_0/2$

\therefore The current flowing through the potentiometer will be

$$I = \frac{V}{R_1 + R_0/2} = \frac{2V}{2R_1 + R_0}$$

The voltage V_1 taken from the potentiometer will be the product of current I and resistance R_1 ,

$$V_1 = I R_1 = \left(\frac{2V}{2R_1 + R_0} \right) \times R_1$$

EXAMPLE 3.10

EXAMPLE 3.10

Substituting for R_1 , we have a

$$V_1 = \frac{2V}{2 \left(\frac{R_0 \times R}{R_0 + 2R} \right) + R_0} \times \frac{R_0 \times R}{R_0 + 2R}$$

$$V_1 = \frac{2VR}{2R + R_0 + 2R}$$

$$\text{or } V_1 = \frac{2VR}{R_0 + 4R}$$

SUMMARY

1. *Current* through a given area of a conductor is the net charge passing per unit time through the area.
2. To maintain a steady current, we must have a closed circuit in which an external agency moves electric charge from lower to higher potential energy. The work done per unit charge by the source in taking the charge from lower to higher potential energy (i.e., from one terminal of the source to the other) is called the electromotive force, or *emf*, of the source. Note that the emf is not a force; it is the voltage difference between the two terminals of a source in open circuit.
3. *Ohm's law*: The electric current I flowing through a substance is proportional to the voltage V across its ends, i.e., $V \propto I$ or $V = RI$, where R is called the *resistance* of the substance. The unit of resistance is ohm: $1\Omega = 1 \text{ V A}^{-1}$.
4. The *resistance* R of a conductor depends on its length l and constant cross-sectional area A through the relation,

$$R = \frac{\rho l}{A}$$

where ρ , called *resistivity* is a property of the material and depends on temperature and pressure.

5. *Electrical resistivity* of substances varies over a very wide range. Metals have low resistivity, in the range of $10^{-8} \Omega \text{ m}$ to $10^{-6} \Omega \text{ m}$. Insulators like glass and rubber have 10^{22} to 10^{24} times greater resistivity. Semiconductors like Si and Ge lie roughly in the middle range of resistivity on a logarithmic scale.
6. In most substances, the carriers of current are electrons; in some cases, for example, ionic crystals and electrolytic liquids, positive and negative ions carry the electric current.
7. *Current density* \mathbf{j} gives the amount of charge flowing per second per unit area normal to the flow,

$$\mathbf{j} = nq \mathbf{v}_d$$

where n is the number density (number per unit volume) of charge carriers each of charge q , and \mathbf{v}_d is the *drift velocity* of the charge carriers. For electrons $q = -e$. If \mathbf{j} is normal to a cross-sectional area \mathbf{A} and is constant over the area, the magnitude of the current I through the area is $nev_d A$.

8. Using $E = V/l$, $I = nev_d A$, and Ohm's law, one obtains

$$\frac{eE}{m} = \rho \frac{ne^2}{m} v_d$$

The proportionality between the force eE on the electrons in a metal due to the external field E and the drift velocity v_d (not acceleration) can be understood, if we assume that the electrons suffer collisions with ions in the metal, which deflect them randomly. If such collisions occur on an average at a time interval τ ,

$$v_d = a\tau = eE\tau/m$$

where a is the acceleration of the electron. This gives

$$\rho = \frac{m}{ne^2\tau}$$

9. In the temperature range in which resistivity increases linearly with temperature, the *temperature coefficient of resistivity* α is defined as the fractional increase in resistivity per unit increase in temperature.

10. Ohm's law is obeyed by many substances, but it is not a fundamental law of nature. It fails if

- (a) V depends on I non-linearly.
- (b) the relation between V and I depends on the sign of V for the same absolute value of V .
- (c) The relation between V and I is non-unique.

An example of (a) is when ρ increases with I (even if temperature is kept fixed). A rectifier combines features (a) and (b). GaAs shows the feature (c).

11. When a source of emf ε is connected to an external resistance R , the voltage V_{ext} across R is given by

$$V_{ext} = IR = \frac{\varepsilon}{R + r} R$$

where r is the *internal resistance* of the source.

12. (a) Total resistance R of n resistors connected in *series* is given by

$$R = R_1 + R_2 + \dots + R_n$$

- (b) Total resistance R of n resistors connected in *parallel* is given by

$$\frac{1}{R} = \frac{1}{R_1} + \frac{1}{R_2} + \dots + \frac{1}{R_n}$$

13. *Kirchhoff's Rules* –

- (a) *Junction Rule*: At any junction of circuit elements, the sum of currents entering the junction must equal the sum of currents leaving it.
- (b) *Loop Rule*: The algebraic sum of changes in potential around any closed loop must be zero.

14. The *Wheatstone bridge* is an arrangement of four resistances – R_1 , R_2 , R_3 , R_4 as shown in the text. The null-point condition is given by

$$\frac{R_1}{R_2} = \frac{R_3}{R_4}$$

using which the value of one resistance can be determined, knowing the other three resistances.

15. The *potentiometer* is a device to compare potential differences. Since the method involves a condition of *no current flow*, the device can be used to measure potential difference; internal resistance of a cell and compare emf's of two sources.

Physics

Physical Quantity	Symbol	Dimensions	Unit	Remark
Electric current	I	[A]	A	SI base unit
Charge	Q, q	[T A]	C	
Voltage, Electric potential difference	V	[M L ² T ⁻³ A ⁻¹]	V	Work/charge
Electromotive force	ε	[M L ² T ⁻³ A ⁻¹]	V	Work/charge
Resistance	R	[M L ² T ⁻³ A ⁻²]	Ω	$R = V/I$
Resistivity	ρ	[M ³ L ⁻³ T ⁻³ A ⁻²]	$\Omega \text{ m}$	$R = \rho l/A$
Electrical conductivity	σ	[M ⁻¹ L ⁻³ T ³ A ²]	S	$\sigma = 1/\rho$
Electric field	\mathbf{E}	[M L T ⁻³ A ⁻¹]	V m ⁻¹	$\frac{\text{Electric force}}{\text{charge}}$
Drift speed	v_d	[L T ⁻¹]	m s ⁻¹	$v_d = \frac{e E \tau}{m}$
Relaxation time	τ	[T]	s	
Current density	\mathbf{j}	[L ⁻² A]	A m ⁻²	current/area
Mobility	μ	[M L ³ T ⁻⁴ A ⁻¹]	m ² V ⁻¹ s ⁻¹	v_d / E

POINTS TO PONDER

1. Current is a scalar although we represent current with an arrow. Currents do not obey the law of vector addition. That current is a scalar also follows from its definition. The current I through an area of cross-section is given by the scalar product of two vectors:

$$I = \mathbf{j} \cdot \Delta \mathbf{s}$$

where \mathbf{j} and $\Delta \mathbf{s}$ are vectors.
2. Refer to $V-I$ curves of a resistor and a diode as drawn in the text. A resistor obeys Ohm's law while a diode does not. The assertion that $V = IR$ is a statement of Ohm's law is not true. This equation defines resistance and it may be applied to all conducting devices whether they obey Ohm's law or not. The Ohm's law asserts that the plot of I versus V is linear i.e., R is independent of V .
Equation $\mathbf{E} = \rho \mathbf{j}$ leads to another statement of Ohm's law, i.e., a conducting material obeys Ohm's law when the resistivity of the material does not depend on the magnitude and direction of applied electric field.
3. Homogeneous conductors like silver or semiconductors like pure germanium or germanium containing impurities obey Ohm's law within some range of electric field values. If the field becomes too strong, there are departures from Ohm's law in all cases.
4. Motion of conduction electrons in electric field \mathbf{E} is the sum of (i) motion due to random collisions and (ii) that due to \mathbf{E} . The motion

due to random collisions averages to zero and does not contribute to v_d (Chapter 11, Textbook of Class XI). v_d , thus is only due to applied electric field on the electron.

5. The relation $\mathbf{j} = \rho \mathbf{v}$ should be applied to each type of charge carriers separately. In a conducting wire, the total current and charge density arises from both positive and negative charges:

$$\mathbf{j} = \rho_+ \mathbf{v}_+ + \rho_- \mathbf{v}_-$$

$$\rho = \rho_+ + \rho_-$$

Now in a neutral wire carrying electric current,

$$\rho_+ = -\rho_-$$

Further, $v_+ \sim 0$ which gives

$$\rho = 0$$

$$\mathbf{j} = \rho_- \mathbf{v}$$

Thus, the relation $\mathbf{j} = \rho \mathbf{v}$ does not apply to the total current charge density.

6. Kirchhoff's junction rule is based on conservation of charge and the outgoing currents add up and are equal to incoming current at a junction. Bending or reorienting the wire does not change the validity of Kirchhoff's junction rule.

EXERCISES

- 3.1** The storage battery of a car has an emf of 12 V. If the internal resistance of the battery is 0.4Ω , what is the maximum current that can be drawn from the battery?
- 3.2** A battery of emf 10 V and internal resistance 3Ω is connected to a resistor. If the current in the circuit is 0.5 A, what is the resistance of the resistor? What is the terminal voltage of the battery when the circuit is closed?
- 3.3**
 - (a) Three resistors 1Ω , 2Ω , and 3Ω are combined in series. What is the total resistance of the combination?
 - (b) If the combination is connected to a battery of emf 12 V and negligible internal resistance, obtain the potential drop across each resistor.
- 3.4**
 - (a) Three resistors 2Ω , 4Ω and 5Ω are combined in parallel. What is the total resistance of the combination?
 - (b) If the combination is connected to a battery of emf 20 V and negligible internal resistance, determine the current through each resistor, and the total current drawn from the battery.
- 3.5** At room temperature (27.0°C) the resistance of a heating element is 100Ω . What is the temperature of the element if the resistance is found to be 117Ω , given that the temperature coefficient of the material of the resistor is $1.70 \times 10^{-4} \text{ }^\circ\text{C}^{-1}$.
- 3.6** A negligibly small current is passed through a wire of length 15 m and uniform cross-section $6.0 \times 10^{-7} \text{ m}^2$, and its resistance is measured to be 5.0Ω . What is the resistivity of the material at the temperature of the experiment?
- 3.7** A silver wire has a resistance of 2.1Ω at 27.5°C , and a resistance of 2.7Ω at 100°C . Determine the temperature coefficient of resistivity of silver.
- 3.8** A heating element using nichrome connected to a 230 V supply draws an initial current of 3.2 A which settles after a few seconds to

a steady value of 2.8 A. What is the steady temperature of the heating element if the room temperature is 27.0 °C? Temperature coefficient of resistance of nichrome averaged over the temperature range involved is $1.70 \times 10^{-4} \text{ }^{\circ}\text{C}^{-1}$.

- 3.9** Determine the current in each branch of the network shown in Fig. 3.30:

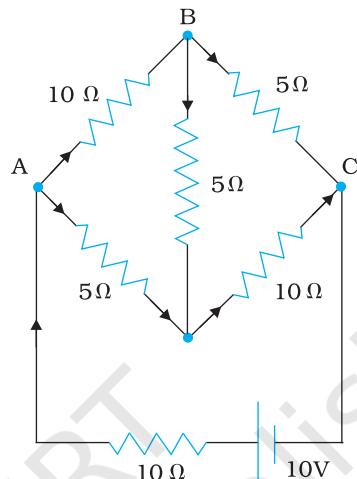


FIGURE 3.30

- 3.10** (a) In a metre bridge [Fig. 3.27], the balance point is found to be at 39.5 cm from the end A, when the resistor Y is of 12.5 Ω. Determine the resistance of X. Why are the connections between resistors in a Wheatstone or meter bridge made of thick copper strips?
 (b) Determine the balance point of the bridge above if X and Y are interchanged.
 (c) What happens if the galvanometer and cell are interchanged at the balance point of the bridge? Would the galvanometer show any current?
- 3.11** A storage battery of emf 8.0 V and internal resistance 0.5 Ω is being charged by a 120 V dc supply using a series resistor of 15.5 Ω. What is the terminal voltage of the battery during charging? What is the purpose of having a series resistor in the charging circuit?
- 3.12** In a potentiometer arrangement, a cell of emf 1.25 V gives a balance point at 35.0 cm length of the wire. If the cell is replaced by another cell and the balance point shifts to 63.0 cm, what is the emf of the second cell?
- 3.13** The number density of free electrons in a copper conductor estimated in Example 3.1 is $8.5 \times 10^{28} \text{ m}^{-3}$. How long does an electron take to drift from one end of a wire 3.0 m long to its other end? The area of cross-section of the wire is $2.0 \times 10^{-6} \text{ m}^2$ and it is carrying a current of 3.0 A.

ADDITIONAL EXERCISES

- 3.14** The earth's surface has a negative surface charge density of 10^{-9} C m^{-2} . The potential difference of 400 kV between the top of the atmosphere and the surface results (due to the low conductivity of the lower atmosphere) in a current of only 1800 A over the entire globe. If there were no mechanism of sustaining atmospheric electric

field, how much time (roughly) would be required to neutralise the earth's surface? (This never happens in practice because there is a mechanism to replenish electric charges, namely the continual thunderstorms and lightning in different parts of the globe). (Radius of earth = 6.37×10^6 m.)

- 3.15** (a) Six lead-acid type of secondary cells each of emf 2.0 V and internal resistance $0.015\ \Omega$ are joined in series to provide a supply to a resistance of $8.5\ \Omega$. What are the current drawn from the supply and its terminal voltage?
 (b) A secondary cell after long use has an emf of 1.9 V and a large internal resistance of $380\ \Omega$. What maximum current can be drawn from the cell? Could the cell drive the starting motor of a car?
- 3.16** Two wires of equal length, one of aluminium and the other of copper have the same resistance. Which of the two wires is lighter? Hence explain why aluminium wires are preferred for overhead power cables. ($\rho_{Al} = 2.63 \times 10^{-8}\ \Omega\ m$, $\rho_{Cu} = 1.72 \times 10^{-8}\ \Omega\ m$, Relative density of Al = 2.7, of Cu = 8.9.)
- 3.17** What conclusion can you draw from the following observations on a resistor made of alloy manganin?

Current A	Voltage V	Current A	Voltage V
0.2	3.94	3.0	59.2
0.4	7.87	4.0	78.8
0.6	11.8	5.0	98.6
0.8	15.7	6.0	118.5
1.0	19.7	7.0	138.2
2.0	39.4	8.0	158.0

- 3.18** Answer the following questions:
- (a) A steady current flows in a metallic conductor of non-uniform cross-section. Which of these quantities is constant along the conductor: current, current density, electric field, drift speed?
 (b) Is Ohm's law universally applicable for all conducting elements? If not, give examples of elements which do not obey Ohm's law.
 (c) A low voltage supply from which one needs high currents must have very low internal resistance. Why?
 (d) A high tension (HT) supply of, say, 6 kV must have a very large internal resistance. Why?
- 3.19** Choose the correct alternative:
- (a) Alloys of metals usually have (greater/less) resistivity than that of their constituent metals.
 (b) Alloys usually have much (lower/higher) temperature coefficients of resistance than pure metals.
 (c) The resistivity of the alloy manganin is nearly independent of/ increases rapidly with increase of temperature.
 (d) The resistivity of a typical insulator (e.g., amber) is greater than that of a metal by a factor of the order of $(10^{22}/10^{23})$.
- 3.20** (a) Given n resistors each of resistance R , how will you combine them to get the (i) maximum (ii) minimum effective resistance? What is the ratio of the maximum to minimum resistance?
 (b) Given the resistances of $1\ \Omega$, $2\ \Omega$, $3\ \Omega$, how will be combine them to get an equivalent resistance of (i) $(11/3)\ \Omega$ (ii) $(11/5)\ \Omega$, (iii) $6\ \Omega$, (iv) $(6/11)\ \Omega$?
 (c) Determine the equivalent resistance of networks shown in Fig. 3.31.

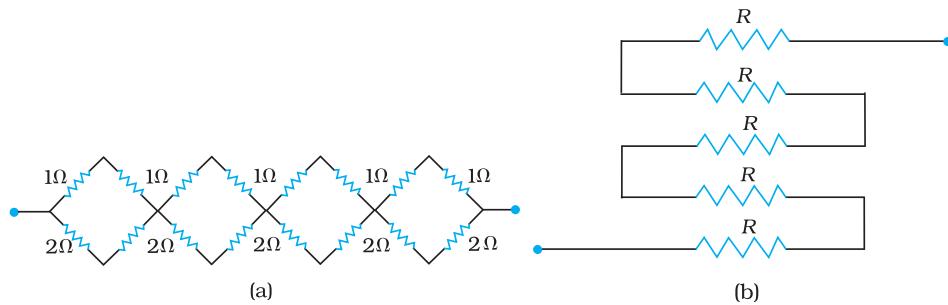


FIGURE 3.31

3.21 Determine the current drawn from a 12V supply with internal resistance 0.5Ω by the infinite network shown in Fig. 3.32. Each resistor has 1Ω resistance.

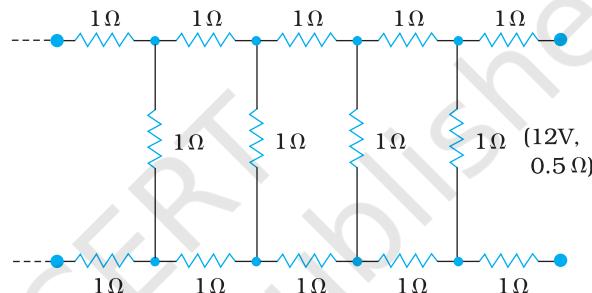


FIGURE 3.32

3.22 Figure 3.33 shows a potentiometer with a cell of 2.0 V and internal resistance $0.40\ \Omega$ maintaining a potential drop across the resistor wire AB. A standard cell which maintains a constant emf of 1.02 V (for very moderate currents upto a few mA) gives a balance point at 67.3 cm length of the wire. To ensure very low currents drawn from the standard cell, a very high resistance of $600\ k\Omega$ is put in series with it, which is shorted close to the balance point. The standard cell is then replaced by a cell of unknown emf ε and the balance point found similarly, turns out to be at 82.3 cm length of the wire.

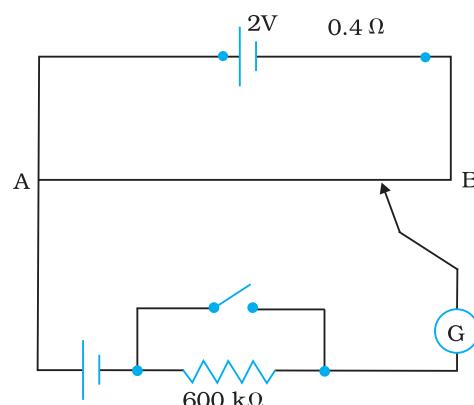


FIGURE 3.33

- (a) What is the value ε ?
 (b) What purpose does the high resistance of $600 \text{ k}\Omega$ have?

- (c) Is the balance point affected by this high resistance?
- (d) Is the balance point affected by the internal resistance of the driver cell?
- (e) Would the method work in the above situation if the driver cell of the potentiometer had an emf of 1.0V instead of 2.0V?
- (f) Would the circuit work well for determining an extremely small emf, say of the order of a few mV (such as the typical emf of a thermo-couple)? If not, how will you modify the circuit?

3.23 Figure 3.34 shows a potentiometer circuit for comparison of two resistances. The balance point with a standard resistor $R = 10.0 \Omega$ is found to be 58.3 cm, while that with the unknown resistance X is 68.5 cm. Determine the value of X . What might you do if you failed to find a balance point with the given cell of emf ε ?

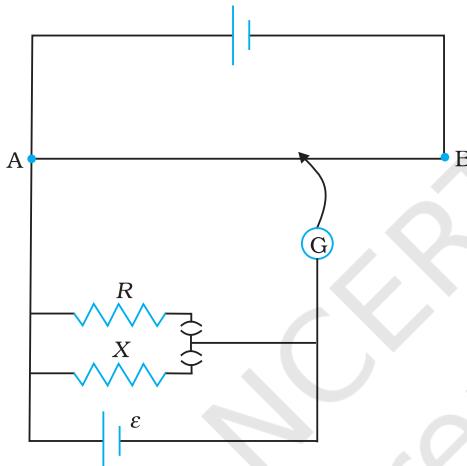


FIGURE 3.34

3.24 Figure 3.35 shows a 2.0 V potentiometer used for the determination of internal resistance of a 1.5 V cell. The balance point of the cell in open circuit is 76.3 cm. When a resistor of 9.5Ω is used in the external circuit of the cell, the balance point shifts to 64.8 cm length of the potentiometer wire. Determine the internal resistance of the cell.

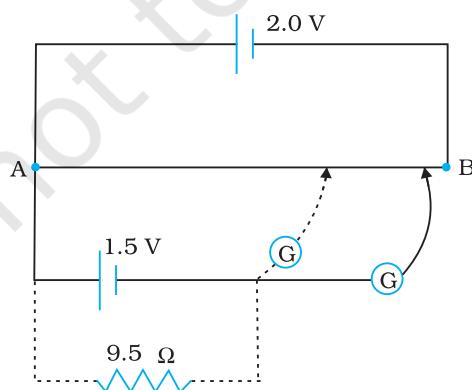


FIGURE 3.35

Chapter Four

MOVING CHARGES AND MAGNETISM

4.1 INTRODUCTION

Both Electricity and Magnetism have been known for more than 2000 years. However, it was only about 200 years ago, in 1820, that it was realised that they were intimately related*. During a lecture demonstration in the summer of 1820, the Danish physicist Hans Christian Oersted noticed that a current in a straight wire caused a noticeable deflection in a nearby magnetic compass needle. He investigated this phenomenon. He found that the alignment of the needle is tangential to an imaginary circle which has the straight wire as its centre and has its plane perpendicular to the wire. This situation is depicted in Fig. 4.1(a). It is noticeable when the current is large and the needle sufficiently close to the wire so that the earth's magnetic field may be ignored. Reversing the direction of the current reverses the orientation of the needle [Fig. 4.1(b)]. The deflection increases on increasing the current or bringing the needle closer to the wire. Iron filings sprinkled around the wire arrange themselves in concentric circles with the wire as the centre [Fig. 4.1(c)]. Oersted concluded that *moving charges or currents produced a magnetic field in the surrounding space.*

Following this there was intense experimentation. In 1864, the laws obeyed by electricity and magnetism were unified and formulated by

* See the box in Chapter 1, Page 3.

James Maxwell who then realised that light was electromagnetic waves. Radio waves were discovered by Hertz, and produced by J.C.Bose and G. Marconi by the end of the 19th century. A remarkable scientific and technological progress has taken place in the 20th century. This is due to our increased understanding of electromagnetism and the invention of devices for production, amplification, transmission and detection of electromagnetic waves.

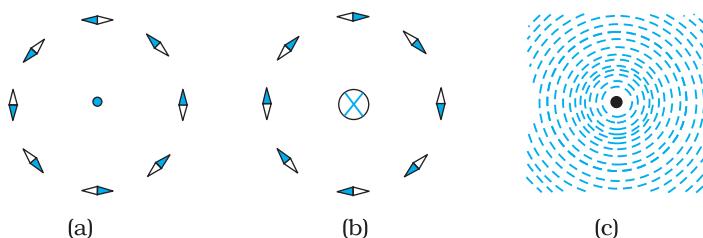


FIGURE 4.1 The magnetic field due to a straight long current-carrying wire. The wire is perpendicular to the plane of the paper. A ring of compass needles surrounds the wire. The orientation of the needles is shown when (a) the current emerges out of the plane of the paper, (b) the current moves into the plane of the paper. (c) The arrangement of iron filings around the wire. The darkened ends of the needle represent north poles. The effect of the earth's magnetic field is neglected.

In this chapter, we will see how magnetic field exerts forces on moving charged particles, like electrons, protons, and current-carrying wires. We shall also learn how currents produce magnetic fields. We shall see how particles can be accelerated to very high energies in a cyclotron. We shall study how currents and voltages are detected by a galvanometer.

In this and subsequent Chapter on magnetism, we adopt the following convention: A current or a field (electric or magnetic) emerging out of the plane of the paper is depicted by a dot (\odot). A current or a field going into the plane of the paper is depicted by a cross (\otimes)*. Figures. 4.1(a) and 4.1(b) correspond to these two situations, respectively.

4.2 MAGNETIC FORCE

4.2.1 Sources and fields

Before we introduce the concept of a magnetic field \mathbf{B} , we shall recapitulate what we have learnt in Chapter 1 about the electric field \mathbf{E} . We have seen that the interaction between two charges can be considered in two stages. The charge Q , the source of the field, produces an electric field \mathbf{E} , where



Hans Christian Oersted (1777–1851) Danish physicist and chemist, professor at Copenhagen. He observed that a compass needle suffers a deflection when placed near a wire carrying an electric current. This discovery gave the first empirical evidence of a connection between electric and magnetic phenomena.

HANS CHRISTIAN OERSTED (1777–1851)

* A dot appears like the tip of an arrow pointed at you, a cross is like the feathered tail of an arrow moving away from you.

Physics

HENDRIK ANTOON LORENTZ (1853 – 1928)



Hendrik Antoon Lorentz (1853 – 1928) Dutch theoretical physicist, professor at Leiden. He investigated the relationship between electricity, magnetism, and mechanics. In order to explain the observed effect of magnetic fields on emitters of light (Zeeman effect), he postulated the existence of electric charges in the atom, for which he was awarded the Nobel Prize in 1902. He derived a set of transformation equations (known after him, as Lorentz transformation equations) by some tangled mathematical arguments, but he was not aware that these equations hinge on a new concept of space and time.

$$\mathbf{E} = Q \hat{\mathbf{r}} / (4\pi\epsilon_0)r^2 \quad (4.1)$$

where $\hat{\mathbf{r}}$ is unit vector along \mathbf{r} , and the field \mathbf{E} is a vector field. A charge q interacts with this field and experiences a force \mathbf{F} given by

$$\mathbf{F} = q \mathbf{E} = q Q \hat{\mathbf{r}} / (4\pi\epsilon_0) r^2 \quad (4.2)$$

As pointed out in the Chapter 1, the field \mathbf{E} is not just an artefact but has a physical role. It can convey energy and momentum and is not established instantaneously but takes finite time to propagate. The concept of a field was specially stressed by Faraday and was incorporated by Maxwell in his unification of electricity and magnetism. In addition to depending on each point in space, it can also vary with time, i.e., be a function of time. In our discussions in this chapter, we will assume that the fields do not change with time.

The field at a particular point can be due to one or more charges. If there are more charges the fields add vectorially. You have already learnt in Chapter 1 that this is called the principle of superposition. Once the field is known, the force on a test charge is given by Eq. (4.2).

Just as static charges produce an electric field, the currents or moving charges produce (in addition) a magnetic field, denoted by $\mathbf{B}(\mathbf{r})$, again a vector field. It has several basic properties identical to the electric field. It is defined at each point in space (and can in addition depend on time). Experimentally, it is found to obey the principle of superposition: *the magnetic field of several sources is the vector addition of magnetic field of each individual source.*

4.2.2 Magnetic Field, Lorentz Force

Let us suppose that there is a point charge q (moving with a velocity \mathbf{v} and, located at \mathbf{r} at a given time t) in presence of both the electric field $\mathbf{E}(\mathbf{r})$ and the magnetic field $\mathbf{B}(\mathbf{r})$. The force on an electric charge q due to both of them can be written as

$$\mathbf{F} = q [\mathbf{E}(\mathbf{r}) + \mathbf{v} \times \mathbf{B}(\mathbf{r})] \equiv \mathbf{F}_{\text{electric}} + \mathbf{F}_{\text{magnetic}} \quad (4.3)$$

This force was given first by H.A. Lorentz based on the extensive experiments of Ampere and others. It is called the *Lorentz force*. You have already studied in detail the force due to the electric field. If we look at the interaction with the magnetic field, we find the following features.

- (i) It depends on q , \mathbf{v} and \mathbf{B} (charge of the particle, the velocity and the magnetic field). *Force on a negative charge is opposite to that on a positive charge.*
- (ii) The magnetic force $q [\mathbf{v} \times \mathbf{B}]$ includes a vector product of velocity and magnetic field. The vector product makes the force due to magnetic

Moving Charges and Magnetism

field vanish (become zero) if velocity and magnetic field are parallel or anti-parallel. The force acts in a (sideways) direction perpendicular to both the velocity and the magnetic field.

Its direction is given by the screw rule or right hand rule for vector (or cross) product as illustrated in Fig. 4.2.

- (iii) The magnetic force is zero if charge is not moving (as then $|\mathbf{v}| = 0$). Only a moving charge feels the magnetic force.

The expression for the magnetic force helps us to define the unit of the magnetic field, if one takes q , \mathbf{F} and \mathbf{v} , all to be unity in the force equation $\mathbf{F} = q [\mathbf{v} \times \mathbf{B}] = q v B \sin \theta \hat{\mathbf{n}}$, where θ is the angle between \mathbf{v} and \mathbf{B} [see Fig. 4.2 (a)]. The magnitude of magnetic field B is 1 SI unit, when the force acting on a unit charge (1 C), moving perpendicular to \mathbf{B} with a speed 1 m/s, is one newton.

Dimensionally, we have $[B] = [F/qv]$ and the unit of \mathbf{B} are Newton second / (coulomb metre). This unit is called *tesla* (T) named after Nikola Tesla (1856 – 1943). Tesla is a rather large unit. A smaller unit (non-SI) called *gauss* ($= 10^{-4}$ tesla) is also often used. The earth's magnetic field is about 3.6×10^{-5} T. Table 4.1 lists magnetic fields over a wide range in the universe.

TABLE 4.1 ORDER OF MAGNITUDES OF MAGNETIC FIELDS IN A VARIETY OF PHYSICAL SITUATIONS

Physical situation	Magnitude of \mathbf{B} (in tesla)
Surface of a neutron star	10^8
Typical large field in a laboratory	1
Near a small bar magnet	10^{-2}
On the earth's surface	10^{-5}
Human nerve fibre	10^{-10}
Interstellar space	10^{-12}

4.2.3 Magnetic force on a current-carrying conductor

We can extend the analysis for force due to magnetic field on a single moving charge to a straight rod carrying current. Consider a rod of a uniform cross-sectional area A and length l . We shall assume one kind of mobile carriers as in a conductor (here electrons). Let the number density of these mobile charge carriers in it be n . Then the total number of mobile charge carriers in it is nlA . For a steady current I in this conducting rod, we may assume that each mobile carrier has an average

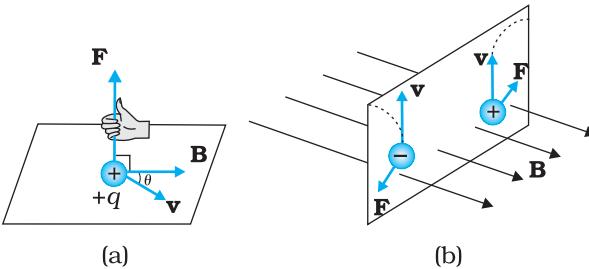


FIGURE 4.2 The direction of the magnetic force acting on a charged particle. (a) The force on a positively charged particle with velocity \mathbf{v} and making an angle θ with the magnetic field \mathbf{B} is given by the right-hand rule. (b) A moving charged particle q is deflected in an opposite sense to $-q$ in the presence of magnetic field.

Physics

drift velocity \mathbf{v}_d (see Chapter 3). In the presence of an external magnetic field \mathbf{B} , the force on these carriers is:

$$\mathbf{F} = (nlA)q \mathbf{v}_d \times \mathbf{B}$$

where q is the value of the charge on a carrier. Now $nq\mathbf{v}_d$ is the current density \mathbf{j} and $|(nq\mathbf{v}_d)| A$ is the current I (see Chapter 3 for the discussion of current and current density). Thus,

$$\begin{aligned}\mathbf{F} &= [(nq\mathbf{v}_d)l A] \times \mathbf{B} = [\mathbf{j}Al] \times \mathbf{B} \\ &= I\mathbf{l} \times \mathbf{B}\end{aligned}\quad (4.4)$$

where \mathbf{l} is a vector of magnitude l , the length of the rod, and with a direction identical to the current I . Note that the current I is not a vector. In the last step leading to Eq. (4.4), we have transferred the vector sign from \mathbf{j} to \mathbf{l} .

Equation (4.4) holds for a straight rod. In this equation, \mathbf{B} is the external magnetic field. It is not the field produced by the current-carrying rod. If the wire has an arbitrary shape we can calculate the Lorentz force on it by considering it as a collection of linear strips $d\mathbf{l}_j$ and summing

$$\mathbf{F} = \sum_j Id\mathbf{l}_j \times \mathbf{B}$$

This summation can be converted to an integral in most cases.

ON PERMITTIVITY AND PERMEABILITY

In the universal law of gravitation, we say that any two point masses exert a force on each other which is proportional to the product of the masses m_1 , m_2 and inversely proportional to the square of the distance r between them. We write it as $F = Gm_1m_2/r^2$ where G is the universal constant of gravitation. Similarly in Coulomb's law of electrostatics we write the force between two point charges q_1 , q_2 , separated by a distance r as $F = kq_1q_2/r^2$ where k is a constant of proportionality. In SI units, k is taken as $1/4\pi\epsilon$ where ϵ is the permittivity of the medium. Also in magnetism, we get another constant, which in SI units, is taken as $\mu/4\pi$ where μ is the permeability of the medium.

Although G , ϵ and μ arise as proportionality constants, there is a difference between gravitational force and electromagnetic force. While the gravitational force does not depend on the intervening medium, the electromagnetic force depends on the medium between the two charges or magnets. Hence while G is a universal constant, ϵ and μ depend on the medium. They have different values for different media. The product $\epsilon\mu$ turns out to be related to the speed v of electromagnetic radiation in the medium through $\epsilon\mu = 1/v^2$.

Electric permittivity ϵ is a physical quantity that describes how an electric field affects and is affected by a medium. It is determined by the ability of a material to polarise in response to an applied field, and thereby to cancel, partially, the field inside the material. Similarly, magnetic permeability μ is the ability of a substance to acquire magnetisation in magnetic fields. It is a measure of the extent to which magnetic field can penetrate matter.

EXAMPLE 4.1

Example 4.1 A straight wire of mass 200 g and length 1.5 m carries a current of 2 A. It is suspended in mid-air by a uniform horizontal magnetic field \mathbf{B} (Fig. 4.3). What is the magnitude of the magnetic field?

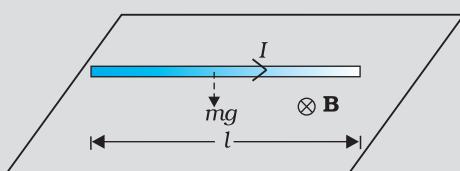


FIGURE 4.3

Solution From Eq. (4.4), we find that there is an upward force \mathbf{F} , of magnitude IlB . For mid-air suspension, this must be balanced by the force due to gravity:

$$mg = IlB$$

$$B = \frac{mg}{Il}$$

$$= \frac{0.2 \times 9.8}{2 \times 1.5} = 0.65 \text{ T}$$

Note that it would have been sufficient to specify m/l , the mass per unit length of the wire. The earth's magnetic field is approximately $4 \times 10^{-5} \text{ T}$ and we have ignored it.

Example 4.2 If the magnetic field is parallel to the positive y -axis and the charged particle is moving along the positive x -axis (Fig. 4.4), which way would the Lorentz force be for (a) an electron (negative charge), (b) a proton (positive charge).

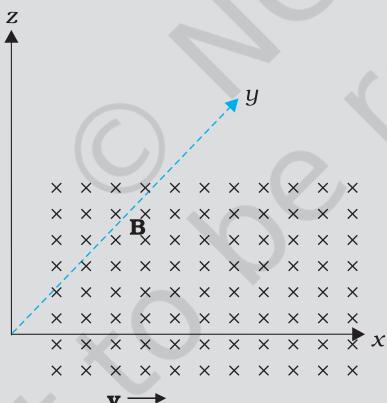


FIGURE 4.4

Solution The velocity \mathbf{v} of particle is along the x -axis, while \mathbf{B} , the magnetic field is along the y -axis, so $\mathbf{v} \times \mathbf{B}$ is along the z -axis (screw rule or right-hand thumb rule). So, (a) for electron it will be along $-z$ axis. (b) for a positive charge (proton) the force is along $+z$ axis.



Charged particles moving in a magnetic field.
Interactive demonstration:
<http://www.phys.hawaii.edu/~teb/optics/java/partmagn/index.html>

EXAMPLE 4.1

EXAMPLE 4.2

4.3 MOTION IN A MAGNETIC FIELD

We will now consider, in greater detail, the motion of a charge moving in a magnetic field. We have learnt in Mechanics (see Class XI book, Chapter 6) that a force on a particle does work if the force has a component along (or opposed to) the direction of motion of the particle. In the case of motion

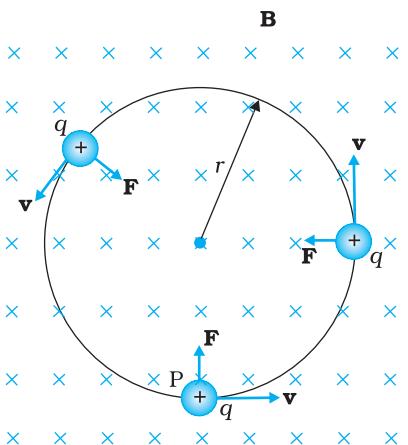


FIGURE 4.5 Circular motion

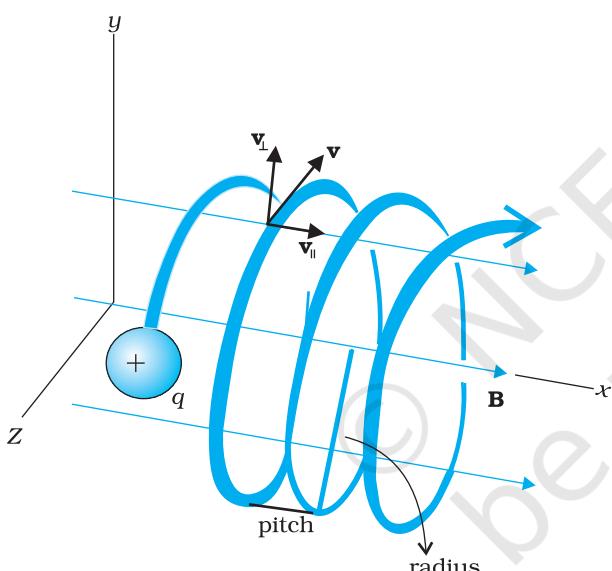


FIGURE 4.6 Helical motion

the larger is the radius and bigger the circle described. If ω is the angular frequency, then $v = \omega r$. So,

$$\omega = 2\pi v = qB/m \quad [4.6(a)]$$

which is independent of the velocity or energy. Here v is the frequency of rotation. The independence of v from energy has important application in the design of a cyclotron (see Section 4.4.2).

The time taken for one revolution is $T = 2\pi/\omega = 1/v$. If there is a component of the velocity parallel to the magnetic field (denoted by $v_{||}$), it will make the particle move along the field and the path of the particle would be a helical one (Fig. 4.6). The distance moved along the magnetic field in one rotation is called pitch p . Using Eq. [4.6 (a)], we have

$$p = v_{||} T = 2\pi m v_{||} / qB \quad [4.6(b)]$$

The radius of the circular component of motion is called the *radius of the helix*.

of a charge in a magnetic field, the magnetic force is perpendicular to the velocity of the particle. So no work is done and no change in the magnitude of the velocity is produced (though the direction of momentum may be changed). [Notice that this is unlike the force due to an electric field, $q\mathbf{E}$, which can have a component parallel (or antiparallel) to motion and thus can transfer energy in addition to momentum.]

We shall consider motion of a charged particle in a *uniform* magnetic field. First consider the case of \mathbf{v} perpendicular to \mathbf{B} . The perpendicular force, $q\mathbf{v} \times \mathbf{B}$, acts as a centripetal force and produces a circular motion perpendicular to the magnetic field. *The particle will describe a circle if \mathbf{v} and \mathbf{B} are perpendicular to each other* (Fig. 4.5).

If velocity has a component along \mathbf{B} , this component remains unchanged as the motion along the magnetic field will

not be affected by the magnetic field. The motion in a plane perpendicular to \mathbf{B} is as before a circular one, thereby producing a *helical motion* (Fig. 4.6).

You have already learnt in earlier classes (See Class XI, Chapter 4) that if r is the radius of the circular path of a particle, then a force of $m v^2 / r$, acts perpendicular to the path towards the centre of the circle, and is called the centripetal force. If the velocity \mathbf{v} is perpendicular to the magnetic field \mathbf{B} , the magnetic force is perpendicular to both \mathbf{v} and \mathbf{B} and acts like a centripetal force. It has a magnitude $q v B$. Equating the two expressions for centripetal force,

$$m v^2 / r = q v B, \text{ which gives}$$

$$r = m v / qB \quad (4.5)$$

for the radius of the circle described by the charged particle. The larger the momentum,

Example 4.3 What is the radius of the path of an electron (mass 9×10^{-31} kg and charge 1.6×10^{-19} C) moving at a speed of 3×10^7 m/s in a magnetic field of 6×10^{-4} T perpendicular to it? What is its frequency? Calculate its energy in keV. ($1 \text{ eV} = 1.6 \times 10^{-19}$ J).

Solution Using Eq. (4.5) we find

$$r = m v / (qB) = 9 \times 10^{-31} \text{ kg} \times 3 \times 10^7 \text{ m s}^{-1} / (1.6 \times 10^{-19} \text{ C} \times 6 \times 10^{-4} \text{ T}) \\ = 26 \times 10^{-2} \text{ m} = 26 \text{ cm}$$

$$v = v / (2 \pi r) = 2 \times 10^6 \text{ s}^{-1} = 2 \times 10^6 \text{ Hz} = 2 \text{ MHz.}$$

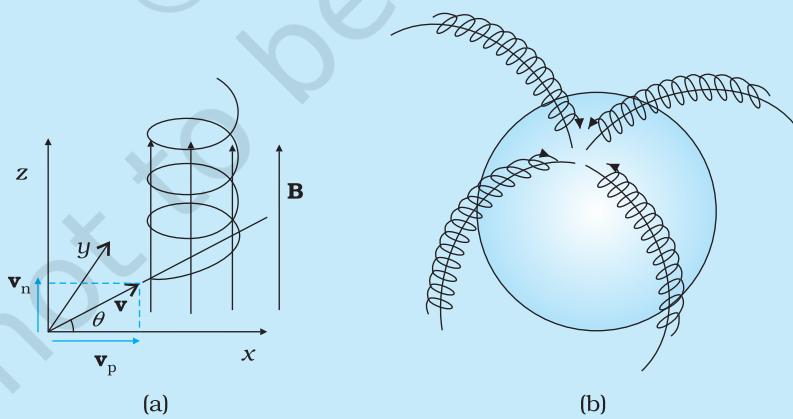
$$E = (\frac{1}{2}) m v^2 = (\frac{1}{2}) 9 \times 10^{-31} \text{ kg} \times 9 \times 10^{14} \text{ m}^2/\text{s}^2 = 40.5 \times 10^{-17} \text{ J} \\ \approx 4 \times 10^{-16} \text{ J} = 2.5 \text{ keV.}$$

EXAMPLE 4.3

HELICAL MOTION OF CHARGED PARTICLES AND AURORA BOREALIS

In polar regions like Alaska and Northern Canada, a splendid display of colours is seen in the sky. The appearance of dancing green pink lights is fascinating, and equally puzzling. An explanation of this natural phenomenon is now found in physics, in terms of what we have studied here.

Consider a charged particle of mass m and charge q , entering a region of magnetic field \mathbf{B} with an initial velocity \mathbf{v} . Let this velocity have a component \mathbf{v}_p parallel to the magnetic field and a component \mathbf{v}_n normal to it. There is no force on a charged particle in the direction of the field. Hence the particle continues to travel with the velocity \mathbf{v}_p parallel to the field. The normal component \mathbf{v}_n of the particle results in a Lorentz force $(\mathbf{v}_n \times \mathbf{B})$ which is perpendicular to both \mathbf{v}_n and \mathbf{B} . As seen in Section 4.3.1 the particle thus has a tendency to perform a circular motion in a plane perpendicular to the magnetic field. When this is coupled with the velocity parallel to the field, the resulting trajectory will be a helix along the magnetic field line, as shown in Figure (a) here. Even if the field line bends, the helically moving particle is trapped and guided to move around the field line. Since the Lorentz force is normal to the velocity of each point, the field does no work on the particle and the magnitude of velocity remains the same.



During a solar flare, a large number of electrons and protons are ejected from the sun. Some of them get trapped in the earth's magnetic field and move in helical paths along the field lines. The field lines come closer to each other near the magnetic poles; see figure (b). Hence the density of charges increases near the poles. These particles collide with atoms and molecules of the atmosphere. Excited oxygen atoms emit green light and excited nitrogen atoms emit pink light. This phenomenon is called *Aurora Borealis* in physics.

4.4 MOTION IN COMBINED ELECTRIC AND MAGNETIC FIELDS

4.4.1 Velocity selector

You know that a charge q moving with velocity \mathbf{v} in presence of both electric and magnetic fields experiences a force given by Eq. (4.3), that is,

$$\mathbf{F} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B}) = \mathbf{F}_E + \mathbf{F}_B$$

We shall consider the simple case in which electric and magnetic fields are perpendicular to each other and also perpendicular to the velocity of the particle, as shown in Fig. 4.7. We have,

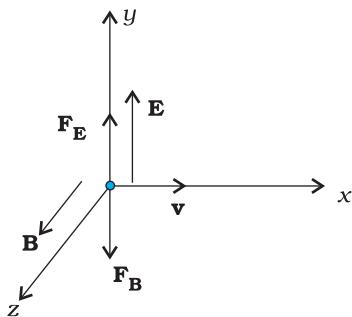


FIGURE 4.7

$$\mathbf{F}_E = q\mathbf{E} = qE\hat{\mathbf{j}}, \mathbf{F}_B = q\mathbf{v} \times \mathbf{B}, = q(v\hat{\mathbf{i}} \times B\hat{\mathbf{k}}) = -qB\hat{\mathbf{j}}$$

$$\text{Therefore, } \mathbf{F} = q(E - vB)\hat{\mathbf{j}}.$$

Thus, electric and magnetic forces are in opposite directions as shown in the figure. Suppose, we adjust the value of \mathbf{E} and \mathbf{B} such that magnitudes of the two forces are equal. Then, total force on the charge is zero and the charge will move in the fields undeflected. This happens when,

$$qE = qvB \quad \text{or} \quad v = \frac{E}{B} \quad (4.7)$$

This condition can be used to select charged particles of a particular velocity out of a beam containing charges moving with different speeds (irrespective of their charge and mass). The crossed E and B fields, therefore, serve as a *velocity selector*. Only particles with speed E/B pass undeflected through the region of crossed fields. This method was employed by J. J. Thomson in 1897 to measure the charge to mass ratio (e/m) of an electron. The principle is also employed in Mass Spectrometer – a device that separates charged particles, usually ions, according to their charge to mass ratio.

4.4.2 Cyclotron

The cyclotron is a machine to accelerate charged particles or ions to high energies. It was invented by E.O. Lawrence and M.S. Livingston in 1934 to investigate nuclear structure. The cyclotron uses both electric and magnetic fields in combination to increase the energy of charged particles. As the fields are perpendicular to each other they are called *crossed fields*. Cyclotron uses the fact that the frequency of revolution of the charged particle in a magnetic field is independent of its energy. The particles move most of the time inside two semicircular disc-like metal containers, D_1 and D_2 , which are called *dees* as they look like the letter D. Figure 4.8 shows a schematic view of the cyclotron. Inside the metal boxes the particle is shielded and is not acted on by the electric field. The magnetic field, however, acts on the particle and makes it go round in a circular path inside a dee. Every time the particle moves from one dee to another it is acted upon by the electric field. The sign of the electric field is changed alternately in tune with the circular motion of the particle. This ensures that the particle is always accelerated by the electric field. Each time the acceleration increases the energy of the particle. As energy

Moving Charges and Magnetism

increases, the radius of the circular path increases. So the path is a spiral one.

The whole assembly is evacuated to minimise collisions between the ions and the air molecules. A high frequency alternating voltage is applied to the dees. In the sketch shown in Fig. 4.8, positive ions or positively charged particles (e.g., protons) are released at the centre P. They move in a semi-circular path in one of the dees and arrive in the gap between the dees in a time interval $T/2$; where T , the period of revolution, is given by Eq. (4.6),

$$T = \frac{1}{v_c} = \frac{2\pi m}{qB}$$

or $v_c = \frac{qB}{2\pi m}$ (4.8)

This frequency is called the *cyclotron frequency* for obvious reasons and is denoted by v_c .

The frequency v_a of the applied voltage is adjusted so that the polarity of the dees is reversed in the same time that it takes the ions to complete one half of the revolution. The requirement $v_a = v_c$ is called the *resonance condition*. The phase of the supply is adjusted so that when the positive ions arrive at the edge of D_1 , D_2 is at a lower potential and the ions are accelerated across the gap. Inside the dees the particles travel in a region free of the electric field. The increase in their kinetic energy is qV each time they cross from one dee to another (V refers to the voltage across the dees at that time). From Eq. (4.5), it is clear that the radius of their path goes on increasing each time their kinetic energy increases. The ions are repeatedly accelerated across the dees until they have the required energy to have a radius approximately that of the dees. They are then deflected by a magnetic field and leave the system via an exit slit. From Eq. (4.5) we have,

$$v = \frac{qBR}{m} \quad (4.9)$$

where R is the radius of the trajectory at exit, and equals the radius of a dee.

Hence, the kinetic energy of the ions is,

$$\frac{1}{2}mv^2 = \frac{q^2B^2R^2}{2m} \quad (4.10)$$

The operation of the cyclotron is based on the fact that the time for one revolution of an ion is independent of its speed or radius of its orbit. The cyclotron is used to bombard nuclei with energetic particles, so accelerated by it, and study

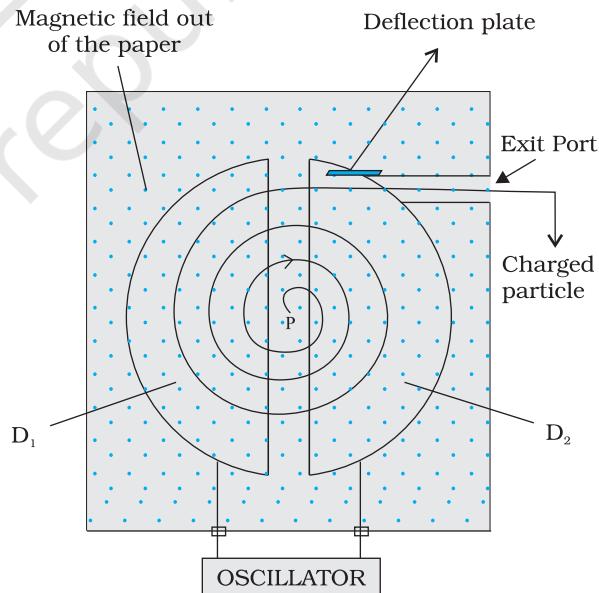


FIGURE 4.8 A schematic sketch of the cyclotron. There is a source of charged particles or ions at P which move in a circular fashion in the dees, D_1 and D_2 , on account of a uniform perpendicular magnetic field B . An alternating voltage source accelerates these ions to high speeds. The ions are eventually 'extracted' at the exit port.

Physics

the resulting nuclear reactions. It is also used to implant ions into solids and modify their properties or even synthesise new materials. It is used in hospitals to produce radioactive substances which can be used in diagnosis and treatment.

EXAMPLE 4.4

Example 4.4 A cyclotron's oscillator frequency is 10 MHz. What should be the operating magnetic field for accelerating protons? If the radius of its 'dees' is 60 cm, what is the kinetic energy (in MeV) of the proton beam produced by the accelerator.

$$(e = 1.60 \times 10^{-19} \text{ C}, m_p = 1.67 \times 10^{-27} \text{ kg}, 1 \text{ MeV} = 1.6 \times 10^{-13} \text{ J}).$$

Solution The oscillator frequency should be same as proton's cyclotron frequency.

Using Eqs. (4.5) and [4.6(a)] we have

$$B = 2\pi m v / q = 6.3 \times 1.67 \times 10^{-27} \times 10^7 / (1.6 \times 10^{-19}) = 0.66 \text{ T}$$

Final velocity of protons is

$$v = r \times 2\pi f = 0.6 \text{ m} \times 6.3 \times 10^7 = 3.78 \times 10^7 \text{ m/s.}$$

$$E = \frac{1}{2} mv^2 = 1.67 \times 10^{-27} \times 14.3 \times 10^{14} / (2 \times 1.6 \times 10^{-13}) = 7 \text{ MeV.}$$

ACCELERATORS IN INDIA

India has been an early entrant in the area of accelerator-based research. The vision of Dr. Meghnath Saha created a 37" Cyclotron in the Saha Institute of Nuclear Physics in Kolkata in 1953. This was soon followed by a series of Cockcroft-Walton type of accelerators established in Tata Institute of Fundamental Research (TIFR), Mumbai, Aligarh Muslim University (AMU), Aligarh, Bose Institute, Kolkata and Andhra University, Waltair.

The sixties saw the commissioning of a number of Van de Graaff accelerators: a 5.5 MV terminal machine in Bhabha Atomic Research Centre (BARC), Mumbai (1963); a 2 MV terminal machine in Indian Institute of Technology (IIT), Kanpur; a 400 kV terminal machine in Banaras Hindu University (BHU), Varanasi; and Punjabi University, Patiala. One 66 cm Cyclotron donated by the Rochester University of USA was commissioned in Panjab University, Chandigarh. A small electron accelerator was also established in University of Pune, Pune.

In a major initiative taken in the seventies and eighties, a Variable Energy Cyclotron was built indigenously in Variable Energy Cyclotron Centre (VECC), Kolkata; 2 MV Tandem Van de Graaff accelerator was developed and built in BARC and a 14 MV Tandem Pelletron accelerator was installed in TIFR.

This was soon followed by a 15 MV Tandem Pelletron established by University Grants Commission (UGC), as an inter-university facility in Inter-University Accelerator Centre (IUAC), New Delhi; a 3 MV Tandem Pelletron in Institute of Physics, Bhubaneshwar; and two 1.7 MV Tandetrons in Atomic Minerals Directorate for Exploration and Research, Hyderabad and Indira Gandhi Centre for Atomic Research, Kalpakkam. Both TIFR and IUAC are augmenting their facilities with the addition of superconducting LINAC modules to accelerate the ions to higher energies.

Besides these ion accelerators, the Department of Atomic Energy (DAE) has developed many electron accelerators. A 2 GeV Synchrotron Radiation Source is being built in Raja Ramanna Centre for Advanced Technologies, Indore.

The Department of Atomic Energy is considering Accelerator Driven Systems (ADS) for power production and fissile material breeding as future options.

4.5 MAGNETIC FIELD DUE TO A CURRENT ELEMENT, BIOT-SAVART LAW

All magnetic fields that we know are due to currents (or moving charges) and due to intrinsic magnetic moments of particles. Here, we shall study the relation between current and the magnetic field it produces.

It is given by the Biot-Savart's law. Figure 4.9 shows a finite conductor XY carrying current I . Consider an infinitesimal element $d\mathbf{l}$ of the conductor. The magnetic field $d\mathbf{B}$ due to this element is to be determined at a point P which is at a distance r from it. Let θ be the angle between $d\mathbf{l}$ and the displacement vector \mathbf{r} . According to Biot-Savart's law, the magnitude of the magnetic field $d\mathbf{B}$ is proportional to the current I , the element length $|d\mathbf{l}|$, and inversely proportional to the square of the distance r . Its direction* is perpendicular to the plane containing $d\mathbf{l}$ and \mathbf{r} .

Thus, in vector notation,

$$\begin{aligned} d\mathbf{B} &\propto \frac{I d\mathbf{l} \times \mathbf{r}}{r^3} \\ &= \frac{\mu_0}{4\pi} \frac{I d\mathbf{l} \times \mathbf{r}}{r^3} \end{aligned} \quad [4.11(a)]$$

where $\mu_0/4\pi$ is a constant of proportionality. The above expression holds when the medium is vacuum.

The magnitude of this field is,

$$|d\mathbf{B}| = \frac{\mu_0}{4\pi} \frac{I d\mathbf{l} \sin \theta}{r^2} \quad [4.11(b)]$$

where we have used the property of cross-product. Equation [4.11 (a)] constitutes our basic equation for the magnetic field. The proportionality constant in SI units has the exact value,

$$\frac{\mu_0}{4\pi} = 10^{-7} \text{ Tm/A} \quad [4.11(c)]$$

We call μ_0 the *permeability of free space* (or vacuum).

The Biot-Savart law for the magnetic field has certain similarities as well as differences with the Coulomb's law for the electrostatic field. Some of these are:

- (i) Both are long range, since both depend inversely on the square of distance from the source to the point of interest. The principle of superposition applies to both fields. [In this connection, note that the magnetic field is *linear* in the source $I d\mathbf{l}$ just as the electrostatic field is linear in its source: the electric charge.]
- (ii) The electrostatic field is produced by a scalar source, namely, the electric charge. The magnetic field is produced by a vector source $I d\mathbf{l}$.

* The sense of $d\mathbf{l} \times \mathbf{r}$ is also given by the *Right Hand Screw rule*: Look at the plane containing vectors $d\mathbf{l}$ and \mathbf{r} . Imagine moving from the first vector towards second vector. If the movement is anticlockwise, the resultant is towards you. If it is clockwise, the resultant is away from you.

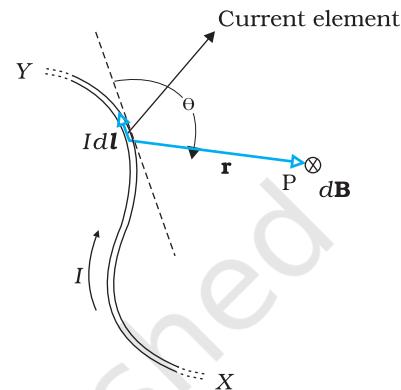


FIGURE 4.9 Illustration of the Biot-Savart law. The current element $I d\mathbf{l}$ produces a field $d\mathbf{B}$ at a distance r . The \otimes sign indicates that the field is perpendicular to the plane of this page and directed into it.

- (iii) The electrostatic field is along the displacement vector joining the source and the field point. The magnetic field is perpendicular to the plane containing the displacement vector \mathbf{r} and the current element $I d\mathbf{l}$.
- (iv) There is an angle dependence in the Biot-Savart law which is not present in the electrostatic case. In Fig. 4.9, the magnetic field at any point in the direction of $d\mathbf{l}$ (the dashed line) is zero. Along this line, $\theta = 0$, $\sin \theta = 0$ and from Eq. [4.11(a)], $|d\mathbf{B}| = 0$.

There is an interesting relation between ϵ_0 , the permittivity of free space; μ_0 , the permeability of free space; and c , the speed of light in vacuum:

$$\epsilon_0 \mu_0 = (4\pi \epsilon_0) \left(\frac{\mu_0}{4\pi} \right) = \left(\frac{1}{9 \times 10^9} \right) (10^{-7}) = \frac{1}{(3 \times 10^8)^2} = \frac{1}{c^2}$$

We will discuss this connection further in Chapter 8 on the electromagnetic waves. Since the speed of light in vacuum is constant, the product $\mu_0 \epsilon_0$ is fixed in magnitude. Choosing the value of either ϵ_0 or μ_0 , fixes the value of the other. In SI units, μ_0 is fixed to be equal to $4\pi \times 10^{-7}$ in magnitude.

Example 4.5 An element $\Delta \mathbf{l} = \Delta x \hat{\mathbf{i}}$ is placed at the origin and carries a large current $I = 10$ A (Fig. 4.10). What is the magnetic field on the y -axis at a distance of 0.5 m. $\Delta x = 1$ cm.

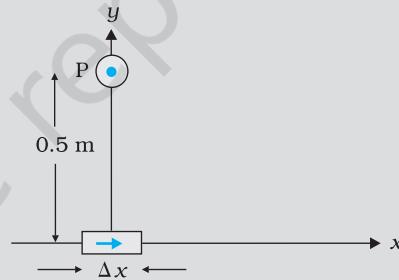


FIGURE 4.10

Solution

$$|d\mathbf{B}| = \frac{\mu_0}{4\pi} \frac{I d\mathbf{l} \sin \theta}{r^2} \quad [\text{using Eq. (4.11)}]$$

$$d\mathbf{l} = \Delta x \hat{\mathbf{i}} = 10^{-2} \text{ m} \hat{\mathbf{i}}, \quad I = 10 \text{ A}, \quad r = 0.5 \text{ m} = y, \quad \mu_0 / 4\pi = 10^{-7} \frac{\text{T m}}{\text{A}}$$

$$\theta = 90^\circ; \sin \theta = 1$$

$$|d\mathbf{B}| = \frac{10^{-7} \times 10 \times 10^{-2}}{25 \times 10^{-2}} = 4 \times 10^{-8} \text{ T}$$

The direction of the field is in the $+z$ -direction. This is so since,

$$d\mathbf{l} \times \mathbf{r} = \Delta x \hat{\mathbf{i}} \times y \hat{\mathbf{j}} = y \Delta x (\hat{\mathbf{i}} \times \hat{\mathbf{j}}) = y \Delta x \hat{\mathbf{k}}$$

We remind you of the following cyclic property of cross-products,

$$\hat{\mathbf{i}} \times \hat{\mathbf{j}} = \hat{\mathbf{k}}; \hat{\mathbf{j}} \times \hat{\mathbf{k}} = \hat{\mathbf{i}}; \hat{\mathbf{k}} \times \hat{\mathbf{i}} = \hat{\mathbf{j}}$$

Note that the field is small in magnitude.

In the next section, we shall use the Biot-Savart law to calculate the magnetic field due to a circular loop.

4.6 MAGNETIC FIELD ON THE AXIS OF A CIRCULAR CURRENT LOOP

In this section, we shall evaluate the magnetic field due to a circular coil along its axis. The evaluation entails summing up the effect of infinitesimal current elements (Idl) mentioned in the previous section.

We assume that the current I is steady and that the evaluation is carried out in free space (i.e., vacuum).

Figure 4.11 depicts a circular loop carrying a steady current I . The loop is placed in the y - z plane with its centre at the origin O and has a radius R . The x -axis is the axis of the loop. We wish to calculate the magnetic field at the point P on this axis. Let x be the distance of P from the centre O of the loop.

Consider a conducting element dl of the loop. This is shown in Fig. 4.11. The magnitude dB of the magnetic field due to dl is given by the Biot-Savart law [Eq. 4.11(a)],

$$dB = \frac{\mu_0}{4\pi} \frac{I|dl \times r|}{r^3} \quad (4.12)$$

Now $r^2 = x^2 + R^2$. Further, any element of the loop will be perpendicular to the displacement vector from the element to the axial point. For example, the element dl in Fig. 4.11 is in the y - z plane whereas the displacement vector r from dl to the axial point P is in the x - y plane. Hence $|dl \times r| = r dl$. Thus,

$$dB = \frac{\mu_0}{4\pi} \frac{Idl}{(x^2 + R^2)} \quad (4.13)$$

The direction of $d\mathbf{B}$ is shown in Fig. 4.11. It is perpendicular to the plane formed by dl and r . It has an x -component $d\mathbf{B}_x$ and a component perpendicular to x -axis, $d\mathbf{B}_\perp$. When the components perpendicular to the x -axis are summed over, they cancel out and we obtain a null result. For example, the $d\mathbf{B}_\perp$ component due to dl is cancelled by the contribution due to the diametrically opposite dl element, shown in Fig. 4.11. Thus, only the x -component survives. The net contribution along x -direction can be obtained by integrating $dB_x = dB \cos \theta$ over the loop. For Fig. 4.11,

$$\cos \theta = \frac{R}{(x^2 + R^2)^{1/2}} \quad (4.14)$$

From Eqs. (4.13) and (4.14),

$$dB_x = \frac{\mu_0 I dl}{4\pi} \frac{R}{(x^2 + R^2)^{3/2}}$$

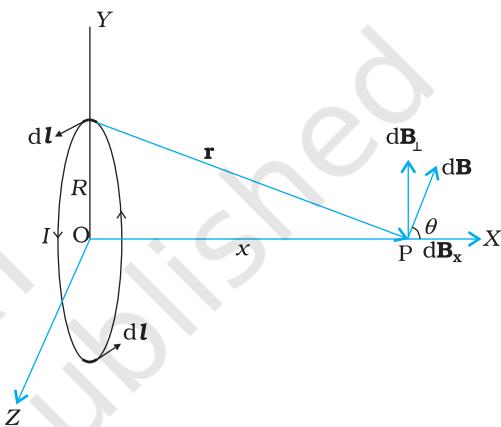


FIGURE 4.11 Magnetic field on the axis of a current carrying circular loop of radius R . Shown are the magnetic field $d\mathbf{B}$ (due to a line element dl) and its components along and perpendicular to the axis.

The summation of elements dl over the loop yields $2\pi R$, the circumference of the loop. Thus, the magnetic field at P due to entire circular loop is

$$\mathbf{B} = B_x \hat{\mathbf{i}} = \frac{\mu_0 I R^2}{2(x^2 + R^2)^{3/2}} \hat{\mathbf{i}} \quad (4.15)$$

As a special case of the above result, we may obtain the field at the centre of the loop. Here $x = 0$, and we obtain,

$$\mathbf{B}_0 = \frac{\mu_0 I}{2R} \hat{\mathbf{i}} \quad (4.16)$$

The magnetic field lines due to a circular wire form closed loops and are shown in Fig. 4.12. The direction of the magnetic field is given by (another) *right-hand thumb rule* stated below:

Curl the palm of your right hand around the circular wire with the fingers pointing in the direction of the current. The right-hand thumb gives the direction of the magnetic field.

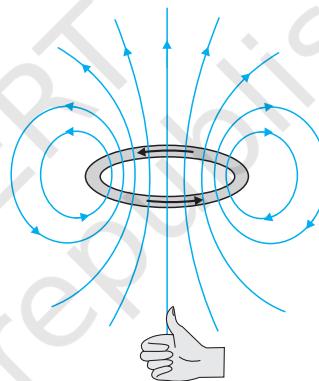


FIGURE 4.12 The magnetic field lines for a current loop. The direction of the field is given by the right-hand thumb rule described in the text. The upper side of the loop may be thought of as the north pole and the lower side as the south pole of a magnet.

Example 4.6 A straight wire carrying a current of 12 A is bent into a semi-circular arc of radius 2.0 cm as shown in Fig. 4.13(a). Consider the magnetic field \mathbf{B} at the centre of the arc. (a) What is the magnetic field due to the straight segments? (b) In what way the contribution to \mathbf{B} from the semicircle differs from that of a circular loop and in what way does it resemble? (c) Would your answer be different if the wire were bent into a semi-circular arc of the same radius but in the opposite way as shown in Fig. 4.13(b)?

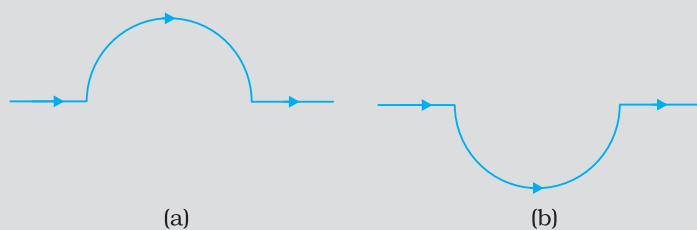


FIGURE 4.13

Solution

- (a) $d\mathbf{l}$ and \mathbf{r} for each element of the straight segments are parallel. Therefore, $d\mathbf{l} \times \mathbf{r} = 0$. Straight segments do not contribute to $|\mathbf{B}|$.
- (b) For all segments of the semicircular arc, $d\mathbf{l} \times \mathbf{r}$ are all parallel to each other (into the plane of the paper). All such contributions add up in magnitude. Hence direction of \mathbf{B} for a semicircular arc is given by the right-hand rule and magnitude is half that of a circular loop. Thus \mathbf{B} is 1.9×10^{-4} T normal to the plane of the paper going into it.
- (c) Same magnitude of \mathbf{B} but opposite in direction to that in (b).

EXAMPLE 4.6

Example 4.7 Consider a tightly wound 100 turn coil of radius 10 cm, carrying a current of 1 A. What is the magnitude of the magnetic field at the centre of the coil?

Solution Since the coil is tightly wound, we may take each circular element to have the same radius $R = 10$ cm = 0.1 m. The number of turns $N = 100$. The magnitude of the magnetic field is,

$$B = \frac{\mu_0 NI}{2R} = \frac{4\pi \times 10^{-7} \times 10^2 \times 1}{2 \times 10^{-1}} = 2\pi \times 10^{-4} = 6.28 \times 10^{-4}$$
 T

EXAMPLE 4.7

4.7 AMPERE'S CIRCUITAL LAW

There is an alternative and appealing way in which the Biot-Savart law may be expressed. Ampere's circuital law considers an open surface with a boundary (Fig. 4.14). The surface has current passing through it. We consider the boundary to be made up of a number of small line elements. Consider one such element of length dl . We take the value of the tangential component of the magnetic field, B_t , at this element and multiply it by the length of that element dl . [Note: $B_t dl = \mathbf{B} \cdot d\mathbf{l}$]. All such products are added together. We consider the limit as the lengths of elements get smaller and their number gets larger. The sum then tends to an integral. Ampere's law states that this integral is equal to μ_0 times the total current passing through the surface, i.e.,

$$\oint \mathbf{B} \cdot d\mathbf{l} = \mu_0 I \quad [4.17(a)]$$

where I is the total current through the surface. The integral is taken over the closed loop coinciding with the boundary C of the surface. The relation above involves a sign-convention, given by the right-hand rule. Let the fingers of the right-hand be curled in the sense the boundary is traversed in the loop integral $\oint \mathbf{B} \cdot d\mathbf{l}$. Then the direction of the thumb gives the sense in which the current I is regarded as positive.

For several applications, a much simplified version of Eq. [4.17(a)] proves sufficient. We shall assume that, in such cases, it is possible to choose the loop (called an *amperian loop*) such that at each point of the loop, either

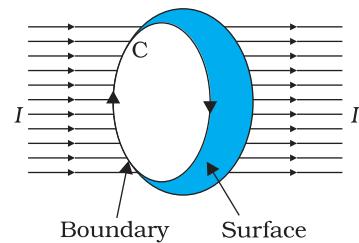


FIGURE 4.14



Andre Ampere (1775 – 1836) Andre Marie Ampere was a French physicist, mathematician and chemist who founded the science of electrodynamics. Ampere was a child prodigy who mastered advanced mathematics by the age of 12. Ampere grasped the significance of Oersted's discovery. He carried out a large series of experiments to explore the relationship between current electricity and magnetism. These investigations culminated in 1827 with the publication of the 'Mathematical Theory of Electrodynamic Phenomena Deduced Solely from Experiments'. He hypothesised that *all* magnetic phenomena are due to circulating electric currents. Ampere was humble and absent-minded. He once forgot an invitation to dine with the Emperor Napoleon. He died of pneumonia at the age of 61. His gravestone bears the epitaph: *Tandem Felix* (Happy at last).

ANDRE AMPERE (1775 – 1836)

- (i) **B** is tangential to the loop and is a non-zero constant **B**, or
- (ii) **B** is normal to the loop, or
- (iii) **B** vanishes.

Now, let L be the length (part) of the loop for which **B** is tangential. Let I_e be the current enclosed by the loop. Then, Eq. (4.17) reduces to,

$$BL = \mu_0 I_e \quad [4.17(b)]$$

When there is a system with a symmetry such as for a *straight infinite current-carrying wire* in Fig. 4.15, the Ampere's law enables an easy evaluation of the magnetic field, much the same way Gauss' law helps in determination of the electric field. This is exhibited in the Example 4.9 below. The boundary of the loop chosen is a circle and magnetic field is tangential to the circumference of the circle. The law gives, for the left hand side of Eq. [4.17 (b)], $B \cdot 2\pi r$. We find that the magnetic field at a distance r outside the wire is *tangential* and given by

$$\begin{aligned} B \times 2\pi r &= \mu_0 I, \\ B &= \mu_0 I / (2\pi r) \end{aligned} \quad (4.18)$$

The above result for the infinite wire is interesting from several points of view.

- (i) It implies that the field at every point on a circle of radius r , (with the wire along the axis), is same in magnitude. In other words, the magnetic field possesses what is called a *cylindrical symmetry*. The field that normally can depend on three coordinates depends only on one: r . Whenever there is symmetry, the solutions simplify.
- (ii) The field direction at any point on this circle is tangential to it. Thus, the lines of constant magnitude of magnetic field form concentric circles. Notice now, in Fig. 4.1(c), the iron filings form concentric circles. These lines called *magnetic field lines* form closed loops. This is unlike the electrostatic field lines which originate from positive charges and end at negative charges. The expression for the magnetic field of a straight wire provides a theoretical justification to Oersted's experiments.
- (iii) Another interesting point to note is that even though the wire is infinite, the field due to it at a nonzero distance is *not infinite*. It tends to blow up only when we come very close to the wire. The field is directly proportional to the current and inversely proportional to the distance from the (infinitely long) current source.

- (iv) There exists a simple rule to determine the direction of the magnetic field due to a long wire. This rule, called the *right-hand rule**, is:

Grasp the wire in your right hand with your extended thumb pointing in the direction of the current. Your fingers will curl around in the direction of the magnetic field.

Ampere's circuital law is not new in content from Biot-Savart law. Both relate the magnetic field and the current, and both express the same physical consequences of a steady electrical current. Ampere's law is to Biot-Savart law, what Gauss's law is to Coulomb's law. Both, Ampere's and Gauss's law relate a physical quantity on the periphery or boundary (magnetic or electric field) to another physical quantity, namely, the source, in the interior (current or charge). We also note that Ampere's circuital law holds for steady currents which do not fluctuate with time. The following example will help us understand what is meant by the term *enclosed current*.

Example 4.8 Figure 4.15 shows a long straight wire of a circular cross-section (radius a) carrying steady current I . The current I is uniformly distributed across this cross-section. Calculate the magnetic field in the region $r < a$ and $r > a$.

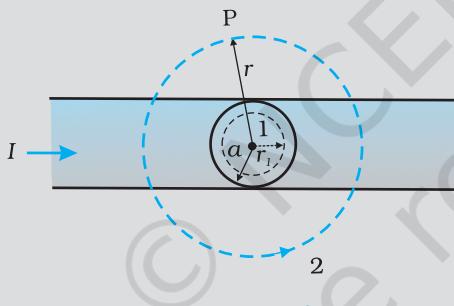


FIGURE 4.15

Solution (a) Consider the case $r > a$. The Amperian loop, labelled 2, is a circle concentric with the cross-section. For this loop,

$$L = 2\pi r$$

$$I_e = \text{Current enclosed by the loop} = I$$

The result is the familiar expression for a long straight wire

$$B(2\pi r) = \mu_0 I$$

$$B = \frac{\mu_0 I}{2\pi r} \quad [4.19(a)]$$

$$B \propto \frac{1}{r} \quad (r > a)$$

(b) Consider the case $r < a$. The Amperian loop is a circle labelled 1. For this loop, taking the radius of the circle to be r ,

$$L = 2\pi r$$

EXAMPLE 4.8

* Note that there are *two distinct* right-hand rules: One which gives the direction of \mathbf{B} on the axis of current-loop and the other which gives direction of \mathbf{B} for a straight conducting wire. Fingers and thumb play different roles in the two.

EXAMPLE 4.8

Now the current enclosed I_e is not I , but is less than this value. Since the current distribution is uniform, the current enclosed is,

$$I_e = I \left(\frac{\pi r^2}{\pi a^2} \right) = \frac{Ir^2}{a^2}$$

Using Ampere's law, $B(2\pi r) = \mu_0 \frac{Ir^2}{a^2}$

$$B = \left(\frac{\mu_0 I}{2\pi a^2} \right) r \quad [4.19(b)]$$

$B \propto r \quad (r < a)$

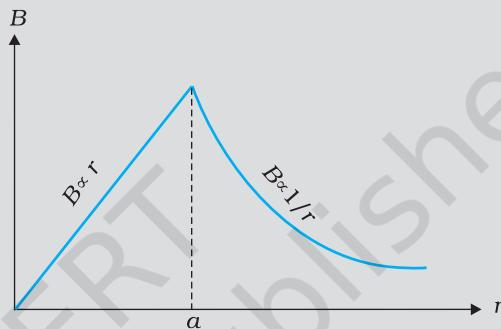


FIGURE 4.16

Figure (4.16) shows a plot of the magnitude of \mathbf{B} with distance r from the centre of the wire. The direction of the field is tangential to the respective circular loop (1 or 2) and given by the right-hand rule described earlier in this section.

This example possesses the required symmetry so that Ampere's law can be applied readily.

It should be noted that while Ampere's circuital law holds for any loop, it may not always facilitate an evaluation of the magnetic field in every case. For example, for the case of the circular loop discussed in Section 4.6, it cannot be applied to extract the simple expression $B = \mu_0 I / 2R$ [Eq. (4.16)] for the field at the centre of the loop. However, there exists a large number of situations of high symmetry where the law can be conveniently applied. We shall use it in the next section to calculate the magnetic field produced by two commonly used and very useful magnetic systems: the *solenoid* and the *toroid*.

4.8 THE SOLENOID AND THE TOROID

The solenoid and the toroid are two pieces of equipment which generate magnetic fields. The television uses the solenoid to generate magnetic fields needed. The synchrotron uses a combination of both to generate the high magnetic fields required. In both, solenoid and toroid, we come across a situation of high symmetry where Ampere's law can be conveniently applied.

4.8.1 The solenoid

We shall discuss a long solenoid. By long solenoid we mean that the solenoid's length is large compared to its radius. It consists of a long wire wound in the form of a helix where the neighbouring turns are closely spaced. So each turn can be regarded as a circular loop. The net magnetic field is the vector sum of the fields due to all the turns. Enamelled wires are used for winding so that turns are insulated from each other.

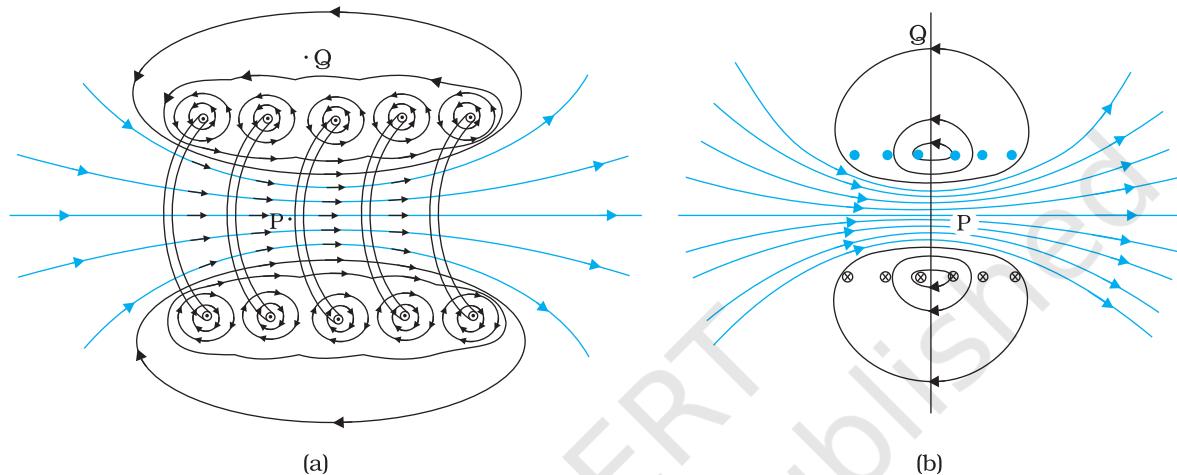


FIGURE 4.17 (a) The magnetic field due to a section of the solenoid which has been stretched out for clarity. Only the exterior semi-circular part is shown. Notice how the circular loops between neighbouring turns tend to cancel.
 (b) The magnetic field of a finite solenoid.

Figure 4.17 displays the magnetic field lines for a finite solenoid. We show a section of this solenoid in an enlarged manner in Fig. 4.17(a). Figure 4.17(b) shows the entire finite solenoid with its magnetic field. In Fig. 4.17(a), it is clear from the circular loops that the field between two neighbouring turns vanishes. In Fig. 4.17(b), we see that the field at the interior mid-point P is uniform, strong and along the axis of the solenoid. The field at the exterior mid-point Q is weak and moreover is along the axis of the solenoid with no perpendicular or normal component. As the solenoid is made longer it appears like a long cylindrical metal sheet. Figure 4.18 represents this idealised picture. The field outside the solenoid approaches zero. We shall assume that the field outside is zero. The field inside becomes everywhere parallel to the axis.

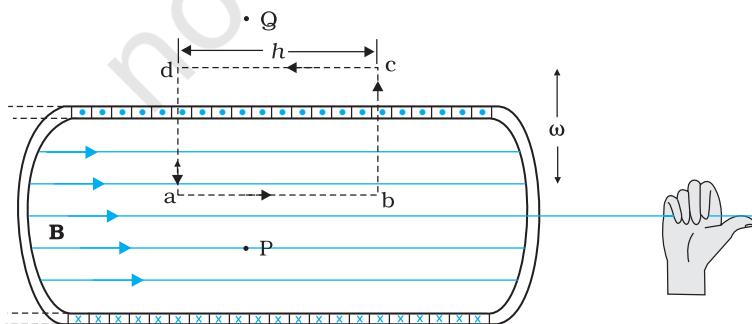


FIGURE 4.18 The magnetic field of a very long solenoid. We consider a rectangular Amperian loop abcd to determine the field.

Consider a rectangular Amperian loop abcd. Along cd the field is zero as argued above. Along transverse sections bc and ad, the field component is zero. Thus, these two sections make no contribution. Let the field along ab be B . Thus, the relevant length of the Amperian loop is, $L = h$.

Let n be the number of turns per unit length, then the total number of turns is nh . The enclosed current is, $I_e = I(nh)$, where I is the current in the solenoid. From Ampere's circuital law [Eq. 4.17 (b)]

$$\begin{aligned} BL &= \mu_0 I_e \quad B h = \mu_0 I (n h) \\ B &= \mu_0 n I \end{aligned} \quad (4.20)$$

The direction of the field is given by the right-hand rule. The solenoid is commonly used to obtain a uniform magnetic field. We shall see in the next chapter that a large field is possible by inserting a soft iron core inside the solenoid.

4.8.2 The toroid

The toroid is a hollow circular ring on which a large number of turns of a wire are closely wound. It can be viewed as a solenoid which has been bent into a circular shape to close on itself. It is shown in Fig. 4.19(a) carrying a current I . We shall see that the magnetic field in the open space inside (point P) and exterior to the toroid (point Q) is zero. The field \mathbf{B} inside the toroid is constant in magnitude for the ideal toroid of closely wound turns.

Figure 4.19(b) shows a sectional view of the toroid. The direction of the magnetic field inside is clockwise as per the right-hand thumb rule for circular loops. Three circular Amperian loops 1, 2 and 3 are shown by dashed lines. By symmetry, the magnetic field should be tangential to each of them and constant in magnitude for a given loop. The circular areas bounded by loops 2 and 3 both cut the toroid: so that each turn of current carrying wire is cut once by the loop 2 and twice by the loop 3.

Let the magnetic field along loop 1 be B_1 in magnitude. Then in Ampere's circuital law [Eq. 4.17(a)], $L = 2\pi r_1$. However, the loop encloses no current, so $I_e = 0$. Thus,

$$B_1 (2\pi r_1) = \mu_0 (0), \quad B_1 = 0$$

Thus, the magnetic field at any point P in the open space inside the toroid is zero.

We shall now show that magnetic field at Q is likewise zero. Let the magnetic field along loop 3 be B_3 . Once again from Ampere's law $L = 2\pi r_3$. However, from the sectional cut, we see that the current coming out of the plane of the paper is cancelled exactly by the current going into it. Thus, $I_e = 0$, and $B_3 = 0$. Let the magnetic field inside the solenoid

be B . We shall now consider the magnetic field at S. Once again we employ Ampere's law in the form of Eq. [4.17 (a)]. We find, $L = 2\pi r$.

The current enclosed I_e is (for N turns of toroidal coil) NI .

$$B (2\pi r) = \mu_0 NI$$

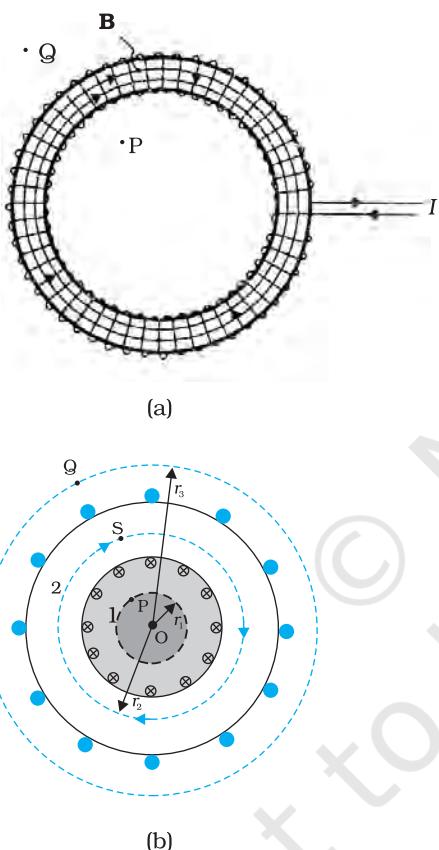


FIGURE 4.19 (a) A toroid carrying a current I . (b) A sectional view of the toroid. The magnetic field can be obtained at an arbitrary distance r from the centre O of the toroid by Ampere's circuital law. The dashed lines labelled 1, 2 and 3 are three circular Amperian loops.

Moving Charges and Magnetism

$$B = \frac{\mu_0 N I}{2\pi r} \quad (4.21)$$

We shall now compare the two results: for a toroid and solenoid. We re-express Eq. (4.21) to make the comparison easier with the solenoid result given in Eq. (4.20). Let r be the average radius of the toroid and n be the number of turns per unit length. Then

$$\begin{aligned} N &= 2\pi r n = (\text{average perimeter of the toroid}) \\ &\quad \times \text{number of turns per unit length} \end{aligned}$$

and thus,

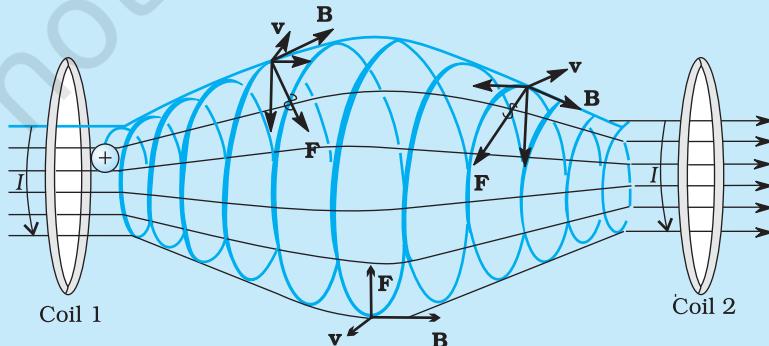
$$B = \mu_0 n I, \quad (4.22)$$

i.e., the result for the solenoid!

In an ideal toroid the coils are circular. In reality the turns of the toroidal coil form a helix and there is always a small magnetic field external to the toroid.

MAGNETIC CONFINEMENT

We have seen in Section 4.3 (see also the box on helical motion of charged particles earlier in this chapter) that orbits of charged particles are helical. If the magnetic field is non-uniform, but does not change much during one circular orbit, then the radius of the helix will decrease as it enters stronger magnetic field and the radius will increase when it enters weaker magnetic fields. We consider two solenoids at a distance from each other, enclosed in an evacuated container (see figure below where we have not shown the container). Charged particles moving in the region between the two solenoids will start with a small radius. The radius will increase as field decreases and the radius will decrease again as field due to the second solenoid takes over. The solenoids act as a mirror or reflector. [See the direction of \mathbf{F} as the particle approaches coil 2 in the figure. It has a horizontal component against the forward motion.] This makes the particles turn back when they approach the solenoid. Such an arrangement will act like *magnetic bottle* or magnetic container. The particles will never touch the sides of the container. Such magnetic bottles are of great use in confining the high energy plasma in fusion experiments. The plasma will destroy any other form of material container because of its high temperature. Another useful container is a toroid. Toroids are expected to play a key role in the *tokamak*, an equipment for plasma confinement in fusion power reactors. There is an international collaboration called the *International Thermonuclear Experimental Reactor* (ITER), being set up in France, for achieving controlled fusion, of which India is a collaborating nation. For details of ITER collaboration and the project, you may visit <http://www.iter.org>.



EXAMPLE 4.9

Example 4.9 A solenoid of length 0.5 m has a radius of 1 cm and is made up of 500 turns. It carries a current of 5 A. What is the magnitude of the magnetic field inside the solenoid?

Solution The number of turns per unit length is,

$$n = \frac{500}{0.5} = 1000 \text{ turns/m}$$

The length $l = 0.5 \text{ m}$ and radius $r = 0.01 \text{ m}$. Thus, $l/a = 50$ i.e., $l \gg a$. Hence, we can use the *long* solenoid formula, namely, Eq. (4.20)

$$\begin{aligned} B &= \mu_0 n I \\ &= 4\pi \times 10^{-7} \times 10^3 \times 5 \\ &= 6.28 \times 10^{-3} \text{ T} \end{aligned}$$

4.9 FORCE BETWEEN TWO PARALLEL CURRENTS, THE AMPERE

We have learnt that there exists a magnetic field due to a conductor carrying a current which obeys the Biot-Savart law. Further, we have

learnt that an external magnetic field will exert a force on a current-carrying conductor. This follows from the Lorentz force formula. Thus, it is logical to expect that two current-carrying conductors placed near each other will exert (magnetic) forces on each other. In the period 1820-25, Ampere studied the nature of this magnetic force and its dependence on the magnitude of the current, on the shape and size of the conductors as well as the distances between the conductors. In this section, we shall take the simple example of two parallel current-carrying conductors, which will perhaps help us to appreciate Ampere's painstaking work.

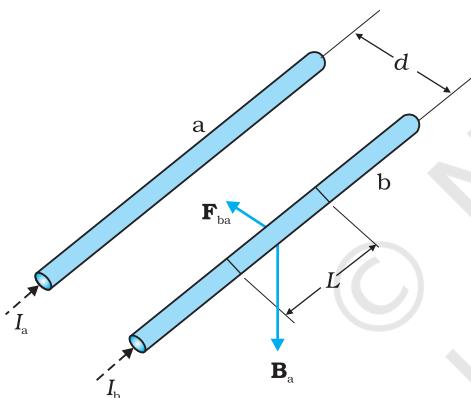


FIGURE 4.20 Two long straight parallel conductors carrying steady currents I_a and I_b and separated by a distance d . \mathbf{B}_a is the magnetic field set up by conductor 'a' at conductor 'b'.

are placed horizontally). Its magnitude is given by Eq. [4.19(a)] or from Ampere's circuital law,

$$B_a = \frac{\mu_0 I_a}{2\pi d}$$

The conductor 'b' carrying a current I_b will experience a sideways force due to the field \mathbf{B}_a . The direction of this force is towards the conductor 'a' (Verify this). We label this force as \mathbf{F}_{ba} , the force on a segment L of 'b' due to 'a'. The magnitude of this force is given by Eq. (4.4),

Moving Charges and Magnetism

$$\begin{aligned} F_{ba} &= I_b L B_a \\ &= \frac{\mu_0 I_a I_b}{2\pi d} L \end{aligned} \quad (4.23)$$

It is of course possible to compute the force on 'a' due to 'b'. From considerations similar to above we can find the force \mathbf{F}_{ab} , on a segment of length L of 'a' due to the current in 'b'. It is equal in magnitude to F_{ba} , and directed towards 'b'. Thus,

$$\mathbf{F}_{ba} = -\mathbf{F}_{ab} \quad (4.24)$$

Note that this is consistent with Newton's third Law. Thus, at least for parallel conductors and steady currents, we have shown that the Biot-Savart law and the Lorentz force yield results in accordance with Newton's third Law*.

We have seen from above that currents flowing in the same direction attract each other. One can show that oppositely directed currents repel each other. Thus,

Parallel currents attract, and antiparallel currents repel.

This rule is the opposite of what we find in electrostatics. Like (same sign) charges repel each other, but like (parallel) currents attract each other.

Let f_{ba} represent the magnitude of the force \mathbf{F}_{ba} per unit length. Then, from Eq. (4.23),

$$f_{ba} = \frac{\mu_0 I_a I_b}{2\pi d} \quad (4.25)$$

The above expression is used to define the ampere (A), which is one of the seven SI base units.

The *ampere* is the value of that steady current which, when maintained in each of the two very long, straight, parallel conductors of negligible cross-section, and placed one metre apart in vacuum, would produce on each of these conductors a force equal to 2×10^{-7} newtons per metre of length.

This definition of the ampere was adopted in 1946. It is a theoretical definition. In practice one must eliminate the effect of the earth's magnetic field and substitute very long wires by multturn coils of appropriate geometries. An instrument called the current balance is used to measure this mechanical force.

The SI unit of charge, namely, the coulomb, can now be defined in terms of the ampere.

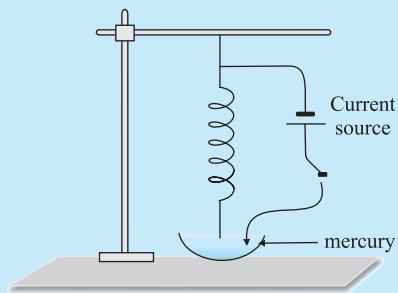
When a steady current of 1A is set up in a conductor, the quantity of charge that flows through its cross-section in 1s is one coulomb (1C).

* It turns out that when we have time-dependent currents and/or charges in motion, Newton's third law may not hold for forces between charges and/or conductors. An essential consequence of the Newton's third law in mechanics is conservation of momentum of an isolated system. This, however, holds even for the case of time-dependent situations with electromagnetic fields, provided the momentum carried by fields is also taken into account.

ROGET'S SPIRAL FOR ATTRACTION BETWEEN PARALLEL CURRENTS

Magnetic effects are generally smaller than electric effects. As a consequence, the force between currents is rather small, because of the smallness of the factor μ . Hence it is difficult to demonstrate attraction or repulsion between currents. Thus for 5 A current in each wire at a separation of 1 cm, the force per metre would be 5×10^{-4} N, which is about 50 mg weight. It would be like pulling a wire by a string going over a pulley to which a 50 mg weight is attached. The displacement of the wire would be quite unnoticeable.

With the use of a soft spring, we can increase the effective length of the parallel current and by using mercury, we can make the displacement of even a few mm observable very dramatically. You will also need a constant-current supply giving a constant current of about 5 A.



Take a soft spring whose natural period of oscillations is about 0.5 – 1 s. Hang it vertically and attach a pointed tip to its lower end, as shown in the figure here. Take some mercury in a dish and adjust the spring such that the tip is just above the mercury surface. Take the DC current source, connect one of its terminals to the upper end of the spring, and dip the other terminal in mercury. If the tip of the spring touches mercury, the circuit is completed through mercury.

Let the DC source be put off to begin with. Let the tip be adjusted so that it just touches the mercury surface. Switch on the constant current supply, and watch the fascinating outcome. The spring shrinks with a jerk, the tip comes out of mercury (just by a mm or so), the circuit is broken, the current stops, the spring relaxes and tries to come back to its original position, the tip again touches mercury establishing a current in the circuit, and the cycle continues with tick, tick, tick, In the beginning, you may require some small adjustments to get a good effect.

Keep your face away from mercury vapours as they are poisonous. Do not inhale mercury vapours for long.

EXAMPLE 4.10

Example 4.10 The horizontal component of the earth's magnetic field at a certain place is 3.0×10^{-5} T and the direction of the field is from the geographic south to the geographic north. A very long straight conductor is carrying a steady current of 1 A. What is the force per unit length on it when it is placed on a horizontal table and the direction of the current is (a) east to west; (b) south to north?

Solution $\mathbf{F} = I\mathbf{l} \times \mathbf{B}$

$$F = ILB \sin\theta$$

The force per unit length is

$$f = F/l = IB \sin\theta$$

(a) When the current is flowing from east to west,

$$\theta = 90^\circ$$

Hence,

$$f = IB \\ = 1 \times 3 \times 10^{-5} = 3 \times 10^{-5} \text{ N m}^{-1}$$

EXAMPLE 4.10

This is larger than the value $2 \times 10^{-7} \text{ Nm}^{-1}$ quoted in the definition of the ampere. Hence it is important to eliminate the effect of the earth's magnetic field and other stray fields while standardising the ampere.

The direction of the force is downwards. This direction may be obtained by the directional property of cross product of vectors.

- (b) When the current is flowing from south to north,

$$\theta = 0^\circ$$

$$f = 0$$

Hence there is no force on the conductor.

4.10 TORQUE ON CURRENT LOOP, MAGNETIC DIPOLE

4.10.1 Torque on a rectangular current loop in a uniform magnetic field

We now show that a rectangular loop carrying a steady current I and placed in a uniform magnetic field experiences a torque. It does not experience a net force. This behaviour is analogous to that of electric dipole in a uniform electric field (Section 1.10).

We first consider the simple case when the rectangular loop is placed such that the uniform magnetic field \mathbf{B} is in the plane of the loop. This is illustrated in Fig. 4.21(a).

The field exerts no force on the two arms AD and BC of the loop. It is perpendicular to the arm AB of the loop and exerts a force \mathbf{F}_1 on it which is directed into the plane of the loop. Its magnitude is,

$$F_1 = I b B$$

Similarly it exerts a force \mathbf{F}_2 on the arm CD and \mathbf{F}_2 is directed out of the plane of the paper.

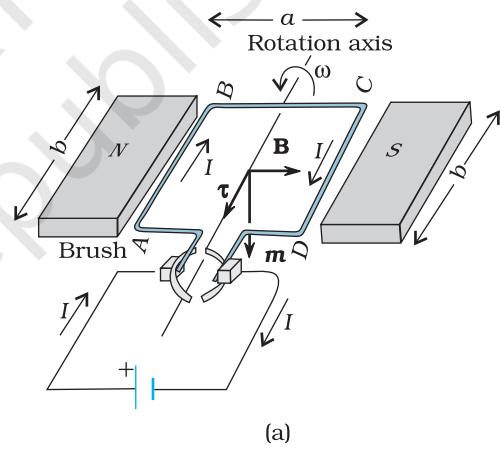
$$F_2 = I b B = F_1$$

Thus, the *net force* on the loop is zero. There is a torque on the loop due to the pair of forces \mathbf{F}_1 and \mathbf{F}_2 . Figure 4.21(b) shows a view of the loop from the AD end. It shows that the torque on the loop tends to rotate it anti-clockwise. This torque is (in magnitude),

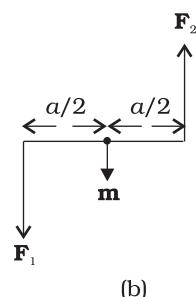
$$\begin{aligned} \tau &= F_1 \frac{a}{2} + F_2 \frac{a}{2} \\ &= IbB \frac{a}{2} + IbB \frac{a}{2} = I(ab)B \\ &= IA B \end{aligned} \quad (4.26)$$

where $A = ab$ is the area of the rectangle.

We next consider the case when the plane of the loop, is not along the magnetic field, but makes an angle with it. We take the angle between the field and the normal to



(a)



(b)

FIGURE 4.21 (a) A rectangular current-carrying coil in uniform magnetic field. The magnetic moment \mathbf{m} points downwards. The torque τ is along the axis and tends to rotate the coil anticlockwise. (b) The couple acting on the coil.

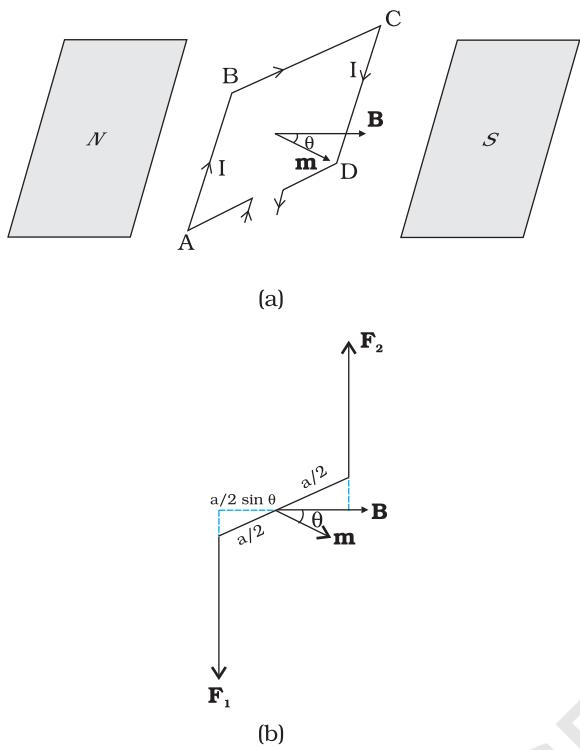


FIGURE 4.22 (a) The area vector of the loop ABCD makes an arbitrary angle θ with the magnetic field. (b) Top view of the loop. The forces \mathbf{F}_1 and \mathbf{F}_2 acting on the arms AB and CD are indicated.

can be expressed as vector product of the magnetic moment of the coil and the magnetic field. We define the *magnetic moment* of the current loop as,

$$\mathbf{m} = IA \quad (4.28)$$

where the direction of the area vector \mathbf{A} is given by the right-hand thumb rule and is directed into the plane of the paper in Fig. 4.21. Then as the angle between \mathbf{m} and \mathbf{B} is θ , Eqs. (4.26) and (4.27) can be expressed by one expression

$$\tau = \mathbf{m} \times \mathbf{B} \quad (4.29)$$

This is analogous to the electrostatic case (Electric dipole of dipole moment \mathbf{p}_e in an electric field \mathbf{E}).

$$\tau = \mathbf{p}_e \times \mathbf{E}$$

As is clear from Eq. (4.28), the dimensions of the magnetic moment are $[A][L^2]$ and its unit is Am^2 .

From Eq. (4.29), we see that the torque τ vanishes when \mathbf{m} is either parallel or antiparallel to the magnetic field \mathbf{B} . This indicates a state of equilibrium as there is no torque on the coil (this also applies to any object with a magnetic moment \mathbf{m}). When \mathbf{m} and \mathbf{B} are parallel the

the coil to be angle θ (The previous case corresponds to $\theta = \pi/2$). Figure 4.22 illustrates this general case.

The forces on the arms BC and DA are equal, opposite, and act along the axis of the coil, which connects the centres of mass of BC and DA. Being collinear along the axis they cancel each other, resulting in no net force or torque. The forces on arms AB and CD are \mathbf{F}_1 and \mathbf{F}_2 . They too are equal and opposite, with magnitude,

$$F_1 = F_2 = IbB$$

But they are not collinear! This results in a couple as before. The torque is, however, less than the earlier case when plane of loop was along the magnetic field. This is because the perpendicular distance between the forces of the couple has decreased. Figure 4.22(b) is a view of the arrangement from the AD end and it illustrates these two forces constituting a couple. The magnitude of the torque on the loop is,

$$\begin{aligned} \tau &= F_1 \frac{a}{2} \sin \theta + F_2 \frac{a}{2} \sin \theta \\ &= I ab B \sin \theta \\ &= I A B \sin \theta \end{aligned} \quad (4.27)$$

As $B \rightarrow 0$, the perpendicular distance between the forces of the couple also approaches zero. This makes the forces collinear and the net force and torque zero. The torques in Eqs. (4.26) and (4.27)

equilibrium is a stable one. Any small rotation of the coil produces a torque which brings it back to its original position. When they are antiparallel, the equilibrium is unstable as any rotation produces a torque which increases with the amount of rotation. The presence of this torque is also the reason why a small magnet or any magnetic dipole aligns itself with the external magnetic field.

If the loop has N closely wound turns, the expression for torque, Eq. (4.29), still holds, with

$$\mathbf{m} = NIA \quad (4.30)$$

Example 4.11 A 100 turn closely wound circular coil of radius 10 cm carries a current of 3.2 A. (a) What is the field at the centre of the coil? (b) What is the magnetic moment of this coil?

The coil is placed in a vertical plane and is free to rotate about a horizontal axis which coincides with its diameter. A uniform magnetic field of 2 T in the horizontal direction exists such that initially the axis of the coil is in the direction of the field. The coil rotates through an angle of 90° under the influence of the magnetic field. (c) What are the magnitudes of the torques on the coil in the initial and final position? (d) What is the angular speed acquired by the coil when it has rotated by 90° ? The moment of inertia of the coil is 0.1 kg m^2 .

Solution

(a) From Eq. (4.16)

$$B = \frac{\mu_0 NI}{2R}$$

Here, $N = 100$; $I = 3.2 \text{ A}$, and $R = 0.1 \text{ m}$. Hence,

$$B = \frac{4\pi \times 10^{-7} \times 10^2 \times 3.2}{2 \times 10^{-1}} = \frac{4 \times 10^{-5} \times 10}{2 \times 10^{-1}} \quad (\text{using } \pi \times 3.2 = 10) \\ = 2 \times 10^{-3} \text{ T}$$

The direction is given by the right-hand thumb rule.

(b) The magnetic moment is given by Eq. (4.30),

$$m = NIA = NI\pi r^2 = 100 \times 3.2 \times 3.14 \times 10^{-2} = 10 \text{ A m}^2$$

The direction is once again given by the right hand thumb rule.

(c) $\tau = |\mathbf{m} \times \mathbf{B}| \quad [\text{from Eq. (4.29)}]$

$$= mB \sin \theta$$

Initially, $\theta = 0$. Thus, initial torque $\tau_i = 0$. Finally, $\theta = \pi/2$ (or 90°).

Thus, final torque $\tau_f = mB = 10 \times 2 = 20 \text{ N m}$.

(d) From Newton's second law,

$$\mathcal{I} \frac{d\omega}{dt} = mB \sin \theta$$

where \mathcal{I} is the moment of inertia of the coil. From chain rule,

$$\frac{d\omega}{dt} = \frac{d\omega}{d\theta} \frac{d\theta}{dt} = \frac{d\omega}{d\theta} \omega$$

Using this,

$$\mathcal{I} \omega d\omega = mB \sin \theta d\theta$$

EXAMPLE 4.11

Integrating from $\theta = 0$ to $\theta = \pi/2$,

$$\oint_0^{\omega_f} \omega \, d\omega = mB \int_0^{\pi/2} \sin\theta \, d\theta$$

$$\oint \frac{\omega^2}{2} = -mB \cos\theta \Big|_0^{\pi/2} = mB$$

$$\omega_f = \left(\frac{2mB}{\oint} \right)^{1/2} = \left(\frac{2 \times 20}{10^{-1}} \right)^{1/2} = 20 \text{ s}^{-1}$$

EXAMPLE 4.12
Example 4.12

- (a) A current-carrying circular loop lies on a smooth horizontal plane. Can a uniform magnetic field be set up in such a manner that the loop turns around itself (i.e., turns about the vertical axis).
- (b) A current-carrying circular loop is located in a uniform external magnetic field. If the loop is free to turn, what is its orientation of stable equilibrium? Show that in this orientation, the flux of the total field (external field + field produced by the loop) is maximum.
- (c) A loop of irregular shape carrying current is located in an external magnetic field. If the wire is flexible, why does it change to a circular shape?

Solution

- (a) No, because that would require τ to be in the vertical direction. But $\tau = I \mathbf{A} \times \mathbf{B}$, and since \mathbf{A} of the horizontal loop is in the vertical direction, τ would be in the plane of the loop for any \mathbf{B} .
- (b) Orientation of stable equilibrium is one where the area vector \mathbf{A} of the loop is in the direction of external magnetic field. In this orientation, the magnetic field produced by the loop is in the same direction as external field, both normal to the plane of the loop, thus giving rise to maximum flux of the total field.
- (c) It assumes circular shape with its plane normal to the field to maximize flux, since for a given perimeter, a circle encloses greater area than any other shape.

4.10.2 Circular current loop as a magnetic dipole

In this section, we shall consider the elementary magnetic element: the current loop. We shall show that the magnetic field (at large distances) due to current in a circular current loop is very similar in behavior to the electric field of an electric dipole. In Section 4.6, we have evaluated the magnetic field on the axis of a circular loop, of a radius R , carrying a steady current I . The magnitude of this field is [(Eq. (4.15)],

$$B = \frac{\mu_0 I R^2}{2(x^2 + R^2)^{3/2}}$$

and its direction is along the axis and given by the right-hand thumb rule (Fig. 4.12). Here, x is the distance along the axis from the centre of the loop. For $x \gg R$, we may drop the R^2 term in the denominator. Thus,

Moving Charges and Magnetism

$$B = \frac{\mu_0 R^2}{2x^3}$$

Note that the area of the loop $A = \pi R^2$. Thus,

$$B = \frac{\mu_0 IA}{2\pi x^3}$$

As earlier, we define the magnetic moment \mathbf{m} to have a magnitude IA , $\mathbf{m} = I\mathbf{A}$. Hence,

$$\begin{aligned}\mathbf{B} &\approx \frac{\mu_0 \mathbf{m}}{2\pi x^3} \\ &= \frac{\mu_0}{4\pi} \frac{2\mathbf{m}}{x^3}\end{aligned}\quad [4.31(a)]$$

The expression of Eq. [4.31(a)] is very similar to an expression obtained earlier for the electric field of a dipole. The similarity may be seen if we substitute,

$$\mu_0 \rightarrow 1/\epsilon_0$$

$$\mathbf{m} \rightarrow \mathbf{p}_e \text{ (electrostatic dipole)}$$

$$\mathbf{B} \rightarrow \mathbf{E} \text{ (electrostatic field)}$$

We then obtain,

$$\mathbf{E} = \frac{2\mathbf{p}_e}{4\pi \epsilon_0 x^3}$$

which is precisely the field for an electric dipole at a point on its axis, considered in Chapter 1, Section 1.10 [Eq. (1.20)].

It can be shown that the above analogy can be carried further. We had found in Chapter 1 that the electric field on the perpendicular bisector of the dipole is given by [See Eq.(1.21)],

$$\mathbf{E} \approx \frac{\mathbf{p}_e}{4\pi \epsilon_0 x^3}$$

where x is the distance from the dipole. If we replace $\mathbf{p} \rightarrow \mathbf{m}$ and $\mu_0 \rightarrow 1/\epsilon_0$ in the above expression, we obtain the result for \mathbf{B} for a point *in the plane of the loop* at a distance x from the centre. For $x \gg R$,

$$\mathbf{B} \approx \frac{\mu_0 \mathbf{m}}{4\pi x^3}; \quad x \gg R \quad [4.31(b)]$$

The results given by Eqs. [4.31(a)] and [4.31(b)] become exact for a *point* magnetic dipole.

The results obtained above can be shown to apply to any planar loop: a planar current loop is equivalent to a magnetic dipole of dipole moment $\mathbf{m} = I\mathbf{A}$, which is the analogue of electric dipole moment \mathbf{p} . Note, however, a fundamental difference: an electric dipole is built up of two elementary units — the charges (or electric monopoles). In magnetism, a magnetic dipole (or a current loop) is the most elementary element. The equivalent of electric charges, i.e., magnetic monopoles, are not known to exist.

We have shown that a current loop (i) produces a magnetic field (see Fig. 4.12) and behaves like a magnetic dipole at large distances, and

(ii) is subject to torque like a magnetic needle. This led Ampere to suggest that all magnetism is due to circulating currents. This seems to be partly true and no magnetic monopoles have been seen so far. However, elementary particles such as an electron or a proton also carry an *intrinsic* magnetic moment, not accounted by circulating currents.

4.10.3 The magnetic dipole moment of a revolving electron

In Chapter 12 we shall read about the Bohr model of the hydrogen atom. You may perhaps have heard of this model which was proposed by the

Danish physicist Niels Bohr in 1911 and was a stepping stone to a new kind of mechanics, namely, quantum mechanics. In the Bohr model, the electron (a negatively charged particle) revolves around a positively charged nucleus much as a planet revolves around the sun. The force in the former case is electrostatic (Coulomb force) while it is gravitational for the planet-Sun case. We show this Bohr picture of the electron in Fig. 4.23.

The electron of charge ($-e$) ($e = +1.6 \times 10^{-19} \text{ C}$) performs uniform circular motion around a stationary heavy nucleus of charge $+Ze$. This constitutes a current I , where,

$$I = \frac{e}{T} \quad (4.32)$$

and T is the time period of revolution. Let r be the orbital radius of the electron, and v the orbital speed. Then,

$$T = \frac{2\pi r}{v} \quad (4.33)$$

Substituting in Eq. (4.32), we have $I = ev/2\pi r$.

There will be a magnetic moment, usually denoted by μ_l , associated with this circulating current. From Eq. (4.28) its magnitude is, $\mu_l = I\pi r^2 = evr/2$.

The direction of this magnetic moment is into the plane of the paper in Fig. 4.23. [This follows from the right-hand rule discussed earlier and the fact that the negatively charged electron is moving anti-clockwise, leading to a clockwise current.] Multiplying and dividing the right-hand side of the above expression by the electron mass m_e , we have,

$$\begin{aligned} \mu_l &= \frac{e}{2m_e} (m_e v r) \\ &= \frac{e}{2m_e} l \end{aligned} \quad [4.34(a)]$$

Here, l is the magnitude of the angular momentum of the electron about the central nucleus ("orbital" angular momentum). Vectorially,

$$\boldsymbol{\mu} = -\frac{e}{2m_e} \mathbf{l} \quad [4.34(b)]$$

The negative sign indicates that the angular momentum of the electron is opposite in direction to the magnetic moment. Instead of electron with

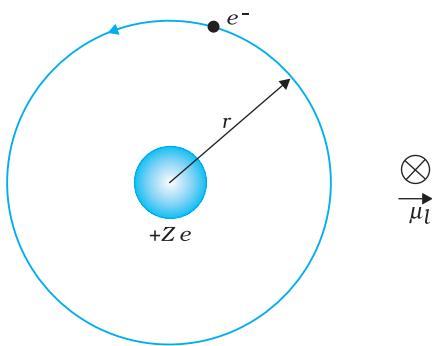


FIGURE 4.23 In the Bohr model of hydrogen-like atoms, the negatively charged electron is revolving with uniform speed around a centrally placed positively charged ($+Ze$) nucleus. The uniform circular motion of the electron constitutes a current. The direction of the magnetic moment is into the plane of the paper and is indicated separately by \otimes .

electron with

Moving Charges and Magnetism

charge ($-e$), if we had taken a particle with charge ($+q$), the angular momentum and magnetic moment would be in the same direction. The ratio

$$\frac{\mu_l}{l} = \frac{e}{2m_e} \quad (4.35)$$

is called the *gyromagnetic ratio* and is a constant. Its value is $8.8 \times 10^{10} \text{ C/kg}$ for an electron, which has been verified by experiments.

The fact that even at an atomic level there is a magnetic moment, confirms Ampere's bold hypothesis of atomic magnetic moments. This according to Ampere, would help one to explain the magnetic properties of materials. Can one assign a value to this atomic dipole moment? The answer is Yes. One can do so within the Bohr model. Bohr hypothesised that the angular momentum assumes a discrete set of values, namely,

$$l = \frac{n\hbar}{2\pi} \quad (4.36)$$

where n is a natural number, $n = 1, 2, 3, \dots$ and \hbar is a constant named after Max Planck (Planck's constant) with a value $\hbar = 6.626 \times 10^{-34} \text{ J s}$. This condition of discreteness is called the *Bohr quantisation condition*. We shall discuss it in detail in Chapter 12. Our aim here is merely to use it to calculate the elementary dipole moment. Take the value $n = 1$, we have from Eq. (4.34) that,

$$\begin{aligned} (\mu_l)_{\min} &= \frac{e}{4\pi m_e} \hbar \\ &= \frac{1.60 \times 10^{-19} \times 6.63 \times 10^{-34}}{4 \times 3.14 \times 9.11 \times 10^{-31}} \\ &= 9.27 \times 10^{-24} \text{ Am}^2 \end{aligned} \quad (4.37)$$

where the subscript 'min' stands for minimum. This value is called the *Bohr magneton*.

Any charge in uniform circular motion would have an associated magnetic moment given by an expression similar to Eq. (4.34). This dipole moment is labelled as the *orbital magnetic moment*. Hence the subscript 'l' in μ_l . Besides the orbital moment, the electron has an *intrinsic magnetic moment*, which has the same numerical value as given in Eq. (4.37). It is called the *spin magnetic moment*. But we hasten to add that it is not as though the electron is spinning. The electron is an elementary particle and it does not have an axis to spin around like a top or our earth. Nevertheless it does possess this *intrinsic magnetic moment*. The microscopic roots of magnetism in iron and other materials can be traced back to this intrinsic spin magnetic moment.

4.11 THE MOVING COIL GALVANOMETER

Currents and voltages in circuits have been discussed extensively in Chapters 3. But how do we measure them? How do we claim that current in a circuit is 1.5 A or the voltage drop across a resistor is 1.2 V? Figure 4.24 exhibits a very useful instrument for this purpose: the *moving*



Conversion of galvanometer into ammeter and voltmeter:
www.citycollegiate.com/galvanometer_XIIa.htm

Physics

coil galvanometer (MCG). It is a device whose principle can be understood on the basis of our discussion in Section 4.10.

The galvanometer consists of a coil, with many turns, free to rotate about a fixed axis (Fig. 4.24), in a uniform radial magnetic field. There is a cylindrical soft iron core which not only makes the field radial but also increases the strength of the magnetic field. When a current flows through the coil, a torque acts on it. This torque is given by Eq. (4.26) to be

$$\tau = NIAB$$

where the symbols have their usual meaning. Since the field is radial by design, we have taken $\sin \theta = 1$ in the above expression for the torque. The magnetic torque $NIAB$ tends to rotate the coil. A spring S_p provides a counter torque $k\phi$ that balances the magnetic torque $NIAB$; resulting in a steady angular deflection ϕ . In equilibrium

$$k\phi = NIAB$$

where k is the torsional constant of the spring; i.e. the restoring torque per unit twist. The deflection ϕ is indicated on the scale by a pointer attached to the spring. We have

$$\phi = \left(\frac{NAB}{k} \right) I \quad (4.38)$$

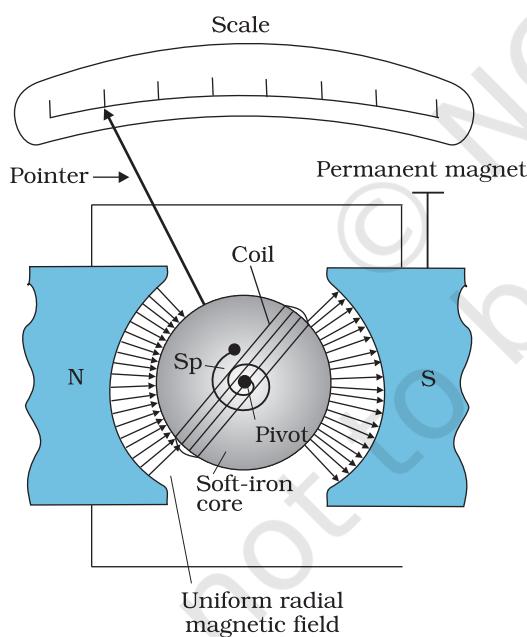


FIGURE 4.24 The moving coil galvanometer. Its elements are described in the text. Depending on the requirement, this device can be used as a current detector or for measuring the value of the current (ammeter) or voltage (voltmeter).

The quantity in brackets is a constant for a given galvanometer.

The galvanometer can be used in a number of ways. It can be used as a detector to check if a current is flowing in the circuit. We have come across this usage in the Wheatstone's bridge arrangement. In this usage the neutral position of the pointer (when no current is flowing through the galvanometer) is in the middle of the scale and not at the left end as shown in Fig. 4.24. Depending on the direction of the current, the pointer deflection is either to the right or the left.

The galvanometer cannot as such be used as an ammeter to measure the value of the current in a given circuit. This is for two reasons: (i) Galvanometer is a very sensitive device, it gives a full-scale deflection for a current of the order of μA . (ii) For measuring currents, the galvanometer has to be connected in series, and as it has a large resistance, this will change the value of the current in the circuit. To overcome these difficulties, one attaches a small resistance r_s , called *shunt resistance*, in parallel with the galvanometer coil; so that most of the current passes through the shunt. The resistance of this arrangement is,

$$R_G r_s / (R_G + r_s) \approx r_s \quad \text{if } R_G \gg r_s$$

If r_s has small value, in relation to the resistance of the rest of the circuit R_c , the effect of introducing the measuring instrument is also small and negligible. This

Moving Charges and Magnetism

arrangement is schematically shown in Fig. 4.25. The scale of this ammeter is calibrated and then graduated to read off the current value with ease. We define the *current sensitivity of the galvanometer as the deflection per unit current*. From Eq. (4.38) this current sensitivity is,

$$\frac{\phi}{I} = \frac{NAB}{k} \quad (4.39)$$

A convenient way for the manufacturer to increase the sensitivity is to increase the number of turns N . We choose galvanometers having sensitivities of value, required by our experiment.

The galvanometer can also be used as a voltmeter to measure the voltage across a given section of the circuit. For this it must be connected *in parallel* with that section of the circuit. Further, it must draw a very small current, otherwise the voltage measurement will disturb the original set up by an amount which is very large. Usually we like to keep the disturbance due to the measuring device below one per cent. To ensure this, a large resistance R is connected *in series* with the galvanometer. This arrangement is schematically depicted in Fig. 4.26. Note that the resistance of the voltmeter is now,

$$R_G + R \approx R : \text{large}$$

The scale of the voltmeter is calibrated to read off the voltage value with ease. We define the *voltage sensitivity as the deflection per unit voltage*. From Eq. (4.38),

$$\frac{\phi}{V} = \left(\frac{NAB}{k} \right) \frac{I}{V} = \left(\frac{NAB}{k} \right) \frac{1}{R} \quad (4.40)$$

An interesting point to note is that increasing the current sensitivity may not necessarily increase the voltage sensitivity. Let us take Eq. (4.39) which provides a measure of current sensitivity. If $N \rightarrow 2N$, i.e., we double the number of turns, then

$$\frac{\phi}{I} \rightarrow 2 \frac{\phi}{I}$$

Thus, the current sensitivity doubles. However, the resistance of the galvanometer is also likely to double, since it is proportional to the length of the wire. In Eq. (4.40), $N \rightarrow 2N$, and $R \rightarrow 2R$, thus the voltage sensitivity,

$$\frac{\phi}{V} \rightarrow \frac{\phi}{V}$$

remains unchanged. So in general, the modification needed for conversion of a galvanometer to an ammeter will be different from what is needed for converting it into a voltmeter.

Example 4.13 In the circuit (Fig. 4.27) the current is to be measured. What is the value of the current if the ammeter shown (a) is a galvanometer with a resistance $R_G = 60.00 \Omega$; (b) is a galvanometer described in (a) but converted to an ammeter by a shunt resistance $r_s = 0.02 \Omega$; (c) is an ideal ammeter with zero resistance?

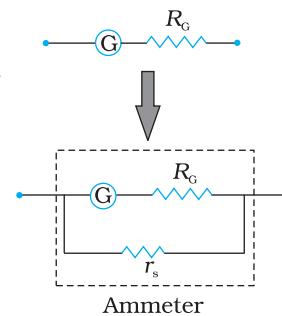


FIGURE 4.25
Conversion of a galvanometer (G) to an ammeter by the introduction of a shunt resistance r_s of very small value in parallel.

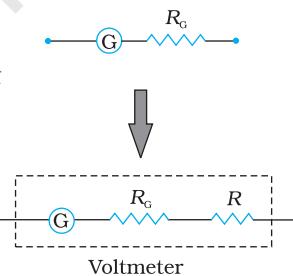


FIGURE 4.26
Conversion of a galvanometer (G) to a voltmeter by the introduction of a resistance R of large value in series.

EXAMPLE 4.13

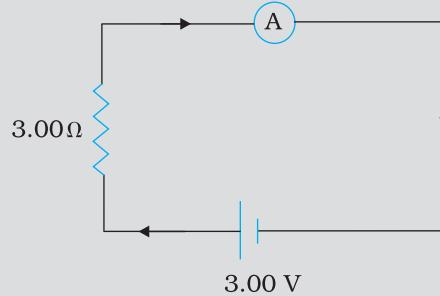


FIGURE 4.27

Solution

(a) Total resistance in the circuit is,

$$R_G + 3 = 63 \Omega. \text{ Hence, } I = 3/63 = 0.048 \text{ A.}$$

(b) Resistance of the galvanometer converted to an ammeter is,

$$\frac{R_G r_s}{R_G + r_s} = \frac{60 \Omega \times 0.02 \Omega}{(60 + 0.02) \Omega} \approx 0.02 \Omega$$

Total resistance in the circuit is,

$$0.02 \Omega + 3 \Omega = 3.02 \Omega. \text{ Hence, } I = 3/3.02 = 0.99 \text{ A.}$$

(c) For the ideal ammeter with zero resistance,

$$I = 3/3 = 1.00 \text{ A}$$

SUMMARY

- The total force on a charge q moving with velocity \mathbf{v} in the presence of magnetic and electric fields \mathbf{B} and \mathbf{E} , respectively is called the *Lorentz force*. It is given by the expression:

$$\mathbf{F} = q(\mathbf{v} \times \mathbf{B} + \mathbf{E})$$

The magnetic force $q(\mathbf{v} \times \mathbf{B})$ is normal to \mathbf{v} and work done by it is zero.
- A straight conductor of length l and carrying a steady current I experiences a force \mathbf{F} in a uniform external magnetic field \mathbf{B} ,

$$\mathbf{F} = I\mathbf{l} \times \mathbf{B}$$

where $|\mathbf{l}| = l$ and the direction of \mathbf{l} is given by the direction of the current.
- In a uniform magnetic field \mathbf{B} , a charge q executes a circular orbit in a plane normal to \mathbf{B} . Its frequency of uniform circular motion is called the *cyclotron frequency* and is given by:

$$v_c = \frac{qB}{2\pi m}$$

This frequency is independent of the particle's speed and radius. This fact is exploited in a machine, the cyclotron, which is used to accelerate charged particles.

- The *Biot-Savart* law asserts that the magnetic field $d\mathbf{B}$ due to an element $d\mathbf{l}$ carrying a steady current I at a point P at a distance r from the current element is:

$$d\mathbf{B} = \frac{\mu_0}{4\pi} I \frac{d\mathbf{l} \times \mathbf{r}}{r^3}$$

Moving Charges and Magnetism

To obtain the total field at P, we must integrate this vector expression over the entire length of the conductor.

5. The magnitude of the magnetic field due to a circular coil of radius R carrying a current I at an axial distance x from the centre is

$$B = \frac{\mu_0 I R^2}{2(x^2 + R^2)^{3/2}}$$

At the center this reduces to

$$B = \frac{\mu_0 I}{2R}$$

6. *Ampere's Circuital Law:* Let an open surface S be bounded by a loop

C. Then the Ampere's law states that $\oint_C \mathbf{B} \cdot d\mathbf{l} = \mu_0 I$ where I refers to

the current passing through S. The sign of I is determined from the right-hand rule. We have discussed a simplified form of this law. If \mathbf{B} is directed along the tangent to every point on the perimeter L of a closed curve and is constant in magnitude along perimeter then,

$$BL = \mu_0 I_e$$

where I_e is the net current enclosed by the closed circuit.

7. The magnitude of the magnetic field at a distance R from a long, straight wire carrying a current I is given by:

$$B = \frac{\mu_0 I}{2\pi R}$$

The field lines are circles concentric with the wire.

8. The magnitude of the field B inside a long solenoid carrying a current I is

$$B = \mu_0 n I$$

where n is the number of turns per unit length. For a toroid one obtains,

$$B = \frac{\mu_0 N I}{2\pi r}$$

where N is the total number of turns and r is the average radius.

9. Parallel currents attract and anti-parallel currents repel.

10. A planar loop carrying a current I , having N closely wound turns, and an area A possesses a magnetic moment \mathbf{m} where,

$$\mathbf{m} = N I \mathbf{A}$$

and the direction of \mathbf{m} is given by the right-hand thumb rule : curl the palm of your right hand along the loop with the fingers pointing in the direction of the current. The thumb sticking out gives the direction of \mathbf{m} (and \mathbf{A})

When this loop is placed in a uniform magnetic field \mathbf{B} , the force \mathbf{F} on it is: $F = 0$

And the torque on it is,

$$\tau = \mathbf{m} \times \mathbf{B}$$

In a moving coil galvanometer, this torque is balanced by a counter-torque due to a spring, yielding

$$k\phi = NIAB$$

Physics

where ϕ is the equilibrium deflection and k the torsion constant of the spring.

11. An electron moving around the central nucleus has a magnetic moment μ_l given by:

$$\mu_l = \frac{e}{2m} l$$

where l is the magnitude of the angular momentum of the circulating electron about the central nucleus. The smallest value of μ_l is called the Bohr magneton μ_B and it is $\mu_B = 9.27 \times 10^{-24}$ J/T

12. A moving coil galvanometer can be converted into a ammeter by introducing a shunt resistance r_s , of small value in parallel. It can be converted into a voltmeter by introducing a resistance of a large value in series.

Physical Quantity	Symbol	Nature	Dimensions	Units	Remarks
Permeability of free space	μ_0	Scalar	$[MLT^{-2}A^{-2}]$	$T m A^{-1}$	$4\pi \times 10^{-7} T m A^{-1}$
Magnetic Field	B	Vector	$[M T^{-2}A^{-1}]$	T (tesla)	
Magnetic Moment	m	Vector	$[L^2A]$	$A m^2$ or J/T	
Torsion Constant	k	Scalar	$[M L^2T^{-2}]$	$N m rad^{-1}$	Appears in MCG

POINTS TO PONDER

- Electrostatic field lines originate at a positive charge and terminate at a negative charge or fade at infinity. Magnetic field lines always form closed loops.
- The discussion in this Chapter holds only for steady currents which do not vary with time.
When currents vary with time Newton's third law is valid only if momentum carried by the electromagnetic field is taken into account.
- Recall the expression for the Lorentz force,

$$\mathbf{F} = q(\mathbf{v} \times \mathbf{B} + \mathbf{E})$$

This velocity dependent force has occupied the attention of some of the greatest scientific thinkers. If one switches to a frame with instantaneous velocity \mathbf{v} , the magnetic part of the force vanishes. The motion of the charged particle is then explained by arguing that there exists an appropriate electric field in the new frame. We shall not discuss the details of this mechanism. However, we stress that the resolution of this paradox implies that electricity and magnetism are linked phenomena (*electromagnetism*) and that the Lorentz force expression *does not imply* a universal preferred frame of reference in nature.

- Ampere's Circuital law is not independent of the Biot-Savart law. It can be derived from the Biot-Savart law. Its relationship to the Biot-Savart law is similar to the relationship between Gauss's law and Coulomb's law.

EXERCISES

- 4.1** A circular coil of wire consisting of 100 turns, each of radius 8.0 cm carries a current of 0.40 A. What is the magnitude of the magnetic field **B** at the centre of the coil?
- 4.2** A long straight wire carries a current of 35 A. What is the magnitude of the field **B** at a point 20 cm from the wire?
- 4.3** A long straight wire in the horizontal plane carries a current of 50 A in north to south direction. Give the magnitude and direction of **B** at a point 2.5 m east of the wire.
- 4.4** A horizontal overhead power line carries a current of 90 A in east to west direction. What is the magnitude and direction of the magnetic field due to the current 1.5 m below the line?
- 4.5** What is the magnitude of magnetic force per unit length on a wire carrying a current of 8 A and making an angle of 30° with the direction of a uniform magnetic field of 0.15 T?
- 4.6** A 3.0 cm wire carrying a current of 10 A is placed inside a solenoid perpendicular to its axis. The magnetic field inside the solenoid is given to be 0.27 T. What is the magnetic force on the wire?
- 4.7** Two long and parallel straight wires A and B carrying currents of 8.0 A and 5.0 A in the same direction are separated by a distance of 4.0 cm. Estimate the force on a 10 cm section of wire A.
- 4.8** A closely wound solenoid 80 cm long has 5 layers of windings of 400 turns each. The diameter of the solenoid is 1.8 cm. If the current carried is 8.0 A, estimate the magnitude of **B** inside the solenoid near its centre.
- 4.9** A square coil of side 10 cm consists of 20 turns and carries a current of 12 A. The coil is suspended vertically and the normal to the plane of the coil makes an angle of 30° with the direction of a uniform horizontal magnetic field of magnitude 0.80 T. What is the magnitude of torque experienced by the coil?
- 4.10** Two moving coil meters, M_1 and M_2 have the following particulars:
 $R_1 = 10 \Omega$, $N_1 = 30$,
 $A_1 = 3.6 \times 10^{-3} \text{ m}^2$, $B_1 = 0.25 \text{ T}$
 $R_2 = 14 \Omega$, $N_2 = 42$,
 $A_2 = 1.8 \times 10^{-3} \text{ m}^2$, $B_2 = 0.50 \text{ T}$
(The spring constants are identical for the two meters). Determine the ratio of (a) current sensitivity and (b) voltage sensitivity of M_2 and M_1 .
- 4.11** In a chamber, a uniform magnetic field of 6.5 G ($1 \text{ G} = 10^{-4} \text{ T}$) is maintained. An electron is shot into the field with a speed of $4.8 \times 10^6 \text{ m s}^{-1}$ normal to the field. Explain why the path of the electron is a circle. Determine the radius of the circular orbit. ($e = 1.5 \times 10^{-19} \text{ C}$, $m_e = 9.1 \times 10^{-31} \text{ kg}$)
- 4.12** In Exercise 4.11 obtain the frequency of revolution of the electron in its circular orbit. Does the answer depend on the speed of the electron? Explain.
- 4.13** (a) A circular coil of 30 turns and radius 8.0 cm carrying a current of 6.0 A is suspended vertically in a uniform horizontal magnetic field of magnitude 1.0 T. The field lines make an angle of 60°

with the normal of the coil. Calculate the magnitude of the counter torque that must be applied to prevent the coil from turning.

- (b) Would your answer change, if the circular coil in (a) were replaced by a planar coil of some irregular shape that encloses the same area? (All other particulars are also unaltered.)

ADDITIONAL EXERCISES

- 4.14** Two concentric circular coils X and Y of radii 16 cm and 10 cm, respectively, lie in the same vertical plane containing the north to south direction. Coil X has 20 turns and carries a current of 16 A; coil Y has 25 turns and carries a current of 18 A. The sense of the current in X is anticlockwise, and clockwise in Y, for an observer looking at the coils facing west. Give the magnitude and direction of the net magnetic field due to the coils at their centre.

- 4.15** A magnetic field of 100 G ($1 \text{ G} = 10^{-4} \text{ T}$) is required which is uniform in a region of linear dimension about 10 cm and area of cross-section about 10^{-3} m^2 . The maximum current-carrying capacity of a given coil of wire is 15 A and the number of turns per unit length that can be wound round a core is at most $1000 \text{ turns m}^{-1}$. Suggest some appropriate design particulars of a solenoid for the required purpose. Assume the core is not ferromagnetic.

- 4.16** For a circular coil of radius R and N turns carrying current I , the magnitude of the magnetic field at a point on its axis at a distance x from its centre is given by,

$$B = \frac{\mu_0 I R^2 N}{2(x^2 + R^2)^{3/2}}$$

- (a) Show that this reduces to the familiar result for field at the centre of the coil.
(b) Consider two parallel co-axial circular coils of equal radius R , and number of turns N , carrying equal currents in the same direction, and separated by a distance R . Show that the field on the axis around the mid-point between the coils is uniform over a distance that is small as compared to R , and is given by,

$$B = 0.72 \frac{\mu_0 N I}{R}, \text{ approximately.}$$

[Such an arrangement to produce a nearly uniform magnetic field over a small region is known as *Helmholtz coils*.]

- 4.17** A toroid has a core (non-ferromagnetic) of inner radius 25 cm and outer radius 26 cm, around which 3500 turns of a wire are wound. If the current in the wire is 11 A, what is the magnetic field (a) outside the toroid, (b) inside the core of the toroid, and (c) in the empty space surrounded by the toroid.

- 4.18** Answer the following questions:

- (a) A magnetic field that varies in magnitude from point to point but has a constant direction (east to west) is set up in a chamber. A charged particle enters the chamber and travels undeflected

Moving Charges and Magnetism

along a straight path with constant speed. What can you say about the initial velocity of the particle?

- (b) A charged particle enters an environment of a strong and non-uniform magnetic field varying from point to point both in magnitude and direction, and comes out of it following a complicated trajectory. Would its final speed equal the initial speed if it suffered no collisions with the environment?
- (c) An electron travelling west to east enters a chamber having a uniform electrostatic field in north to south direction. Specify the direction in which a uniform magnetic field should be set up to prevent the electron from deflecting from its straight line path.
- 4.19** An electron emitted by a heated cathode and accelerated through a potential difference of 2.0 kV, enters a region with uniform magnetic field of 0.15 T. Determine the trajectory of the electron if the field (a) is transverse to its initial velocity, (b) makes an angle of 30° with the initial velocity.
- 4.20** A magnetic field set up using Helmholtz coils (described in Exercise 4.16) is uniform in a small region and has a magnitude of 0.75 T. In the same region, a uniform electrostatic field is maintained in a direction normal to the common axis of the coils. A narrow beam of (single species) charged particles all accelerated through 15 kV enters this region in a direction perpendicular to both the axis of the coils and the electrostatic field. If the beam remains undeflected when the electrostatic field is $9.0 \times 10^{-5} \text{ V m}^{-1}$, make a simple guess as to what the beam contains. Why is the answer not unique?
- 4.21** A straight horizontal conducting rod of length 0.45 m and mass 60 g is suspended by two vertical wires at its ends. A current of 5.0 A is set up in the rod through the wires.
- What magnetic field should be set up normal to the conductor in order that the tension in the wires is zero?
 - What will be the total tension in the wires if the direction of current is reversed keeping the magnetic field same as before? (Ignore the mass of the wires.) $g = 9.8 \text{ m s}^{-2}$.
- 4.22** The wires which connect the battery of an automobile to its starting motor carry a current of 300 A (for a short time). What is the force per unit length between the wires if they are 70 cm long and 1.5 cm apart? Is the force attractive or repulsive?
- 4.23** A uniform magnetic field of 1.5 T exists in a cylindrical region of radius 10.0 cm, its direction parallel to the axis along east to west. A wire carrying current of 7.0 A in the north to south direction passes through this region. What is the magnitude and direction of the force on the wire if,
- the wire intersects the axis,
 - the wire is turned from N-S to northeast-northwest direction,
 - the wire in the N-S direction is lowered from the axis by a distance of 6.0 cm?
- 4.24** A uniform magnetic field of 3000 G is established along the positive z-direction. A rectangular loop of sides 10 cm and 5 cm carries a current of 12 A. What is the torque on the loop in the different cases shown in Fig. 4.28? What is the force on each case? Which case corresponds to stable equilibrium?

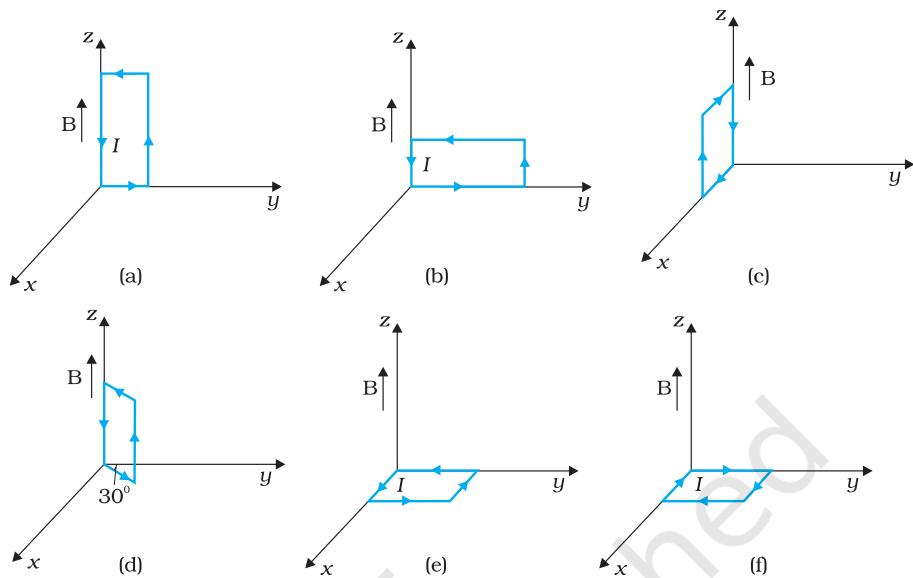


FIGURE 4.28

- 4.25** A circular coil of 20 turns and radius 10 cm is placed in a uniform magnetic field of 0.10 T normal to the plane of the coil. If the current in the coil is 5.0 A, what is the
 (a) total torque on the coil,
 (b) total force on the coil,
 (c) average force on each electron in the coil due to the magnetic field?

(The coil is made of copper wire of cross-sectional area 10^{-5} m^2 , and the free electron density in copper is given to be about 10^{29} m^{-3} .)

- 4.26** A solenoid 60 cm long and of radius 4.0 cm has 3 layers of windings of 300 turns each. A 2.0 cm long wire of mass 2.5 g lies inside the solenoid (near its centre) normal to its axis; both the wire and the axis of the solenoid are in the horizontal plane. The wire is connected through two leads parallel to the axis of the solenoid to an external battery which supplies a current of 6.0 A in the wire. What value of current (with appropriate sense of circulation) in the windings of the solenoid can support the weight of the wire? $g = 9.8 \text{ m s}^{-2}$.

- 4.27** A galvanometer coil has a resistance of 12Ω and the metre shows full scale deflection for a current of 3 mA. How will you convert the metre into a voltmeter of range 0 to 18 V?

- 4.28** A galvanometer coil has a resistance of 15Ω and the metre shows full scale deflection for a current of 4 mA. How will you convert the metre into an ammeter of range 0 to 6 A?

Chapter Five

MAGNETISM AND MATTER



5.1 INTRODUCTION

Magnetic phenomena are universal in nature. Vast, distant galaxies, the tiny invisible atoms, men and beasts all are permeated through and through with a host of magnetic fields from a variety of sources. The earth magnetism predates human evolution. The word magnet is derived from the name of an island in Greece called *magnesia* where magnetic ore deposits were found, as early as 600 BC. Shepherds on this island complained that their wooden shoes (which had nails) at times stayed struck to the ground. Their iron-tipped rods were similarly affected. This attractive property of magnets made it difficult for them to move around.

The directional property of magnets was also known since ancient times. A thin long piece of a magnet, when suspended freely, pointed in the north-south direction. A similar effect was observed when it was placed on a piece of cork which was then allowed to float in still water. The name *lodestone* (or *loadstone*) given to a naturally occurring ore of iron-magnetite means leading stone. The technological exploitation of this property is generally credited to the Chinese. Chinese texts dating 400 BC mention the use of magnetic needles for navigation on ships. Caravans crossing the Gobi desert also employed magnetic needles.

A Chinese legend narrates the tale of the victory of the emperor Huang-ti about four thousand years ago, which he owed to his craftsmen (whom

Physics

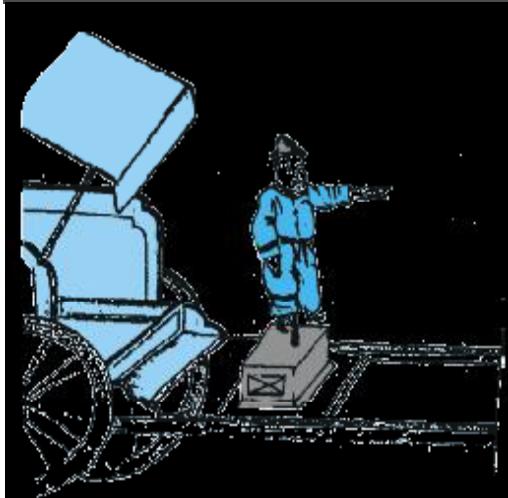


FIGURE 5.1 The arm of the statuette mounted on the chariot always points south. This is an artist of the earliest known compasses, thousands of years old.

nowadays you would call engineers). These built a chariot on which they placed a magnetic figure with arms outstretched. Figure 5.1 is an artist description of this chariot. The figure swiveled around so that the finger of the statuette on it always pointed south. With this chariot, Huang-ti to attack the enemy from the rear in thick fog, and to defeat them.

In the previous chapter we have learned that moving charges or electric currents produce magnetic fields. This discovery, which was made in the early part of the nineteenth century is credited to Oersted, Ampere, Biot and Savart, among others.

In the present chapter, we take a look at magnetism as a subject in its own right.

Some of the commonly known ideas regarding magnetism are:

- (i) The earth behaves as a magnet with the magnetic field pointing approximately from the geographic south to the north.
- (ii) When a bar magnet is freely suspended, it points in the north-south direction. The tip which points to the geographic north is called the *north pole* and the tip which points to the geographic south is called the *south pole* of the magnet.
- (iii) There is a repulsive force when north poles (or south poles) of two magnets are brought close together. Conversely, there is an attractive force between the north pole of one magnet and the south pole of the other.
- (iv) We cannot isolate the north, or south pole of a magnet. If a bar magnet is broken into two halves, we get two similar bar magnets with somewhat weaker properties. Unlike electric charges, isolated magnetic north and south poles known as *magnetic monopoles* do not exist.
- (v) It is possible to make magnets out of iron and its alloys.

We begin with a description of a bar magnet and its behaviour in an external magnetic field. We describe Gauss follow it up with an account of the earth how materials can be classified on the basis of their magnetic properties. We describe para-, dia-, and ferromagnetism. We conclude with a section on electromagnets and permanent magnets.

5.2 THE BAR MAGNET

One of the earliest childhood memories of the famous physicist Albert Einstein was that of a magnet gifted to him by a relative. Einstein was fascinated, and played endlessly with it. He wondered how the magnet could affect objects such as nails or pins placed away from it and not in any way connected to it by a spring or string.

We begin our study by examining iron filings sprinkled on a sheet of glass placed over a short bar magnet. The arrangement of iron filings is shown in Fig. 5.2.

The pattern of iron filings suggests that the magnet has two poles similar to the positive and negative charge of an electric dipole. As mentioned in the introductory section, one pole is designated the *North pole* and the other, the *South pole*. When suspended freely, these poles point approximately towards the geographic north and south poles, respectively. A similar pattern of iron filings is observed around a current carrying solenoid.

5.2.1 The magnetic field lines

The pattern of iron filings permits us to plot the magnetic field lines*. This is shown both for the bar-magnet and the current-carrying solenoid in Fig. 5.3. For comparison refer to the Chapter 1, Figure 1.17(d). Electric field lines of an electric dipole are also displayed in Fig. 5.3(c). The magnetic field lines are a visual and intuitive realisation of the magnetic field. Their properties are:

- The magnetic field lines of a magnet (or a solenoid) form continuous closed loops. This is unlike the electric dipole where these field lines begin from a positive charge and end on the negative charge or escape to infinity.
- The tangent to the field line at a given point represents the direction of the net magnetic field \mathbf{B} at that point.

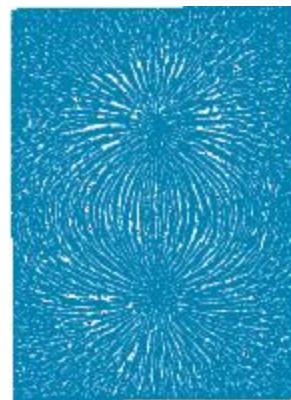


FIGURE 5.2 The arrangement of iron filings surrounding a bar magnet. The pattern mimics magnetic field lines. The pattern suggests that the bar magnet is a magnetic dipole.

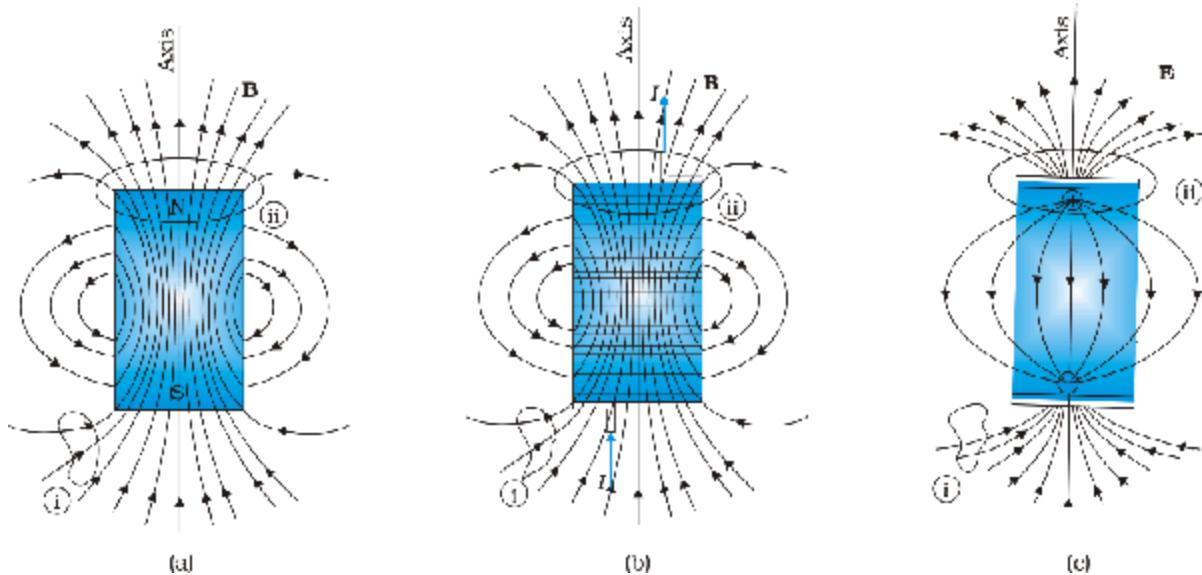


FIGURE 5.3 The field lines of (a) a bar magnet, (b) a current-carrying finite solenoid and (c) electric dipole. At large distances, the field lines are very similar. The curves labelled ① and ② are closed Gaussian surfaces.

* In some textbooks the magnetic field lines are called *magnetic lines of force*. This nomenclature is avoided since it can be confusing. Unlike electrostatics the field lines in magnetism do not indicate the direction of the force on a (moving) charge.

■ Physics

(iii) The larger the number of field lines crossing per unit area, the stronger is the magnitude of the magnetic field B . In Fig. 5.3(a), B is larger around region (ii) than in region (i).

(iv) The magnetic field lines do not intersect, for if they did, the direction of the magnetic field would not be unique at the point of intersection.

One can plot the magnetic field lines in a variety of ways. One way is to place a small magnetic compass needle at various positions and note its orientation. This gives us an idea of the magnetic field direction at various points in space.

5.2.2 Bar magnet as an equivalent solenoid

In the previous chapter, we have explained how a current loop acts as a magnetic dipole (Section 4.10). We mentioned Ampere all magnetic phenomena can be explained in terms of circulating currents.

Recall that the magnetic dipole moment m associated with a current loop was defined to be $m = NIA$ where N is the number of turns in the loop, I the current and A the area vector (Eq. 4.30).

The resemblance of magnetic field lines for a bar magnet and a solenoid suggest that a bar magnet may be thought of as a large number of circulating currents in analogy with a solenoid. Cutting a bar magnet in half is like cutting a solenoid. We get two smaller solenoids with weaker magnetic properties. The field lines remain continuous, emerging from one face of the solenoid and entering into the other face. One can test this analogy by moving a small compass needle in the neighbourhood of a bar magnet and a current-carrying finite solenoid and noting that the deflections of the needle are similar in both cases.

To make this analogy more firm we calculate the axial field of a finite solenoid depicted in Fig. 5.4 (a). We shall demonstrate that at large distances this axial field resembles that of a bar magnet.

Let the solenoid of Fig. 5.4(a) consists of n turns per unit length. Let its length be $2l$ and radius a . We can evaluate the axial field at a point P , at a distance r from the centre O

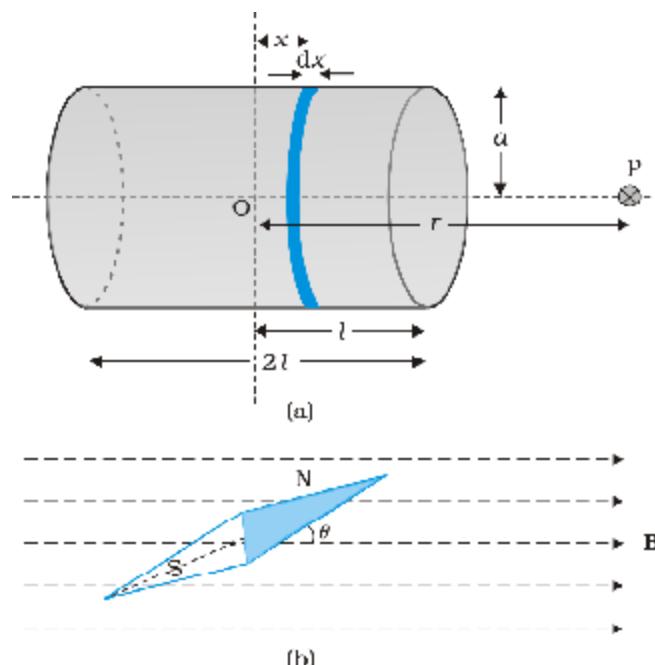


FIGURE 5.4 Calculation of (a) The axial field of a finite solenoid in order to demonstrate its similarity to that of a bar magnet. (b) A magnetic needle in a uniform magnetic field B . The arrangement may be used to determine either B or the magnetic moment m of the needle.

of the solenoid. To do this, consider a circular element of thickness dx of the solenoid at a distance x from its centre. It consists of $n dx$ turns. Let I be the current in the solenoid. In Section 4.6 of the previous chapter we have calculated the magnetic field on the axis of a circular current loop. From Eq. (4.13), the magnitude of the field at point P due to the circular element is

$$dB = \frac{\mu_0 n dx I a^2}{2[(r - x)^2 + a^2]^{3/2}}$$

The magnitude of the total field is obtained by summing over all the elements $x = -l$ to $x = +l$. Thus,

$$B = \frac{\mu_0 n l a^2}{2} \int_{-l}^l \frac{dx}{[(r - x)^2 + a^2]^{3/2}}$$

This integration can be done by trigonometric substitutions. This exercise, however, is not necessary for our purpose. Note that the range of x is from $-l$ to $+l$. Consider the far axial field of the solenoid, i.e., $r \gg a$ and $r \gg l$. Then the denominator is approximated by

$$[(r - x)^2 + a^2]^{3/2} \approx r^3$$

$$\begin{aligned} \text{and } B &= \frac{\mu_0 n l a^2}{2r^3} \int_{-l}^l dx \\ &= \frac{\mu_0 n l}{2} \frac{2l a^2}{r^3} \end{aligned} \tag{5.1}$$

Note that the magnitude of the magnetic moment of the solenoid is, $m = n(2l)I(\pi a^2)$ (area). Thus,

$$B = \frac{\mu_0}{4\pi} \frac{2m}{r^3} \tag{5.2}$$

This is also the far axial magnetic field of a bar magnet which one may obtain experimentally. Thus, a bar magnet and a solenoid produce similar magnetic fields. The magnetic moment of a bar magnet is thus equal to the magnetic moment of an equivalent solenoid that produces the same magnetic field.

Some textbooks assign a *magnetic charge* (also called *pole strength*) $+q_m$ to the north pole and $-q_m$ to the south pole of a bar magnet of length $2l$, and magnetic moment $q_m(2l)$. The field strength due to q_m at a distance r from it is given by $\mu_0 q_m / 4\pi r^2$. The magnetic field due to the bar magnet is then obtained, both for the axial and the equatorial case, in a manner analogous to that of an electric dipole (Chapter 1). The method is simple and appealing. However, *magnetic monopoles do not exist*, and we have avoided this approach for that reason.

5.2.3 The dipole in a uniform magnetic field

The pattern of iron filings, i.e., the magnetic field lines gives us an approximate idea of the magnetic field B . We may at times be required to determine the magnitude of B accurately. This is done by placing a small compass needle of known magnetic moment m and moment of inertia I and allowing it to oscillate in the magnetic field. This arrangement is shown in Fig. 5.4(b).

The torque on the needle is [see Eq. (4.29)],

$$\tau = m \times B \tag{5.3}$$

Physics

In magnitude $\tau = mB \sin \theta$

Here τ is restoring torque and θ is the angle between m and B .

Therefore, in equilibrium $I \frac{d^2\theta}{dt^2} = -mB \sin \theta$

Negative sign with $mB \sin \theta$ implies that restoring torque is in opposition to deflecting torque. For small values of θ in radians, we approximate $\sin \theta \approx \theta$ and get

$$I \frac{d^2\theta}{dt^2} \approx mB \theta$$

$$\text{or, } \frac{d^2\theta}{dt^2} = -\frac{mB}{I} \theta$$

This represents a simple harmonic motion. The square of the angular frequency is $\omega^2 = mB/I$ and the time period is,

$$T = 2\pi \sqrt{\frac{I}{mB}} \quad (5.4)$$

$$\text{or } B = \frac{4\pi^2 I}{m T^2} \quad (5.5)$$

An expression for magnetic potential energy can also be obtained on lines similar to electrostatic potential energy.

The magnetic potential energy U_m is given by

$$\begin{aligned} U_m &= \int \tau(\theta) d\theta \\ &= \int mB \sin \theta = -mB \cos \theta \\ &= -m \cdot B \end{aligned} \quad (5.6)$$

We have emphasised in Chapter 2 that the zero of potential energy can be fixed at one

zero means fixing the zero of potential energy at $\theta = 90^\circ$, i.e., when the needle is perpendicular to the field. Equation (5.6) shows that potential energy is minimum ($= -mB$) at $\theta = 0^\circ$ (most stable position) and maximum ($= +mB$) at $\theta = 180^\circ$ (most unstable position).

Example 5.1 In Fig. 5.4(b), the magnetic needle has magnetic moment 6.7×10^{-6} Am 2 and moment of inertia $I = 7.5 \times 10^{-6}$ kg m 2 . It performs 10 complete oscillations in 6.70 s. What is the magnitude of the magnetic field?

Solution The time period of oscillation is,

$$T = \frac{6.70}{10} = 0.67 \text{ s}$$

From Eq. (5.5)

$$\begin{aligned} B &= \frac{4\pi^2 I}{m T^2} \\ &= \frac{4 \times (3.14)^2 \times 7.5 \times 10^{-6}}{6.7 \times 10^{-6} \times (0.67)} \\ &= 0.01 \text{ T} \end{aligned}$$

Example 5.2 A short bar magnet placed with its axis at 30° with an external field of 800 G experiences a torque of 0.016 Nm. (a) What is the magnetic moment of the magnet? (b) What is the work done in moving it from its most stable to most unstable position? (c) The bar magnet is replaced by a solenoid of cross-sectional area $2 \times 10^{-3} \text{ m}^2$ and 1000 turns, but of the same magnetic moment. Determine the current flowing through the solenoid.

Solution

(a) From Eq. (5.3), $\tau = m B \sin \theta$, $\theta = 30^\circ$, hence $\sin \theta = 1/2$.

$$\text{Thus, } 0.016 = m \times (800 \times 10 \text{ T}) \times (1/2)$$

$$m = 160 \times 2/800 = 0.40 \text{ A m}^2$$

(b) From Eq. (5.6), the most stable position is $\theta = 0^\circ$ and the most unstable position is $\theta = 180^\circ$. Work done is given by

$$W = U_m(\theta = 180^\circ) - U_m(\theta = 0^\circ) \\ = 2 m B = 2 \times 0.40 \times 800 \times 10 = 0.064 \text{ J}$$

(c) From Eq. (4.30), $m_s = NIA$. From part (a), $m_s = 0.40 \text{ A m}^2$

$$0.40 = 1000 \times I \times 2 \times 10$$

$$I = 0.40 \times 10^4 / (1000 \times 2) = 2 \text{ A}$$

EXAMPLE 5.2

Example 5.3

- (a) What happens if a bar magnet is cut into two pieces: (i) transverse to its length, (ii) along its length?
- (b) A magnetised needle in a uniform magnetic field experiences a torque but no net force. An iron nail near a bar magnet, however, experiences a force of attraction in addition to a torque. Why?
- (c) Must every magnetic configuration have a north pole and a south pole? What about the field due to a toroid?
- (d) Two identical looking iron bars A and B are given, one of which is definitely known to be magnetised. (We do not know which one.) How would one ascertain whether or not both are magnetised? If only one is magnetised, how does one ascertain which one? [Use nothing else but the bars A and B.]

Solution

- (a) In either case, one gets two magnets, each with a north and south pole.
- (b) No force if the field is uniform. The iron nail experiences a non-uniform field due to the bar magnet. There is induced magnetic moment in the nail, therefore, it experiences both force and torque. The net force is attractive because the induced south pole (say) in the nail is closer to the north pole of magnet than induced north pole.
- (c) Not necessarily. True only if the source of the field has a net non-zero magnetic moment. This is not so for a toroid or even for a straight infinite conductor.
- (d) Try to bring different ends of the bars closer. A repulsive force in some situation establishes that both are magnetised. If it is always attractive, then one of them is not magnetised. In a bar magnet the intensity of the magnetic field is the strongest at the two ends (poles) and weakest at the central region. This fact may be used to determine whether A or B is the magnet. In this case, to see which

EXAMPLE 5.3

■ Physics

EXAMPLE 5.3

one of the two bars is a magnet, pick up one, (say, A) and lower one of its ends; first on one of the ends of the other (say, B), and then on the middle of B. If you notice that in the middle of B, A experiences no force, then B is magnetised. If you do not notice any change from the end to the middle of B, then A is magnetised.

5.2.4 The electrostatic analog

Comparison of Eqs. (5.2), (5.3) and (5.6) with the corresponding equations for electric dipole (Chapter 1), suggests that magnetic field at large distances due to a bar magnet of magnetic moment m can be obtained from the equation for electric field due to an electric dipole of dipole moment p , by making the following replacements:

$$E \rightarrow B, p \rightarrow m, \frac{1}{4\pi\epsilon_0} \rightarrow \frac{\mu_0}{4\pi}$$

In particular, we can write down the equatorial field (B_E) of a bar magnet at a distance r , for $r \gg l$, where l is the size of the magnet:

$$B_E = -\frac{\mu_0 m}{4\pi r^3} \quad (5.7)$$

Likewise, the axial field (B_A) of a bar magnet for $r \gg l$ is:

$$B_A = \frac{\mu_0}{4\pi} \frac{2m}{r^3} \quad (5.8)$$

Equation (5.8) is just Eq. (5.2) in the vector form. Table 5.1 summarises the analogy between electric and magnetic dipoles.

TABLE 5.1 THE DIPOLE ANALOGY

	Electrostatics	Magnetism
Dipole moment	$1/\epsilon_0$	μ_0
Equatorial Field for a short dipole	p	m
Axial Field for a short dipole	$/4\pi\epsilon_0 r^3$	$\mu_0 m / 4\pi r^3$
External Field: torque	$2p/4\pi\epsilon_0 r^3$	$\mu_0 2m / 4\pi r^3$
External Field: Energy	$p \times E$	$m \times B$
	$\cdot E$	$\cdot B$

EXAMPLE 5.4

Example 5.4 What is the magnitude of the equatorial and axial fields due to a bar magnet of length 5.0 cm at a distance of 50 cm from its mid-point? The magnetic moment of the bar magnet is 0.40 A m^2 , the same as in Example 5.2.

Solution From Eq. (5.7)

$$B_E = \frac{\mu_0 m}{4\pi r^3} = \frac{10^{-7} \times 0.4}{(0.5)^3} = \frac{10^{-7} \times 0.4}{0.125} = 3.2 \times 10^{-7} \text{ T}$$

$$\text{From Eq. (5.8), } B_A = \frac{\mu_0 2m}{4\pi r^3} = 6.4 \times 10^{-7} \text{ T}$$

Example 5.5 Figure 5.5 shows a small magnetised needle P placed at a point O. The arrow shows the direction of its magnetic moment. The other arrows show different positions (and orientations of the magnetic moment) of another identical magnetised needle Q.

- In which configuration the system is not in equilibrium?
- In which configuration is the system in (i) stable, and (ii) unstable equilibrium?
- Which configuration corresponds to the lowest potential energy among all the configurations shown?

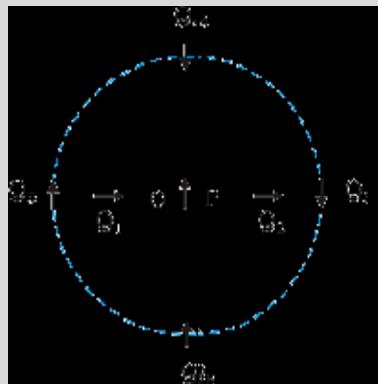


FIGURE 5.5

Solution

Potential energy of the configuration arises due to the potential energy of one dipole (say, Q) in the magnetic field due to other (P). Use the result that the field due to P is given by the expression [Eqs. (5.7) and (5.8)]:

$$B_p = -\frac{\mu_0}{4\pi} \frac{m_p}{r^3} \quad (\text{on the normal bisector})$$

$$B_p = \frac{\mu_0}{4\pi} \frac{2m_p}{r^3} \quad (\text{on the axis})$$

where m_p is the magnetic moment of the dipole P.

Equilibrium is stable when m_Q is parallel to B_p , and unstable when it is anti-parallel to B_p .

For instance for the configuration Q_3 for which Q is along the perpendicular bisector of the dipole P, the magnetic moment of Q is parallel to the magnetic field at the position 3. Hence Q_3 is stable.

Thus,

- PQ_1 and PQ_2
- (i) PQ_3 , PQ_6 (stable); (ii) PQ_5 , PQ_4 (unstable)
- PQ_6

EXAMPLE 5.5

5.3 MAGNETISM AND GAUSS'S LAW

In Chapter 1, we studied Gauss

see that for a closed surface represented by ①, the number of lines leaving the surface is equal to the number of lines entering it. This is consistent with the fact that no net charge is enclosed by the surface. However, in the same figure, for the closed surface ②, there is a net outward flux, since it does include a net (positive) charge.

Physics

KARL FRIEDRICH GAUSS (1777 – 1855)



Karl Friedrich Gauss (1777 – 1855) was a child prodigy and was gifted in mathematics, physics, engineering, astronomy and even land surveying. The properties of numbers fascinated him, and in his work he anticipated major mathematical development of later times. Along with Wilhelm Welser, he built the first electric telegraph in 1833. His mathematical theory of curved surface laid the foundation for the later work of Riemann.

The situation is radically different for magnetic fields which are continuous and form closed loops. Examine the Gaussian surfaces represented by ① or ⑪ in Fig 5.3(a) or Fig. 5.3(b). Both cases visually demonstrate that the number of magnetic field lines leaving the surface is balanced by the number of lines entering it. The *net magnetic flux is zero for both the surfaces*. This is true for any closed surface.

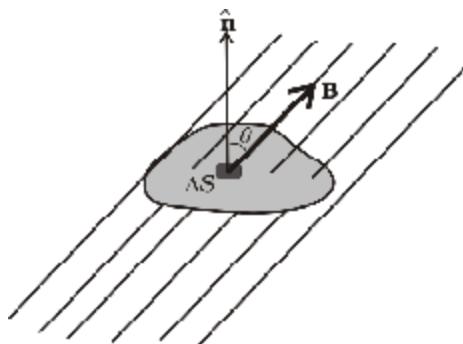


FIGURE 5.6

Consider a small vector area element ΔS of a closed surface S as in Fig. 5.6. The magnetic flux through ΔS is defined as $\Delta\phi_B = \mathbf{B} \cdot \Delta \mathbf{S}$, where \mathbf{B} is the field at ΔS . We divide S into many small area elements and calculate the individual flux through each. Then, the net flux ϕ_B is,

$$\phi_B = \sum_{\text{'all'}} \Delta\phi_B = \sum_{\text{'all'}} \mathbf{B} \cdot \Delta \mathbf{S} = 0 \quad (5.9)$$

where

with the Gauss
in that case is given by

$$\sum E \cdot \Delta S = \frac{q}{\epsilon_0}$$

where q is the electric charge enclosed by the surface.

The difference between the Gauss electrostatics is a reflection of the fact that isolated magnetic poles (also called monopoles) are not known to exist. There are no sources or sinks of \mathbf{B} ; the simplest magnetic element is a dipole or a current loop. All magnetic phenomena can be explained in terms of an arrangement of dipoles and/or current loops.

Thus, Gauss

The net magnetic flux through any closed surface is zero.

EXAMPLE 5.6

Example 5.6 Many of the diagrams given in Fig. 5.7 show magnetic field lines (thick lines in the figure) wrongly. Point out what is wrong with them. Some of them may describe electrostatic field lines correctly. Point out which ones.

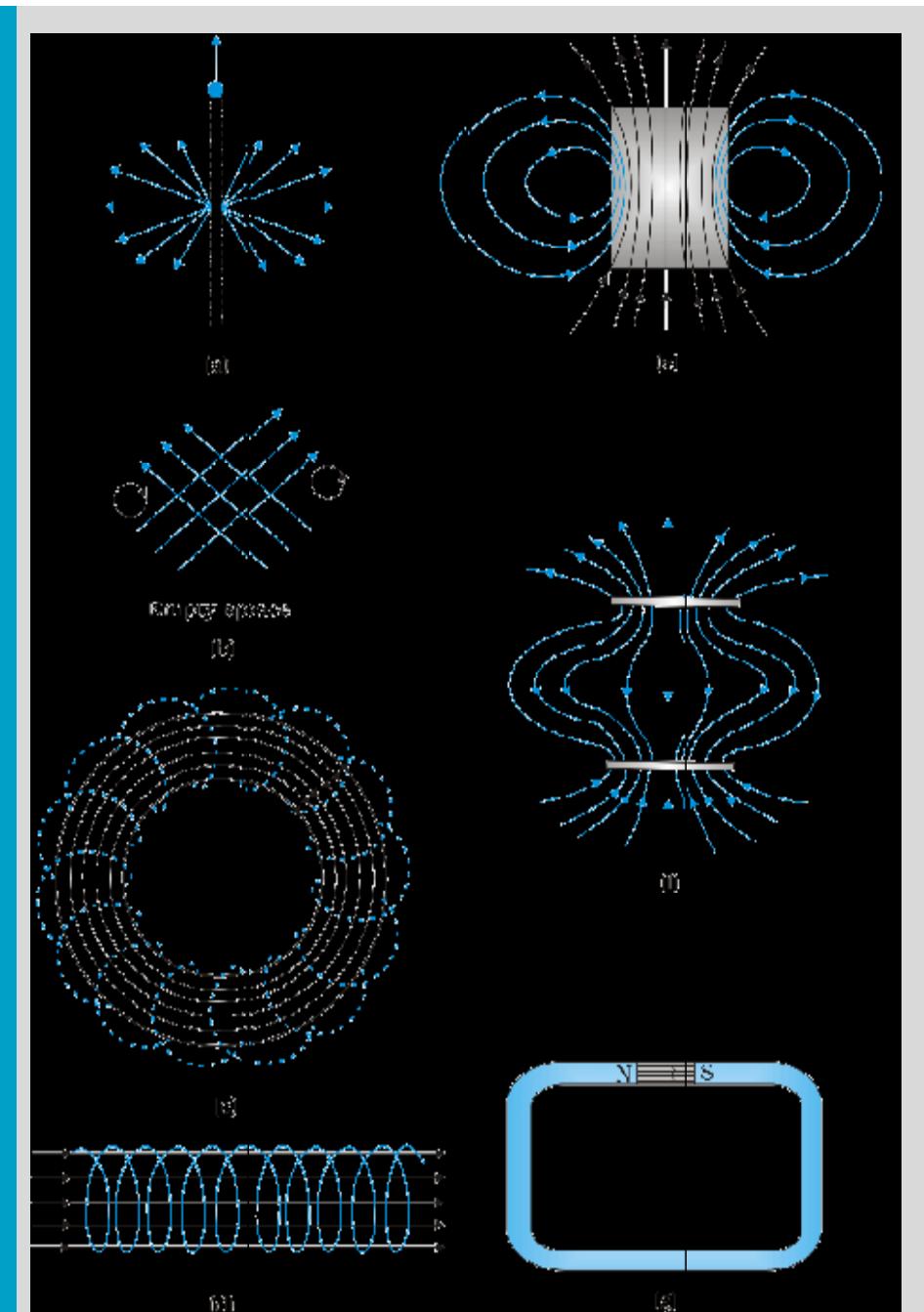


FIGURE 5.7

Solution

- (a) *Wrong.* Magnetic field lines can never emanate from a point, as shown in figure. Over any closed surface, the net flux of B must always be zero, i.e., pictorially as many field lines should seem to enter the surface as the number of lines leaving it. The field lines shown, in fact, represent electric field of a long positively charged wire. The correct magnetic field lines are circling the straight conductor, as described in Chapter 4.

EXAMPLE 5.6

Physics

EXAMPLE 5.6

- (b) *Wrong.* Magnetic field lines (like electric field lines) can never cross each other, because otherwise the direction of field at the point of intersection is ambiguous. There is further error in the figure. Magnetostatic field lines can never form closed loops around empty space. A closed loop of static magnetic field line must enclose a region across which a current is passing. By contrast, electrostatic field lines can never form closed loops, neither in empty space, nor when the loop encloses charges.
- (c) *Right.* Magnetic lines are completely confined within a toroid. Nothing wrong here in field lines forming closed loops, since each loop encloses a region across which a current passes. Note, for clarity of figure, only a few field lines within the toroid have been shown. Actually, the entire region enclosed by the windings contains magnetic field.
- (d) *Wrong.* Field lines due to a solenoid at its ends and outside cannot be so completely straight and confined; such a thing violates Ampere eventually to form closed loops.
- (e) *Right.* These are field lines outside and inside a bar magnet. Note carefully the direction of field lines inside. Not all field lines emanate out of a north pole (or converge into a south pole). Around both the N-pole, and the S-pole, the net flux of the field is zero.
- (f) *Wrong.* These field lines cannot possibly represent a magnetic field. Look at the upper region. All the field lines seem to emanate out of the shaded plate. The net flux through a surface surrounding the shaded plate is not zero. This is impossible for a magnetic field. The given field lines, in fact, show the electrostatic field lines around a positively charged upper plate and a negatively charged lower plate. The difference between Fig. [5.7(e) and (f)] should be carefully grasped.
- (g) *Wrong.* Magnetic field lines between two pole pieces cannot be precisely straight at the ends. Some fringing of lines is inevitable. Otherwise, Ampere field lines.

EXAMPLE 5.7

Example 5.7

- (a) Magnetic field lines show the direction (at every point) along which a small magnetised needle aligns (at the point). Do the magnetic field lines also represent the *lines of force* on a moving charged particle at every point?
- (b) Magnetic field lines can be entirely confined within the core of a toroid, but not within a straight solenoid. Why?
- (c) If magnetic monopoles existed, how would the Gauss magnetism be modified?
- (d) Does a bar magnet exert a torque on itself due to its own field? Does one element of a current-carrying wire exert a force on another element of the *same wire*?
- (e) Magnetic field arises due to charges in motion. Can a system have magnetic moments even though its net charge is zero?

Solution

- (a) *No.* The magnetic force is always normal to \mathbf{B} (remember magnetic force = $qv \times \mathbf{B}$). It is misleading to call *magnetic field lines* as *lines of force*.

- (b) If field lines were entirely confined between two ends of a straight solenoid, the flux through the cross-section at each end would be non-zero. But the flux of field B through any closed surface must always be zero. For a toroid, this difficulty is absent because it has no
- (c) Gauss B through any closed surface is always zero $\oint_S \mathbf{B} \cdot d\mathbf{s} = 0$.
If monopoles existed, the right hand side would be equal to the monopole (magnetic charge) q_m enclosed by S . [Analogous to Gauss $\int_S \mathbf{B} \cdot d\mathbf{s} = \mu_0 q_m$ where q_m is the (monopole) magnetic charge enclosed by S .]
- (d) No. There is no force or torque on an element due to the field produced by that element itself. But there is a force (or torque) on an element of the same wire. (For the special case of a straight wire, this force is zero.)
- (e) Yes. The average of the charge in the system may be zero. Yet, the mean of the magnetic moments due to various current loops may not be zero. We will come across such examples in connection with paramagnetic material where atoms have net dipole moment through their net charge is zero.

EXAMPLE 5.7

5.4 THE EARTH'S MAGNETISM

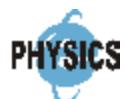
Earlier we have referred to the magnetic field of the earth. The strength of the earth

its value being of the order of 10^{-5} T.

What causes the earth to have a magnetic field is not clear. Originally the magnetic field was thought of as arising from a giant bar magnet placed approximately along the axis of rotation of the earth and deep in the interior. However, this simplistic picture is certainly not correct. The magnetic field is now thought to arise due to electrical currents produced by convective motion of metallic fluids (consisting mostly of molten iron and nickel) in the outer core of the earth. This is known as the *dynamo effect*.

The magnetic field lines of the earth resemble that of a (hypothetical) magnetic dipole located at the centre of the earth. The axis of the dipole does not coincide with the axis of rotation of the earth but is presently tilted by approximately 11.3° with respect to the later. In this way of looking at it, the magnetic poles are located where the magnetic field lines due to the dipole enter or leave the earth. The location of the north magnetic pole is at a latitude of 79.74° N and a longitude of 71.8° W, a place somewhere in north Canada. The magnetic south pole is at 79.74° S, 108.22° E in the Antarctica.

The pole near the geographic north pole of the earth is called the *north magnetic pole*. Likewise, the pole near the geographic south pole is called



Geomagnetic field frequently asked questions
<http://www.ngdc.noaa.gov/seg/geomag/>

Physics

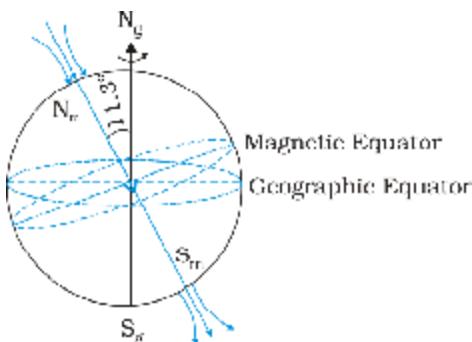


FIGURE 5.8 The earth as a giant magnetic dipole.

EXAMPLE 5.8

Example 5.8 The earth
0.4 G. Estimate the earth

Solution From Eq. (5.7), the equatorial magnetic field is,

$$B_E = \frac{\mu_0 m}{4\pi r^3}$$

We are given that $B_E \sim 0.4$ G = 4×10^{-5} T. For r , we take the radius of the earth 6.4×10^6 m. Hence,

$$m = \frac{4 \times 10^{-5} \times (6.4 \times 10^6)^3}{\mu_0 / 4\pi} = 4 \times 10^2 \times (6.4 \times 10^6)^3 \quad (\mu_0 / 4\pi = 10) \\ = 1.05 \times 10^{23} \text{ A m}^2$$

This is close to the value 8×10^{22} A m² quoted in geomagnetic texts.

5.4.1 Magnetic declination and dip

Consider a point on the earth

the longitude circle determines the geographic north-south direction, the

line of longitude towards the north pole being the direction of true north. The vertical plane containing the longitude circle and the axis of rotation of the earth is called the *geographic meridian*. In a similar way, one can define *magnetic meridian* of a place as the vertical plane which passes through the imaginary line joining the magnetic north and the south poles. This plane would intersect the surface of the earth in a longitude like circle. A magnetic needle, which is free to swing horizontally, would then lie in the magnetic meridian and the north pole of the needle would point towards the magnetic north pole. Since the line joining the magnetic poles is tilted with respect to the geographic axis of the earth, the magnetic meridian at a point makes angle with the geographic meridian. This, then, is the angle between the true geographic north and the north shown by a compass needle. This angle is called the *magnetic declination* or simply *declination* (Fig. 5.9).

The declination is greater at higher latitudes and smaller near the equator. The declination in India is small, it being

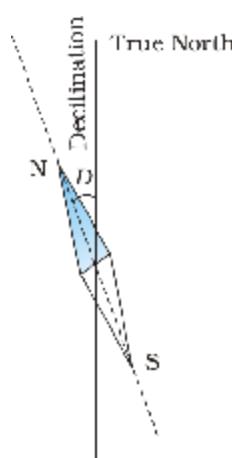


FIGURE 5.9 A magnetic needle free to move in horizontal plane, points toward the magnetic north-south direction.

$0^\circ 41' E$ at Delhi and $0^\circ 58' W$ at Mumbai. Thus, at both these places a magnetic needle shows the true north quite accurately.

There is one more quantity of interest. If a magnetic needle is perfectly balanced about a horizontal axis so that it can swing in a plane of the magnetic meridian, the needle would make an angle with the horizontal (Fig. 5.10). This is known as the *angle of dip* (also known as *inclination*). Thus, dip is the angle that the total magnetic field B_E of the earth makes with the surface of the earth. Figure 5.11 shows the magnetic meridian plane at a point P on the surface of the earth. The plane is a section through the earth. The total magnetic field at P can be resolved into a horizontal component H_E and a vertical component Z_E . The angle that B_E makes with H_E is the angle of dip, I .

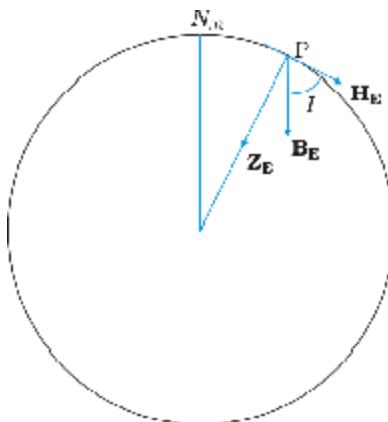


FIGURE 5.10 The circle is a section through the earth containing the magnetic meridian. The angle between B_E and the horizontal component H_E is the angle of dip.

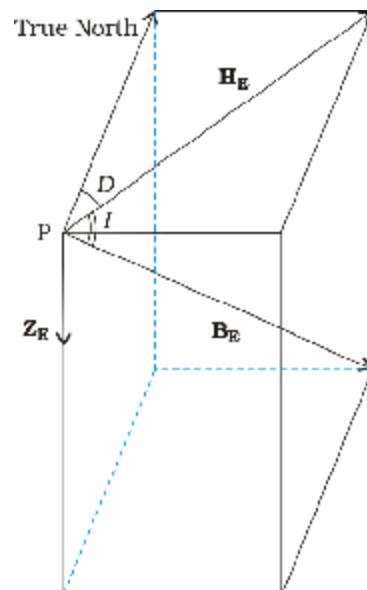


FIGURE 5.11 The earth magnetic field, B_E , its horizontal and vertical components, H_E and Z_E . Also shown are the declination, D and the inclination or angle of dip, I .

In most of the northern hemisphere, the north pole of the dip needle tilts downwards. Likewise in most of the southern hemisphere, the south pole of the dip needle tilts downwards.

To describe the magnetic field of the earth at a point on its surface, we need to specify three quantities, viz., the declination D , the angle of dip or the inclination I and the horizontal component of the earth H_E . These are known as the *elements of the earth*.

Representing the vertical component by Z_E , we have

$$Z_E = B_E \sin I \quad [5.10(a)]$$

$$H_E = B_E \cos I \quad [5.10(b)]$$

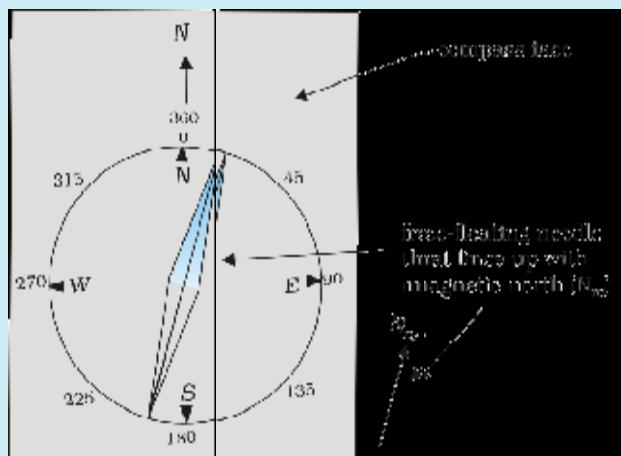
which gives,

$$\tan I = \frac{Z_E}{H_E} \quad [5.10(c)]$$

Physics

WHAT HAPPENS TO MY COMPASS NEEDLES AT THE POLES?

A compass needle consists of a magnetic needle which floats on a pivotal point. When the compass is held level, it points along the direction of the horizontal component of the earth magnetic field at the location. Thus, the compass needle would stay along the magnetic meridian of the place. In some places on the earth there are deposits of magnetic minerals which cause the compass needle to deviate from the magnetic meridian. Knowing the magnetic declination at a place allows us to correct the compass to determine the direction of true north.



So what happens if we take our compass to the magnetic pole? At the poles, the magnetic field lines are converging or diverging vertically so that the horizontal component is negligible. If the needle is only capable of moving in a horizontal plane, it can point along any direction, rendering it useless as a direction finder. What one needs in such a case is a *dip needle* which is a compass pivoted to move in a vertical plane containing the magnetic field of the earth. The needle of the compass then shows the angle which the magnetic field makes with the vertical. At the magnetic poles such a needle will point straight down.

EXAMPLE 5.9

Example 5.9 In the magnetic meridian of a certain place, the horizontal component of the earth dip angle is 60° . What is the magnetic field of the earth at this location?

Solution

It is given that $H_E = 0.26$ G. From Fig. 5.11, we have

$$\cos 60^\circ = \frac{H_E}{B_E}$$

$$B_E = \frac{H_E}{\cos 60^\circ}$$

$$= \frac{0.26}{(1/2)} = 0.52 \text{ G}$$

EARTH'S MAGNETIC FIELD

It must not be assumed that there is a giant bar magnet deep inside the earth which is causing the earth

it is highly unlikely that a large solid block of iron stretches from the magnetic north pole to the magnetic south pole. The earth

nickel are responsible for earth

which has no molten core, has no magnetic field, Venus has a slower rate of rotation, and a weaker magnetic field, while Jupiter, which has the fastest rotation rate among planets, has a fairly strong magnetic field. However, the precise mode of these circulating currents and the energy needed to sustain them are not very well understood. These are several open questions which form an important area of continuing research.

The variation of the earth

study. Charged particles emitted by the sun flow towards the earth and beyond, in a stream called the solar wind. Their motion is affected by the earth

affect the pattern of the earth

quite different from that in other regions of the earth.

The variation of earth

term variations taking place over centuries and long term variations taking place over a period of a million years. In a span of 240 years from 1580 to 1820 AD, over which records are available, the magnetic declination at London has been found to change by 3.5° , suggesting that the magnetic poles inside the earth change position with time. On the scale of a million years, the earth

contains iron, and basalt is emitted during volcanic activity. The little iron magnets inside it align themselves parallel to the magnetic field at that place as the basalt cools and solidifies. Geological studies of basalt containing such pieces of magnetised region have provided evidence for the change of direction of earth

5.5 MAGNETISATION AND MAGNETIC INTENSITY

The earth abounds with a bewildering variety of elements and compounds. In addition, we have been synthesising new alloys, compounds and even elements. One would like to classify the magnetic properties of these substances. In the present section, we define and explain certain terms which will help us to carry out this exercise.

We have seen that a circulating electron in an atom has a magnetic moment. In a bulk material, these moments add up vectorially and they can give a net magnetic moment which is non-zero. We define *magnetisation M* of a sample to be equal to its net magnetic moment per unit volume:

$$M = \frac{m_{net}}{V} \quad (5.11)$$

M is a vector with dimensions L A and is measured in units of A m .

Consider a long solenoid of n turns per unit length and carrying a current I . The magnetic field in the interior of the solenoid was shown to be given by

■ Physics

$$B_0 = \mu_0 n l \quad (5.12)$$

If the interior of the solenoid is filled with a material with non-zero magnetisation, the field inside the solenoid will be greater than B_0 . The net B field in the interior of the solenoid may be expressed as

$$B = B_0 + B_m \quad (5.13)$$

where B_m is the field contributed by the material core. It turns out that this additional field B_m is proportional to the magnetisation M of the material and is expressed as

$$B_m = \mu_0 M \quad (5.14)$$

where μ_0 is the same constant (permittivity of vacuum) that appears in Biot-Savart

It is convenient to introduce another vector field H , called the *magnetic intensity*, which is defined by

$$H = \frac{B}{\mu_0} - M \quad (5.15)$$

where H has the same dimensions as M and is measured in units of $A m^{-1}$. Thus, the total magnetic field B is written as

$$B = \mu_0 (H + M) \quad (5.16)$$

We repeat our defining procedure. We have partitioned the contribution to the total magnetic field inside the sample into two parts: *one*, due to external factors such as the current in the solenoid. This is represented by H . The *other* is due to the specific nature of the magnetic material, namely M . The latter quantity can be influenced by external factors. This influence is mathematically expressed as

$$M = \chi H \quad (5.17)$$

where χ , a dimensionless quantity, is appropriately called the *magnetic susceptibility*. It is a measure of how a magnetic material responds to an external field. Table 5.2 lists χ for some elements. It is small and positive for materials, which are called *paramagnetic*. It is small and negative for materials, which are termed *diamagnetic*. In the latter case M and H are opposite in direction. From Eqs. (5.16) and (5.17) we obtain,

$$B = \mu_0 (1 + \chi) H \quad (5.18)$$

$$\begin{aligned} &= \mu_0 \mu_r H \\ &= \mu H \end{aligned} \quad (5.19)$$

where $\mu_r = 1 + \chi$, is a dimensionless quantity called the *relative magnetic permeability* of the substance. It is the analog of the dielectric constant in electrostatics. The *magnetic permeability* of the substance is μ and it has the same dimensions and units as μ_0 :

$$\mu = \mu_0 \mu_r = \mu_0 (1 + \chi).$$

The three quantities χ , μ_r and μ are interrelated and only one of them is independent. Given one, the other two may be easily determined.

Magnetism and Matter

TABLE 5.2 MAGNETIC SUSCEPTIBILITY OF SOME ELEMENTS AT 300 K

Diamagnetic substance	χ	Paramagnetic substance	χ
Bismuth		Aluminium	2.3×10^{-9}
Copper		Calcium	1.9×10^{-9}
Diamond		Chromium	2.7×10^{-9}
Gold		Lithium	2.1×10^{-9}
Lead		Magnesium	1.2×10^{-9}
Mercury		Niobium	2.6×10^{-9}
Nitrogen (STP)		Oxygen (STP)	2.1×10^{-9}
Silver		Platinum	2.9×10^{-9}
Silicon		Tungsten	6.8×10^{-9}

Example 5.10 A solenoid has a core of a material with relative permeability 400. The windings of the solenoid are insulated from the core and carry a current of 2A. If the number of turns is 1000 per metre, calculate (a) H , (b) M , (c) B and (d) the magnetising current I_m .

Solution

(a) The field H is dependent of the material of the core, and is

$$H = nl = 1000 \times 2.0 = 2 \times 10^3 \text{ A/m.}$$

(b) The magnetic field B is given by

$$\begin{aligned} B &= \mu_r \mu_0 H \\ &= 400 \times 4\pi \times 10^{-7} \text{ N/A}^2 \times 2 \times 10^3 \text{ A/m} \\ &= 1.0 \text{ T} \end{aligned}$$

(c) Magnetisation is given by

$$\begin{aligned} M &= (B - \mu_0 H) / \mu_0 \\ &= (\mu_r \mu_0 H - \mu_0 H) / \mu_0 = \mu_r H = 399 \times H \\ &\approx 8 \times 10^5 \text{ A/m} \end{aligned}$$

(d) The magnetising current I_m is the additional current that needs to be passed through the windings of the solenoid in the absence of the core which would give a B value as in the presence of the core. Thus $B = \mu_r n_0 (I + I_m)$. Using $I = 2\text{A}$, $B = 1\text{T}$, we get $I_m = 794\text{A}$.

EXAMPLE 5.10

5.6 MAGNETIC PROPERTIES OF MATERIALS

The discussion in the previous section helps us to classify materials as diamagnetic, paramagnetic or ferromagnetic. In terms of the susceptibility χ , a material is diamagnetic if χ is negative, para- if χ is positive and small, and ferro- if χ is large and positive.

A glance at Table 5.3 gives one a better feeling for these materials. Here ε is a small positive number introduced to quantify paramagnetic materials. Next, we describe these materials in some detail.

TABLE 5.3

Diamagnetic	Paramagnetic	Ferromagnetic
$\leq \chi < 0$	$0 < \chi < \varepsilon$	$\chi \gg 1$
$0 \leq \mu_r < 1$	$1 < \mu_r < 1 + \varepsilon$	$\mu_r \gg 1$
$\mu < \mu_0$	$\mu > \mu_0$	$\mu \gg \mu_0$

5.6.1 Diamagnetism

Diamagnetic substances are those which have tendency to move from stronger to the weaker part of the external magnetic field. In other words, unlike the way a magnet attracts metals like iron, it would repel a diamagnetic substance.

Figure 5.12(a) shows a bar of diamagnetic material placed in an external magnetic field. The field lines are repelled or expelled and the field inside the material is reduced. In most cases, as is evident from Table 5.2, this reduction is slight, being one part in 10^5 . When placed in a non-uniform magnetic field, the bar will tend to move from high to low field.

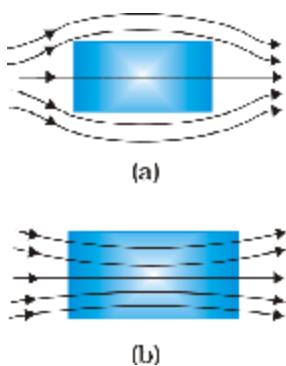


FIGURE 5.12
Behaviour of
magnetic field lines
near a
(a) diamagnetic,
(b) paramagnetic
substance.

The simplest explanation for diamagnetism is as follows. Electrons in an atom orbiting around nucleus possess orbital angular momentum. These orbiting electrons are equivalent to current-carrying loop and thus possess orbital magnetic moment. Diamagnetic substances are the ones in which resultant magnetic moment in an atom is zero. When magnetic field is applied, those electrons having orbital magnetic moment in the same direction slow down and those in the opposite direction speed up. This happens due to induced current in accordance with Lenz you will study in Chapter 6. Thus, the substance develops a net magnetic moment in direction opposite to that of the applied field and hence repulsion.

Some diamagnetic materials are bismuth, copper, lead, silicon, nitrogen (at STP), water and sodium chloride. Diamagnetism is present in all the substances. However, the effect is so weak in most cases that it gets shifted by other effects like paramagnetism, ferromagnetism, etc.

The most exotic diamagnetic materials are *superconductors*. These are metals, cooled to very low temperatures which exhibits both *perfect conductivity* and *perfect diamagnetism*. Here the field lines are completely expelled! $\chi = -\mu_r = 0$. A superconductor repels a magnet and (by Newton

diamagnetism in superconductors is called the *Meissner effect*, after the name of its discoverer. Superconducting magnets can be gainfully exploited in variety of situations, for example, for running magnetically levitated superfast trains.

5.6.2 Paramagnetism

Paramagnetic substances are those which get weakly magnetised when placed in an external magnetic field. They have tendency to move from a region of weak magnetic field to strong magnetic field, i.e., they get weakly attracted to a magnet.

Magnetism and Matter

The individual atoms (or ions or molecules) of a paramagnetic material possess a permanent magnetic dipole moment of their own. On account of the ceaseless random thermal motion of the atoms, no net magnetisation is seen. In the presence of an external field B_0 , which is strong enough, and at low temperatures, the individual atomic dipole moment can be made to align and point in the same direction as B_0 . Figure 5.12(b) shows a bar of paramagnetic material placed in an external field. The field lines get concentrated inside the material, and the field inside is enhanced. In most cases, as is evident from Table 5.2, this enhancement is slight, being one part in 10^5 . When placed in a non-uniform magnetic field, the bar will tend to move from weak field to strong.

Some paramagnetic materials are aluminium, sodium, calcium, oxygen (at STP) and copper chloride. Experimentally, one finds that the magnetisation of a paramagnetic material is inversely proportional to the absolute temperature T ,

$$M = C \frac{B_0}{T} \quad [5.20(a)]$$

or equivalently, using Eqs. (5.12) and (5.17)

$$\chi = C \frac{\mu_0}{T} \quad [5.20(b)]$$

This is known as *Curie* , after its discoverer Pierre Curie (1859-1906). The constant C is called *Curie* . Thus, for a paramagnetic material both χ and μ_r depend not only on the material, but also (in a simple fashion) on the sample temperature. As the field is increased or the temperature is lowered, the magnetisation increases until it reaches the saturation value M_s , at which point all the dipoles are perfectly aligned with the field. Beyond this, Curie longer valid.

5.6.3 Ferromagnetism

Ferromagnetic substances are those which get strongly magnetised when placed in an external magnetic field. They have strong tendency to move from a region of weak magnetic field to strong magnetic field, i.e., they get strongly attracted to a magnet.

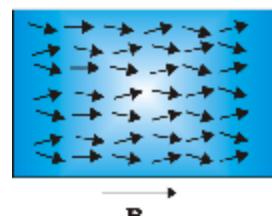
The individual atoms (or ions or molecules) in a ferromagnetic material possess a dipole moment as in a paramagnetic material. However, they interact with one another in such a way that they spontaneously align themselves in a common direction over a macroscopic volume called *domain*. The explanation of this cooperative effect requires quantum mechanics and is beyond the scope of this textbook. Each domain has a net magnetisation. Typical domain size is 1 mm and the domain contains about 10^{11} atoms. In the first instant, the magnetisation varies randomly from domain to domain and there is no bulk magnetisation. This is shown in Fig. 5.13(a). When we apply an external magnetic field B_0 , the domains orient themselves in the direction of B_0 and simultaneously the domain oriented in the direction of B_0 grow in size. This existence of domains and their motion in B_0 are not speculations. One may observe this under a microscope after sprinkling a liquid suspension of powdered



Magnetic materials, domain, etc.:
<http://www.ndt-ed.org/EducationResources/CommunityCollege/MagParticlePhysics/MagneticMats.htm>



(a)



(b)

FIGURE 5.13
(a) Randomly oriented domains,
(b) Aligned domains.

Physics

ferromagnetic substance of samples. This motion of suspension can be observed. Figure 5.12(b) shows the situation when the domains have aligned and amalgamated to form a single

Thus, in a ferromagnetic material the field lines are highly concentrated. In non-uniform magnetic field, the sample tends to move towards the region of high field. We may wonder as to what happens when the external field is removed. In some ferromagnetic materials the magnetisation persists. Such materials are called *hard* magnetic materials or *hard ferromagnets*. Alnico, an alloy of iron, aluminium, nickel, cobalt and copper, is one such material. The naturally occurring lodestone is another. Such materials form permanent magnets to be used among other things as a compass needle. On the other hand, there is a class of ferromagnetic materials in which the magnetisation disappears on removal of the external field. Soft iron is one such material. Appropriately enough, such materials are called *soft ferromagnetic materials*. There are a number of elements, which are ferromagnetic: iron, cobalt, nickel, gadolinium, etc. The relative magnetic permeability is >1000 !

The ferromagnetic property depends on temperature. At high enough temperature, a ferromagnet becomes a paramagnet. The domain structure disintegrates with temperature. This disappearance of magnetisation with temperature is gradual. It is a phase transition reminding us of the melting of a solid crystal. The temperature of transition from ferromagnetic to paramagnetism is called the *Curie temperature* T_c . Table 5.4 lists the Curie temperature of certain ferromagnets. The susceptibility above the Curie temperature, i.e., in the paramagnetic phase is described by,

$$\chi = \frac{C}{T - T_c} \quad (T > T_c) \quad (5.21)$$

TABLE 5.4 CURIE TEMPERATURE T_c OF SOME FERROMAGNETIC MATERIALS

Material	T_c (K)
Cobalt	1394
Iron	1043
Fe_2O_3	893
Nickel	631
Gadolinium	317

EXAMPLE 5.11

Example 5.11 A domain in ferromagnetic iron is in the form of a cube of side length $1\mu\text{m}$. Estimate the number of iron atoms in the domain and the maximum possible dipole moment and magnetisation of the domain. The molecular mass of iron is 55 g/mole and its density is 7.9 g/cm^3 . Assume that each iron atom has a dipole moment of $9.27 \times 10^{-29} \text{ A m}^2$.



EXAMPLE 5.11

Solution The volume of the cubic domain is

$$V = (10 \text{ m})^3 = 10 \text{ m}^3 = 10 \text{ cm}^3$$

Its mass is volume \times density $= 7.9 \text{ g cm}^{-3} \times 10 \text{ cm}^3 = 7.9 \times 10 \text{ g}$

It is given that Avagadro number (6.023×10^{23}) of iron atoms have a mass of 55 g. Hence, the number of atoms in the domain is

$$N = \frac{7.9 \times 10^{-12} \times 6.023 \times 10^{23}}{55}$$

$$= 8.65 \times 10^{10} \text{ atoms}$$

The maximum possible dipole moment m_{\max} is achieved for the (unrealistic) case when all the atomic moments are perfectly aligned.

Thus,

$$m_{\max} = (8.65 \times 10^{10}) \times (9.27 \times 10^{-27})$$

$$= 8.0 \times 10^{-17} \text{ A m}^2$$

The consequent magnetisation is

$$M_{\max} = m_{\max} / \text{Domain volume}$$

$$= 8.0 \times 10^{-17} \text{ Am}^2 / 10 \text{ cm}^3$$

$$= 8.0 \times 10^5 \text{ Am}$$

The relation between B and H in ferromagnetic materials is complex. It is often not linear and it depends on the magnetic history of the sample. Figure 5.14 depicts the behaviour of the material as we take it through one cycle of magnetisation. Let the material be unmagnetised initially. We place it in a solenoid and increase the current through the solenoid. The magnetic field B in the material rises and saturates as depicted in the curve Oa. This behaviour represents the alignment and merger of domains until no further enhancement is possible. It is pointless to increase the current (and hence the magnetic intensity H) beyond this. Next, we decrease H and reduce it to zero. At $H = 0$, $B \neq 0$. This is represented by the curve ab. The value of B at $H = 0$ is called *retentivity* or *remanence*. In Fig. 5.14, $B_R \sim 1.2 \text{ T}$, where the subscript R denotes retentivity. The domains are not completely randomised even though the external driving field has been removed. Next, the current in the solenoid is reversed and slowly increased. Certain domains are flipped until the net field inside stands nullified. This is represented by the curve bc. The value of H at c is called *coercivity*. In Fig. 5.14 $H_c \sim 100 \text{ A/m}$. As the reversed current is increased in magnitude, we once again obtain saturation. The curve cd depicts this. The saturated magnetic field $B_s \sim 1.5 \text{ T}$. Next, the current is reduced (curve de) and reversed (curve ea). The cycle repeats itself. Note that the curve Oa does not retrace itself as H is reduced. For a given value of H , B is not unique but depends on previous history of the sample. This phenomenon is called *hysteresis*. The word *hysteresis* means *lagging behind* (and not

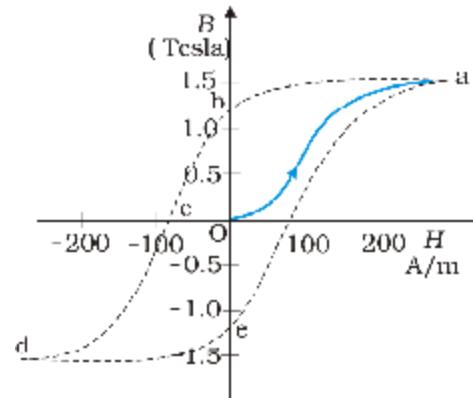


FIGURE 5.14 The magnetic hysteresis loop is the B-H curve for ferromagnetic materials.

5.7 PERMANENT MAGNETS AND ELECTROMAGNETS

Substances which at room temperature retain their ferromagnetic property for a long period of time are called *permanent magnets*. Permanent

Physics



FIGURE 5.15 A blacksmith forging a permanent magnet by striking a red-hot rod of iron kept in the north-south direction with a hammer. The sketch is recreated from an illustration in *De Magnete*, a work published in 1600 and authored by William Gilbert, the court physician to Queen Elizabeth of England.

magnets can be made in a variety of ways. One can hold an iron rod in the north-south direction and hammer it repeatedly. The method is illustrated in Fig. 5.15. The illustration is from a 400 year old book to emphasise that the making of permanent magnets is an *old art*. One can also hold a steel rod and stroke it with one end of a bar magnet a large number of times, always in the same sense to make a permanent magnet.

An efficient way to make a permanent magnet is to place a ferromagnetic rod in a solenoid and pass a current. The magnetic field of the solenoid magnetises the rod.

The hysteresis curve (Fig. 5.14) allows us to select suitable materials for permanent magnets. The material should have high retentivity so that the magnet is strong and high coercivity so that the magnetisation is not erased by stray magnetic fields, temperature fluctuations or minor mechanical damage. Further, the material should have a high permeability. Steel is one-favoured choice. It has a slightly smaller retentivity than soft iron but this is outweighed by the much smaller coercivity of soft iron. Other suitable materials for permanent magnets are alnico, cobalt steel and ticonal.

Core of electromagnets are made of ferromagnetic materials which have high permeability and low retentivity. Soft iron is a suitable material for electromagnets. On placing a soft iron rod in a solenoid and passing a current, we increase the magnetism of the solenoid by a thousand fold. When we switch off the solenoid current, the magnetism is effectively switched off since the soft iron core has a low retentivity. The arrangement is shown in Fig. 5.16.

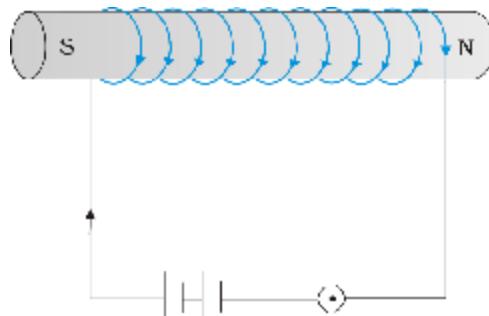


FIGURE 5.16 A soft iron core in solenoid acts as an electromagnet.

In certain applications, the material goes through an ac cycle of magnetisation for a long period. This is the case in transformer cores and telephone diaphragms. The hysteresis curve of such materials must be narrow. The energy dissipated and the heating will consequently be small. The material must have a high resistivity to lower eddy current losses. You will study about eddy currents in Chapter 6.

Electromagnets are used in electric bells, loudspeakers and telephone diaphragms. Giant electromagnets are used in cranes to lift machinery, and bulk quantities of iron and steel.

MAPPING INDIA'S MAGNETIC FIELD

Because of its practical application in prospecting, communication, and navigation, the magnetic field of the earth is mapped by most nations with an accuracy comparable to geographical mapping. In India over a dozen observatories exist, extending from Trivandrum (now Thiruvananthapuram) in the south to Gulmarg in the north. These observatories work under the aegis of the Indian Institute of Geomagnetism (IIG), in Colaba, Mumbai. The IIG grew out of the Colaba and Alibag observatories and was formally established in 1971. The IIG monitors (via its nation-wide observatories), the geomagnetic fields and fluctuations on land, and under the ocean and in space. Its services are used by the Oil and Natural Gas Corporation Ltd. (ONGC), the National Institute of Oceanography (NIO) and the Indian Space Research Organisation (ISRO). It is a part of the world-wide network which ceaselessly updates the geomagnetic data. Now India has a permanent station called Gangotri.

SUMMARY

1. The science of magnetism is old. It has been known since ancient times that magnetic materials tend to point in the north-south direction; like magnetic poles repel and unlike ones attract; and cutting a bar magnet in two leads to two smaller magnets. Magnetic poles cannot be isolated.
2. When a bar magnet of dipole moment m is placed in a uniform magnetic field B ,
 - (a) the force on it is zero,
 - (b) the torque on it is $m \times B$,
 - (c) its potential energy is $\cdot B$, where we choose the zero of energy at the orientation when m is perpendicular to B .
3. Consider a bar magnet of size l and magnetic moment m , at a distance r from its mid-point, where $r \gg l$, the magnetic field B due to this bar is,

$$B = \frac{\mu_0 m}{2\pi r^3} \quad (\text{along axis})$$

$$= \frac{\mu_0 m}{4\pi r^3} \quad (\text{along equator})$$

4. Gauss
any closed surface is zero

$$\phi_B = \sum_{\substack{\text{all area} \\ \text{elements } \Delta S}} B \cdot \Delta S = 0$$

5. The earth
dipole located at the centre of the earth. The pole near the geographic north pole of the earth is called the north magnetic pole. Similarly, the pole near the geographic south pole is called the south magnetic pole. This dipole is aligned making a small angle with the rotation axis of the earth. The magnitude of the field on the earth

$$\approx 4 \times 10^{-5} \text{ T.}$$

Physics

6. Three quantities are needed to specify the magnetic field of the earth on its surface and the magnetic dip. These are known as the elements of the earth magnetic field.
 7. Consider a material placed in an external magnetic field B_0 . The magnetic intensity is defined as,
- $$H = \frac{B_0}{\mu_0}$$
- The magnetisation M of the material is its dipole moment per unit volume. The magnetic field B in the material is,
- $$B = \mu_0 (H + M)$$
8. For a linear material $M = \chi H$. So that $B = \mu H$ and χ is called the magnetic susceptibility of the material. The three quantities, χ , the relative magnetic permeability μ_r , and the magnetic permeability μ are related as follows:
- $$\mu = \mu_0 \mu_r$$
- $$\mu_r = 1 + \chi$$
9. Magnetic materials are broadly classified as: diamagnetic, paramagnetic, and ferromagnetic. For diamagnetic materials χ is negative and small and for paramagnetic materials it is positive and small. Ferromagnetic materials have large χ and are characterised by non-linear relation between B and H . They show the property of hysteresis.
 10. Substances, which at room temperature, retain their ferromagnetic property for a long period of time are called permanent magnets.

Physical quantity	Symbol	Nature	Dimensions	Units	Remarks
Permeability of free space	μ_0	Scalar	[MLT ⁻¹ A]	T m A	$\mu_0/4\pi = 10^{-7}$
Magnetic field, Magnetic induction, Magnetic flux density	B	Vector	[MT ⁻¹ A]	T (tesla)	10^4 G (gauss) = 1 T
Magnetic moment	m	Vector	[L ² A]	A m ²	
Magnetic flux	ϕ_B	Scalar	[ML ² T ⁻¹ A]	W (weber)	$W = T m^2$
Magnetisation	M	Vector	[L ⁻¹ A]	A m	$\frac{\text{Magnetic moment}}{\text{Volume}}$
Magnetic intensity Magnetic field strength	H	Vector	[L ⁻¹ A]	A m	$B = \mu_0 (H + M)$
Magnetic susceptibility	χ	Scalar	-	-	$M = \chi H$
Relative magnetic permeability	μ_r	Scalar	-	-	$B = \mu_0 \mu_r H$
Magnetic permeability	μ	Scalar	[MLT ⁻¹ A]	T m A N A	$\mu = \mu_0 \mu_r$ $B = \mu H$

POINTS TO PONDER

1. A satisfactory understanding of magnetic phenomenon in terms of moving charges/currents was arrived at after 1800 AD. But technological exploitation of the directional properties of magnets predates this scientific understanding by two thousand years. Thus, scientific understanding is not a necessary condition for engineering applications. Ideally, science and engineering go hand-in-hand, one leading and assisting the other in tandem.
2. Magnetic monopoles do not exist. If you slice a magnet in half, you get two smaller magnets. On the other hand, isolated positive and negative charges exist. There exists a smallest unit of charge, for example, the electronic charge with value $|e| = 1.6 \times 10^{-19}$ C. All other charges are integral multiples of this smallest unit charge. In other words, charge is quantised. We do not know why magnetic monopoles do not exist or why electric charge is quantised.
3. A consequence of the fact that magnetic monopoles do not exist is that the magnetic field lines are continuous and form closed loops. In contrast, the electrostatic lines of force begin on a positive charge and terminate on the negative charge (or fade out at infinity).
4. The earth
earth
are responsible for the earth
sustains this current, and why the earth
million years or so, we do not know.
5. A minuscule difference in the value of χ , the magnetic susceptibility, yields radically different behaviour: diamagnetic versus paramagnetic. For diamagnetic materials $\chi = -10^{-6}$ whereas $\chi = +10^{-3}$ for paramagnetic materials.
6. There exists a perfect diamagnet, namely, a superconductor. This is a metal at very low temperatures. In this case $\chi = -1$, $\mu_r = 0$, $\mu = 0$. The external magnetic field is totally expelled. Interestingly, this material is also a perfect conductor. However, there exists no classical theory which ties these two properties together. A quantum-mechanical theory by Bardeen, Cooper, and Schrieffer (BCS theory) explains these effects. The BCS theory was proposed in 1957 and was eventually recognised by a Nobel Prize in 1970.
7. The phenomenon of magnetic hysteresis is reminiscent of similar behaviour concerning the elastic properties of materials. Strain may not be proportional to stress; here H and B (or M) are not linearly related. The stress-strain curve exhibits hysteresis and area enclosed by it represents the energy dissipated per unit volume. A similar interpretation can be given to the B - H magnetic hysteresis curve.
8. Diamagnetism is universal. It is present in all materials. But it is weak and hard to detect if the substance is para- or ferromagnetic.
9. We have classified materials as diamagnetic, paramagnetic, and ferromagnetic. However, there exist additional types of magnetic material such as ferrimagnetic, anti-ferromagnetic, spin glass, etc. with properties which are exotic and mysterious.

EXERCISES

- 5.1 Answer the following questions regarding earth
- (a) A vector needs three quantities for its specification. Name the three independent quantities conventionally used to specify the earth
 - (b) The angle of dip at a location in southern India is about 18° . Would you expect a greater or smaller dip angle in Britain?
 - (c) If you made a map of magnetic field lines at Melbourne in Australia, would the lines seem to go into the ground or come out of the ground?
 - (d) In which direction would a compass free to move in the vertical plane point to, if located right on the geomagnetic north or south pole?
 - (e) The earth due to a dipole of magnetic moment $8 \times 10^{22} \text{ J T}$ located at its centre. Check the order of magnitude of this number in some way.
 - (f) Geologists claim that besides the main magnetic N-S poles, there are several local poles on the earth directions. How is such a thing possible at all?
- 5.2 Answer the following questions:
- (a) The earth Does it also change with time? If so, on what time scale does it change appreciably?
 - (b) The earth regard this as a source of the earth
 - (c) The charged currents in the outer conducting regions of the earth What might be the these currents?
 - (d) The earth may have even reversed the direction of its field several times during its history of 4 to 5 billion years. How can geologists know about the earth
 - (e) The earth large distances (greater than about 30,000 km). What agencies may be responsible for this distortion?
 - (f) Interstellar space has an extremely weak magnetic field of the order of 10^{-10} T . Can such a weak field be of any significant consequence? Explain.
- [Note: Exercise 5.2 is meant mainly to arouse your curiosity. Answers to some questions above are tentative or unknown. Brief answers wherever possible are given at the end. For details, you should consult a good text on geomagnetism.]
- 5.3 A short bar magnet placed with its axis at 30° with a uniform external magnetic field of 0.25 T experiences a torque of magnitude equal to $4.5 \times 10^{-3} \text{ J}$. What is the magnitude of magnetic moment of the magnet?
- 5.4 A short bar magnet of magnetic moment $m = 0.32 \text{ JT}^{-1}$ is placed in a uniform magnetic field of 0.15 T . If the bar is free to rotate in the plane of the field, which orientation would correspond to its (a) stable, and (b) unstable equilibrium? What is the potential energy of the magnet in each case?

Magnetism and Matter

- 5.5 A closely wound solenoid of 800 turns and area of cross section $2.5 \times 10^{-2} \text{ m}^2$ carries a current of 3.0 A. Explain the sense in which the solenoid acts like a bar magnet. What is its associated magnetic moment?
- 5.6 If the solenoid in Exercise 5.5 is free to turn about the vertical direction and a uniform horizontal magnetic field of 0.25 T is applied, what is the magnitude of torque on the solenoid when its axis makes an angle of 30° with the direction of applied field?
- 5.7 A bar magnet of magnetic moment 1.5 J T^{-1} lies aligned with the direction of a uniform magnetic field of 0.22 T.
- What is the amount of work required by an external torque to turn the magnet so as to align its magnetic moment: (i) normal to the field direction, (ii) opposite to the field direction?
 - What is the torque on the magnet in cases (i) and (ii)?
- 5.8 A closely wound solenoid of 2000 turns and area of cross-section $1.6 \times 10^{-2} \text{ m}^2$, carrying a current of 4.0 A, is suspended through its centre allowing it to turn in a horizontal plane.
- What is the magnetic moment associated with the solenoid?
 - What is the force and torque on the solenoid if a uniform horizontal magnetic field of $7.5 \times 10^{-2} \text{ T}$ is set up at an angle of 30° with the axis of the solenoid?
- 5.9 A circular coil of 16 turns and radius 10 cm carrying a current of 0.75 A rests with its plane normal to an external field of magnitude $5.0 \times 10^{-2} \text{ T}$. The coil is free to turn about an axis in its plane perpendicular to the field direction. When the coil is turned slightly and released, it oscillates about its stable equilibrium with a frequency of 2.0 s^{-1} . What is the moment of inertia of the coil about its axis of rotation?
- 5.10 A magnetic needle free to rotate in a vertical plane parallel to the magnetic meridian has its north tip pointing down at 22° with the horizontal. The horizontal component of the earth at the place is known to be 0.35 G. Determine the magnitude of the earth
- 5.11 At a certain location in Africa, a compass points 12° west of the geographic north. The north tip of the magnetic needle of a dip circle placed in the plane of magnetic meridian points 60° above the horizontal. The horizontal component of the earth to be 0.16 G. Specify the direction and magnitude of the earth at the location.
- 5.12 A short bar magnet has a magnetic moment of 0.48 J T^{-1} . Give the direction and magnitude of the magnetic field produced by the magnet at a distance of 10 cm from the centre of the magnet on (a) the axis, (b) the equatorial lines (normal bisector) of the magnet.
- 5.13 A short bar magnet placed in a horizontal plane has its axis aligned along the magnetic north-south direction. Null points are found on the axis of the magnet at 14 cm from the centre of the magnet. The earth zero. What is the total magnetic field on the normal bisector of the magnet at the same distance as the null centre of the magnet? (At *null points*, field due to a magnet is equal and opposite to the horizontal component of earth
- 5.14 If the bar magnet in exercise 5.13 is turned around by 180° , where will the new null points be located?

■ Physics

- 5.15 A short bar magnet of magnetic movement 5.25×10^{-6} JT⁻¹ is placed with its axis perpendicular to the earth distance from the centre of the magnet, the resultant field is inclined at 45° with earth
Magnitude of the earth
Ignore the length of the magnet in comparison to the distances involved.

ADDITIONAL EXERCISES

- 5.16 Answer the following questions:
- Why does a paramagnetic sample display greater magnetisation (for the same magnetising field) when cooled?
 - Why is diamagnetism, in contrast, almost independent of temperature?
 - If a toroid uses bismuth for its core, will the field in the core be (slightly) greater or (slightly) less than when the core is empty?
 - Is the permeability of a ferromagnetic material independent of the magnetic field? If not, is it more for lower or higher fields?
 - Magnetic field lines are always nearly normal to the surface of a ferromagnet at every point. (This fact is analogous to the static electric field lines being normal to the surface of a conductor at every point.) Why?
 - Would the maximum possible magnetisation of a paramagnetic sample be of the same order of magnitude as the magnetisation of a ferromagnet?
- 5.17 Answer the following questions:
- Explain qualitatively on the basis of domain picture the irreversibility in the magnetisation curve of a ferromagnet.
 - The hysteresis loop of a soft iron piece has a much smaller area than that of a carbon steel piece. If the material is to go through repeated cycles of magnetisation, which piece will dissipate greater heat energy?
 - a device for storing memory?
statement.
 - What kind of ferromagnetic material is used for coating magnetic tapes in a cassette player, or for building modern computer?
 - A certain region of space is to be shielded from magnetic fields.
Suggest a method.
- 5.18 A long straight horizontal cable carries a current of 2.5 A in the direction 10° south of west to 10° north of east. The magnetic meridian of the place happens to be 10° west of the geographic meridian. The earth is zero. Locate the line of neutral points (ignore the thickness of the cable)? (At *neutral points*, magnetic field due to a current-carrying cable is equal and opposite to the horizontal component of earth magnetic field.)
- 5.19 A telephone cable at a place has four long straight horizontal wires carrying a current of 1.0 A in the same direction east to west. The

Magnetism and Matter

earth

35° . The magnetic declination is nearly zero. What are the resultant magnetic fields at points 4.0 cm below the cable?

- 5.20 A compass needle free to turn in a horizontal plane is placed at the centre of circular coil of 30 turns and radius 12 cm. The coil is in a vertical plane making an angle of 45° with the magnetic meridian. When the current in the coil is 0.35 A, the needle points west to east.
- Determine the horizontal component of the earth at the location.
 - The current in the coil is reversed, and the coil is rotated about its vertical axis by an angle of 90° in the anticlockwise sense looking from above. Predict the direction of the needle. Take the magnetic declination at the places to be zero.
- 5.21 A magnetic dipole is under the influence of two magnetic fields. The angle between the field directions is 60° , and one of the fields has a magnitude of 1.2×10^{-6} T. If the dipole comes to stable equilibrium at an angle of 15° with this field, what is the magnitude of the other field?
- 5.22 A monoenergetic (18 keV) electron beam initially in the horizontal direction is subjected to a horizontal magnetic field of 0.04 G normal to the initial direction. Estimate the up or down deflection of the beam over a distance of 30 cm ($m_e = 9.11 \times 10^{-31}$ kg). [Note: Data in this exercise are so chosen that the answer will give you an idea of the effect of earth from the electron gun to the screen in a TV set.]
- 5.23 A sample of paramagnetic salt contains 2.0×10^{24} atomic dipoles each of dipole moment 1.5×10^{-10} J T. The sample is placed under a homogeneous magnetic field of 0.64 T, and cooled to a temperature of 4.2 K. The degree of magnetic saturation achieved is equal to 15%. What is the total dipole moment of the sample for a magnetic field of 0.98 T and a temperature of 2.8 K? (Assume Curie
- 5.24 A Rowland ring of mean radius 15 cm has 3500 turns of wire wound on a ferromagnetic core of relative permeability 800. What is the magnetic field B in the core for a magnetising current of 1.2 A?
- 5.25 The magnetic moment vectors μ_s and μ_l associated with the intrinsic spin angular momentum S and orbital angular momentum I, respectively, of an electron are predicted by quantum theory (and verified experimentally to a high accuracy) to be given by:
$$\mu_s = e/m) S,$$

$$\mu_l = e/2m)I$$
- Which of these relations is in accordance with the result expected classically? Outline the derivation of the classical result.

Chapter Six

ELECTROMAGNETIC INDUCTION



6.1 INTRODUCTION

Electricity and magnetism were considered separate and unrelated phenomena for a long time. In the early decades of the nineteenth century, experiments on electric current by Oersted, Ampere and a few others established the fact that electricity and magnetism are inter-related. They found that moving electric charges produce magnetic fields. For example, an electric current deflects a magnetic compass needle placed in its vicinity. This naturally raises the questions like: Is the converse effect possible? Can moving magnets produce electric currents? Does the nature permit such a relation between electricity and magnetism? The answer is resounding yes! The experiments of Michael Faraday in England and Joseph Henry in USA, conducted around 1830, demonstrated conclusively that electric currents were induced in closed coils when subjected to changing magnetic fields. In this chapter, we will study the phenomena associated with changing magnetic fields and understand the underlying principles. The phenomenon in which electric current is generated by varying magnetic fields is appropriately called *electromagnetic induction*.

When Faraday first made public his discovery that relative motion between a bar magnet and a wire loop produced a small current in the latter, he was asked, "What is the use of it?" His reply was: "What is the use of a new born baby?" The phenomenon of electromagnetic induction

is not merely of theoretical or academic interest but also of practical utility. Imagine a world where there is no electricity – no electric lights, no trains, no telephones and no personal computers. The pioneering experiments of Faraday and Henry have led directly to the development of modern day generators and transformers. Today's civilisation owes its progress to a great extent to the discovery of electromagnetic induction.

6.2 THE EXPERIMENTS OF FARADAY AND HENRY

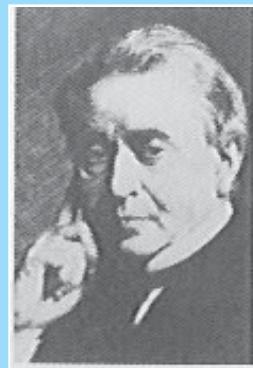
The discovery and understanding of electromagnetic induction are based on a long series of experiments carried out by Faraday and Henry. We shall now describe some of these experiments.

Experiment 6.1

Figure 6.1 shows a coil C_1 * connected to a galvanometer G . When the North-pole of a bar magnet is pushed towards the coil, the pointer in the galvanometer deflects, indicating the presence of electric current in the coil. The deflection lasts as long as the bar magnet is in motion. The galvanometer does not show any deflection when the magnet is held stationary. When the magnet is pulled away from the coil, the galvanometer shows deflection in the opposite direction, which indicates reversal of the current's direction. Moreover, when the South-pole of the bar magnet is moved towards or away from the coil, the deflections in the galvanometer are opposite to that observed with the North-pole for similar movements. Further, the deflection (and hence current) is found to be larger when the magnet is pushed towards or pulled away from the coil faster. Instead, when the bar magnet is held fixed and the coil C_1 is moved towards or away from the magnet, the same effects are observed. It shows that *it is the relative motion between the magnet and the coil that is responsible for generation (induction) of electric current in the coil*.

Experiment 6.2

In Fig. 6.2 the bar magnet is replaced by a second coil C_2 connected to a battery. The steady current in the coil C_2 produces a steady magnetic field. As coil C_2 is



Joseph Henry [1797 – 1878] American experimental physicist professor at Princeton University and first director of the Smithsonian Institution. He made important improvements in electromagnets by winding coils of insulated wire around iron pole pieces and invented an electromagnetic motor and a new, efficient telegraph. He discovered self-induction and investigated how currents in one circuit induce currents in another.

JOSEPH HENRY (1797 – 1878)

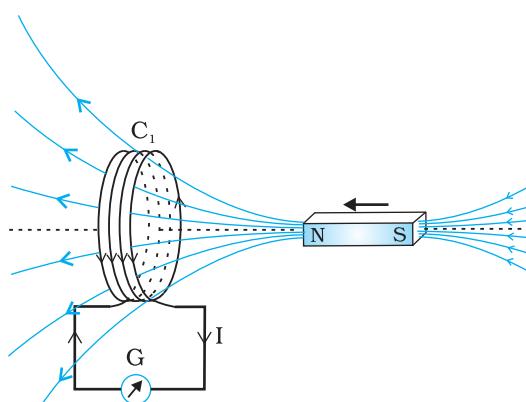


FIGURE 6.1 When the bar magnet is pushed towards the coil, the pointer in the galvanometer G deflects.

* Wherever the term 'coil or 'loop' is used, it is assumed that they are made up of conducting material and are prepared using wires which are coated with insulating material.

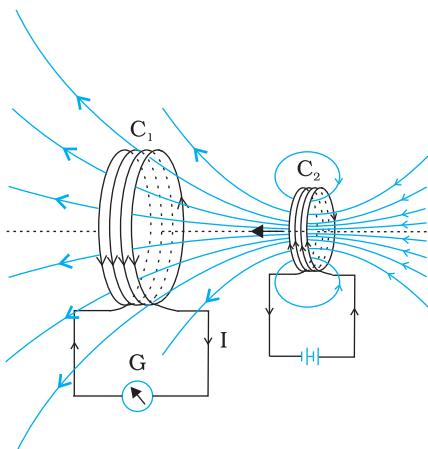


FIGURE 6.2 Current is induced in coil C_1 due to motion of the current carrying coil C_2 .

Interactive animation on Faraday's experiments and Lenz's law:
<http://micro.magnet.fsu.edu/electromagnet/java/faraday>



moved towards the coil C_1 , the galvanometer shows a deflection. This indicates that electric current is induced in coil C_1 . When C_2 is moved away, the galvanometer shows a deflection again, but this time in the opposite direction. The deflection lasts as long as coil C_2 is in motion. When the coil C_2 is held fixed and C_1 is moved, the same effects are observed. Again, it is the relative motion between the coils that induces the electric current.

Experiment 6.3

The above two experiments involved relative motion between a magnet and a coil and between two coils, respectively. Through another experiment, Faraday showed that this relative motion is not an absolute requirement. Figure 6.3 shows two coils C_1 and C_2 held stationary. Coil C_1 is connected to galvanometer G while the second coil C_2 is connected to a battery through a tapping key K .

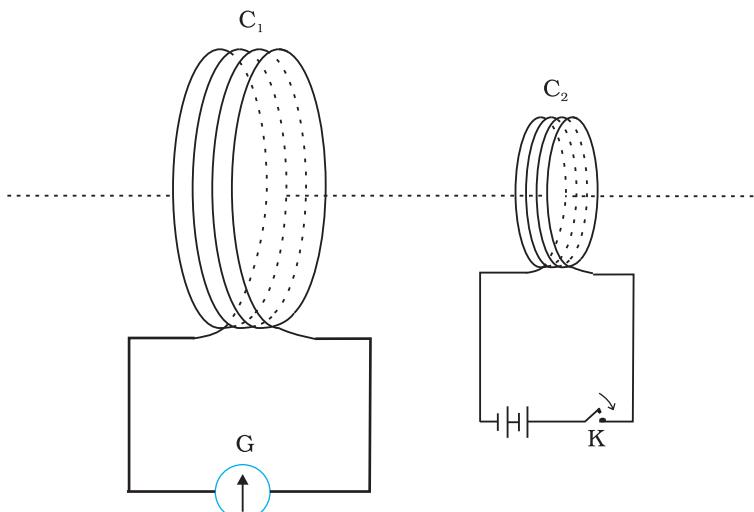


FIGURE 6.3 Experimental set-up for Experiment 6.3.

It is observed that the galvanometer shows a momentary deflection when the tapping key K is pressed. The pointer in the galvanometer returns to zero immediately. If the key is held pressed continuously, there is no deflection in the galvanometer. When the key is released, a momentary deflection is observed again, but in the opposite direction. It is also observed that the deflection increases dramatically when an iron rod is inserted into the coils along their axis.

6.3 MAGNETIC FLUX

Faraday's great insight lay in discovering a simple mathematical relation to explain the series of experiments he carried out on electromagnetic induction. However, before we state and appreciate his laws, we must get familiar with the notion of magnetic flux, Φ_B . Magnetic flux is defined in the same way as electric flux is defined in Chapter 1. Magnetic flux through

a plane of area A placed in a uniform magnetic field \mathbf{B} (Fig. 6.4) can be written as

$$\Phi_B = \mathbf{B} \cdot \mathbf{A} = BA \cos \theta \quad (6.1)$$

where θ is angle between \mathbf{B} and \mathbf{A} . The notion of the area as a vector has been discussed earlier in Chapter 1. Equation (6.1) can be extended to curved surfaces and nonuniform fields.

If the magnetic field has different magnitudes and directions at various parts of a surface as shown in Fig. 6.5, then the magnetic flux through the surface is given by

$$\Phi_B = \mathbf{B}_1 \cdot d\mathbf{A}_1 + \mathbf{B}_2 \cdot d\mathbf{A}_2 + \dots = \sum_{\text{all}} \mathbf{B}_i \cdot d\mathbf{A}_i \quad (6.2)$$

where 'all' stands for summation over all the area elements $d\mathbf{A}_i$ comprising the surface and \mathbf{B}_i is the magnetic field at the area element $d\mathbf{A}_i$. The SI unit of magnetic flux is weber (Wb) or tesla meter squared ($T m^2$). Magnetic flux is a scalar quantity.

6.4 FARADAY'S LAW OF INDUCTION

From the experimental observations, Faraday arrived at a conclusion that an emf is induced in a coil when magnetic flux through the coil changes with time. Experimental observations discussed in Section 6.2 can be explained using this concept.

The motion of a magnet towards or away from coil C_1 in Experiment 6.1 and moving a current-carrying coil C_2 towards or away from coil C_1 in Experiment 6.2, change the magnetic flux associated with coil C_1 . The change in magnetic flux induces emf in coil C_1 . It was this induced emf which caused electric current to flow in coil C_1 and through the galvanometer. A plausible explanation for the observations of Experiment 6.3 is as follows: When the tapping key K is pressed, the current in coil C_2 (and the resulting magnetic field) rises from zero to a maximum value in a short time. Consequently, the magnetic flux through the neighbouring coil C_1 also increases. It is the change in magnetic flux through coil C_1 that produces an induced emf in coil C_1 . When the key is held pressed, current in coil C_2 is constant. Therefore, there is no change in the magnetic flux through coil C_1 and the current in coil C_1 drops to zero. When the key is released, the current in C_2 and the resulting magnetic field decreases from the maximum value to zero in a short time. This results in a decrease in magnetic flux through coil C_1 and hence again induces an electric current in coil C_1 *. The common point in all these observations is that the time rate of change of magnetic flux through a circuit induces emf in it. Faraday stated experimental observations in the form of a law called *Faraday's law of electromagnetic induction*. The law is stated below.

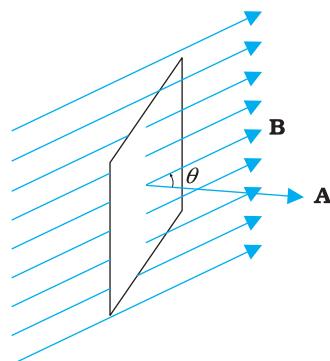


FIGURE 6.4 A plane of surface area \mathbf{A} placed in a uniform magnetic field \mathbf{B} .

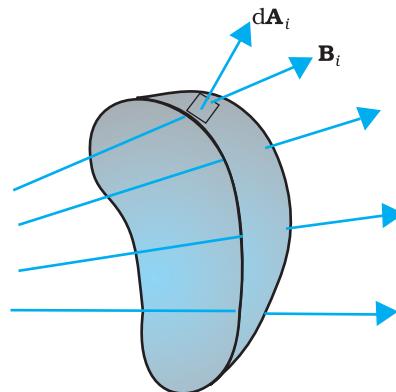


FIGURE 6.5 Magnetic field \mathbf{B}_i at the i^{th} area element. $d\mathbf{A}_i$ represents area vector of the i^{th} area element.

* Note that sensitive electrical instruments in the vicinity of an electromagnet can be damaged due to the induced emfs (and the resulting currents) when the electromagnet is turned on or off.

Physics

MICHAEL FARADAY (1791–1867)



Michael Faraday [1791–1867] Faraday made numerous contributions to science, viz., the discovery of electromagnetic induction, the laws of electrolysis, benzene, and the fact that the plane of polarisation is rotated in an electric field. He is also credited with the invention of the electric motor, the electric generator and the transformer. He is widely regarded as the greatest experimental scientist of the nineteenth century.

EXAMPLE 6.1

Example 6.1 Consider Experiment 6.2. (a) What would you do to obtain a large deflection of the galvanometer? (b) How would you demonstrate the presence of an induced current in the absence of a galvanometer?

Solution

- To obtain a large deflection, one or more of the following steps can be taken: (i) Use a rod made of soft iron inside the coil C_2 , (ii) Connect the coil to a powerful battery, and (iii) Move the arrangement rapidly towards the test coil C_1 .
- Replace the galvanometer by a small bulb, the kind one finds in a small torch light. The relative motion between the two coils will cause the bulb to glow and thus demonstrate the presence of an induced current.

In experimental physics one must learn to innovate. Michael Faraday who is ranked as one of the best experimentalists ever, was legendary for his innovative skills.

EXAMPLE 6.2

Example 6.2 A square loop of side 10 cm and resistance 0.5Ω is placed vertically in the east-west plane. A uniform magnetic field of 0.10 T is set up across the plane in the north-east direction. The magnetic field is decreased to zero in 0.70 s at a steady rate. Determine the magnitudes of induced emf and current during this time-interval.

The magnitude of the induced emf in a circuit is equal to the time rate of change of magnetic flux through the circuit.

Mathematically, the induced emf is given by

$$\varepsilon = -\frac{d\Phi_B}{dt} \quad (6.3)$$

The negative sign indicates the direction of ε and hence the direction of current in a closed loop. This will be discussed in detail in the next section.

In the case of a closely wound coil of N turns, change of flux associated with each turn, is the same. Therefore, the expression for the total induced emf is given by

$$\varepsilon = -N \frac{d\Phi_B}{dt} \quad (6.4)$$

The induced emf can be increased by increasing the number of turns N of a closed coil.

From Eqs. (6.1) and (6.2), we see that the flux can be varied by changing any one or more of the terms \mathbf{B} , \mathbf{A} and θ . In Experiments 6.1 and 6.2 in Section 6.2, the flux is changed by varying \mathbf{B} . The flux can also be altered by changing the shape of a coil (that is, by shrinking it or stretching it) in a magnetic field, or rotating a coil in a magnetic field such that the angle θ between \mathbf{B} and \mathbf{A} changes. In these cases too, an emf is induced in the respective coils.

Solution The angle θ made by the area vector of the coil with the magnetic field is 45° . From Eq. (6.1), the initial magnetic flux is

$$\Phi = BA \cos \theta$$

$$= \frac{0.1 \times 10^{-2}}{\sqrt{2}} \text{ Wb}$$

Final flux, $\Phi_{\min} = 0$

The change in flux is brought about in 0.70 s. From Eq. (6.3), the magnitude of the induced emf is given by

$$\varepsilon = \frac{|\Delta \Phi_B|}{\Delta t} = \frac{|(\Phi - 0)|}{\Delta t} = \frac{10^{-3}}{\sqrt{2} \times 0.7} = 1.0 \text{ mV}$$

And the magnitude of the current is

$$I = \frac{\varepsilon}{R} = \frac{10^{-3} \text{ V}}{0.5 \Omega} = 2 \text{ mA}$$

Note that the earth's magnetic field also produces a flux through the loop. But it is a steady field (which does not change within the time span of the experiment) and hence does not induce any emf.

EXAMPLE 6.2

Example 6.3

A circular coil of radius 10 cm, 500 turns and resistance 2Ω is placed with its plane perpendicular to the horizontal component of the earth's magnetic field. It is rotated about its vertical diameter through 180° in 0.25 s. Estimate the magnitudes of the emf and current induced in the coil. Horizontal component of the earth's magnetic field at the place is $3.0 \times 10^{-5} \text{ T}$.

Solution

Initial flux through the coil,

$$\begin{aligned}\Phi_{B \text{ (initial)}} &= BA \cos \theta \\ &= 3.0 \times 10^{-5} \times (\pi \times 10^{-2}) \times \cos 0^\circ \\ &= 3\pi \times 10^{-7} \text{ Wb}\end{aligned}$$

Final flux after the rotation,

$$\begin{aligned}\Phi_{B \text{ (final)}} &= 3.0 \times 10^{-5} \times (\pi \times 10^{-2}) \times \cos 180^\circ \\ &= -3\pi \times 10^{-7} \text{ Wb}\end{aligned}$$

Therefore, estimated value of the induced emf is,

$$\begin{aligned}\varepsilon &= N \frac{\Delta \Phi}{\Delta t} \\ &= 500 \times (6\pi \times 10^{-7}) / 0.25 \\ &= 3.8 \times 10^{-3} \text{ V}\end{aligned}$$

$$I = \varepsilon / R = 1.9 \times 10^{-3} \text{ A}$$

Note that the magnitudes of ε and I are the estimated values. Their instantaneous values are different and depend upon the speed of rotation at the particular instant.

EXAMPLE 6.3

6.5 LENZ'S LAW AND CONSERVATION OF ENERGY

In 1834, German physicist Heinrich Friedrich Lenz (1804–1865) deduced a rule, known as *Lenz's law* which gives the polarity of the induced emf in a clear and concise fashion. The statement of the law is:

The polarity of induced emf is such that it tends to produce a current which opposes the change in magnetic flux that produced it.

The negative sign shown in Eq. (6.3) represents this effect. We can understand Lenz's law by examining Experiment 6.1 in Section 6.2.1. In Fig. 6.1, we see that the North-pole of a bar magnet is being pushed towards the closed coil. As the North-pole of the bar magnet moves towards the coil, the magnetic flux through the coil increases. Hence current is induced in the coil in such a direction that it opposes the increase in flux. This is possible only if the current in the coil is in a counter-clockwise direction with respect to an observer situated on the side of the magnet. Note that magnetic moment associated with this current has North polarity towards the North-pole of the approaching magnet. Similarly, if the North-pole of the magnet is being withdrawn from the coil, the magnetic flux through the coil will decrease. To counter this decrease in magnetic flux, the induced current in the coil flows in clockwise direction and its South-pole faces the receding North-pole of the bar magnet. This would result in an attractive force which opposes the motion of the magnet and the corresponding decrease in flux.

What will happen if an open circuit is used in place of the closed loop in the above example? In this case too, an emf is induced across the open ends of the circuit. The direction of the induced emf can be found using Lenz's law. Consider Figs. 6.6 (a) and (b). They provide an easier way to understand the direction of induced currents. Note that the direction shown by \curvearrowleft and \curvearrowright indicate the directions of the induced currents.

A little reflection on this matter should convince us on the correctness of Lenz's law. Suppose that the induced current was in the direction opposite to the one depicted in Fig. 6.6(a). In that case, the South-pole due to the induced current will face the approaching North-pole of the magnet. The bar magnet will then be attracted towards the coil at an ever increasing acceleration. A gentle push on the magnet will initiate the process and its velocity and kinetic energy will continuously increase without expending any energy. If this can happen, one could construct a perpetual-motion machine by a suitable arrangement. This violates the law of conservation of energy and hence can not happen.

Now consider the correct case shown in Fig. 6.6(a). In this situation, the bar magnet experiences a repulsive force due to the induced current. Therefore, a person has to do work in moving the magnet.

Where does the energy spent by the person go? This energy is dissipated by Joule heating produced by the induced current.

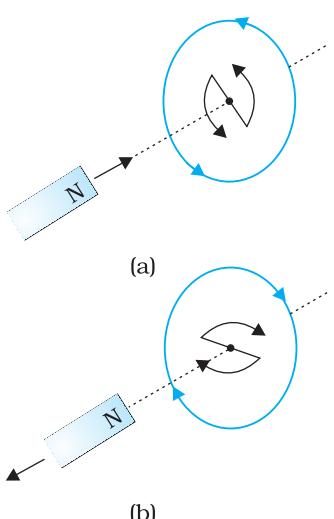


FIGURE 6.6
Illustration of
Lenz's law.

Example 6.4

Figure 6.7 shows planar loops of different shapes moving out of or into a region of a magnetic field which is directed normal to the plane of the loop away from the reader. Determine the direction of induced current in each loop using Lenz's law.

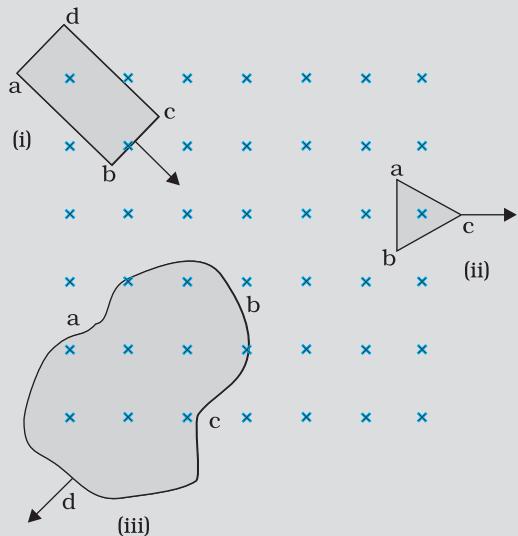


FIGURE 6.7

Solution

- The magnetic flux through the rectangular loop abcd increases, due to the motion of the loop into the region of magnetic field. The induced current must flow along the path bcdab so that it opposes the increasing flux.
- Due to the outward motion, magnetic flux through the triangular loop abc decreases due to which the induced current flows along bacb, so as to oppose the change in flux.
- As the magnetic flux decreases due to motion of the irregular shaped loop abcd out of the region of magnetic field, the induced current flows along cdabc, so as to oppose change in flux.
Note that there are no induced current as long as the loops are completely inside or outside the region of the magnetic field.

EXAMPLE 6.4

Example 6.5

- A closed loop is held stationary in the magnetic field between the north and south poles of two permanent magnets held fixed. Can we hope to generate current in the loop by using very strong magnets?
- A closed loop moves normal to the constant electric field between the plates of a large capacitor. Is a current induced in the loop
 - when it is wholly inside the region between the capacitor plates
 - when it is partially outside the plates of the capacitor? The electric field is normal to the plane of the loop.
- A rectangular loop and a circular loop are moving out of a uniform magnetic field region (Fig. 6.8) to a field-free region with a *constant velocity v*. In which loop do you expect the induced emf to be constant *during the passage out of the field region*? The field is normal to the loops.

EXAMPLE 6.5

EXAMPLE 6.5

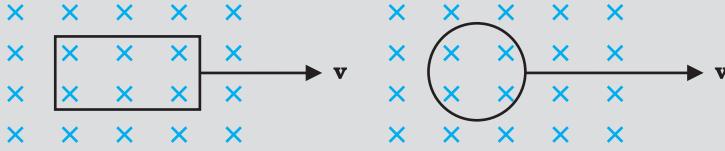


FIGURE 6.8

- (d) Predict the polarity of the capacitor in the situation described by Fig. 6.9.

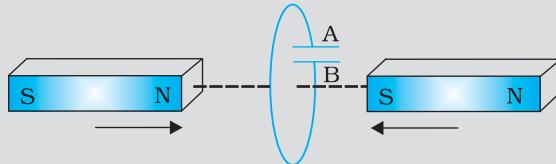


FIGURE 6.9

Solution

- No. However strong the magnet may be, current can be induced only by changing the magnetic flux through the loop.
- No current is induced in either case. Current can not be induced by changing the electric flux.
- The induced emf is expected to be constant only in the case of the rectangular loop. In the case of circular loop, the rate of change of area of the loop during its passage out of the field region is not constant, hence induced emf will vary accordingly.
- The polarity of plate 'A' will be positive with respect to plate 'B' in the capacitor.

6.6 MOTIONAL ELECTROMOTIVE FORCE

Let us consider a straight conductor moving in a uniform and time-independent magnetic field. Figure 6.10 shows a rectangular conductor PQRS in which the conductor PQ is free to move. The rod PQ is moved

towards the left with a constant velocity \mathbf{v} as shown in the figure. Assume that there is no loss of energy due to friction. PQRS forms a closed circuit enclosing an area that changes as PQ moves. It is placed in a uniform magnetic field \mathbf{B} which is perpendicular to the plane of this system. If the length $RQ = x$ and $RS = l$, the magnetic flux Φ_B enclosed by the loop PQRS will be

$$\Phi_B = Blx$$

Since x is changing with time, the rate of change of flux Φ_B will induce an emf given by:

$$\begin{aligned} \varepsilon &= -\frac{d\Phi_B}{dt} = -\frac{d}{dt}(Blx) \\ &= -Bl \frac{dx}{dt} = Blv \end{aligned} \quad (6.5)$$

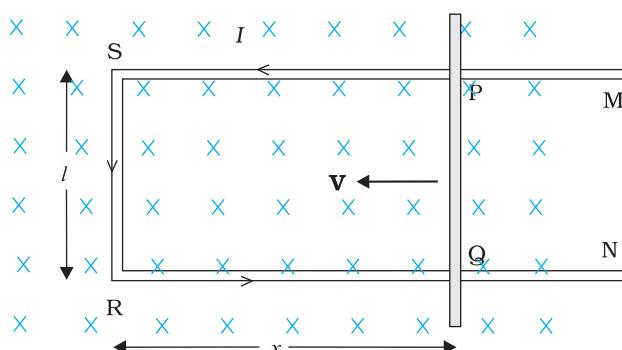


FIGURE 6.10 The arm PQ is moved to the left side, thus decreasing the area of the rectangular loop. This movement induces a current I as shown.

Electromagnetic Induction

where we have used $dx/dt = -v$ which is the speed of the conductor PQ. The induced emf $B\ell v$ is called *motional emf*. Thus, we are able to produce induced emf by moving a conductor instead of varying the magnetic field, that is, by changing the magnetic flux enclosed by the circuit.

It is also possible to explain the motional emf expression in Eq. (6.5) by invoking the Lorentz force acting on the free charge carriers of conductor PQ. Consider any arbitrary charge q in the conductor PQ. When the rod moves with speed v , the charge will also be moving with speed v in the magnetic field \mathbf{B} . The Lorentz force on this charge is qvB in magnitude, and its direction is towards Q. All charges experience the same force, in magnitude and direction, irrespective of their position in the rod PQ. The work done in moving the charge from P to Q is,

$$W = qvBl$$

Since emf is the work done per unit charge,

$$\epsilon = \frac{W}{q}$$

$$= Blv$$

This equation gives emf induced across the rod PQ and is identical to Eq. (6.5). We stress that our presentation is not wholly rigorous. But it does help us to understand the basis of Faraday's law when the conductor is moving in a uniform and time-independent magnetic field.

On the other hand, it is not obvious how an emf is induced when a conductor is stationary and the magnetic field is changing – a fact which Faraday verified by numerous experiments. In the case of a stationary conductor, the force on its charges is given by

$$\mathbf{F} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B}) = q\mathbf{E} \quad (6.6)$$

since $\mathbf{v} = 0$. Thus, any force on the charge must arise from the electric field term \mathbf{E} alone. Therefore, to explain the existence of induced emf or induced current, we must assume that a time-varying magnetic field generates an electric field. However, we hasten to add that electric fields produced by static electric charges have properties different from those produced by time-varying magnetic fields. In Chapter 4, we learnt that charges in motion (current) can exert force/torque on a stationary magnet. Conversely, a bar magnet in motion (or more generally, a changing magnetic field) can exert a force on the stationary charge. This is the fundamental significance of the Faraday's discovery. Electricity and magnetism are related.

Example 6.6 A metallic rod of 1 m length is rotated with a frequency of 50 rev/s, with one end hinged at the centre and the other end at the circumference of a circular metallic ring of radius 1 m, about an axis passing through the centre and perpendicular to the plane of the ring (Fig. 6.11). A constant and uniform magnetic field of 1 T parallel to the axis is present everywhere. What is the emf between the centre and the metallic ring?



Interactive animation on motional emf:
<http://ngsir.netfirms.com/englishhtm/Induction.htm>
http://webphysics.davidson.edu/physlet_resources/bu_semester2/index.html

EXAMPLE 6.6

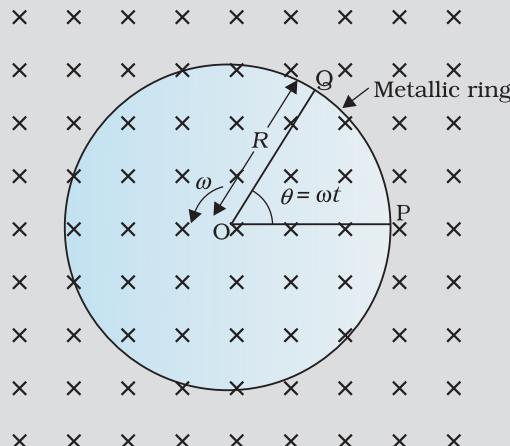


FIGURE 6.11

Solution

Method I

As the rod is rotated, free electrons in the rod move towards the outer end due to Lorentz force and get distributed over the ring. Thus, the resulting separation of charges produces an emf across the ends of the rod. At a certain value of emf, there is no more flow of electrons and a steady state is reached. Using Eq. (6.5), the magnitude of the emf generated across a length dr of the rod as it moves at right angles to the magnetic field is given by

$d\varepsilon = Bv dr$. Hence,

$$\varepsilon = \int d\varepsilon = \int_0^R Bv dr = \int_0^R B\omega r dr = \frac{B\omega R^2}{2}$$

Note that we have used $v = \omega r$. This gives

$$\begin{aligned} \varepsilon &= \frac{1}{2} \times 1.0 \times 2\pi \times 50 \times (1^2) \\ &= 157 \text{ V} \end{aligned}$$

Method II

To calculate the emf, we can imagine a closed loop OPQ in which point O and P are connected with a resistor R and OQ is the rotating rod. The potential difference across the resistor is then equal to the induced emf and equals $B \times (\text{rate of change of area of loop})$. If θ is the angle between the rod and the radius of the circle at P at time t , the area of the sector OPQ is given by

$$\pi R^2 \times \frac{\theta}{2\pi} = \frac{1}{2} R^2 \theta$$

where R is the radius of the circle. Hence, the induced emf is

$$\varepsilon = B \times \frac{d}{dt} \left[\frac{1}{2} R^2 \theta \right] = \frac{1}{2} BR^2 \frac{d\theta}{dt} = \frac{B\omega R^2}{2}$$

[Note: $\frac{d\theta}{dt} = \omega = 2\pi v$]

This expression is identical to the expression obtained by Method I and we get the same value of ε .

Example 6.7

A wheel with 10 metallic spokes each 0.5 m long is rotated with a speed of 120 rev/min in a plane normal to the horizontal component of earth's magnetic field H_E at a place. If $H_E = 0.4$ G at the place, what is the induced emf between the axle and the rim of the wheel? Note that 1 G = 10^{-4} T.

Solution

$$\begin{aligned}\text{Induced emf} &= (1/2) \omega B R^2 \\ &= (1/2) \times 4\pi \times 0.4 \times 10^{-4} \times (0.5)^2 \\ &= 6.28 \times 10^{-5} \text{ V}\end{aligned}$$

The number of spokes is immaterial because the emf's across the spokes are *in parallel*.

EXAMPLE 6.7

6.7 ENERGY CONSIDERATION: A QUANTITATIVE STUDY

In Section 6.5, we discussed qualitatively that Lenz's law is consistent with the law of conservation of energy. Now we shall explore this aspect further with a concrete example.

Let r be the resistance of movable arm PQ of the rectangular conductor shown in Fig. 6.10. We assume that the remaining arms QR, RS and SP have negligible resistances compared to r . Thus, the overall resistance of the rectangular loop is r and this does not change as PQ is moved. The current I in the loop is,

$$\begin{aligned}I &= \frac{\mathcal{E}}{r} \\ &= \frac{Blv}{r} \quad (6.7)\end{aligned}$$

On account of the presence of the magnetic field, there will be a force on the arm PQ. This force $I(\mathbf{l} \times \mathbf{B})$, is directed outwards in the direction opposite to the velocity of the rod. The magnitude of this force is,

$$F = IlB = \frac{B^2 l^2 v}{r}$$

where we have used Eq. (6.7). Note that this force arises due to drift velocity of charges (responsible for current) along the rod and the consequent Lorentz force acting on them.

Alternatively, the arm PQ is being pushed with a constant speed v , the power required to do this is,

$$\begin{aligned}P &= Fv \\ &= \frac{B^2 l^2 v^2}{r} \quad (6.8)\end{aligned}$$

The agent that does this work is mechanical. Where does this mechanical energy go? The answer is: it is dissipated as Joule heat, and is given by

$$P_J = I^2 r = \left(\frac{Blv}{r}\right)^2 r = \frac{B^2 l^2 v^2}{r}$$

which is identical to Eq. (6.8).

Thus, mechanical energy which was needed to move the arm PQ is converted into electrical energy (the induced emf) and then to thermal energy.

There is an interesting relationship between the charge flow through the circuit and the change in the magnetic flux. From Faraday's law, we have learnt that the magnitude of the induced emf is,

$$|\mathcal{E}| = \frac{\Delta \Phi_B}{\Delta t}$$

However,

$$|\mathcal{E}| = Ir = \frac{\Delta Q}{\Delta t} r$$

Thus,

$$\Delta Q = \frac{\Delta \Phi_B}{r}$$

Example 6.8 Refer to Fig. 6.12(a). The arm PQ of the rectangular conductor is moved from $x = 0$, outwards. The uniform magnetic field is perpendicular to the plane and extends from $x = 0$ to $x = b$ and is zero for $x > b$. Only the arm PQ possesses substantial resistance r . Consider the situation when the arm PQ is pulled outwards from $x = 0$ to $x = 2b$, and is then moved back to $x = 0$ with constant speed v . Obtain expressions for the flux, the induced emf, the force necessary to pull the arm and the power dissipated as Joule heat. Sketch the variation of these quantities with distance.

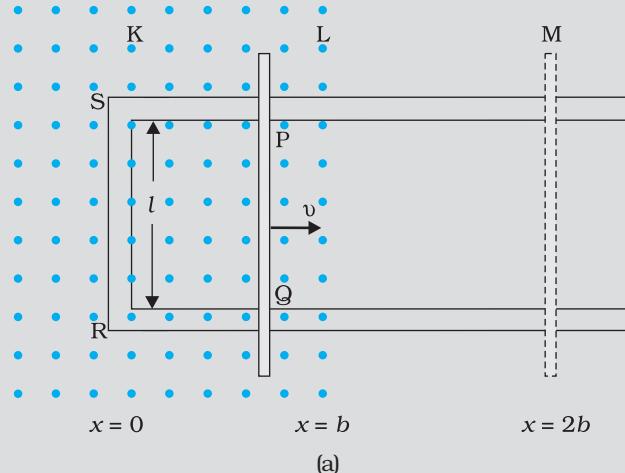


FIGURE 6.12

Solution Let us first consider the forward motion from $x = 0$ to $x = 2b$. The flux Φ_B linked with the circuit SPQR is

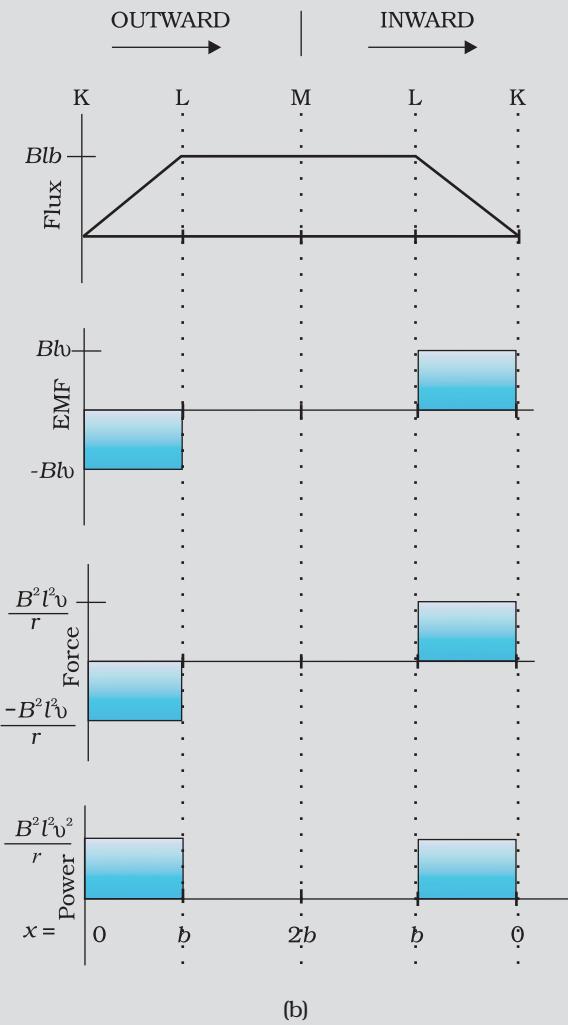
$$\begin{aligned}\Phi_B &= Blx & 0 \leq x < b \\ &= Blb & b \leq x < 2b\end{aligned}$$

The induced emf is,

$$\begin{aligned}\mathcal{E} &= -\frac{d\Phi_B}{dt} \\ &= -Blv & 0 \leq x < b \\ &= 0 & b \leq x < 2b\end{aligned}$$

When the induced emf is non-zero, the current I is (in magnitude)

$$I = \frac{Blv}{r}$$



(b)

FIGURE 6.12

The force required to keep the arm PQ in constant motion is IlB . Its direction is to the left. In magnitude

$$\begin{aligned} F &= \frac{B^2 l^2 v}{r} & 0 \leq x < b \\ &= 0 & b \leq x < 2b \end{aligned}$$

The Joule heating loss is

$$\begin{aligned} P_J &= I^2 r \\ &= \frac{B^2 l^2 v^2}{r} & 0 \leq x < b \\ &= 0 & b \leq x < 2b \end{aligned}$$

One obtains similar expressions for the inward motion from $x = 2b$ to $x = 0$. One can appreciate the whole process by examining the sketch of various quantities displayed in Fig. 6.12(b).

6.8 EDDY CURRENTS

So far we have studied the electric currents induced in well defined paths in conductors like circular loops. Even when bulk pieces of conductors are subjected to changing magnetic flux, induced currents are produced in them. However, their flow patterns resemble swirling eddies in water. This effect was discovered by physicist Foucault (1819-1868) and these currents are called *eddy currents*.

Consider the apparatus shown in Fig. 6.13. A copper plate is allowed to swing like a simple pendulum between the pole pieces of a strong magnet. It is found that the motion is damped and in a little while the plate comes to a halt in the magnetic field. We can explain this phenomenon on the basis of electromagnetic induction. Magnetic flux associated with the plate keeps on changing as the plate moves in and out of the region between magnetic poles. The flux change induces eddy currents in the plate. Directions of eddy currents are opposite when the plate swings into the region between the poles and when it swings out of the region.

If rectangular slots are made in the copper plate as shown in Fig. 6.14, area available to the flow of eddy currents is less. Thus, the pendulum plate with holes or slots reduces electromagnetic damping and the plate swings more freely. Note that magnetic moments of the induced currents (which oppose the motion) depend upon the area enclosed by the currents (recall equation $\mathbf{m} = IA$ in Chapter 4).

This fact is helpful in reducing eddy currents in the metallic cores of transformers, electric motors and other such devices in which a coil is to be wound over metallic core. Eddy currents are undesirable since they heat up the core and dissipate electrical energy in the form of heat. Eddy currents are minimised by using laminations of metal to make a metal core. The laminations are separated by an insulating material like lacquer. The plane of the laminations must be arranged parallel to the magnetic field, so that they cut across the eddy current paths. This arrangement reduces the strength of the eddy currents. Since the dissipation of electrical energy into heat depends on the square of the strength of electric current, heat loss is substantially reduced.

Eddy currents are used to advantage in certain applications like:

- Magnetic braking in trains:** Strong electromagnets are situated above the rails in some electrically powered trains. When the electromagnets are activated, the eddy currents induced in the rails oppose the motion of the train. As there are no mechanical linkages, the braking effect is smooth.
- Electromagnetic damping:** Certain galvanometers have a fixed core made of nonmagnetic metallic material. When the coil oscillates, the eddy currents generated in the core oppose the motion and bring the coil to rest quickly.

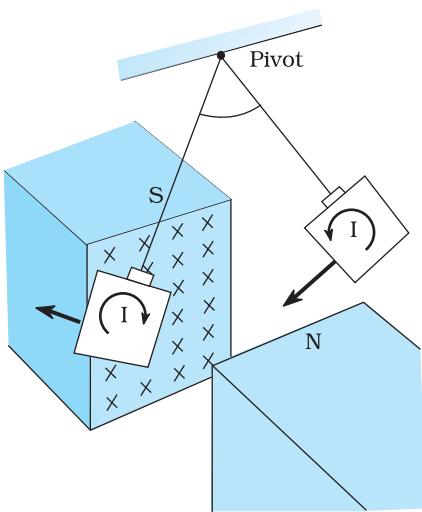


FIGURE 6.13 Eddy currents are generated in the copper plate, while entering and leaving the region of magnetic field.

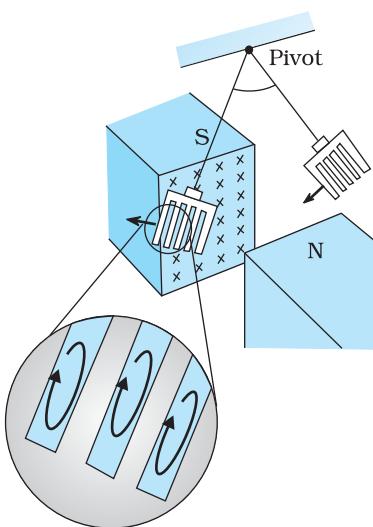


FIGURE 6.14 Cutting slots in the copper plate reduces the effect of eddy currents.

- (iii) *Induction furnace*: Induction furnace can be used to produce high temperatures and can be utilised to prepare alloys, by melting the constituent metals. A high frequency alternating current is passed through a coil which surrounds the metals to be melted. The eddy currents generated in the metals produce high temperatures sufficient to melt it.
- (iv) *Electric power meters*: The shiny metal disc in the electric power meter (analogue type) rotates due to the eddy currents. Electric currents are induced in the disc by magnetic fields produced by sinusoidally varying currents in a coil.
You can observe the rotating shiny disc in the power meter of your house.

ELECTROMAGNETIC DAMPING

Take two hollow thin cylindrical pipes of equal internal diameters made of aluminium and PVC, respectively. Fix them vertically with clamps on retort stands. Take a small cylindrical magnet having diameter slightly smaller than the inner diameter of the pipes and drop it through each pipe in such a way that the magnet does not touch the sides of the pipes during its fall. You will observe that the magnet dropped through the PVC pipe takes the same time to come out of the pipe as it would take when dropped through the same height without the pipe. Note the time it takes to come out of the pipe in each case. You will see that the magnet takes much longer time in the case of aluminium pipe. Why is it so? It is due to the eddy currents that are generated in the aluminium pipe which oppose the change in magnetic flux, i.e., the motion of the magnet. The retarding force due to the eddy currents inhibits the motion of the magnet. Such phenomena are referred to as *electromagnetic damping*. Note that eddy currents are not generated in PVC pipe as its material is an insulator whereas aluminium is a conductor.

6.9 INDUCTANCE

An electric current can be induced in a coil by flux change produced by another coil in its vicinity or flux change produced by the same coil. These two situations are described separately in the next two sub-sections. However, in both the cases, the flux through a coil is proportional to the current. That is, $\Phi_B \propto I$.

Further, if the geometry of the coil does not vary with time then,

$$\frac{d\Phi_B}{dt} \propto \frac{dl}{dt}$$

For a closely wound coil of N turns, the same magnetic flux is linked with all the turns. When the flux Φ_B through the coil changes, each turn contributes to the induced emf. Therefore, a term called *flux linkage* is used which is equal to $N\Phi_B$ for a closely wound coil and in such a case

$$N\Phi_B \propto I$$

The constant of proportionality, in this relation, is called *inductance*. We shall see that inductance depends only on the geometry of the coil

and intrinsic material properties. This aspect is akin to capacitance which for a parallel plate capacitor depends on the plate area and plate separation (geometry) and the dielectric constant K of the intervening medium (intrinsic material property).

Inductance is a scalar quantity. It has the dimensions of $[M L^2 T^{-2} A^{-2}]$ given by the dimensions of flux divided by the dimensions of current. The SI unit of inductance is *henry* and is denoted by H. It is named in honour of Joseph Henry who discovered electromagnetic induction in USA, independently of Faraday in England.

6.9.1 Mutual inductance

Consider Fig. 6.15 which shows two long co-axial solenoids each of length l . We denote the radius of the inner solenoid S_1 by r_1 and the number of turns per unit length by n_1 . The corresponding quantities for the outer solenoid S_2 are r_2 and n_2 , respectively. Let N_1 and N_2 be the total number of turns of coils S_1 and S_2 , respectively.

When a current I_2 is set up through S_2 , it in turn sets up a magnetic flux through S_1 . Let us denote it by Φ_1 . The corresponding flux linkage with solenoid S_1 is

$$N_1 \Phi_1 = M_{12} I_2 \quad (6.9)$$

M_{12} is called the *mutual inductance* of solenoid S_1 with respect to solenoid S_2 . It is also referred to as the *coefficient of mutual induction*.

For these simple co-axial solenoids it is possible to calculate M_{12} . The magnetic field due to the current I_2 in S_2 is $\mu_0 n_2 I_2$. The resulting flux linkage with coil S_1 is,

$$\begin{aligned} N_1 \Phi_1 &= (n_1 l) (\pi r_1^2) (\mu_0 n_2 I_2) \\ &= \mu_0 n_1 n_2 \pi r_1^2 l I_2 \end{aligned} \quad (6.10)$$

where $n_1 l$ is the total number of turns in solenoid S_1 . Thus, from Eq. (6.9) and Eq. (6.10),

$$M_{12} = \mu_0 n_1 n_2 \pi r_1^2 l \quad (6.11)$$

Note that we neglected the edge effects and considered the magnetic field $\mu_0 n_2 I_2$ to be uniform throughout the length and width of the solenoid S_2 . This is a good approximation keeping in mind that the solenoid is long, implying $l \gg r_2$.

We now consider the reverse case. A current I_1 is passed through the solenoid S_1 and the flux linkage with coil S_2 is,

$$N_2 \Phi_2 = M_{21} I_1 \quad (6.12)$$

M_{21} is called the *mutual inductance* of solenoid S_2 with respect to solenoid S_1 .

The flux due to the current I_1 in S_1 can be assumed to be confined solely inside S_1 since the solenoids are very long. Thus, flux linkage with solenoid S_2 is

$$N_2 \Phi_2 = (n_2 l) (\pi r_1^2) (\mu_0 n_1 I_1)$$

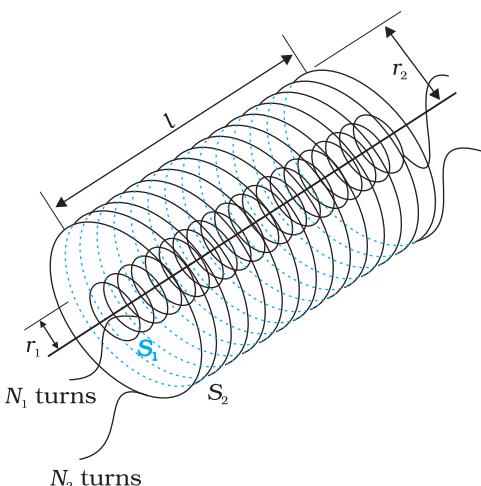


FIGURE 6.15 Two long co-axial solenoids of same length l .

where $n_2 l$ is the total number of turns of S_2 . From Eq. (6.12),

$$M_{21} = \mu_0 n_1 n_2 \pi r_1^2 l \quad (6.13)$$

Using Eq. (6.11) and Eq. (6.12), we get

$$M_{12} = M_{21} = M \text{ (say)} \quad (6.14)$$

We have demonstrated this equality for long co-axial solenoids. However, the relation is far more general. Note that if the inner solenoid was much shorter than (and placed well inside) the outer solenoid, then we could still have calculated the flux linkage $N_1 \Phi_1$ because the inner solenoid is effectively immersed in a uniform magnetic field due to the outer solenoid. In this case, the calculation of M_{12} would be easy. However, it would be extremely difficult to calculate the flux linkage with the outer solenoid as the magnetic field due to the inner solenoid would vary across the length as well as cross section of the outer solenoid. Therefore, the calculation of M_{21} would also be extremely difficult in this case. The equality $M_{12} = M_{21}$ is very useful in such situations.

We explained the above example with air as the medium within the solenoids. Instead, if a medium of relative permeability μ_r had been present, the mutual inductance would be

$$M = \mu_r \mu_0 n_1 n_2 \pi r_1^2 l$$

It is also important to know that the mutual inductance of a pair of coils, solenoids, etc., depends on their separation as well as their relative orientation.

Example 6.9 Two concentric circular coils, one of small radius r_1 and the other of large radius r_2 , such that $r_1 \ll r_2$, are placed co-axially with centres coinciding. Obtain the mutual inductance of the arrangement.

Solution Let a current I_2 flow through the outer circular coil. The field at the centre of the coil is $B_2 = \mu_0 I_2 / 2r_2$. Since the other co-axially placed coil has a very small radius, B_2 may be considered constant over its cross-sectional area. Hence,

$$\begin{aligned} \Phi_1 &= \pi r_1^2 B_2 \\ &= \frac{\mu_0 \pi r_1^2}{2r_2} I_2 \\ &= M_{12} I_2 \end{aligned}$$

Thus,

$$M_{12} = \frac{\mu_0 \pi r_1^2}{2r_2}$$

From Eq. (6.14)

$$M_{12} = M_{21} = \frac{\mu_0 \pi r_1^2}{2r_2}$$

Note that we calculated M_{12} from an approximate value of Φ_1 , assuming the magnetic field B_2 to be uniform over the area πr_1^2 . However, we can accept this value because $r_1 \ll r_2$.

■ Physics

Now, let us recollect Experiment 6.3 in Section 6.2. In that experiment, emf is induced in coil C_1 wherever there was any change in current through coil C_2 . Let Φ_1 be the flux through coil C_1 (say of N_1 turns) when current in coil C_2 is I_2 .

Then, from Eq. (6.9), we have

$$N_1 \Phi_1 = MI_2$$

For currents varying with time,

$$\frac{d(N_1 \Phi_1)}{dt} = \frac{d(MI_2)}{dt}$$

Since induced emf in coil C_1 is given by

$$\varepsilon_1 = -\frac{d(N_1 \Phi_1)}{dt}$$

We get,

$$\varepsilon_1 = -M \frac{dI_2}{dt}$$

It shows that varying current in a coil can induce emf in a neighbouring coil. The magnitude of the induced emf depends upon the rate of change of current and mutual inductance of the two coils.

6.9.2 Self-inductance

In the previous sub-section, we considered the flux in one solenoid due to the current in the other. It is also possible that emf is induced in a single isolated coil due to change of flux through the coil by means of varying the current through the same coil. This phenomenon is called *self-induction*. In this case, flux linkage through a coil of N turns is proportional to the current through the coil and is expressed as

$$N\Phi_B \propto I$$

$$N\Phi_B = LI \quad (6.15)$$

where constant of proportionality L is called *self-inductance* of the coil. It is also called the *coefficient of self-induction* of the coil. When the current is varied, the flux linked with the coil also changes and an emf is induced in the coil. Using Eq. (6.15), the induced emf is given by

$$\varepsilon = -\frac{d(N\Phi_B)}{dt}$$

$$\varepsilon = -L \frac{dI}{dt} \quad (6.16)$$

Thus, the self-induced emf always opposes any change (increase or decrease) of current in the coil.

It is possible to calculate the self-inductance for circuits with simple geometries. Let us calculate the self-inductance of a long solenoid of cross-sectional area A and length l , having n turns per unit length. The magnetic field due to a current I flowing in the solenoid is $B = \mu_0 n I$ (neglecting edge effects, as before). The total flux linked with the solenoid is

$$N\Phi_B = (nl)(\mu_0 n I)(A)$$

$$= \mu_0 n^2 A l I$$

where $n l$ is the total number of turns. Thus, the self-inductance is,

$$\begin{aligned} L &= \frac{N\Phi_B}{I} \\ &= \mu_0 n^2 A l \end{aligned} \tag{6.17}$$

If we fill the inside of the solenoid with a material of relative permeability μ_r (for example soft iron, which has a high value of relative permeability), then,

$$L = \mu_r \mu_0 n^2 A l \tag{6.18}$$

The self-inductance of the coil depends on its geometry and on the permeability of the medium.

The self-induced emf is also called the *back emf* as it opposes any change in the current in a circuit. Physically, the *self-inductance plays the role of inertia*. It is the electromagnetic analogue of mass in mechanics. So, work needs to be done against the back emf (ϵ) in establishing the current. This work done is stored as magnetic potential energy. For the current I at an instant in a circuit, the rate of work done is

$$\frac{dW}{dt} = |\epsilon| I$$

If we ignore the resistive losses and consider only inductive effect, then using Eq. (6.16),

$$\frac{dW}{dt} = L I \frac{dI}{dt}$$

Total amount of work done in establishing the current I is

$$W = \int dW = \int_0^I L I dI$$

Thus, the energy required to build up the current I is,

$$W = \frac{1}{2} L I^2 \tag{6.19}$$

This expression reminds us of $mv^2/2$ for the (mechanical) kinetic energy of a particle of mass m , and shows that L is analogous to m (i.e., L is electrical inertia and opposes growth and decay of current in the circuit).

Consider the general case of currents flowing simultaneously in two nearby coils. The flux linked with one coil will be the sum of two fluxes which exist independently. Equation (6.9) would be modified into

$$N_1 \Phi_1 = M_{11} I_1 + M_{12} I_2$$

where M_{11} represents inductance due to the same coil.

Therefore, using Faraday's law,

$$\epsilon_1 = -M_{11} \frac{dI_1}{dt} - M_{12} \frac{dI_2}{dt}$$

M_{11} is the *self-inductance* and is written as L_1 . Therefore,

$$\varepsilon_1 = -L_1 \frac{dI_1}{dt} - M_{12} \frac{dI_2}{dt}$$

Example 6.10 (a) Obtain the expression for the magnetic energy stored in a solenoid in terms of magnetic field B , area A and length l of the solenoid. (b) How does this magnetic energy compare with the electrostatic energy stored in a capacitor?

Solution

(a) From Eq. (6.19), the magnetic energy is

$$\begin{aligned} U_B &= \frac{1}{2} LI^2 \\ &= \frac{1}{2} L \left(\frac{B}{\mu_0 n} \right)^2 && (\text{since } B = \mu_0 n I, \text{ for a solenoid}) \\ &= \frac{1}{2} (\mu_0 n^2 Al) \left(\frac{B}{\mu_0 n} \right)^2 && [\text{from Eq. (6.17)}] \\ &= \frac{1}{2\mu_0} B^2 Al \end{aligned}$$

(b) The magnetic energy per unit volume is,

$$\begin{aligned} u_B &= \frac{U_B}{V} && (\text{where } V \text{ is volume that contains flux}) \\ &= \frac{U_B}{Al} \\ &= \frac{B^2}{2\mu_0} \end{aligned} \tag{6.20}$$

We have already obtained the relation for the electrostatic energy stored per unit volume in a parallel plate capacitor (refer to Chapter 2, Eq. 2.77),

$$u_E = \frac{1}{2} \epsilon_0 E^2 \tag{2.77}$$

In both the cases energy is proportional to the square of the field strength. Equations (6.20) and (2.77) have been derived for special cases: a solenoid and a parallel plate capacitor, respectively. But they are general and valid for any region of space in which a magnetic field or/and an electric field exist.

Interactive animation on ac generator:
<http://micro.magnet.fsu.edu/electromag/java/generator/ac.html>



EXAMPLE 6.10

6.10 AC GENERATOR

The phenomenon of electromagnetic induction has been technologically exploited in many ways. An exceptionally important application is the generation of alternating currents (ac). The modern ac generator with a typical output capacity of 100 MW is a highly evolved machine. In this section, we shall describe the basic principles behind this machine. The Yugoslav inventor Nicola Tesla is credited with the development of the machine. As was pointed out in Section 6.3, one method to induce an emf

Electromagnetic Induction

or current in a loop is through a change in the loop's orientation or a change in its effective area. As the coil rotates in a magnetic field **B**, the effective area of the loop (the face perpendicular to the field) is $A \cos \theta$, where θ is the angle between **A** and **B**. This method of producing a flux change is the principle of operation of a simple ac generator. An ac generator converts mechanical energy into electrical energy.

The basic elements of an ac generator are shown in Fig. 6.16. It consists of a coil mounted on a rotor shaft. The axis of rotation of the coil is perpendicular to the direction of the magnetic field. The coil (called armature) is mechanically rotated in the uniform magnetic field by some external means. The rotation of the coil causes the magnetic flux through it to change, so an emf is induced in the coil. The ends of the coil are connected to an external circuit by means of slip rings and brushes.

When the coil is rotated with a constant angular speed ω , the angle θ between the magnetic field vector **B** and the area vector **A** of the coil at any instant t is $\theta = \omega t$ (assuming $\theta = 0^\circ$ at $t = 0$). As a result, the effective area of the coil exposed to the magnetic field lines changes with time, and from Eq. (6.1), the flux at any time t is

$$\Phi_B = BA \cos \theta = BA \cos \omega t$$

From Faraday's law, the induced emf for the rotating coil of N turns is then,

$$\varepsilon = -N \frac{d\Phi_B}{dt} = -NBA \frac{d}{dt}(\cos \omega t)$$

Thus, the instantaneous value of the emf is

$$\varepsilon = NBA \omega \sin \omega t \quad (6.21)$$

where $NBA\omega$ is the maximum value of the emf, which occurs when $\sin \omega t = \pm 1$. If we denote $NBA\omega$ as ε_0 , then

$$\varepsilon = \varepsilon_0 \sin \omega t \quad (6.22)$$

Since the value of the sine function varies between +1 and -1, the sign, or polarity of the emf changes with time. Note from Fig. 6.17 that the emf has its extremum value when $\theta = 90^\circ$ or $\theta = 270^\circ$, as the change of flux is greatest at these points.

The direction of the current changes periodically and therefore the current is called *alternating current* (ac). Since $\omega = 2\pi\nu$, Eq (6.22) can be written as

$$\varepsilon = \varepsilon_0 \sin 2\pi \nu t \quad (6.23)$$

where ν is the frequency of revolution of the generator's coil.

Note that Eq. (6.22) and (6.23) give the instantaneous value of the emf and ε varies between $+\varepsilon_0$ and $-\varepsilon_0$ periodically. We shall learn how to determine the time-averaged value for the alternating voltage and current in the next chapter.

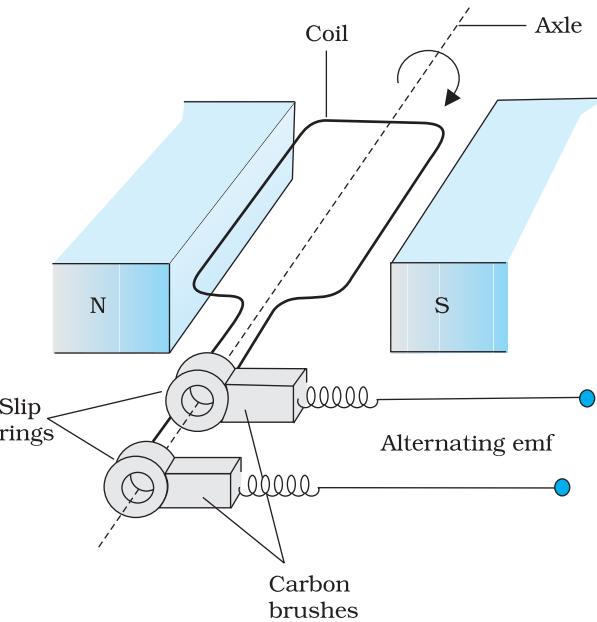


FIGURE 6.16 AC Generator

Physics

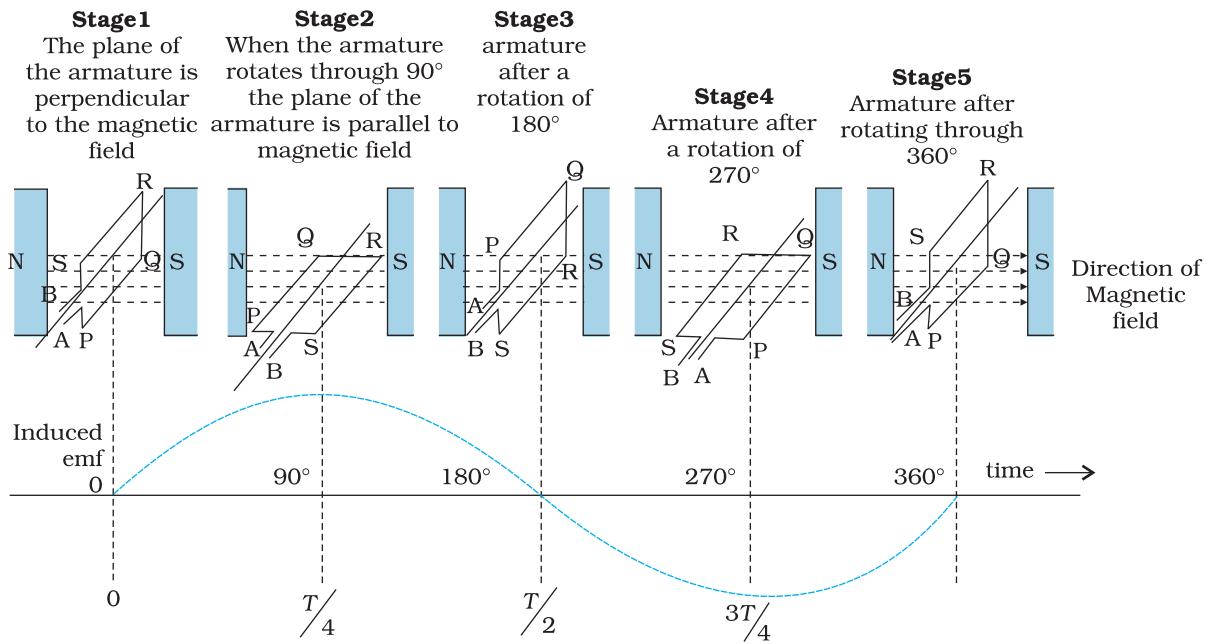


FIGURE 6.17 An alternating emf is generated by a loop of wire rotating in a magnetic field.

In commercial generators, the mechanical energy required for rotation of the armature is provided by water falling from a height, for example, from dams. These are called *hydro-electric generators*. Alternatively, water is heated to produce steam using coal or other sources. The steam at high pressure produces the rotation of the armature. These are called *thermal generators*. Instead of coal, if a nuclear fuel is used, we get *nuclear power generators*. Modern day generators produce electric power as high as 500 MW, i.e., one can light up 5 million 100 W bulbs! In most generators, the coils are held stationary and it is the electromagnets which are rotated. The frequency of rotation is 50 Hz in India. In certain countries such as USA, it is 60 Hz.

EXAMPLE 6.11

Example 6.11 Kamla peddles a stationary bicycle the pedals of the bicycle are attached to a 100 turn coil of area 0.10 m^2 . The coil rotates at half a revolution per second and it is placed in a uniform magnetic field of 0.01 T perpendicular to the axis of rotation of the coil. What is the maximum voltage generated in the coil?

Solution Here $f = 0.5 \text{ Hz}$; $N = 100$, $A = 0.1 \text{ m}^2$ and $B = 0.01 \text{ T}$. Employing Eq. (6.21)

$$\begin{aligned}\varepsilon_0 &= NBA (2\pi v) \\ &= 100 \times 0.01 \times 0.1 \times 2 \times 3.14 \times 0.5 \\ &= 0.314 \text{ V}\end{aligned}$$

The maximum voltage is 0.314 V.

We urge you to explore such alternative possibilities for power generation.

Migration of Birds

The migratory pattern of birds is one of the mysteries in the field of biology, and indeed all of science. For example, every winter birds from Siberia fly unerringly to water spots in the Indian subcontinent. There has been a suggestion that electromagnetic induction may provide a clue to these migratory patterns. The earth's magnetic field has existed throughout evolutionary history. It would be of great benefit to migratory birds to use this field to determine the direction. As far as we know birds contain no ferromagnetic material. So electromagnetic induction seems to be the only reasonable mechanism to determine direction. Consider the optimal case where the magnetic field \mathbf{B} , the velocity of the bird \mathbf{v} , and two relevant points of its anatomy separated by a distance l , all three are mutually perpendicular. From the formula for motional emf, Eq. (6.5),

$$\epsilon = Blv$$

Taking $B = 4 \times 10^{-5} \text{ T}$, $l = 2 \text{ cm}$ wide, and $v = 10 \text{ m/s}$, we obtain

$$\begin{aligned}\epsilon &= 4 \times 10^{-5} \times 2 \times 10^{-2} \times 10 \text{ V} = 8 \times 10^{-6} \text{ V} \\ &= 8 \mu\text{V}\end{aligned}$$

This extremely small potential difference suggests that our hypothesis is of doubtful validity. Certain kinds of fish are able to detect small potential differences. However, in these fish, special cells have been identified which detect small voltage differences. In birds no such cells have been identified. Thus, the migration patterns of birds continues to remain a mystery.

SUMMARY

1. The magnetic flux through a surface of area \mathbf{A} placed in a uniform magnetic field \mathbf{B} is defined as,
$$\Phi_B = \mathbf{B} \cdot \mathbf{A} = BA \cos \theta$$
where θ is the angle between \mathbf{B} and \mathbf{A} .
2. Faraday's laws of induction imply that the emf induced in a coil of N turns is directly related to the rate of change of flux through it,
$$\epsilon = -N \frac{d\Phi_B}{dt}$$
Here Φ_B is the flux linked with one turn of the coil. If the circuit is closed, a current $I = \epsilon/R$ is set up in it, where R is the resistance of the circuit.
3. Lenz's law states that the polarity of the induced emf is such that it tends to produce a current which opposes the change in magnetic flux that produces it. The negative sign in the expression for Faraday's law indicates this fact.
4. When a metal rod of length l is placed normal to a uniform magnetic field B and moved with a velocity v perpendicular to the field, the induced emf (called motional emf) across its ends is
$$\epsilon = Blv$$
5. Changing magnetic fields can set up current loops in nearby metal (any conductor) bodies. They dissipate electrical energy as heat. Such currents are called eddy currents.
6. Inductance is the ratio of the flux-linkage to current. It is equal to $N\Phi/I$.

Physics

7. A changing current in a coil (coil 2) can induce an emf in a nearby coil (coil 1). This relation is given by,

$$\varepsilon_1 = -M_{12} \frac{dI_2}{dt}$$

The quantity M_{12} is called mutual inductance of coil 1 with respect to coil 2. One can similarly define M_{21} . There exists a general equality,

$$M_{12} = M_{21}$$

8. When a current in a coil changes, it induces a back emf in the same coil. The self-induced emf is given by,

$$\varepsilon = -L \frac{dI}{dt}$$

L is the self-inductance of the coil. It is a measure of the inertia of the coil against the change of current through it.

9. The self-inductance of a long solenoid, the core of which consists of a magnetic material of permeability μ_r , is given by

$$L = \mu_r \mu_0 n^2 Al$$

where A is the area of cross-section of the solenoid, l its length and n the number of turns per unit length.

10. In an ac generator, mechanical energy is converted to electrical energy by virtue of electromagnetic induction. If coil of N turn and area A is rotated at v revolutions per second in a uniform magnetic field B , then the motional emf produced is

$$\varepsilon = NBA (2\pi v) \sin (2\pi vt)$$

where we have assumed that at time $t = 0$ s, the coil is perpendicular to the field.

Quantity	Symbol	Units	Dimensions	Equations
Magnetic Flux	Φ_B	Wb (weber)	[$M L^2 T^{-2} A^{-1}$]	$\Phi_B = \mathbf{B} \cdot \mathbf{A}$
EMF	ε	V (volt)	[$M L^2 T^{-3} A^{-1}$]	$\varepsilon = -d(N\Phi_B)/dt$
Mutual Inductance	M	H (henry)	[$M L^2 T^{-2} A^{-2}$]	$\varepsilon_1 = -M_{12} (dI_2 / dt)$
Self Inductance	L	H (henry)	[$M L^2 T^{-2} A^{-2}$]	$\varepsilon = -L (dI / dt)$

POINTS TO PONDER

- Electricity and magnetism are intimately related. In the early part of the nineteenth century, the experiments of Oersted, Ampere and others established that moving charges (currents) produce a magnetic field. Somewhat later, around 1830, the experiments of Faraday and Henry demonstrated that a moving magnet can induce electric current.
- In a closed circuit, electric currents are induced so as to oppose the changing magnetic flux. It is as per the law of conservation of energy. However, in case of an open circuit, an emf is induced across its ends. How is it related to the flux change?
- The motional emf discussed in Section 6.5 can be argued independently from Faraday's law using the Lorentz force on moving charges. However,

even if the charges are stationary [and the $q(\mathbf{v} \times \mathbf{B})$ term of the Lorentz force is not operative], an emf is nevertheless induced in the presence of a time-varying magnetic field. Thus, moving charges in static field and static charges in a time-varying field seem to be symmetric situation for Faraday's law. This gives a tantalising hint on the relevance of the principle of relativity for Faraday's law.

4. The motion of a copper plate is damped when it is allowed to oscillate between the magnetic pole-pieces. How is the damping force, produced by the eddy currents?



EXERCISES

- 6.1** Predict the direction of induced current in the situations described by the following Figs. 6.18(a) to (f).

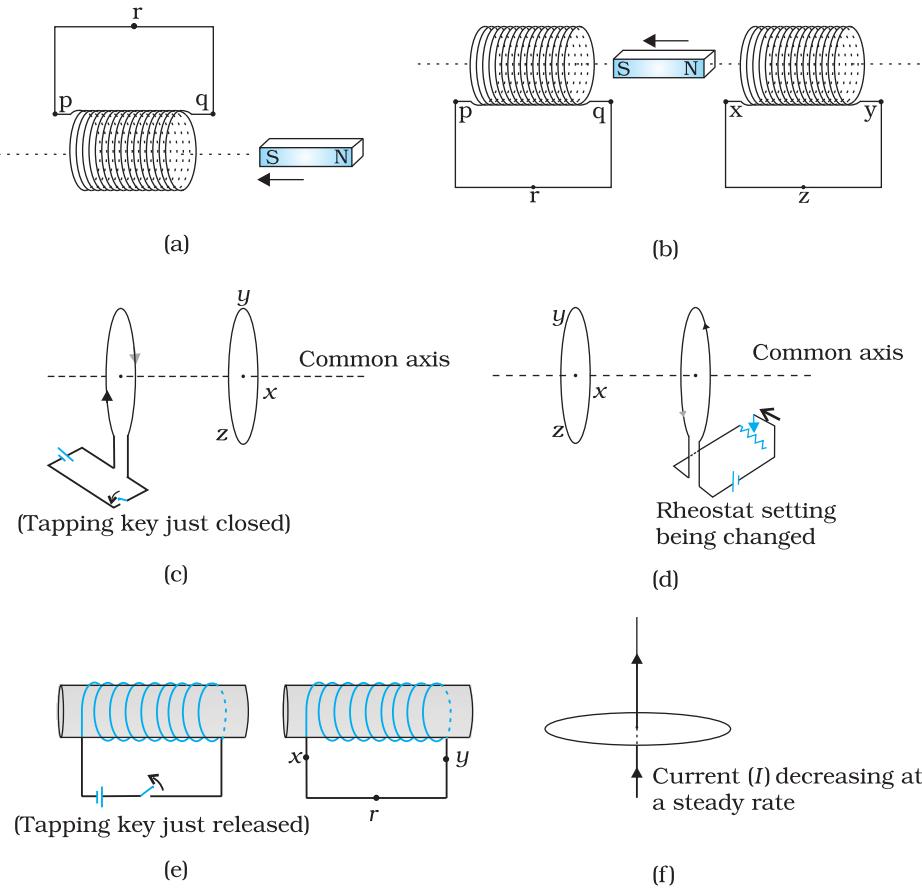


FIGURE 6.18

Physics

- 6.2** Use Lenz's law to determine the direction of induced current in the situations described by Fig. 6.19:
- A wire of irregular shape turning into a circular shape;
 - A circular loop being deformed into a narrow straight wire.

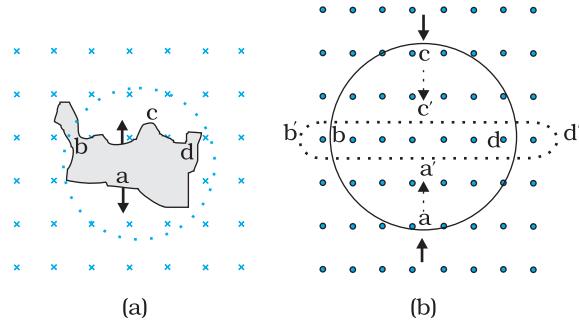


FIGURE 6.19

- 6.3** A long solenoid with 15 turns per cm has a small loop of area 2.0 cm^2 placed inside the solenoid normal to its axis. If the current carried by the solenoid changes steadily from 2.0 A to 4.0 A in 0.1 s, what is the induced emf in the loop while the current is changing?
- 6.4** A rectangular wire loop of sides 8 cm and 2 cm with a small cut is moving out of a region of uniform magnetic field of magnitude 0.3 T directed normal to the loop. What is the emf developed across the cut if the velocity of the loop is 1 cm s^{-1} in a direction normal to the (a) longer side, (b) shorter side of the loop? For how long does the induced voltage last in each case?
- 6.5** A 1.0 m long metallic rod is rotated with an angular frequency of 400 rad s^{-1} about an axis normal to the rod passing through its one end. The other end of the rod is in contact with a circular metallic ring. A constant and uniform magnetic field of 0.5 T parallel to the axis exists everywhere. Calculate the emf developed between the centre and the ring.
- 6.6** A circular coil of radius 8.0 cm and 20 turns is rotated about its vertical diameter with an angular speed of 50 rad s^{-1} in a uniform horizontal magnetic field of magnitude 3.0×10^{-2} T. Obtain the maximum and average emf induced in the coil. If the coil forms a closed loop of resistance 10Ω , calculate the maximum value of current in the coil. Calculate the average power loss due to Joule heating. Where does this power come from?
- 6.7** A horizontal straight wire 10 m long extending from east to west is falling with a speed of 5.0 m s^{-1} , at right angles to the horizontal component of the earth's magnetic field, 0.30×10^{-4} Wb m $^{-2}$.
- What is the instantaneous value of the emf induced in the wire?
 - What is the direction of the emf?
 - Which end of the wire is at the higher electrical potential?
- 6.8** Current in a circuit falls from 5.0 A to 0.0 A in 0.1 s. If an average emf of 200 V induced, give an estimate of the self-inductance of the circuit.
- 6.9** A pair of adjacent coils has a mutual inductance of 1.5 H. If the current in one coil changes from 0 to 20 A in 0.5 s, what is the change of flux linkage with the other coil?
- 6.10** A jet plane is travelling towards west at a speed of 1800 km/h. What is the voltage difference developed between the ends of the wing

having a span of 25 m, if the Earth's magnetic field at the location has a magnitude of 5×10^{-4} T and the dip angle is 30° .

ADDITIONAL EXERCISES

- 6.11** Suppose the loop in Exercise 6.4 is stationary but the current feeding the electromagnet that produces the magnetic field is gradually reduced so that the field decreases from its initial value of 0.3 T at the rate of 0.02 T s^{-1} . If the cut is joined and the loop has a resistance of 1.6Ω , how much power is dissipated by the loop as heat? What is the source of this power?
- 6.12** A square loop of side 12 cm with its sides parallel to X and Y axes is moved with a velocity of 8 cm s^{-1} in the positive x -direction in an environment containing a magnetic field in the positive z -direction. The field is neither uniform in space nor constant in time. It has a gradient of $10^{-3} \text{ T cm}^{-1}$ along the negative x -direction (that is it increases by $10^{-3} \text{ T cm}^{-1}$ as one moves in the negative x -direction), and it is decreasing in time at the rate of 10^{-3} T s^{-1} . Determine the direction and magnitude of the induced current in the loop if its resistance is $4.50 \text{ m}\Omega$.
- 6.13** It is desired to measure the magnitude of field between the poles of a powerful loud speaker magnet. A small flat search coil of area 2 cm^2 with 25 closely wound turns, is positioned normal to the field direction, and then quickly snatched out of the field region. Equivalently, one can give it a quick 90° turn to bring its plane parallel to the field direction. The total charge flown in the coil (measured by a ballistic galvanometer connected to coil) is 7.5 mC . The combined resistance of the coil and the galvanometer is 0.50Ω . Estimate the field strength of magnet.
- 6.14** Figure 6.20 shows a metal rod PQ resting on the smooth rails AB and positioned between the poles of a permanent magnet. The rails, the rod, and the magnetic field are in three mutual perpendicular directions. A galvanometer G connects the rails through a switch K. Length of the rod = 15 cm, $B = 0.50 \text{ T}$, resistance of the closed loop containing the rod = $9.0 \text{ m}\Omega$. Assume the field to be uniform.
- (a) Suppose K is open and the rod is moved with a speed of 12 cm s^{-1} in the direction shown. Give the polarity and magnitude of the induced emf.

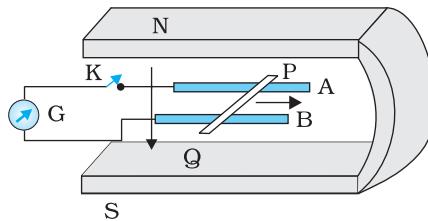


FIGURE 6.20

- (b) Is there an excess charge built up at the ends of the rods when K is open? What if K is closed?
- (c) With K open and the rod moving uniformly, there is *no net force* on the electrons in the rod PQ even though they do

Physics

experience magnetic force due to the motion of the rod. Explain.

- (d) What is the retarding force on the rod when K is closed?
- (e) How much power is required (by an external agent) to keep the rod moving at the same speed ($=12 \text{ cm s}^{-1}$) when K is closed? How much power is required when K is open?
- (f) How much power is dissipated as heat in the closed circuit? What is the source of this power?
- (g) What is the induced emf in the moving rod if the magnetic field is parallel to the rails instead of being perpendicular?

6.15 An air-cored solenoid with length 30 cm, area of cross-section 25 cm^2 and number of turns 500, carries a current of 2.5 A. The current is suddenly switched off in a brief time of 10^{-3} s. How much is the average back emf induced across the ends of the open switch in the circuit? Ignore the variation in magnetic field near the ends of the solenoid.

6.16 (a) Obtain an expression for the mutual inductance between a long straight wire and a square loop of side a as shown in Fig. 6.21.
 (b) Now assume that the straight wire carries a current of 50 A and the loop is moved to the right with a constant velocity, $v = 10 \text{ m/s}$. Calculate the induced emf in the loop at the instant when $x = 0.2 \text{ m}$. Take $a = 0.1 \text{ m}$ and assume that the loop has a large resistance.

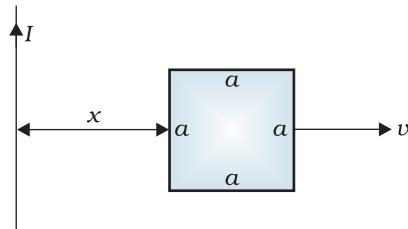


FIGURE 6.21

6.17 A line charge λ per unit length is lodged uniformly onto the rim of a wheel of mass M and radius R . The wheel has light non-conducting spokes and is free to rotate without friction about its axis (Fig. 6.22). A uniform magnetic field extends over a circular region within the rim. It is given by,

$$\begin{aligned}\mathbf{B} &= -B_0 \mathbf{k} & (r \leq a; a < R) \\ &= 0 & (\text{otherwise})\end{aligned}$$

What is the angular velocity of the wheel after the field is suddenly switched off?

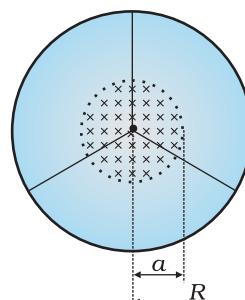


FIGURE 6.22

Chapter Seven

ALTERNATING CURRENT

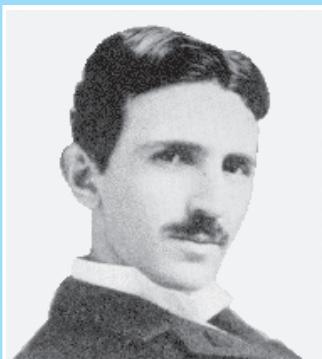


7.1 INTRODUCTION

We have so far considered direct current (dc) sources and circuits with dc sources. These currents do not change direction with time. But voltages and currents that vary with time are very common. The electric mains supply in our homes and offices is a voltage that varies like a sine function with time. Such a voltage is called *alternating voltage* (ac voltage) and the current driven by it in a circuit is called the *alternating current* (ac current)*. Today, most of the electrical devices we use require ac voltage. This is mainly because most of the electrical energy sold by power companies is transmitted and distributed as alternating current. The main reason for preferring use of ac voltage over dc voltage is that ac voltages can be easily and efficiently converted from one voltage to the other by means of transformers. Further, electrical energy can also be transmitted economically over long distances. AC circuits exhibit characteristics which are exploited in many devices of daily use. For example, whenever we tune our radio to a favourite station, we are taking advantage of a special property of ac circuits – one of many that you will study in this chapter.

* The phrases *ac voltage* and *ac current* are contradictory and redundant, respectively, since they mean, literally, *alternating current voltage* and *alternating current current*. Still, the abbreviation *ac* to designate an electrical quantity displaying simple harmonic time dependence has become so universally accepted that we follow others in its use. Further, *voltage* – another phrase commonly used means potential difference between two points.

Physics



Nicola Tesla (1836 – 1943) Yugoslav scientist, inventor and genius. He conceived the idea of the rotating magnetic field, which is the basis of practically all alternating current machinery, and which helped usher in the age of electric power. He also invented among other things the induction motor, the polyphase system of ac power, and the high frequency induction coil (the Tesla coil) used in radio and television sets and other electronic equipment. The SI unit of magnetic field is named in his honour.

NICOLA TESLA (1836 – 1943)

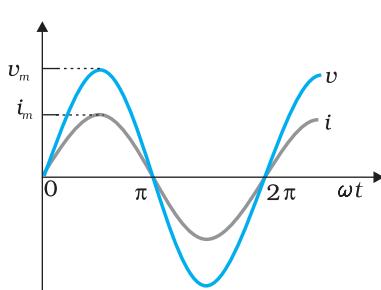


FIGURE 7.2 In a pure resistor, the voltage and current are in phase. The minima, zero and maxima occur at the same respective times.

7.2 AC VOLTAGE APPLIED TO A RESISTOR

Figure 7.1 shows a resistor connected to a source ε of ac voltage. The symbol for an ac source in a circuit diagram is \sim . We consider a source which produces sinusoidally varying potential difference across its terminals. Let this potential difference, also called ac voltage, be given by

$$v = v_m \sin \omega t \quad (7.1)$$

where v_m is the amplitude of the oscillating potential difference and ω is its angular frequency.



FIGURE 7.1 AC voltage applied to a resistor.

To find the value of current through the resistor, we apply Kirchhoff's loop rule $\sum \varepsilon(t) = 0$, to the circuit shown in Fig. 7.1 to get

$$v_m \sin \omega t = i R$$

$$\text{or } i = \frac{v_m}{R} \sin \omega t$$

Since R is a constant, we can write this equation as

$$i = i_m \sin \omega t \quad (7.2)$$

where the current amplitude i_m is given by

$$i_m = \frac{v_m}{R} \quad (7.3)$$

Equation (7.3) is just Ohm's law which for resistors works equally well for both ac and dc voltages. The voltage across a pure resistor and the current through it, given by Eqs. (7.1) and (7.2) are plotted as a function of time in Fig. 7.2. Note, in particular that both v and i reach zero, minimum and maximum values at the same time. Clearly, *the voltage and current are in phase with each other*.

We see that, like the applied voltage, the current varies sinusoidally and has corresponding positive and negative values during each cycle. Thus, the sum of the instantaneous current values over one complete cycle is zero, and the average current is zero. The fact that the average current is zero, however, does

Alternating Current

not mean that the average power consumed is zero and that there is no dissipation of electrical energy. As you know, Joule heating is given by i^2R and depends on i^2 (which is always positive whether i is positive or negative) and not on i . Thus, there is Joule heating and dissipation of electrical energy when an ac current passes through a resistor.

The instantaneous power dissipated in the resistor is

$$p = i^2 R = i_m^2 R \sin^2 \omega t \quad (7.4)$$

The average value of p over a cycle is*

$$\bar{p} = \langle i^2 R \rangle = \langle i_m^2 R \sin^2 \omega t \rangle \quad [7.5(a)]$$

where the bar over a letter (here, p) denotes its average value and $\langle \dots \dots \rangle$ denotes taking average of the quantity inside the bracket. Since, i_m^2 and R are constants,

$$\bar{p} = i_m^2 R \langle \sin^2 \omega t \rangle \quad [7.5(b)]$$

Using the trigonometric identity, $\sin^2 \omega t = 1/2(1 - \cos 2\omega t)$, we have $\langle \sin^2 \omega t \rangle = (1/2)(1 - \langle \cos 2\omega t \rangle)$ and since $\langle \cos 2\omega t \rangle = 0^{**}$, we have,

$$\langle \sin^2 \omega t \rangle = \frac{1}{2}$$

Thus,

$$\bar{p} = \frac{1}{2} i_m^2 R \quad [7.5(c)]$$

To express ac power in the same form as dc power ($P = I^2 R$), a special value of current is defined and used. It is called, *root mean square* (rms) or *effective current* (Fig. 7.3) and is denoted by I_{rms} or I .

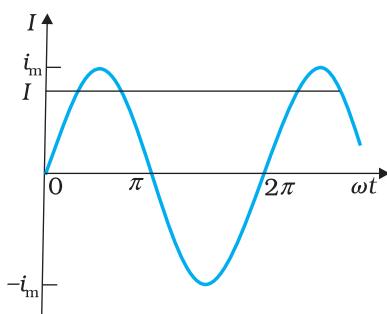
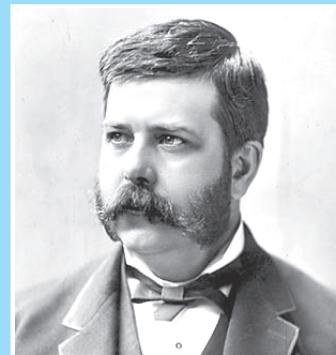


FIGURE 7.3 The rms current I is related to the peak current i_m by $I = i_m / \sqrt{2} = 0.707 i_m$.

* The average value of a function $F(t)$ over a period T is given by $\langle F(t) \rangle = \frac{1}{T} \int_0^T F(t) dt$

** $\langle \cos 2\omega t \rangle = \frac{1}{T} \int_0^T \cos 2\omega t dt = \frac{1}{T} \left[\frac{\sin 2\omega t}{2\omega} \right]_0^T = \frac{1}{2\omega T} [\sin 2\omega T - 0] = 0$



George Westinghouse (1846 – 1914) A leading proponent of the use of alternating current over direct current. Thus, he came into conflict with Thomas Alva Edison, an advocate of direct current. Westinghouse was convinced that the technology of alternating current was the key to the electrical future. He founded the famous Company named after him and enlisted the services of Nicola Tesla and other inventors in the development of alternating current motors and apparatus for the transmission of high tension current, pioneering in large scale lighting.

GEORGE WESTINGHOUSE (1846 – 1914)

■ Physics

It is defined by

$$\begin{aligned} I &= \sqrt{\bar{i}^2} = \sqrt{\frac{1}{2} i_m^2} = \frac{i_m}{\sqrt{2}} \\ &= 0.707 i_m \end{aligned} \quad (7.6)$$

In terms of I , the average power, denoted by P is

$$P = \bar{P} = \frac{1}{2} i_m^2 R = I^2 R \quad (7.7)$$

Similarly, we define the *rms voltage* or *effective voltage* by

$$V = \frac{v_m}{\sqrt{2}} = 0.707 v_m \quad (7.8)$$

From Eq. (7.3), we have

$$v_m = i_m R$$

$$\text{or, } \frac{v_m}{\sqrt{2}} = \frac{i_m}{\sqrt{2}} R$$

$$\text{or, } V = IR \quad (7.9)$$

Equation (7.9) gives the relation between ac current and ac voltage and is similar to that in the dc case. This shows the advantage of introducing the concept of rms values. In terms of rms values, the equation for power [Eq. (7.7)] and relation between current and voltage in ac circuits are essentially the same as those for the dc case.

It is customary to measure and specify rms values for ac quantities. For example, the household line voltage of 220 V is an rms value with a peak voltage of

$$v_m = \sqrt{2} V = (1.414)(220 \text{ V}) = 311 \text{ V}$$

In fact, the I or rms current is the equivalent dc current that would produce the same average power loss as the alternating current. Equation (7.7) can also be written as

$$P = V^2 / R = I V \quad (\text{since } V = IR)$$

Example 7.1 A light bulb is rated at 100W for a 220 V supply. Find
(a) the resistance of the bulb; (b) the peak voltage of the source; and
(c) the rms current through the bulb.

Solution

(a) We are given $P = 100 \text{ W}$ and $V = 220 \text{ V}$. The resistance of the bulb is

$$R = \frac{V^2}{P} = \frac{(220 \text{ V})^2}{100 \text{ W}} = 484 \Omega$$

(b) The peak voltage of the source is

$$v_m = \sqrt{2} V = 311 \text{ V}$$

(c) Since, $P = I V$

$$I = \frac{P}{V} = \frac{100 \text{ W}}{220 \text{ V}} = 0.450 \text{ A}$$

EXAMPLE 7.1

7.3 REPRESENTATION OF AC CURRENT AND VOLTAGE BY ROTATING VECTORS — PHASORS

In the previous section, we learnt that the current through a resistor is in phase with the ac voltage. But this is not so in the case of an inductor, a capacitor or a combination of these circuit elements. In order to show phase relationship between voltage and current in an ac circuit, we use the notion of *phasors*. The analysis of an ac circuit is facilitated by the use of a phasor diagram. A phasor* is a vector which rotates about the origin with angular speed ω , as shown in Fig. 7.4. The vertical components of phasors **V** and **I** represent the sinusoidally varying quantities v and i . The magnitudes of phasors **V** and **I** represent the amplitudes or the peak values v_m and i_m of these oscillating quantities. Figure 7.4(a) shows the voltage and current phasors and their relationship at time t_1 for the case of an ac source connected to a resistor i.e., corresponding to the circuit shown in Fig. 7.1. The projection of voltage and current phasors on vertical axis, i.e., $v_m \sin \omega t$ and $i_m \sin \omega t$, respectively represent the value of voltage and current at that instant. As they rotate with frequency ω , curves in Fig. 7.4(b) are generated. From Fig. 7.4(a) we see that phasors **V** and **I** for the case of a resistor are in the same direction. This is so for all times. This means that the phase angle between the voltage and the current is zero.

7.4 AC VOLTAGE APPLIED TO AN INDUCTOR

Figure 7.5 shows an ac source connected to an inductor. Usually, inductors have appreciable resistance in their windings, but we shall assume that this inductor has negligible resistance. Thus, the circuit is a purely inductive ac circuit. Let the voltage across the source be $v = v_m \sin \omega t$. Using the Kirchhoff's loop rule, $\sum \varepsilon(t) = 0$, and since there is no resistor in the circuit,

$$v - L \frac{di}{dt} = 0 \quad (7.10)$$

where the second term is the self-induced Faraday emf in the inductor; and L is the self-inductance of

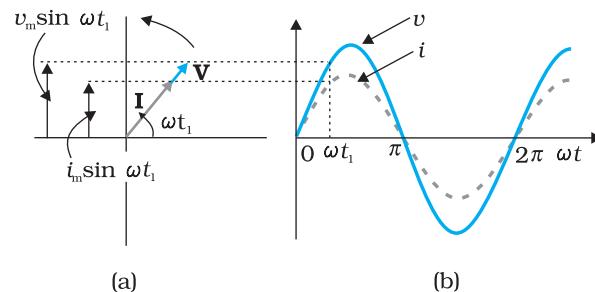


FIGURE 7.4 (a) A phasor diagram for the circuit in Fig. 7.1. (b) Graph of v and i versus ωt .



FIGURE 7.5 An ac source connected to an inductor.

* Though voltage and current in ac circuit are represented by phasors – rotating vectors, they are not vectors themselves. They are scalar quantities. It so happens that the amplitudes and phases of harmonically varying scalars combine mathematically in the same way as do the projections of rotating vectors of corresponding magnitudes and directions. The *rotating vectors* that represent harmonically varying scalar quantities are introduced only to provide us with a simple way of adding these quantities using a rule that we already know.

Physics

Interactive animation on Phasor diagrams of ac circuits containing, R, L, C and RLC series circuits:
<http://www.animations.physics.unsw.edu.au/jw/AC.html>

the inductor. The negative sign follows from Lenz's law (Chapter 6). Combining Eqs. (7.1) and (7.10), we have

$$\frac{di}{dt} = \frac{v}{L} = \frac{v_m}{L} \sin \omega t \quad (7.11)$$

Equation (7.11) implies that the equation for $i(t)$, the current as a function of time, must be such that its slope di/dt is a sinusoidally varying quantity, with the same phase as the source voltage and an amplitude given by v_m/L . To obtain the current, we integrate di/dt with respect to time:

$$\int \frac{di}{dt} dt = \frac{v_m}{L} \int \sin(\omega t) dt$$

and get,

$$i = -\frac{v_m}{\omega L} \cos(\omega t) + \text{constant}$$

The integration constant has the dimension of current and is time-independent. Since the source has an emf which oscillates symmetrically about zero, the current it sustains also oscillates symmetrically about zero, so that no constant or time-independent component of the current exists. Therefore, the integration constant is zero.

Using

$$-\cos(\omega t) = \sin\left(\omega t - \frac{\pi}{2}\right), \text{ we have}$$

$$i = i_m \sin\left(\omega t - \frac{\pi}{2}\right) \quad (7.12)$$

where $i_m = \frac{v_m}{\omega L}$ is the amplitude of the current. The quantity ωL is analogous to the resistance and is called *inductive reactance*, denoted by X_L :

$$X_L = \omega L \quad (7.13)$$

The amplitude of the current is, then

$$i_m = \frac{v_m}{X_L} \quad (7.14)$$

The dimension of inductive reactance is the same as that of resistance and its SI unit is ohm (Ω). The inductive reactance limits the current in a purely inductive circuit in the same way as the resistance limits the current in a purely resistive circuit. The inductive reactance is directly proportional to the inductance and to the frequency of the current.

A comparison of Eqs. (7.1) and (7.12) for the source voltage and the current in an inductor shows that the current lags the voltage by $\pi/2$ or one-quarter (1/4) cycle. Figure 7.6 (a) shows the voltage and the current phasors in the present case at instant t_1 . The current phasor \mathbf{I} is $\pi/2$ behind the voltage phasor \mathbf{V} . When rotated with frequency ω counter-clockwise, they generate the voltage and current given by Eqs. (7.1) and (7.12), respectively and as shown in Fig. 7.6(b).

Alternating Current

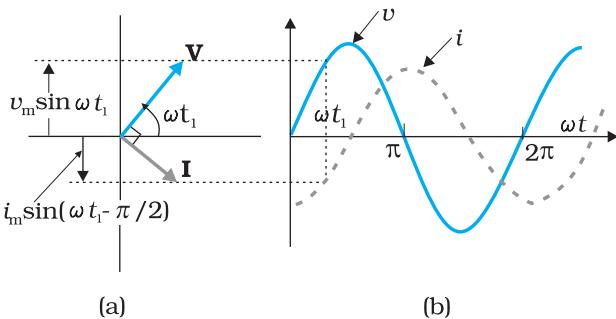


FIGURE 7.6 (a) A Phasor diagram for the circuit in Fig. 7.5.
 (b) Graph of v and i versus ωt .

We see that the current reaches its maximum value later than the voltage by one-fourth of a period $\left[\frac{T}{4} = \frac{\pi/2}{\omega} \right]$. You have seen that an inductor has reactance that limits current similar to resistance in a dc circuit. Does it also consume power like a resistance? Let us try to find out.

The instantaneous power supplied to the inductor is

$$\begin{aligned} p_L &= i v = i_m \sin\left(\omega t - \frac{\pi}{2}\right) \times v_m \sin(\omega t) \\ &= -i_m v_m \cos(\omega t) \sin(\omega t) \\ &= -\frac{i_m v_m}{2} \sin(2\omega t) \end{aligned}$$

So, the average power over a complete cycle is

$$\begin{aligned} P_L &= \left\langle -\frac{i_m v_m}{2} \sin(2\omega t) \right\rangle \\ &= -\frac{i_m v_m}{2} \langle \sin(2\omega t) \rangle = 0, \end{aligned}$$

since the average of $\sin(2\omega t)$ over a complete cycle is zero.

Thus, the *average power supplied to an inductor over one complete cycle is zero*.

Figure 7.7 explains it in detail.

Example 7.2 A pure inductor of 25.0 mH is connected to a source of 220 V. Find the inductive reactance and rms current in the circuit if the frequency of the source is 50 Hz.

Solution The inductive reactance,

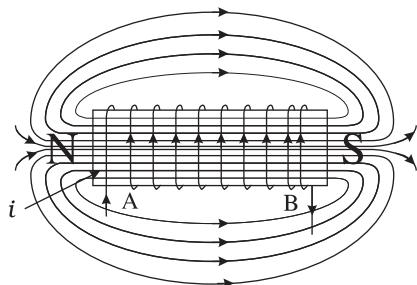
$$\begin{aligned} X_L &= 2\pi f L = 2 \times 3.14 \times 50 \times 25 \times 10^{-3} \text{ W} \\ &= 7.85 \Omega \end{aligned}$$

The rms current in the circuit is

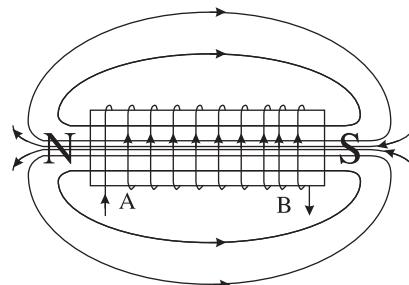
$$I = \frac{V}{X_L} = \frac{220 \text{ V}}{7.85 \Omega} = 28 \text{ A}$$

EXAMPLE 7.2

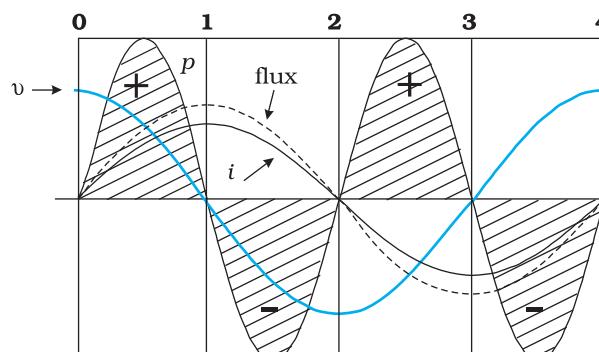
Physics



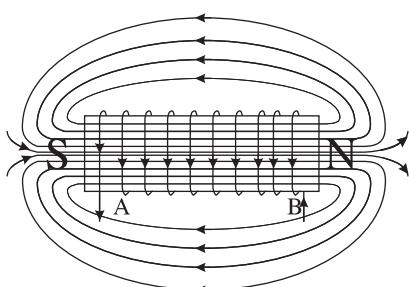
0-1 Current i through the coil entering at A increase from zero to a maximum value. Flux lines are set up i.e., the core gets magnetised. With the polarity shown voltage and current are both positive. So their product p is positive. ENERGY IS ABSORBED FROM THE SOURCE.



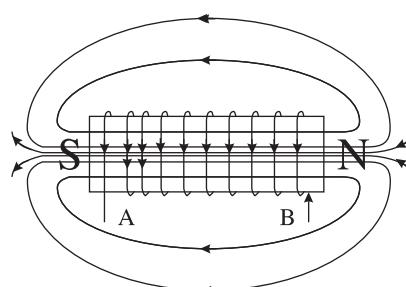
1-2 Current in the coil is still positive but is decreasing. The core gets demagnetised and the net flux becomes zero at the end of a half cycle. The voltage v is negative (since di/dt is negative). The product of voltage and current is negative, and ENERGY IS BEING RETURNED TO SOURCE.



One complete cycle of voltage/current. Note that the current lags the voltage.



2-3 Current i becomes negative i.e., it enters B and comes out of A. Since the direction of current has changed, the polarity of the magnet changes. The current and voltage are both negative. So their product p is positive. ENERGY IS ABSORBED.



3-4 Current i decreases and reaches its zero value at 4 when core is demagnetised and flux is zero. The voltage is positive but the current is negative. The power is, therefore, negative. ENERGY ABSORBED DURING THE 1/4 CYCLE 2-3 IS RETURNED TO THE SOURCE.

7.5 AC VOLTAGE APPLIED TO A CAPACITOR

Figure 7.8 shows an ac source ϵ generating ac voltage $v = v_m \sin \omega t$ connected to a capacitor only, a purely capacitive ac circuit.

When a capacitor is connected to a voltage source in a dc circuit, current will flow for the short time required to charge the capacitor. As charge accumulates on the capacitor plates, the voltage across them increases, opposing the current. That is, a capacitor in a dc circuit will limit or oppose the current as it charges. When the capacitor is fully charged, the current in the circuit falls to zero.

When the capacitor is connected to an ac source, as in Fig. 7.8, it limits or regulates the current, but does not completely prevent the flow of charge. The capacitor is alternately charged and discharged as the current reverses each half cycle. Let q be the charge on the capacitor at any time t . The instantaneous voltage v across the capacitor is

$$v = \frac{q}{C} \quad (7.15)$$

From the Kirchhoff's loop rule, the voltage across the source and the capacitor are equal,

$$v_m \sin \omega t = \frac{q}{C}$$

To find the current, we use the relation $i = \frac{dq}{dt}$

$$i = \frac{d}{dt}(v_m C \sin \omega t) = \omega C v_m \cos(\omega t)$$

Using the relation, $\cos(\omega t) = \sin\left(\omega t + \frac{\pi}{2}\right)$, we have

$$i = i_m \sin\left(\omega t + \frac{\pi}{2}\right) \quad (7.16)$$

where the amplitude of the oscillating current is $i_m = \omega C v_m$. We can rewrite it as

$$i_m = \frac{v_m}{(1/\omega C)}$$

Comparing it to $i_m = v_m/R$ for a purely resistive circuit, we find that $(1/\omega C)$ plays the role of resistance. It is called *capacitive reactance* and is denoted by X_c ,

$$X_c = 1/\omega C \quad (7.17)$$

so that the amplitude of the current is

$$i_m = \frac{v_m}{X_c} \quad (7.18)$$



FIGURE 7.8 An ac source connected to a capacitor.

Physics

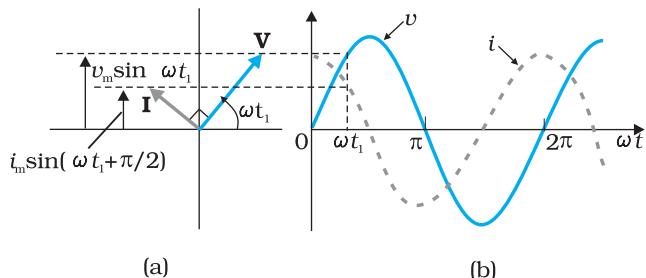


FIGURE 7.9 (a) A Phasor diagram for the circuit in Fig. 7.8. (b) Graph of v and i versus ωt .

Figure 7.9(a) shows the phasor diagram at an instant t_1 . Here the current phasor \mathbf{I} is $\pi/2$ ahead of the voltage phasor \mathbf{V} as they rotate counterclockwise. Figure 7.9(b) shows the variation of voltage and current with time. We see that the current reaches its maximum value earlier than the voltage by one-fourth of a period.

The instantaneous power supplied to the capacitor is

$$\begin{aligned} p_c &= i v = i_m \cos(\omega t) v_m \sin(\omega t) \\ &= i_m v_m \cos(\omega t) \sin(\omega t) \\ &= \frac{i_m v_m}{2} \sin(2\omega t) \end{aligned} \quad (7.19)$$

So, as in the case of an inductor, the average power

$$P_C = \left\langle \frac{i_m v_m}{2} \sin(2\omega t) \right\rangle = \frac{i_m v_m}{2} \langle \sin(2\omega t) \rangle = 0$$

since $\langle \sin(2\omega t) \rangle = 0$ over a complete cycle. Figure 7.10 explains it in detail. Thus, we see that in the case of an inductor, the current lags the voltage by $\pi/2$ and in the case of a capacitor, the current leads the voltage by $\pi/2$.

EXAMPLE 7.3

Example 7.3 A lamp is connected in series with a capacitor. Predict your observations for dc and ac connections. What happens in each case if the capacitance of the capacitor is reduced?

Solution When a dc source is connected to a capacitor, the capacitor gets charged and after charging no current flows in the circuit and the lamp will not glow. There will be no change even if C is reduced. With ac source, the capacitor offers capacitative reactance ($1/\omega C$) and the current flows in the circuit. Consequently, the lamp will shine. Reducing C will increase reactance and the lamp will shine less brightly than before.

EXAMPLE 7.4

Example 7.4 A $15.0 \mu\text{F}$ capacitor is connected to a 220 V , 50 Hz source. Find the capacitive reactance and the current (rms and peak) in the circuit. If the frequency is doubled, what happens to the capacitive reactance and the current?

Solution The capacitive reactance is

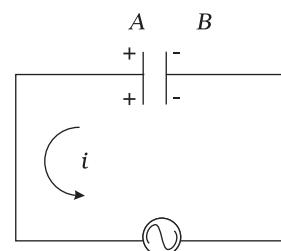
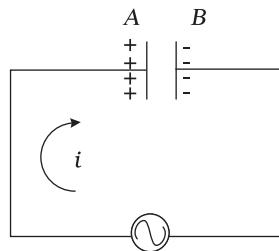
$$X_C = \frac{1}{2\pi f C} = \frac{1}{2\pi(50\text{Hz})(15.0 \times 10^{-6}\text{F})} = 212 \Omega$$

The rms current is

The dimension of capacitive reactance is the same as that of resistance and its SI unit is ohm (Ω). The capacitive reactance limits the amplitude of the current in a purely capacitive circuit in the same way as the resistance limits the current in a purely resistive circuit. But it is inversely proportional to the frequency and the capacitance.

A comparison of Eq. (7.16) with the equation of source voltage, Eq. (7.1) shows that the current is $\pi/2$ ahead of voltage.

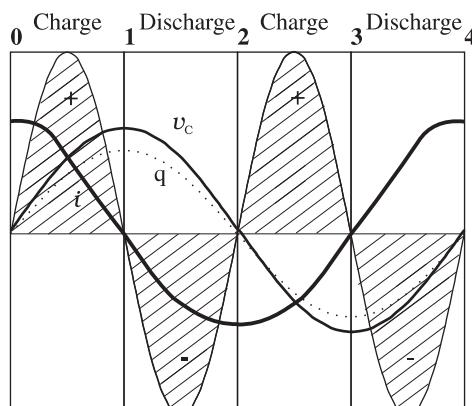
Alternating Current



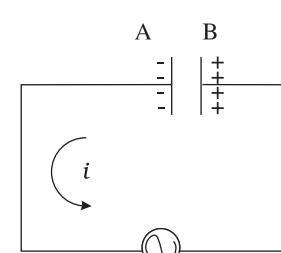
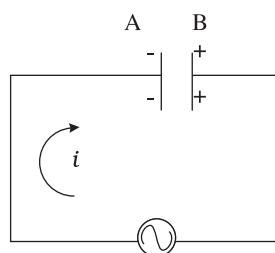
0-1 The current i flows as shown and from the maximum at 0, reaches a zero value at 1. The plate A is charged to positive polarity while negative charge q builds up in B reaching a maximum at 1 until the current becomes zero. The voltage $v_c = q/C$ is in phase with q and reaches maximum value at 1. Current and voltage are both positive. So $p = v_c i$ is positive. ENERGY IS ABSORBED FROM THE SOURCE DURING THIS QUARTER CYCLE AS THE CAPACITOR IS CHARGED.

1-2 The current i reverses its direction. The accumulated charge is depleted i.e., the capacitor is discharged during this quarter cycle. The voltage gets reduced but is still positive. The current is negative. Their product, the power is negative.

THE ENERGY ABSORBED DURING THE 1/4 CYCLE **0-1** IS RETURNED DURING THIS QUARTER.



One complete cycle of voltage/current. Note that the current leads the voltage.



2-3 As i continues to flow from A to B, the capacitor is charged to reversed polarity i.e., the plate B acquires positive and A acquires negative charge. Both the current and the voltage are negative. Their product p is positive. The capacitor ABSORBS ENERGY during this 1/4 cycle.

3-4 The current i reverses its direction at **3** and flows from B to A. The accumulated charge is depleted and the magnitude of the voltage v_c is reduced. v_c becomes zero at **4** when the capacitor is fully discharged. The power is negative. ENERGY ABSORBED DURING **2-3** IS RETURNED TO THE SOURCE. NET ENERGY ABSORBED IS ZERO.

FIGURE 7.10 Charging and discharging of a capacitor.

EXAMPLE 7.4

$$I = \frac{V}{X_C} = \frac{220\text{ V}}{212\ \Omega} = 1.04\text{ A}$$

The peak current is

$$i_m = \sqrt{2}I = (1.41)(1.04\text{ A}) = 1.47\text{ A}$$

This current oscillates between $+1.47\text{ A}$ and -1.47 A , and is ahead of the voltage by $\pi/2$.

If the frequency is doubled, the capacitive reactance is halved and consequently, the current is doubled.

EXAMPLE 7.5

Example 7.5 A light bulb and an open coil inductor are connected to an ac source through a key as shown in Fig. 7.11.

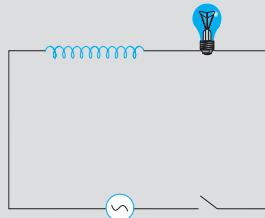


FIGURE 7.11

The switch is closed and after sometime, an iron rod is inserted into the interior of the inductor. The glow of the light bulb (a) increases; (b) decreases; (c) is unchanged, as the iron rod is inserted. Give your answer with reasons.

Solution As the iron rod is inserted, the magnetic field inside the coil magnetizes the iron increasing the magnetic field inside it. Hence, the inductance of the coil increases. Consequently, the inductive reactance of the coil increases. As a result, a larger fraction of the applied ac voltage appears across the inductor, leaving less voltage across the bulb. Therefore, the glow of the light bulb decreases.

7.6 AC VOLTAGE APPLIED TO A SERIES LCR CIRCUIT

Figure 7.12 shows a series LCR circuit connected to an ac source ε . As usual, we take the voltage of the source to be $v = v_m \sin \omega t$.

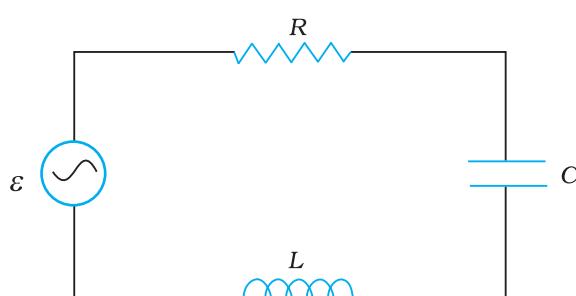


FIGURE 7.12 A series LCR circuit connected to an ac source.

If q is the charge on the capacitor and i the current, at time t , we have, from Kirchhoff's loop rule:

$$L \frac{di}{dt} + iR + \frac{q}{C} = v \quad (7.20)$$

We want to determine the instantaneous current i and its phase relationship to the applied alternating voltage v . We shall solve this problem by two methods. First, we use the technique of phasors and in the second method, we solve Eq. (7.20) analytically to obtain the time-dependence of i .

7.6.1 Phasor-diagram solution

From the circuit shown in Fig. 7.12, we see that the resistor, inductor and capacitor are in series. Therefore, the ac current in each element is the same at any time, having the same amplitude and phase. Let it be

$$i = i_m \sin(\omega t + \phi) \quad (7.21)$$

where ϕ is the phase difference between the voltage across the source and the current in the circuit. On the basis of what we have learnt in the previous sections, we shall construct a phasor diagram for the present case.

Let \mathbf{I} be the phasor representing the current in the circuit as given by Eq. (7.21). Further, let \mathbf{V}_L , \mathbf{V}_R , \mathbf{V}_C , and \mathbf{V} represent the voltage across the inductor, resistor, capacitor and the source, respectively. From previous section, we know that \mathbf{V}_R is parallel to \mathbf{I} , \mathbf{V}_C is $\pi/2$ behind \mathbf{I} and \mathbf{V}_L is $\pi/2$ ahead of \mathbf{I} . \mathbf{V}_L , \mathbf{V}_R , \mathbf{V}_C and \mathbf{I} are shown in Fig. 7.13(a) with appropriate phase-relations.

The length of these phasors or the amplitude of \mathbf{V}_R , \mathbf{V}_C and \mathbf{V}_L are:

$$v_{Rm} = i_m R, v_{Cm} = i_m X_C, v_{Lm} = i_m X_L \quad (7.22)$$

The voltage Equation (7.20) for the circuit can be written as

$$v_L + v_R + v_C = v \quad (7.23)$$

The phasor relation whose vertical component gives the above equation is

$$\mathbf{V}_L + \mathbf{V}_R + \mathbf{V}_C = \mathbf{V} \quad (7.24)$$

This relation is represented in Fig. 7.13(b). Since \mathbf{V}_C and \mathbf{V}_L are always along the same line and in opposite directions, they can be combined into a single phasor $(\mathbf{V}_C + \mathbf{V}_L)$ which has a magnitude $|v_{Cm} - v_{Lm}|$. Since \mathbf{V} is represented as the hypotenuse of a right-triangle whose sides are \mathbf{V}_R and $(\mathbf{V}_C + \mathbf{V}_L)$, the pythagorean theorem gives:

$$v_m^2 = v_{Rm}^2 + (v_{Cm} - v_{Lm})^2$$

Substituting the values of v_{Rm} , v_{Cm} , and v_{Lm} from Eq. (7.22) into the above equation, we have

$$\begin{aligned} v_m^2 &= (i_m R)^2 + (i_m X_C - i_m X_L)^2 \\ &= i_m^2 [R^2 + (X_C - X_L)^2] \end{aligned}$$

$$\text{or, } i_m = \frac{v_m}{\sqrt{R^2 + (X_C - X_L)^2}} \quad [7.25(a)]$$

By analogy to the resistance in a circuit, we introduce the *impedance Z* in an ac circuit:

$$i_m = \frac{v_m}{Z} \quad [7.25(b)]$$

$$\text{where } Z = \sqrt{R^2 + (X_C - X_L)^2} \quad (7.26)$$

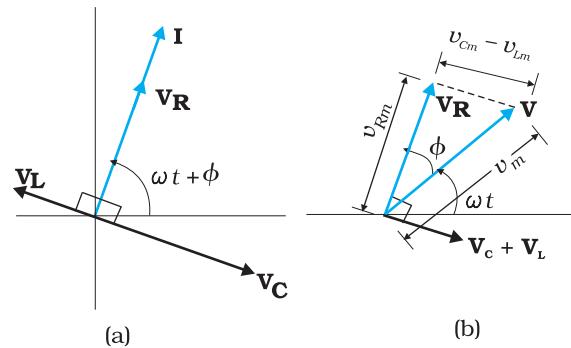


FIGURE 7.13 (a) Relation between the phasors \mathbf{V}_L , \mathbf{V}_R , \mathbf{V}_C , and \mathbf{I} , (b) Relation between the phasors \mathbf{V}_L , \mathbf{V}_R , and $(\mathbf{V}_L + \mathbf{V}_C)$ for the circuit in Fig. 7.11.

Physics

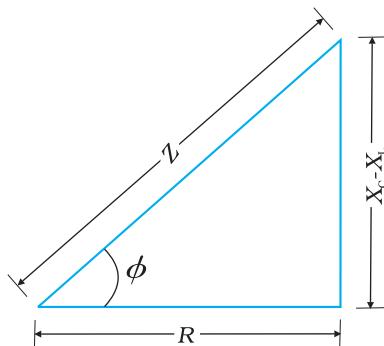


FIGURE 7.14 Impedance diagram.

Since phasor \mathbf{I} is always parallel to phasor \mathbf{V}_R , the phase angle ϕ is the angle between \mathbf{V}_R and \mathbf{V} and can be determined from Fig. 7.14:

$$\tan \phi = \frac{v_{Cm} - v_{Lm}}{v_{Rm}}$$

Using Eq. (7.22), we have

$$\tan \phi = \frac{X_C - X_L}{R} \quad (7.27)$$

Equations (7.26) and (7.27) are graphically shown in Fig. (7.14). This is called *Impedance diagram* which is a right-triangle with Z as its hypotenuse.

Equation 7.25(a) gives the amplitude of the current and Eq. (7.27) gives the phase angle. With these, Eq. (7.21) is completely specified.

If $X_C > X_L$, ϕ is positive and the circuit is predominantly capacitive. Consequently, the current in the circuit leads the source voltage. If $X_C < X_L$, ϕ is negative and the circuit is predominantly inductive. Consequently, the current in the circuit lags the source voltage.

Figure 7.15 shows the phasor diagram and variation of v and i with ωt for the case $X_C > X_L$.

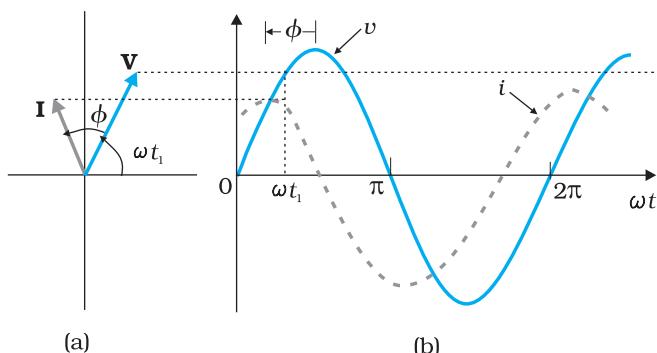


FIGURE 7.15 (a) Phasor diagram of \mathbf{V} and \mathbf{I} . (b) Graphs of v and i versus ωt for a series LCR circuit where $X_C > X_L$.

Thus, we have obtained the amplitude and phase of current for an LCR series circuit using the technique of phasors. But this method of analysing ac circuits suffers from certain disadvantages. First, the phasor diagram say nothing about the initial condition. One can take any arbitrary value of t (say, t_1 , as done throughout this chapter) and draw different phasors which show the relative angle between different phasors. The solution so obtained is called the *steady-state solution*. This is not a general solution. Additionally, we do have a *transient solution* which exists even for $v = 0$. The general solution is the sum of the transient solution and the steady-state

solution. After a sufficiently long time, the effects of the transient solution die out and the behaviour of the circuit is described by the steady-state solution.

7.6.2 Analytical solution

The voltage equation for the circuit is

$$L \frac{di}{dt} + Ri + \frac{q}{C} = v$$

$$= v_m \sin \omega t$$

We know that $i = dq/dt$. Therefore, $di/dt = d^2q/dt^2$. Thus, in terms of q , the voltage equation becomes

Alternating Current

$$L \frac{d^2q}{dt^2} + R \frac{dq}{dt} + \frac{q}{C} = v_m \sin \omega t \quad (7.28)$$

This is like the equation for a forced, damped oscillator, [see Eq. {14.37(b)} in Class XI Physics Textbook]. Let us assume a solution

$$q = q_m \sin (\omega t + \theta) \quad [7.29(a)]$$

$$\text{so that } \frac{dq}{dt} = q_m \omega \cos(\omega t + \theta) \quad [7.29(b)]$$

$$\text{and } \frac{d^2q}{dt^2} = -q_m \omega^2 \sin(\omega t + \theta) \quad [7.29(c)]$$

Substituting these values in Eq. (7.28), we get

$$q_m \omega [R \cos(\omega t + \theta) + (X_C - X_L) \sin(\omega t + \theta)] = v_m \sin \omega t \quad (7.30)$$

where we have used the relation $X_c = 1/\omega C$, $X_L = \omega L$. Multiplying and dividing Eq. (7.30) by $Z = \sqrt{R^2 + (X_C - X_L)^2}$, we have

$$q_m \omega Z \left[\frac{R}{Z} \cos(\omega t + \theta) + \frac{(X_C - X_L)}{Z} \sin(\omega t + \theta) \right] = v_m \sin \omega t \quad (7.31)$$

Now, let $\frac{R}{Z} = \cos \phi$

$$\text{and } \frac{(X_C - X_L)}{Z} = \sin \phi$$

$$\text{so that } \phi = \tan^{-1} \frac{X_C - X_L}{R} \quad (7.32)$$

Substituting this in Eq. (7.31) and simplifying, we get:

$$q_m \omega Z \cos(\omega t + \theta - \phi) = v_m \sin \omega t \quad (7.33)$$

Comparing the two sides of this equation, we see that

$$v_m = q_m \omega Z = i_m Z$$

where

$$i_m = q_m \omega \quad [7.33(a)]$$

$$\text{and } \theta - \phi = -\frac{\pi}{2} \text{ or } \theta = -\frac{\pi}{2} + \phi \quad [7.33(b)]$$

Therefore, the current in the circuit is

$$\begin{aligned} i &= \frac{dq}{dt} = q_m \omega \cos(\omega t + \theta) \\ &= i_m \cos(\omega t + \theta) \\ \text{or } i &= i_m \sin(\omega t + \phi) \end{aligned} \quad (7.34)$$

$$\text{where } i_m = \frac{v_m}{Z} = \frac{v_m}{\sqrt{R^2 + (X_C - X_L)^2}} \quad [7.34(a)]$$

$$\text{and } \phi = \tan^{-1} \frac{X_C - X_L}{R}$$

Physics

Thus, the analytical solution for the amplitude and phase of the current in the circuit agrees with that obtained by the technique of phasors.

7.6.3 Resonance

An interesting characteristic of the series RLC circuit is the phenomenon of resonance. The phenomenon of resonance is common among systems that have a tendency to oscillate at a particular frequency. This frequency is called the system's *natural frequency*. If such a system is driven by an energy source at a frequency that is near the natural frequency, the amplitude of oscillation is found to be large. A familiar example of this phenomenon is a child on a swing. The swing has a natural frequency for swinging back and forth like a pendulum. If the child pulls on the rope at regular intervals and the frequency of the pulls is almost the same as the frequency of swinging, the amplitude of the swinging will be large (Chapter 14, Class XI).

For an RLC circuit driven with voltage of amplitude v_m and frequency ω , we found that the current amplitude is given by

$$i_m = \frac{v_m}{Z} = \frac{v_m}{\sqrt{R^2 + (X_C - X_L)^2}}$$

with $X_c = 1/\omega C$ and $X_L = \omega L$. So if ω is varied, then at a particular frequency ω_0 , $X_c = X_L$, and the impedance is minimum ($Z = \sqrt{R^2 + 0^2} = R$). This frequency is called the *resonant frequency*:

$$X_c = X_L \text{ or } \frac{1}{\omega_0 C} = \omega_0 L$$

$$\text{or } \omega_0 = \frac{1}{\sqrt{LC}} \quad (7.35)$$

At resonant frequency, the current amplitude is maximum; $i_m = v_m/R$.

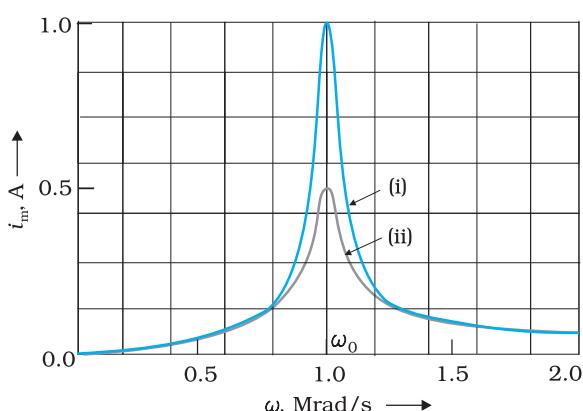


FIGURE 7.16 Variation of i_m with ω for two cases: (i) $R = 100 \Omega$, (ii) $R = 200 \Omega$, $L = 1.00 \text{ mH}$.

Figure 7.16 shows the variation of i_m with ω in a RLC series circuit with $L = 1.00 \text{ mH}$, $C = 1.00 \text{ nF}$ for two values of R : (i) $R = 100 \Omega$ and (ii) $R = 200 \Omega$. For the source applied $v_m = 100 \text{ V}$. ω_0 for this case is $\left(\frac{1}{\sqrt{LC}}\right) = 1.00 \times 10^6 \text{ rad/s}$.

We see that the current amplitude is maximum at the resonant frequency. Since $i_m = v_m / R$ at resonance, the current amplitude for case (i) is twice to that for case (ii).

Resonant circuits have a variety of applications, for example, in the tuning mechanism of a radio or a TV set. The antenna of a radio accepts signals from many broadcasting stations. The signals picked up in the antenna acts as a source in the tuning circuit of the radio, so the circuit can be driven at many frequencies.

Alternating Current

But to hear one particular radio station, we tune the radio. In tuning, we vary the capacitance of a capacitor in the tuning circuit such that the resonant frequency of the circuit becomes nearly equal to the frequency of the radio signal received. When this happens, the amplitude of the current with the frequency of the signal of the particular radio station in the circuit is maximum.

It is important to note that resonance phenomenon is exhibited by a circuit only if both L and C are present in the circuit. Only then do the voltages across L and C cancel each other (both being out of phase) and the current amplitude is v_m/R , the total source voltage appearing across R. This means that we cannot have resonance in a RL or RC circuit.

Sharpness of resonance

The amplitude of the current in the series LCR circuit is given by

$$i_m = \frac{v_m}{\sqrt{R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2}}$$

and is maximum when $\omega = \omega_0 = 1/\sqrt{LC}$. The maximum value is

$$i_m^{\max} = v_m / R.$$

For values of ω other than ω_0 , the amplitude of the current is less than the maximum value. Suppose we choose a value of ω for which the current amplitude is $1/\sqrt{2}$ times its maximum value. At this value, the power dissipated by the circuit becomes half. From the curve in Fig. (7.16), we see that there are two such values of ω , say, ω_1 and ω_2 , one greater and the other smaller than ω_0 and symmetrical about ω_0 . We may write

$$\omega_1 = \omega_0 + \Delta\omega$$

$$\omega_2 = \omega_0 - \Delta\omega$$

The difference $\omega_1 - \omega_2 = 2\Delta\omega$ is often called the *bandwidth* of the circuit. The quantity $(\omega_0 / 2\Delta\omega)$ is regarded as a measure of the sharpness of resonance. The smaller the $\Delta\omega$, the sharper or narrower is the resonance. To get an expression for $\Delta\omega$, we note that the current amplitude i_m is $(1/\sqrt{2})i_m^{\max}$ for $\omega_1 = \omega_0 + \Delta\omega$. Therefore,

$$\text{at } \omega_1, \quad i_m = \frac{v_m}{\sqrt{R^2 + \left(\omega_1 L - \frac{1}{\omega_1 C}\right)^2}}$$

$$= \frac{i_m^{\max}}{\sqrt{2}} = \frac{v_m}{R\sqrt{2}}$$

Physics

$$\text{or } \sqrt{R^2 + \left(\omega_1 L - \frac{1}{\omega_1 C} \right)^2} = R\sqrt{2}$$

$$\text{or } R^2 + \left(\omega_1 L - \frac{1}{\omega_1 C} \right)^2 = 2R^2$$

$$\omega_1 L - \frac{1}{\omega_1 C} = R$$

which may be written as,

$$(\omega_0 + \Delta\omega)L - \frac{1}{(\omega_0 + \Delta\omega)C} = R$$

$$\omega_0 L \left(1 + \frac{\Delta\omega}{\omega_0} \right) - \frac{1}{\omega_0 C \left(1 + \frac{\Delta\omega}{\omega_0} \right)} = R$$

Using $\omega_0^2 = \frac{1}{LC}$ in the second term on the left hand side, we get

$$\omega_0 L \left(1 + \frac{\Delta\omega}{\omega_0} \right) - \frac{\omega_0 L}{\left(1 + \frac{\Delta\omega}{\omega_0} \right)} = R$$

We can approximate $\left(1 + \frac{\Delta\omega}{\omega_0} \right)^{-1}$ as $\left(1 - \frac{\Delta\omega}{\omega_0} \right)$ since $\frac{\Delta\omega}{\omega_0} \ll 1$. Therefore,

$$\omega_0 L \left(1 + \frac{\Delta\omega}{\omega_0} \right) - \omega_0 L \left(1 - \frac{\Delta\omega}{\omega_0} \right) = R$$

$$\text{or } \omega_0 L \frac{2\Delta\omega}{\omega_0} = R$$

$$\Delta\omega = \frac{R}{2L} \quad [7.36(a)]$$

The sharpness of resonance is given by,

$$\frac{\omega_0}{2\Delta\omega} = \frac{\omega_0 L}{R} \quad [7.36(b)]$$

The ratio $\frac{\omega_0 L}{R}$ is also called the *quality factor*, Q of the circuit.

$$Q = \frac{\omega_0 L}{R} \quad [7.36(c)]$$

From Eqs. [7.36 (b)] and [7.36 (c)], we see that $2\Delta\omega = \frac{\omega_0}{Q}$. So, larger the

Alternating Current

value of Q , the smaller is the value of $2\Delta\omega$ or the bandwidth and sharper is the resonance. Using $\omega_0^2 = 1/LC$, Eq. [7.36(c)] can be equivalently expressed as $Q = 1/\omega_0 CR$.

We see from Fig. 7.15, that if the resonance is less sharp, not only is the maximum current less, the circuit is close to resonance for a larger range $\Delta\omega$ of frequencies and the tuning of the circuit will not be good. So, less sharp the resonance, less is the selectivity of the circuit or vice versa. From Eq. (7.36), we see that if quality factor is large, i.e., R is low or L is large, the circuit is more selective.

Example 7.6 A resistor of 200Ω and a capacitor of $15.0 \mu F$ are connected in series to a $220 V$, 50 Hz ac source. (a) Calculate the current in the circuit; (b) Calculate the voltage (rms) across the resistor and the capacitor. Is the algebraic sum of these voltages more than the source voltage? If yes, resolve the paradox.

Solution

Given

$$R = 200 \Omega, C = 15.0 \mu F = 15.0 \times 10^{-6} F$$

$$V = 220 V, \nu = 50 \text{ Hz}$$

(a) In order to calculate the current, we need the impedance of the circuit. It is

$$\begin{aligned} Z &= \sqrt{R^2 + X_C^2} = \sqrt{R^2 + (2\pi\nu C)^{-2}} \\ &= \sqrt{(200 \Omega)^2 + (2 \times 3.14 \times 50 \times 10^{-6} F)^{-2}} \\ &= \sqrt{(200 \Omega)^2 + (212 \Omega)^2} \\ &= 291.5 \Omega \end{aligned}$$

Therefore, the current in the circuit is

$$I = \frac{V}{Z} = \frac{220 V}{291.5 \Omega} = 0.755 A$$

(b) Since the current is the same throughout the circuit, we have

$$V_R = IR = (0.755 A)(200 \Omega) = 151 V$$

$$V_C = IX_C = (0.755 A)(212.3 \Omega) = 160.3 V$$

The algebraic sum of the two voltages, V_R and V_C is $311.3 V$ which is more than the source voltage of $220 V$. How to resolve this paradox? As you have learnt in the text, the two voltages are not in the same phase. Therefore, *they cannot be added like ordinary numbers*. The two voltages are out of phase by ninety degrees. Therefore, the total of these voltages must be obtained using the Pythagorean theorem:

$$\begin{aligned} V_{R+C} &= \sqrt{V_R^2 + V_C^2} \\ &= 220 V \end{aligned}$$

Thus, if the phase difference between two voltages is properly taken into account, the total voltage across the resistor and the capacitor is equal to the voltage of the source.

7.7 POWER IN AC CIRCUIT: THE POWER FACTOR

We have seen that a voltage $v = v_m \sin \omega t$ applied to a series RLC circuit drives a current in the circuit given by $i = i_m \sin(\omega t + \phi)$ where

$$i_m = \frac{v_m}{Z} \quad \text{and} \quad \phi = \tan^{-1} \left(\frac{X_C - X_L}{R} \right)$$

Therefore, the instantaneous power p supplied by the source is

$$\begin{aligned} p &= vi = (v_m \sin \omega t) \times [i_m \sin(\omega t + \phi)] \\ &= \frac{v_m i_m}{2} [\cos \phi - \cos(2\omega t + \phi)] \end{aligned} \quad (7.37)$$

The average power over a cycle is given by the average of the two terms in R.H.S. of Eq. (7.37). It is only the second term which is time-dependent. Its average is zero (the positive half of the cosine cancels the negative half). Therefore,

$$\begin{aligned} P &= \frac{v_m i_m}{2} \cos \phi = \frac{v_m}{\sqrt{2}} \frac{i_m}{\sqrt{2}} \cos \phi \\ &= VI \cos \phi \end{aligned} \quad [7.38(a)]$$

This can also be written as,

$$P = I^2 Z \cos \phi \quad [7.38(b)]$$

So, the average power dissipated depends not only on the voltage and current but also on the cosine of the phase angle ϕ between them. The quantity $\cos \phi$ is called the *power factor*. Let us discuss the following cases:

Case (i) Resistive circuit: If the circuit contains only pure R , it is called *resistive*. In that case $\phi = 0$, $\cos \phi = 1$. There is maximum power dissipation.

Case (ii) Purely inductive or capacitive circuit: If the circuit contains only an inductor or capacitor, we know that the phase difference between voltage and current is $\pi/2$. Therefore, $\cos \phi = 0$, and no power is dissipated even though a current is flowing in the circuit. This current is sometimes referred to as *wattless current*.

Case (iii) LCR series circuit: In an LCR series circuit, power dissipated is given by Eq. (7.38) where $\phi = \tan^{-1}(X_C - X_L)/R$. So, ϕ may be non-zero in a RL or RC or RCL circuit. Even in such cases, power is dissipated only in the resistor.

Case (iv) Power dissipated at resonance in LCR circuit: At resonance $X_C - X_L = 0$, and $\phi = 0$. Therefore, $\cos \phi = 1$ and $P = I^2 Z = I^2 R$. That is, maximum power is dissipated in a circuit (through R) at resonance.

EXAMPLE 7.7

(a) For circuits used for transporting electric power, a low power factor implies large power loss in transmission. Explain.

(b) Power factor can often be improved by the use of a capacitor of appropriate capacitance in the circuit. Explain.

Alternating Current

Solution (a) We know that $P = I V \cos\phi$ where $\cos\phi$ is the power factor. To supply a given power at a given voltage, if $\cos\phi$ is small, we have to increase current accordingly. But this will lead to large power loss ($I^2 R$) in transmission.

(b) Suppose in a circuit, current I lags the voltage by an angle ϕ . Then power factor $\cos\phi = R/Z$.

We can improve the power factor (tending to 1) by making Z tend to R . Let us understand, with the help of a phasor diagram (Fig. 7.17) how this can be achieved. Let us resolve \mathbf{I} into two components. \mathbf{I}_p along

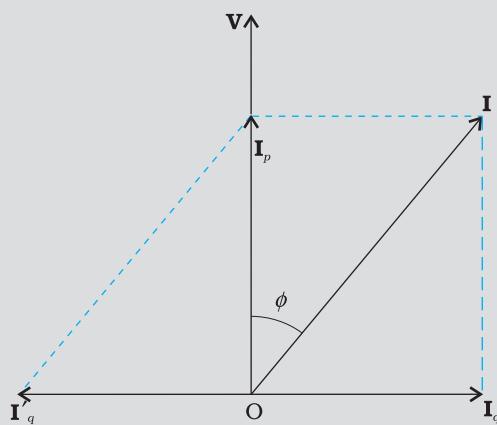


FIGURE 7.17

the applied voltage \mathbf{V} and \mathbf{I}_q perpendicular to the applied voltage. \mathbf{I}_q as you have learnt in Section 7.7, is called the wattless component since corresponding to this component of current, there is no power loss. \mathbf{I}_p is known as the power component because it is in phase with the voltage and corresponds to power loss in the circuit.

It's clear from this analysis that if we want to improve power factor, we must completely neutralize the lagging wattless current \mathbf{I}_q by an equal leading wattless current \mathbf{I}'_q . This can be done by connecting a capacitor of appropriate value in parallel so that \mathbf{I}_q and \mathbf{I}'_q cancel each other and P is effectively $I_p V$.

EXAMPLE 7.7

Example 7.8 A sinusoidal voltage of peak value 283 V and frequency 50 Hz is applied to a series LCR circuit in which $R = 3 \Omega$, $L = 25.48 \text{ mH}$, and $C = 796 \mu\text{F}$. Find (a) the impedance of the circuit; (b) the phase difference between the voltage across the source and the current; (c) the power dissipated in the circuit; and (d) the power factor.

Solution

(a) To find the impedance of the circuit, we first calculate X_L and X_C .

$$X_L = 2\pi fL \\ = 2 \times 3.14 \times 50 \times 25.48 \times 10^{-3} \Omega = 8 \Omega$$

$$X_C = \frac{1}{2\pi fC}$$

EXAMPLE 7.8

EXAMPLE 7.8

$$= \frac{1}{2 \times 3.14 \times 50 \times 796 \times 10^{-6}} = 4\Omega$$

Therefore,

$$Z = \sqrt{R^2 + (X_L - X_C)^2} = \sqrt{3^2 + (8 - 4)^2} \\ = 5 \Omega$$

$$(b) \text{ Phase difference, } \phi = \tan^{-1} \frac{X_C - X_L}{R}$$

$$= \tan^{-1} \left(\frac{4 - 8}{3} \right) = -53.1^\circ$$

Since ϕ is negative, the current in the circuit lags the voltage across the source.

$$(c) \text{ The power dissipated in the circuit is}$$

$$P = I^2 R$$

$$\text{Now, } I = \frac{i_m}{\sqrt{2}} = \frac{1}{\sqrt{2}} \left(\frac{283}{5} \right) = 40A$$

$$\text{Therefore, } P = (40A)^2 \times 3\Omega = 4800W$$

$$(d) \text{ Power factor} = \cos \phi = \cos 53.1^\circ = 0.6$$

Example 7.9 Suppose the frequency of the source in the previous example can be varied. (a) What is the frequency of the source at which resonance occurs? (b) Calculate the impedance, the current, and the power dissipated at the resonant condition.

Solution

(a) The frequency at which the resonance occurs is

$$\omega_0 = \frac{1}{\sqrt{LC}} = \frac{1}{\sqrt{25.48 \times 10^{-3} \times 796 \times 10^{-6}}} \\ = 222.1 \text{ rad/s}$$

$$v_r = \frac{\omega_0}{2\pi} = \frac{222.1}{2 \times 3.14} \text{ Hz} = 35.4 \text{ Hz}$$

(b) The impedance Z at resonant condition is equal to the resistance:

$$Z = R = 3\Omega$$

The rms current at resonance is

$$= \frac{V}{Z} = \frac{V}{R} = \left(\frac{283}{\sqrt{2}} \right) \frac{1}{3} = 66.7A$$

The power dissipated at resonance is

$$P = I^2 \times R = (66.7)^2 \times 3 = 13.35 \text{ kW}$$

You can see that in the present case, power dissipated at resonance is more than the power dissipated in Example 7.8.

EXAMPLE 7.9

Example 7.10 At an airport, a person is made to walk through the doorway of a metal detector, for security reasons. If she/he is carrying anything made of metal, the metal detector emits a sound. On what principle does this detector work?

Solution The metal detector works on the principle of resonance in ac circuits. When you walk through a metal detector, you are, in fact, walking through a coil of many turns. The coil is connected to a capacitor tuned so that the circuit is in resonance. When you walk through with metal in your pocket, the impedance of the circuit changes – resulting in significant change in current in the circuit. This change in current is detected and the electronic circuitry causes a sound to be emitted as an alarm.

EXAMPLE 7.10

7.8 LC OSCILLATIONS

We know that a capacitor and an inductor can store electrical and magnetic energy, respectively. When a capacitor (initially charged) is connected to an inductor, the charge on the capacitor and the current in the circuit exhibit the phenomenon of electrical oscillations similar to oscillations in mechanical systems (Chapter 14, Class XI).

Let a capacitor be charged q_m (at $t = 0$) and connected to an inductor as shown in Fig. 7.18.

The moment the circuit is completed, the charge on the capacitor starts decreasing, giving rise to current in the circuit. Let q and i be the charge and current in the circuit at time t . Since di/dt is positive, the induced emf in L will have polarity as shown, i.e., $v_b < v_a$. According to Kirchhoff's loop rule,

$$\frac{q}{C} - L \frac{di}{dt} = 0 \quad (7.39)$$

$i = -(dq/dt)$ in the present case (as q decreases, i increases). Therefore, Eq. (7.39) becomes:

$$\frac{d^2q}{dt^2} + \frac{1}{LC}q = 0 \quad (7.40)$$

This equation has the form $\frac{d^2x}{dt^2} + \omega_0^2 x = 0$ for a simple harmonic oscillator. The charge, therefore, oscillates with a natural frequency

$$\omega_0 = \frac{1}{\sqrt{LC}} \quad (7.41)$$

and varies sinusoidally with time as

$$q = q_m \cos(\omega_0 t + \phi) \quad (7.42)$$

where q_m is the maximum value of q and ϕ is a phase constant. Since $q = q_m$ at $t = 0$, we have $\cos \phi = 1$ or $\phi = 0$. Therefore, in the present case,

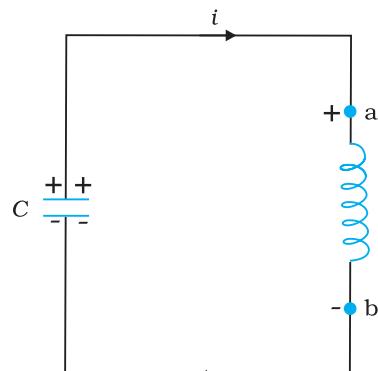


FIGURE 7.18 At the instant shown, the current is increasing so the polarity of induced emf in the inductor is as shown.

Physics

$$q = q_m \cos(\omega_0 t) \quad (7.43)$$

The current $i \left(= -\frac{dq}{dt}\right)$ is given by

$$i = i_m \sin(\omega_0 t) \quad (7.44)$$

where $i_m = \omega_0 q_m$

Let us now try to visualise how this oscillation takes place in the circuit.

Figure 7.19(a) shows a capacitor with initial charge q_m connected to an ideal inductor. The electrical energy stored in the charged capacitor is

$U_E = \frac{1}{2} \frac{q_m^2}{C}$. Since, there is no current in the circuit, energy in the inductor is zero. Thus, the total energy of LC circuit is,

$$U = U_E = \frac{1}{2} \frac{q_m^2}{C}$$

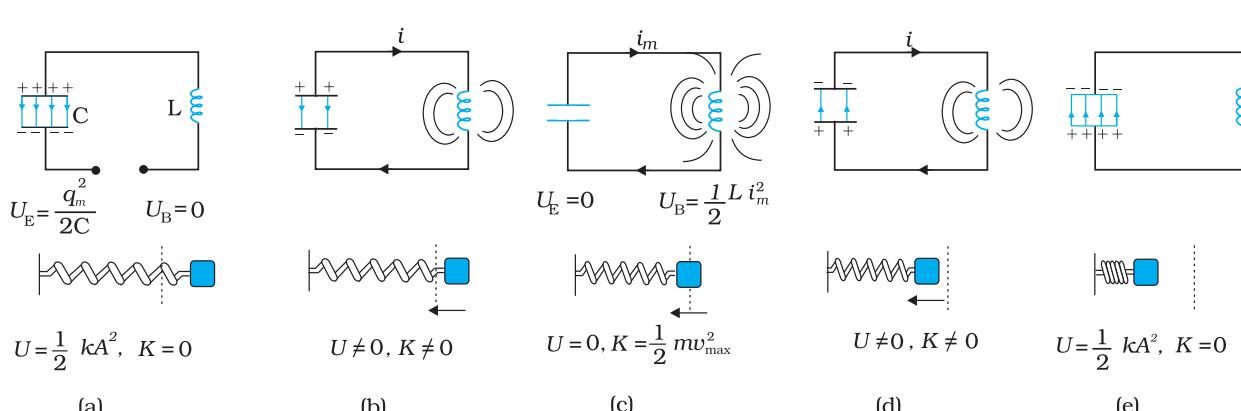


FIGURE 7.19 The oscillations in an LC circuit are analogous to the oscillation of a block at the end of a spring. The figure depicts one-half of a cycle.

At $t = 0$, the switch is closed and the capacitor starts to discharge [Fig. 7.19(b)]. As the current increases, it sets up a magnetic field in the inductor and thereby, some energy gets stored in the inductor in the form of magnetic energy: $U_B = (1/2) L i^2$. As the current reaches its maximum value i_m , (at $t = T/4$) as in Fig. 7.19(c), all the energy is stored in the magnetic field: $U_B = (1/2) L i_m^2$. You can easily check that the maximum electrical energy equals the maximum magnetic energy. The capacitor now has no charge and hence no energy. The current now starts charging the capacitor, as in Fig. 7.19(d). This process continues till the capacitor is fully charged (at $t = T/2$) [Fig. 7.19(e)]. But it is charged with a polarity opposite to its initial state in Fig. 7.19(a). The whole process just described will now repeat itself till the system reverts to its original state. Thus, the energy in the system oscillates between the capacitor and the inductor.

Alternating Current

The *LC* oscillation is similar to the mechanical oscillation of a block attached to a spring. The lower part of each figure in Fig. 7.19 depicts the corresponding stage of a mechanical system (a block attached to a spring). As noted earlier, for a block of a mass m oscillating with frequency ω_0 , the equation is

$$\frac{d^2x}{dt^2} + \omega_0^2 x = 0$$

Here, $\omega_0 = \sqrt{k/m}$, and k is the spring constant. So, x corresponds to q . In case of a mechanical system $F = ma = m(dv/dt) = m(d^2x/dt^2)$. For an electrical system, $\varepsilon = -L(di/dt) = -L(d^2q/dt^2)$. Comparing these two equations, we see that L is analogous to mass m : L is a measure of resistance to change in current. In case of *LC* circuit, $\omega_0 = 1/\sqrt{LC}$ and for mass on a spring, $\omega_0 = \sqrt{k/m}$. So, $1/C$ is analogous to k . The constant k ($=F/x$) tells us the (external) force required to produce a unit displacement whereas $1/C$ ($=V/q$) tells us the potential difference required to store a unit charge. Table 7.1 gives the analogy between mechanical and electrical quantities.

TABLE 7.1 ANALOGIES BETWEEN MECHANICAL AND ELECTRICAL QUANTITIES

Mechanical system	Slectrical system
Mass m	Inductance L
Force constant k	Reciprocal capacitance $1/C$
Displacement x	Charge q
Velocity $v = dx/dt$	Current $i = dq/dt$
Mechanical energy	Electromagnetic energy
$E = \frac{1}{2}kx^2 + \frac{1}{2}mv^2$	$U = \frac{1}{2} \frac{q^2}{C} + \frac{1}{2}Li^2$

Note that the above discussion of *LC* oscillations is not realistic for two reasons:

- (i) Every inductor has some resistance. The effect of this resistance is to introduce a damping effect on the charge and current in the circuit and the oscillations finally die away.
- (ii) Even if the resistance were zero, the total energy of the system would not remain constant. It is radiated away from the system in the form of electromagnetic waves (discussed in the next chapter). In fact, radio and TV transmitters depend on this radiation.

Physics

TWO DIFFERENT PHENOMENA, SAME MATHEMATICAL TREATMENT

You may like to compare the treatment of a forced damped oscillator discussed in Section 14.10 of Class XI physics textbook, with that of an *LCR* circuit when an ac voltage is applied in it. We have already remarked that Eq. [14.37(b)] of Class XI Textbook is exactly similar to Eq. (7.28) here, although they use different symbols and parameters. Let us therefore list the equivalence between different quantities in the two situations:

Forced oscillations

$$m \frac{d^2x}{dt^2} + b \frac{dx}{dt} + kx = F \cos \omega_d t$$

Displacement, x

Time, t

Mass, m

Damping constant, b

Spring constant, k

Driving frequency, ω_d

Natural frequency of oscillations, ω

Amplitude of forced oscillations, A

Amplitude of driving force, F_0

Driven LCR circuit

$$L \frac{d^2q}{dt^2} + R \frac{dq}{dt} + \frac{q}{C} = v_m \sin \omega t$$

Charge on capacitor, q

Time, t

Self inductance, L

Resistance, R

Inverse capacitance, $1/C$

Driving frequency, ω

Natural frequency of LCR circuit, ω_0

Maximum charge stored, q_m

Amplitude of applied voltage, v_m

You must note that since x corresponds to q , the amplitude A (maximum displacement) will correspond to the maximum charge stored, q_m . Equation [14.39 (a)] of Class XI gives the amplitude of oscillations in terms of other parameters, which we reproduce here for convenience:

$$A = \frac{F_0}{\{m^2(\omega^2 - \omega_d^2)^2 + \omega_d^2 b^2\}^{1/2}}$$

Replace each parameter in the above equation by the corresponding electrical quantity, and see what happens. Eliminate L , C , ω , and ω_0 , using $X_L = \omega L$, $X_C = 1/\omega C$, and $\omega^2 = 1/LC$. When you use Eqs. (7.33) and (7.34), you will see that there is a perfect match.

You will come across numerous such situations in physics where diverse physical phenomena are represented by the same mathematical equation. If you have dealt with one of them, and you come across another situation, you may simply replace the corresponding quantities and interpret the result in the new context. We suggest that you may try to find more such parallel situations from different areas of physics. One must, of course, be aware of the differences too.

Example 7.11 Show that in the free oscillations of an *LC* circuit, the sum of energies stored in the capacitor and the inductor is constant in time.

Solution Let q_0 be the initial charge on a capacitor. Let the charged capacitor be connected to an inductor of inductance L . As you have studied in Section 7.8, this *LC* circuit will sustain an oscillation with frequency

$$\omega = 2\pi\nu = \frac{1}{\sqrt{LC}}$$

At an instant t , charge q on the capacitor and the current i are given by:

$$q(t) = q_0 \cos \omega t$$

$$i(t) = -q_0 \omega \sin \omega t$$

Energy stored in the capacitor at time t is

$$U_E = \frac{1}{2} C V^2 = \frac{1}{2} \frac{q^2}{C} = \frac{q_0^2}{2C} \cos^2(\omega t)$$

Energy stored in the inductor at time t is

$$U_M = \frac{1}{2} L i^2$$

$$= \frac{1}{2} L q_0^2 \omega^2 \sin^2(\omega t)$$

$$= \frac{q_0^2}{2C} \sin^2(\omega t) \quad (\because \omega^2 = 1/\sqrt{LC})$$

Sum of energies

$$U_E + U_M = \frac{q_0^2}{2C} [\cos^2 \omega t + \sin^2 \omega t]$$

$$= \frac{q_0^2}{2C}$$

This sum is constant in time as q_0 and C , both are time-independent. Note that it is equal to the initial energy of the capacitor. Why it is so? Think!

EXAMPLE 7.11

7.9 TRANSFORMERS

For many purposes, it is necessary to change (or transform) an alternating voltage from one to another of greater or smaller value. This is done with a device called *transformer* using the principle of mutual induction.

A transformer consists of two sets of coils, insulated from each other. They are wound on a soft-iron core, either one on top of the other as in Fig. 7.20(a) or on separate limbs of the core as in Fig. 7.20(b). One of the coils called the *primary coil* has N_p turns. The other coil is called the *secondary coil*; it has N_s turns. Often the primary coil is the input coil and the secondary coil is the output coil of the transformer.

Physics

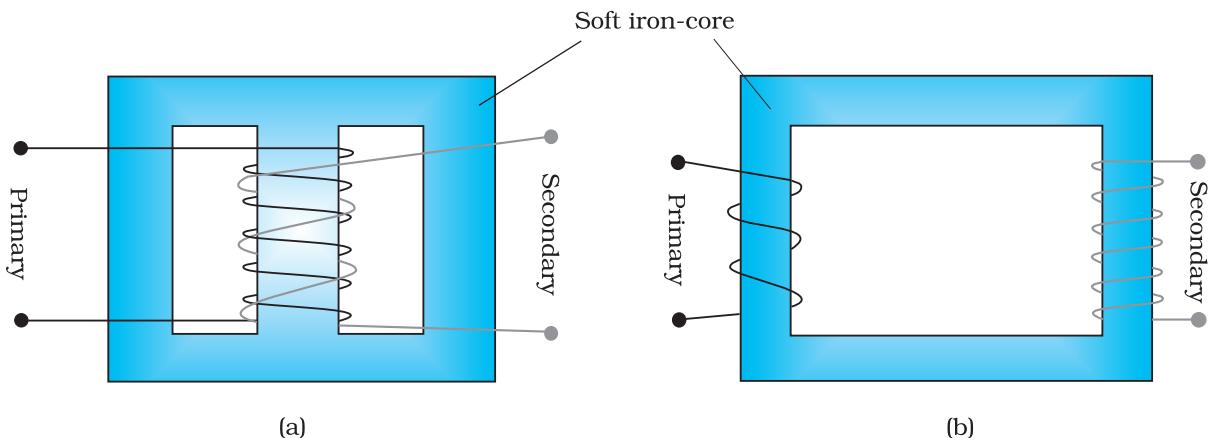


FIGURE 7.20 Two arrangements for winding of primary and secondary coil in a transformer: (a) two coils on top of each other, (b) two coils on separate limbs of the core.

When an alternating voltage is applied to the primary, the resulting current produces an alternating magnetic flux which links the secondary and induces an emf in it. The value of this emf depends on the number of turns in the secondary. We consider an ideal transformer in which the primary has negligible resistance and all the flux in the core links both primary and secondary windings. Let ϕ be the flux in each turn in the core at time t due to current in the primary when a voltage v_p is applied to it.

Then the induced emf or voltage ε_s , in the secondary with N_s turns is

$$\varepsilon_s = -N_s \frac{d\phi}{dt} \quad (7.45)$$

The alternating flux ϕ also induces an emf, called back emf in the primary. This is

$$\varepsilon_p = -N_p \frac{d\phi}{dt} \quad (7.46)$$

But $\varepsilon_p = v_p$. If this were not so, the primary current would be infinite since the primary has zero resistance (as assumed). If the secondary is an open circuit or the current taken from it is small, then to a good approximation

$$\varepsilon_s = v_s$$

where v_s is the voltage across the secondary. Therefore, Eqs. (7.45) and (7.46) can be written as

$$v_s = -N_s \frac{d\phi}{dt} \quad [7.45(a)]$$

$$v_p = -N_p \frac{d\phi}{dt} \quad [7.46(a)]$$

From Eqs. [7.45 (a)] and [7.46 (a)], we have

$$\frac{v_s}{v_p} = \frac{N_s}{N_p} \quad (7.47)$$

Alternating Current

Note that the above relation has been obtained using three assumptions: (i) the primary resistance and current are small; (ii) the same flux links both the primary and the secondary as very little flux escapes from the core, and (iii) the secondary current is small.

If the transformer is assumed to be 100% efficient (no energy losses), the power input is equal to the power output, and since $p = i v$,

$$i_p v_p = i_s v_s \quad (7.48)$$

Although some energy is always lost, this is a good approximation, since a well designed transformer may have an efficiency of more than 95%. Combining Eqs. (7.47) and (7.48), we have

$$\frac{i_p}{i_s} = \frac{v_s}{v_p} = \frac{N_s}{N_p} \quad (7.49)$$

Since i and v both oscillate with the same frequency as the ac source, Eq. (7.49) also gives the ratio of the amplitudes or rms values of corresponding quantities.

Now, we can see how a transformer affects the voltage and current. We have:

$$V_s = \left(\frac{N_s}{N_p} \right) V_p \quad \text{and} \quad I_s = \left(\frac{N_p}{N_s} \right) I_p \quad (7.50)$$

That is, if the secondary coil has a greater number of turns than the primary ($N_s > N_p$), the voltage is stepped up ($V_s > V_p$). This type of arrangement is called a *step-up transformer*. However, in this arrangement, there is less current in the secondary than in the primary ($N_p/N_s < 1$ and $I_s < I_p$). For example, if the primary coil of a transformer has 100 turns and the secondary has 200 turns, $N_s/N_p = 2$ and $N_p/N_s = 1/2$. Thus, a 220V input at 10A will step-up to 440 V output at 5.0 A.

If the secondary coil has less turns than the primary ($N_s < N_p$), we have a *step-down transformer*. In this case, $V_s < V_p$ and $I_s > I_p$. That is, the voltage is stepped down, or reduced, and the current is increased.

The equations obtained above apply to ideal transformers (without any energy losses). But in actual transformers, small energy losses do occur due to the following reasons:

- (i) *Flux Leakage*: There is always some flux leakage; that is, not all of the flux due to primary passes through the secondary due to poor design of the core or the air gaps in the core. It can be reduced by winding the primary and secondary coils one over the other.
- (ii) *Resistance of the windings*: The wire used for the windings has some resistance and so, energy is lost due to heat produced in the wire ($I^2 R$). In high current, low voltage windings, these are minimised by using thick wire.
- (iii) *Eddy currents*: The alternating magnetic flux induces eddy currents in the iron core and causes heating. The effect is reduced by having a laminated core.
- (iv) *Hysteresis*: The magnetisation of the core is repeatedly reversed by the alternating magnetic field. The resulting expenditure of energy in the core appears as heat and is kept to a minimum by using a magnetic material which has a low hysteresis loss.

Physics

The large scale transmission and distribution of electrical energy over long distances is done with the use of transformers. The voltage output of the generator is stepped-up (so that current is reduced and consequently, the I^2R loss is cut down). It is then transmitted over long distances to an area sub-station near the consumers. There the voltage is stepped down. It is further stepped down at distributing sub-stations and utility poles before a power supply of 240 V reaches our homes.

SUMMARY

1. An alternating voltage $v = v_m \sin \omega t$ applied to a resistor R drives a current $i = i_m \sin \omega t$ in the resistor, $i_m = \frac{v_m}{R}$. The current is in phase with the applied voltage.
2. For an alternating current $i = i_m \sin \omega t$ passing through a resistor R , the average power loss P (averaged over a cycle) due to joule heating is $(1/2) i_m^2 R$. To express it in the same form as the dc power ($P = I^2 R$), a special value of current is used. It is called *root mean square (rms) current* and is denoted by I :

$$I = \frac{i_m}{\sqrt{2}} = 0.707 i_m$$

Similarly, the *rms voltage* is defined by

$$V = \frac{v_m}{\sqrt{2}} = 0.707 v_m$$

We have $P = IV = I^2 R$

3. An ac voltage $v = v_m \sin \omega t$ applied to a pure inductor L , drives a current in the inductor $i = i_m \sin (\omega t - \pi/2)$, where $i_m = v_m/X_L$. $X_L = \omega L$ is called *inductive reactance*. The current in the inductor lags the voltage by $\pi/2$. The average power supplied to an inductor over one complete cycle is zero.
4. An ac voltage $v = v_m \sin \omega t$ applied to a capacitor drives a current in the capacitor: $i = i_m \sin (\omega t + \pi/2)$. Here,

$$i_m = \frac{v_m}{X_C}, X_C = \frac{1}{\omega C} \text{ is called } \textit{capacitive reactance}.$$

The current through the capacitor is $\pi/2$ ahead of the applied voltage. As in the case of inductor, the average power supplied to a capacitor over one complete cycle is zero.

5. For a series RLC circuit driven by voltage $v = v_m \sin \omega t$, the current is given by $i = i_m \sin (\omega t + \phi)$

$$\text{where } i_m = \frac{v_m}{\sqrt{R^2 + (X_C - X_L)^2}}$$

$$\text{and } \phi = \tan^{-1} \frac{X_C - X_L}{R}$$

$Z = \sqrt{R^2 + (X_C - X_L)^2}$ is called the *impedance* of the circuit.

Alternating Current

The average power loss over a complete cycle is given by

$$P = V I \cos\phi$$

The term $\cos\phi$ is called the *power factor*.

6. In a purely inductive or capacitive circuit, $\cos\phi = 0$ and no power is dissipated even though a current is flowing in the circuit. In such cases, current is referred to as a *wattless current*.
7. The phase relationship between current and voltage in an ac circuit can be shown conveniently by representing voltage and current by rotating vectors called *phasors*. A phasor is a vector which rotates about the origin with angular speed ω . The magnitude of a phasor represents the amplitude or peak value of the quantity (voltage or current) represented by the phasor.

The analysis of an ac circuit is facilitated by the use of a phasor diagram.

8. An interesting characteristic of a series RLC circuit is the phenomenon of *resonance*. The circuit exhibits resonance, i.e., the amplitude of the current is maximum at the resonant frequency, $\omega_0 = \frac{1}{\sqrt{LC}}$. The *quality factor Q* defined by

$$Q = \frac{\omega_0 L}{R} = \frac{1}{\omega_0 C R}$$
 is an indicator of the sharpness of the resonance,

the higher value of Q indicating sharper peak in the current.

9. A circuit containing an inductor L and a capacitor C (initially charged) with no ac source and no resistors exhibits *free oscillations*. The charge q of the capacitor satisfies the equation of simple harmonic motion:

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

and therefore, the frequency ω of free oscillation is $\omega_0 = \frac{1}{\sqrt{LC}}$. The

energy in the system oscillates between the capacitor and the inductor but their sum or the total energy is constant in time.

10. A transformer consists of an iron core on which are bound a primary coil of N_p turns and a secondary coil of N_s turns. If the primary coil is connected to an ac source, the primary and secondary voltages are related by

$$V_s = \left(\frac{N_s}{N_p} \right) V_p$$

and the currents are related by

$$I_s = \left(\frac{N_p}{N_s} \right) I_p$$

If the secondary coil has a greater number of turns than the primary, the voltage is stepped-up ($V_s > V_p$). This type of arrangement is called a *step-up transformer*. If the secondary coil has turns less than the primary, we have a *step-down transformer*.

Physics

Physical quantity	Symbol	Dimensions	Unit	Remarks
rms voltage	V	$[M L^2 T^{-3} A^{-1}]$	V	$V = \frac{v_m}{\sqrt{2}}$, v_m is the amplitude of the ac voltage.
rms current	I	[A]	A	$I = \frac{i_m}{\sqrt{2}}$, i_m is the amplitude of the ac current.
Reactance: Inductive Capacitive	X_L X_C	$[ML^2 T^{-3} A^{-2}]$ $[ML^2 T^{-3} A^{-2}]$	Ω Ω	$X_L = \omega L$ $X_C = 1/\omega C$
Impedance	Z	$[ML^2 T^{-3} A^{-2}]$	Ω	Depends on elements present in the circuit.
Resonant frequency	ω_r or ω_0	$[T^{-1}]$	Hz	$\omega_0 = \frac{1}{\sqrt{LC}}$ for a series RLC circuit
Quality factor	Q	Dimensionless		$Q = \frac{\omega_0 L}{R} = \frac{1}{\omega_0 C R}$ for a series RLC circuit.
Power factor		Dimensionless		$= \cos\phi$, ϕ is the phase difference between voltage applied and current in the circuit.

POINTS TO PONDER

- When a value is given for ac voltage or current, it is ordinarily the rms value. The voltage across the terminals of an outlet in your room is normally 240 V. This refers to the *rms* value of the voltage. The amplitude of this voltage is

$$v_m = \sqrt{2}V = \sqrt{2}(240) = 340 \text{ V}$$
- The power rating of an element used in ac circuits refers to its average power rating.
- The power consumed in an circuit is never negative.
- Both alternating current and direct current are measured in amperes. But how is the ampere defined for an alternating current? It cannot be derived from the mutual attraction of two parallel wires carrying ac currents, as the dc ampere is derived. An ac current changes direction

Alternating Current

with the source frequency and the attractive force would average to zero. Thus, the ac ampere must be defined in terms of some property that is independent of the direction of the current. Joule heating is such a property, and there is one ampere of *rms* value of alternating current in a circuit if the current produces the same average heating effect as one ampere of dc current would produce under the same conditions.

5. In an ac circuit, while adding voltages across different elements, one should take care of their phases properly. For example, if V_R and V_C are voltages across R and C , respectively in an RC circuit, then the total voltage across RC combination is $V_{RC} = \sqrt{V_R^2 + V_C^2}$ and not $V_R + V_C$ since V_C is $\pi/2$ out of phase of V_R .
6. Though in a phasor diagram, voltage and current are represented by vectors, these quantities are not really vectors themselves. They are scalar quantities. It so happens that the amplitudes and phases of harmonically varying scalars combine mathematically in the same way as do the projections of rotating vectors of corresponding magnitudes and directions. The 'rotating vectors' that represent harmonically varying scalar quantities are introduced only to provide us with a simple way of adding these quantities using a rule that we already know as the law of vector addition.
7. There are no power losses associated with pure capacitances and pure inductances in an ac circuit. The only element that dissipates energy in an ac circuit is the resistive element.
8. In a RLC circuit, resonance phenomenon occur when $X_L = X_C$ or $\omega_0 = \frac{1}{\sqrt{LC}}$. For resonance to occur, the presence of both L and C elements in the circuit is a must. With only one of these (L or C) elements, there is no possibility of voltage cancellation and hence, no resonance is possible.
9. The power factor in a RLC circuit is a measure of how close the circuit is to expending the maximum power.
10. In generators and motors, the roles of input and output are reversed. In a motor, electric energy is the input and mechanical energy is the output. In a generator, mechanical energy is the input and electric energy is the output. Both devices simply transform energy from one form to another.
11. A transformer (step-up) changes a low-voltage into a high-voltage. This does not violate the law of conservation of energy. The current is reduced by the same proportion.
12. The choice of whether the description of an oscillatory motion is by means of sines or cosines or by their linear combinations is unimportant, since changing the zero-time position transforms the one to the other.

EXERCISES

- 7.1** A $100\ \Omega$ resistor is connected to a 220 V , 50 Hz ac supply.
- What is the rms value of current in the circuit?
 - What is the net power consumed over a full cycle?
- 7.2** (a) The peak voltage of an ac supply is 300 V . What is the rms voltage?
(b) The rms value of current in an ac circuit is 10 A . What is the peak current?
- 7.3** A 44 mH inductor is connected to 220 V , 50 Hz ac supply. Determine the rms value of the current in the circuit.
- 7.4** A $60\text{ }\mu\text{F}$ capacitor is connected to a 110 V , 60 Hz ac supply. Determine the rms value of the current in the circuit.
- 7.5** In Exercises 7.3 and 7.4, what is the net power absorbed by each circuit over a complete cycle. Explain your answer.
- 7.6** Obtain the resonant frequency ω_r of a series *LCR* circuit with $L = 2.0\text{ H}$, $C = 32\text{ }\mu\text{F}$ and $R = 10\ \Omega$. What is the *Q*-value of this circuit?
- 7.7** A charged $30\text{ }\mu\text{F}$ capacitor is connected to a 27 mH inductor. What is the angular frequency of free oscillations of the circuit?
- 7.8** Suppose the initial charge on the capacitor in Exercise 7.7 is 6 mC . What is the total energy stored in the circuit initially? What is the total energy at later time?
- 7.9** A series *LCR* circuit with $R = 20\ \Omega$, $L = 1.5\text{ H}$ and $C = 35\text{ }\mu\text{F}$ is connected to a variable-frequency 200 V ac supply. When the frequency of the supply equals the natural frequency of the circuit, what is the average power transferred to the circuit in one complete cycle?
- 7.10** A radio can tune over the frequency range of a portion of MW broadcast band: (800 kHz to 1200 kHz). If its *LC* circuit has an effective inductance of $200\text{ }\mu\text{H}$, what must be the range of its variable capacitor?
[Hint: For tuning, the natural frequency i.e., the frequency of free oscillations of the *LC* circuit should be equal to the frequency of the radiowave.]
- 7.11** Figure 7.21 shows a series *LCR* circuit connected to a variable frequency 230 V source. $L = 5.0\text{ H}$, $C = 80\mu\text{F}$, $R = 40\ \Omega$.

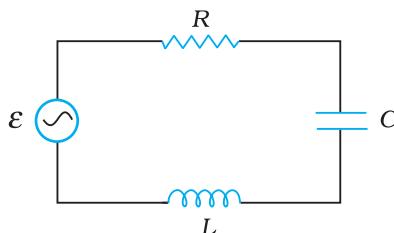


FIGURE 7.21

- Determine the source frequency which drives the circuit in resonance.
- Obtain the impedance of the circuit and the amplitude of current at the resonating frequency.
- Determine the rms potential drops across the three elements of the circuit. Show that the potential drop across the *LC* combination is zero at the resonating frequency.

ADDITIONAL EXERCISES

- 7.12** An *LC* circuit contains a 20 mH inductor and a 50 μF capacitor with an initial charge of 10 mC. The resistance of the circuit is negligible. Let the instant the circuit is closed be $t = 0$.
- What is the total energy stored initially? Is it conserved during *LC* oscillations?
 - What is the natural frequency of the circuit?
 - At what time is the energy stored
(i) completely electrical (i.e., stored in the capacitor)? (ii) completely magnetic (i.e., stored in the inductor)?
 - At what times is the total energy shared equally between the inductor and the capacitor?
 - If a resistor is inserted in the circuit, how much energy is eventually dissipated as heat?
- 7.13** A coil of inductance 0.50 H and resistance 100 Ω is connected to a 240 V, 50 Hz ac supply.
- What is the maximum current in the coil?
 - What is the time lag between the voltage maximum and the current maximum?
- 7.14** Obtain the answers (a) to (b) in Exercise 7.13 if the circuit is connected to a high frequency supply (240 V, 10 kHz). Hence, explain the statement that at very high frequency, an inductor in a circuit nearly amounts to an open circuit. How does an inductor behave in a dc circuit after the steady state?
- 7.15** A 100 μF capacitor in series with a 40 Ω resistance is connected to a 110 V, 60 Hz supply.
- What is the maximum current in the circuit?
 - What is the time lag between the current maximum and the voltage maximum?
- 7.16** Obtain the answers to (a) and (b) in Exercise 7.15 if the circuit is connected to a 110 V, 12 kHz supply? Hence, explain the statement that a capacitor is a conductor at very high frequencies. Compare this behaviour with that of a capacitor in a dc circuit after the steady state.
- 7.17** Keeping the source frequency equal to the resonating frequency of the series *LCR* circuit, if the three elements, *L*, *C* and *R* are arranged in parallel, show that the total current in the parallel *LCR* circuit is minimum at this frequency. Obtain the current rms value in each branch of the circuit for the elements and source specified in Exercise 7.11 for this frequency.
- 7.18** A circuit containing a 80 mH inductor and a 60 μF capacitor in series is connected to a 230 V, 50 Hz supply. The resistance of the circuit is negligible.
- Obtain the current amplitude and rms values.
 - Obtain the rms values of potential drops across each element.
 - What is the average power transferred to the inductor?
 - What is the average power transferred to the capacitor?
 - What is the total average power absorbed by the circuit? ['Average' implies 'averaged over one cycle'.]
- 7.19** Suppose the circuit in Exercise 7.18 has a resistance of 15 Ω . Obtain the average power transferred to each element of the circuit, and the total power absorbed.

Physics

- 7.20** A series *LCR* circuit with $L = 0.12 \text{ H}$, $C = 480 \text{ nF}$, $R = 23 \Omega$ is connected to a 230 V variable frequency supply.
- What is the source frequency for which current amplitude is maximum. Obtain this maximum value.
 - What is the source frequency for which average power absorbed by the circuit is maximum. Obtain the value of this maximum power.
 - For which frequencies of the source is the power transferred to the circuit half the power at resonant frequency? What is the current amplitude at these frequencies?
 - What is the *Q*-factor of the given circuit?
- 7.21** Obtain the resonant frequency and *Q*-factor of a series *LCR* circuit with $L = 3.0 \text{ H}$, $C = 27 \mu\text{F}$, and $R = 7.4 \Omega$. It is desired to improve the sharpness of the resonance of the circuit by reducing its 'full width at half maximum' by a factor of 2. Suggest a suitable way.
- 7.22** Answer the following questions:
- In any ac circuit, is the applied instantaneous voltage equal to the algebraic sum of the instantaneous voltages across the series elements of the circuit? Is the same true for rms voltage?
 - A capacitor is used in the primary circuit of an induction coil.
 - An applied voltage signal consists of a superposition of a dc voltage and an ac voltage of high frequency. The circuit consists of an inductor and a capacitor in series. Show that the dc signal will appear across C and the ac signal across L .
 - A choke coil in series with a lamp is connected to a dc line. The lamp is seen to shine brightly. Insertion of an iron core in the choke causes no change in the lamp's brightness. Predict the corresponding observations if the connection is to an ac line.
 - Why is choke coil needed in the use of fluorescent tubes with ac mains? Why can we not use an ordinary resistor instead of the choke coil?
- 7.23** A power transmission line feeds input power at 2300 V to a step-down transformer with its primary windings having 4000 turns. What should be the number of turns in the secondary in order to get output power at 230 V?
- 7.24** At a hydroelectric power plant, the water pressure head is at a height of 300 m and the water flow available is $100 \text{ m}^3 \text{s}^{-1}$. If the turbine generator efficiency is 60%, estimate the electric power available from the plant ($g = 9.8 \text{ ms}^{-2}$).
- 7.25** A small town with a demand of 800 kW of electric power at 220 V is situated 15 km away from an electric plant generating power at 440 V. The resistance of the two wire line carrying power is 0.5Ω per km. The town gets power from the line through a 4000-220 V step-down transformer at a sub-station in the town.
- Estimate the line power loss in the form of heat.
 - How much power must the plant supply, assuming there is negligible power loss due to leakage?
 - Characterise the step up transformer at the plant.
- 7.26** Do the same exercise as above with the replacement of the earlier transformer by a 40,000-220 V step-down transformer (Neglect, as before, leakage losses though this may not be a good assumption any longer because of the very high voltage transmission involved). Hence, explain why high voltage transmission is preferred?

Chapter Eight

ELECTROMAGNETIC WAVES

8.1 INTRODUCTION

In Chapter 4, we learnt that an electric current produces magnetic field and that two current-carrying wires exert a magnetic force on each other. Further, in Chapter 6, we have seen that a magnetic field changing with time gives rise to an electric field. Is the converse also true? Does an electric field changing with time give rise to a magnetic field? James Clerk Maxwell (1831-1879), argued that this was indeed the case – not only an electric current but also a time-varying electric field generates magnetic field. While applying the Ampere's circuital law to find magnetic field at a point outside a capacitor connected to a time-varying current, Maxwell noticed an inconsistency in the Ampere's circuital law. He suggested the existence of an additional current, called by him, the displacement current to remove this inconsistency.

Maxwell formulated a set of equations involving electric and magnetic fields, and their sources, the charge and current densities. These equations are known as Maxwell's equations. Together with the Lorentz force formula (Chapter 4), they mathematically express all the basic laws of electromagnetism.

The most important prediction to emerge from Maxwell's equations is the existence of electromagnetic waves, which are (coupled) time-varying electric and magnetic fields that propagate in space. The speed of the waves, according to these equations, turned out to be very close to



James Clerk Maxwell (1831 – 1879) Born in Edinburgh, Scotland, was among the greatest physicists of the nineteenth century. He derived the thermal velocity distribution of molecules in a gas and was among the first to obtain reliable estimates of molecular parameters from measurable quantities like viscosity, etc. Maxwell's greatest achievement was the unification of the laws of electricity and magnetism (discovered by Coulomb, Oersted, Ampere and Faraday) into a consistent set of equations now called Maxwell's equations. From these he arrived at the most important conclusion that light is an electromagnetic wave. Interestingly, Maxwell did not agree with the idea (strongly suggested by the Faraday's laws of electrolysis) that electricity was particulate in nature.

JAMES CLERK MAXWELL (1831–1879)

the speed of light (3×10^8 m/s), obtained from optical measurements. This led to the remarkable conclusion that light is an electromagnetic wave. Maxwell's work thus unified the domain of electricity, magnetism and light. Hertz, in 1885, experimentally demonstrated the existence of electromagnetic waves. Its technological use by Marconi and others led in due course to the revolution in communication that we are witnessing today.

In this chapter, we first discuss the need for displacement current and its consequences. Then we present a descriptive account of electromagnetic waves. The broad spectrum of electromagnetic waves, stretching from γ rays (wavelength $\sim 10^{-12}$ m) to long radio waves (wavelength $\sim 10^6$ m) is described. How the electromagnetic waves are sent and received for communication is discussed in Chapter 15.

8.2 DISPLACEMENT CURRENT

We have seen in Chapter 4 that an electrical current produces a magnetic field around it. Maxwell showed that for logical consistency, a changing electric field *must also* produce a magnetic field. This effect is of great importance because it explains the existence of radio waves, gamma rays and visible light, as well as all other forms of electromagnetic waves.

To see how a changing electric field gives rise to a magnetic field, let us consider the process of charging of a capacitor and apply Ampere's circuital law given by (Chapter 4)

$$\oint \mathbf{B} \cdot d\mathbf{l} = \mu_0 i(t) \quad (8.1)$$

to find magnetic field at a point outside the capacitor. Figure 8.1(a) shows a parallel plate capacitor C which is a part of circuit through which a time-dependent current $i(t)$ flows. Let us find the magnetic field at a point such as P, in a region outside the parallel plate capacitor. For this, we consider a plane circular loop of radius r whose plane is perpendicular to the direction of the current-carrying wire, and which is centred symmetrically with respect to the wire [Fig. 8.1(a)]. From symmetry, the magnetic field is directed along the circumference of the circular loop and is the same in magnitude at all points on the loop so that if B is the magnitude of the field, the left side of Eq. (8.1) is $B(2\pi r)$. So we have

$$B(2\pi r) = \mu_0 i(t) \quad (8.2)$$

Electromagnetic Waves

Now, consider a different surface, which has the same boundary. This is a pot like surface [Fig. 8.1(b)] which nowhere touches the current, but has its bottom between the capacitor plates; its mouth is the circular loop mentioned above. Another such surface is shaped like a tiffin box (without the lid) [Fig. 8.1(c)]. On applying Ampere's circuital law to such surfaces with the *same* perimeter, we find that the left hand side of Eq. (8.1) has not changed but the right hand side is zero and *not* $\mu_0 i$, since no current passes through the surface of Fig. 8.1(b) and (c). So we have a *contradiction*; calculated one way, there is a magnetic field at a point P; calculated another way, the magnetic field at P is zero. Since the contradiction arises from our use of Ampere's circuital law, this law must be missing something. The missing term must be such that one gets the same magnetic field at point P, no matter what surface is used.

We can actually guess the missing term by looking carefully at Fig. 8.1(c). Is there anything passing through the surface S *between* the plates of the capacitor? Yes, of course, the electric field! If the plates of the capacitor have an area A, and a total charge Q, the magnitude of the electric field \mathbf{E} between the plates is $(Q/A)/\epsilon_0$ (see Eq. 2.41). The field is perpendicular to the surface S of Fig. 8.1(c). It has the same magnitude over the area A of the capacitor plates, and vanishes outside it. So what is the *electric flux* Φ_E through the surface S? Using Gauss's law, it is

$$\Phi_E = |\mathbf{E}| A = \frac{1}{\epsilon_0} \frac{Q}{A} A = \frac{Q}{\epsilon_0} \quad (8.3)$$

Now if the charge Q on the capacitor plates changes with time, there is a current $i = (dQ/dt)$, so that using Eq. (8.3), we have

$$\frac{d\Phi_E}{dt} = \frac{d}{dt} \left(\frac{Q}{\epsilon_0} \right) = \frac{1}{\epsilon_0} \frac{dQ}{dt}$$

This implies that for consistency,

$$\epsilon_0 \left(\frac{d\Phi_E}{dt} \right) = i \quad (8.4)$$

This is the missing term in Ampere's circuital law. If we generalise this law by adding to the total current carried by conductors through the surface, another term which is ϵ_0 times the rate of change of electric flux through the same surface, the *total* has the same value of current *i* for all surfaces. If this is done, there is no contradiction in the value of B obtained anywhere using the generalised Ampere's law. At the point P is non-zero no matter which surface is used for calculating it. B at a point P outside the plates [Fig. 8.1(a)] is the same as at a point M just inside, as it should be. The current carried by conductors due to flow of charges is called *conduction current*. The current, given by Eq. (8.4), is a new term, and is due to changing electric field (or electric *displacement*, an old term still used sometimes). It is, therefore, called *displacement current* or Maxwell's displacement current. Figure 8.2 shows the electric and magnetic fields inside the parallel plate capacitor discussed above.

The generalisation made by Maxwell then is the following. The source of a magnetic field is not just the conduction electric current due to flowing

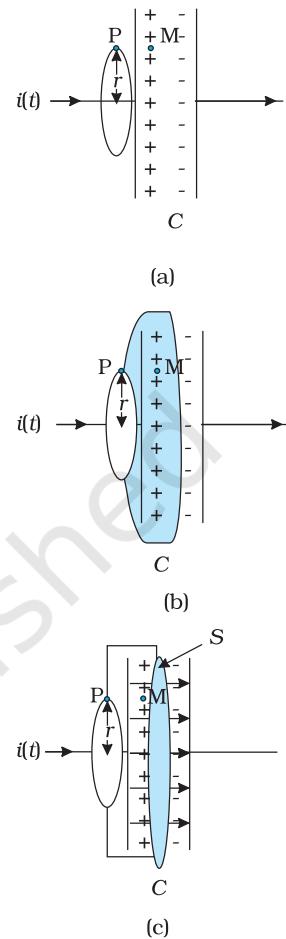


FIGURE 8.1 A parallel plate capacitor C, as part of a circuit through which a time dependent current $i(t)$ flows. (a) a loop of radius r , to determine magnetic field at a point P on the loop; (b) a pot-shaped surface passing through the interior between the capacitor plates with the loop shown in (a) as its rim; (c) a tiffin-shaped surface with the circular loop as its rim and a flat circular bottom S between the capacitor plates. The arrows show uniform electric field between the capacitor plates.

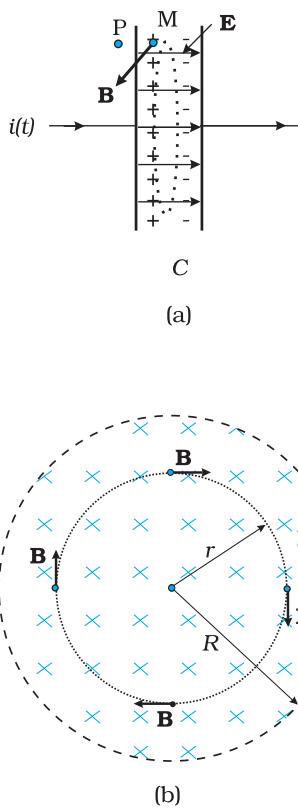


FIGURE 8.2 (a) The electric and magnetic fields **E** and **B** between the capacitor plates, at the point M. (b) A cross sectional view of Fig. (a).

charges, but also the time rate of change of electric field. More precisely, the total current i is the sum of the conduction current denoted by i_c , and the displacement current denoted by $i_d (= \epsilon_0 (\partial \Phi_E / \partial t))$. So we have

$$i = i_e + i_d = i_c + \epsilon_0 \frac{d\Phi_E}{dt} \quad (8.5)$$

In explicit terms, this means that outside the capacitor plates, we have only conduction current $i_c = i$, and no displacement current, i.e., $i_d = 0$. On the other hand, inside the capacitor, there is no conduction current, i.e., $i_c = 0$, and there is only displacement current, so that $i_d = i$.

The generalised (and correct) Ampere's circuital law has the same form as Eq. (8.1), with one difference: "the *total current* passing through any surface of which the closed loop is the perimeter" is the sum of the conduction current and the displacement current. The generalised law is

$$\oint \mathbf{B} \cdot d\mathbf{l} = \mu_0 i_c + \mu_0 \epsilon_0 \frac{d\Phi_E}{dt} \quad (8.6)$$

and is known as Ampere-Maxwell law.

In all respects, the displacement current has the same physical effects as the conduction current. In some cases, for example, steady electric fields in a conducting wire, the displacement current may be zero since the electric field **E** does not change with time. In other cases, for example, the charging capacitor above, both conduction and displacement currents may be present in different regions of space. In most of the cases, they both may be present in the same region of space, as there exist no perfectly conducting or perfectly insulating medium. Most interestingly, there may be large regions of space where there is no conduction current, but there is only a displacement current due to time-varying electric fields. In such a region, we expect a magnetic field, though there is no (conduction) current source nearby! The prediction of such a displacement current can be verified experimentally. For example, a magnetic field (say at point M) between the plates of the capacitor in Fig. 8.2(a) can be measured and is seen to be the same as that just outside (at P).

The displacement current has (literally) far reaching consequences. One thing we immediately notice is that the laws of electricity and magnetism are now more symmetrical*. Faraday's law of induction states that there is an induced emf *equal to the rate of change of magnetic flux*. Now, since the emf between two points 1 and 2 is the work done per unit charge in taking it from 1 to 2, the existence of an emf implies the existence of an electric field. So, we can rephrase Faraday's law of electromagnetic induction by saying that a *magnetic field*, changing with time, gives rise to an *electric field*. Then, the fact that an *electric field* changing with time gives rise to a *magnetic field*, is the symmetrical counterpart, and is

* They are still not perfectly symmetrical; there are no known sources of magnetic field (magnetic monopoles) analogous to electric charges which are sources of electric field.

a consequence of the displacement current being a source of a magnetic field. Thus, time-dependent electric and magnetic fields give rise to each other! Faraday's law of electromagnetic induction and Ampere-Maxwell law give a quantitative expression of this statement, with the current being the total current, as in Eq. (8.5). One very important consequence of this symmetry is the existence of electromagnetic waves, which we discuss qualitatively in the next section.

MAXWELL'S EQUATIONS

1. $\oint \mathbf{E} \cdot d\mathbf{A} = Q / \epsilon_0$ (Gauss's Law for electricity)
2. $\oint \mathbf{B} \cdot d\mathbf{A} = 0$ (Gauss's Law for magnetism)
3. $\oint \mathbf{E} \cdot d\mathbf{l} = -\frac{d\Phi_B}{dt}$ (Faraday's Law)
4. $\oint \mathbf{B} \cdot d\mathbf{l} = \mu_0 i_c + \mu_0 \epsilon_0 \frac{d\Phi_E}{dt}$ (Ampere – Maxwell Law)

Example 8.1 A parallel plate capacitor with circular plates of radius 1 m has a capacitance of 1 nF. At $t = 0$, it is connected for charging in series with a resistor $R = 1 \text{ M}\Omega$ across a 2V battery (Fig. 8.3). Calculate the magnetic field at a point P, halfway between the centre and the periphery of the plates, after $t = 10^{-3} \text{ s}$. (The charge on the capacitor at time t is $q(t) = CV [1 - \exp(-t/\tau)]$, where the time constant τ is equal to CR)

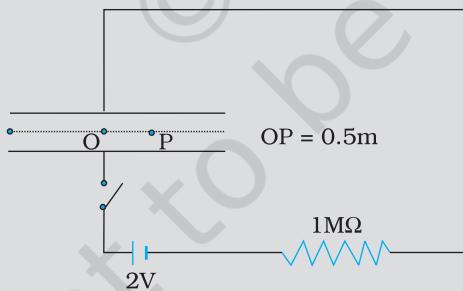


FIGURE 8.3

Solution The time constant of the CR circuit is $\tau = CR = 10^{-3} \text{ s}$. Then, we have

$$\begin{aligned} q(t) &= CV [1 - \exp(-t/\tau)] \\ &= 2 \times 10^{-9} [1 - \exp(-t/10^{-3})] \end{aligned}$$

The electric field in between the plates at time t is

$$E = \frac{q(t)}{\epsilon_0 A} = \frac{q}{\pi \epsilon_0} ; A = \pi (1)^2 \text{ m}^2 = \text{area of the plates.}$$

Consider now a circular loop of radius $(1/2) \text{ m}$ parallel to the plates passing through P. The magnetic field \mathbf{B} at all points on the loop is

EXAMPLE 8.1

along the loop and of the same value.

The flux Φ_E through this loop is

$$\Phi_E = E \times \text{area of the loop}$$

$$= E \times \pi \times \left(\frac{1}{2}\right)^2 = \frac{\pi E}{4} = \frac{q}{4\epsilon_0}$$

The displacement current

$$i_d = \epsilon_0 \frac{d\Phi_E}{dt} = \frac{1}{4} \frac{dq}{dt} = 0.5 \times 10^{-6} \exp(-1)$$

at $t = 10^{-3}$ s. Now, applying Ampere-Maxwell law to the loop, we get

$$B \times 2\pi \times \left(\frac{1}{2}\right) = \mu_0 (i_c + i_d) = \mu_0 (0 + i_d) = 0.5 \times 10^{-6} \mu_0 \exp(-1)$$

$$\text{or, } B = 0.74 \times 10^{-13} \text{ T}$$

8.3 ELECTROMAGNETIC WAVES

8.3.1 Sources of electromagnetic waves

How are electromagnetic waves produced? Neither stationary charges nor charges in uniform motion (steady currents) can be sources of electromagnetic waves. The former produces only electrostatic fields, while the latter produces magnetic fields that, however, do not vary with time. It is an important result of Maxwell's theory that accelerated charges radiate electromagnetic waves. The proof of this basic result is beyond the scope of this book, but we can accept it on the basis of rough, qualitative reasoning. Consider a charge oscillating with some frequency. (An oscillating charge is an example of accelerating charge.) This produces an oscillating electric field in space, which produces an oscillating magnetic field, which in turn, is a source of oscillating electric field, and so on. The oscillating electric and magnetic fields thus regenerate each other, so to speak, as the wave propagates through the space. The frequency of the electromagnetic wave naturally equals the frequency of oscillation of the charge. The energy associated with the propagating wave comes at the expense of the energy of the source – the accelerated charge.

From the preceding discussion, it might appear easy to test the prediction that light is an electromagnetic wave. We might think that all we needed to do was to set up an ac circuit in which the current oscillate at the frequency of visible light, say, yellow light. But, alas, that is not possible. The frequency of yellow light is about 6×10^{14} Hz, while the frequency that we get even with modern electronic circuits is hardly about 10^{11} Hz. This is why the experimental demonstration of electromagnetic wave had to come in the low frequency region (the radio wave region), as in the Hertz's experiment (1887).

Hertz's successful experimental test of Maxwell's theory created a sensation and sparked off other important works in this field. Two important achievements in this connection deserve mention. Seven years after Hertz, Jagdish Chandra Bose, working at Calcutta (now Kolkata),

succeeded in producing and observing electromagnetic waves of much shorter wavelength (25 mm to 5 mm). His experiment, like that of Hertz's, was confined to the laboratory.

At around the same time, Guglielmo Marconi in Italy followed Hertz's work and succeeded in transmitting electromagnetic waves over distances of many kilometres. Marconi's experiment marks the beginning of the field of communication using electromagnetic waves.

8.3.2 Nature of electromagnetic waves

It can be shown from Maxwell's equations that electric and magnetic fields in an electromagnetic wave are perpendicular to each other, and to the direction of propagation. It appears reasonable, say from our discussion of the displacement current. Consider Fig. 8.2. The electric field inside the plates of the capacitor is directed perpendicular to the plates. The magnetic field this gives rise to via the displacement current is along the perimeter of a circle parallel to the capacitor plates. So \mathbf{B} and \mathbf{E} are perpendicular in this case. This is a general feature.

In Fig. 8.4, we show a typical example of a plane electromagnetic wave propagating along the z -direction (the fields are shown as a function of the z coordinate, at a given time t). The electric field E_x is along the x -axis, and varies sinusoidally with z , at a given time. The magnetic field B_y is along the y -axis, and again varies sinusoidally with z . The electric and magnetic fields E_x and B_y are perpendicular to each other, and to the direction z of propagation. We can write E_x and B_y as follows:

$$E_x = E_0 \sin(kz - \omega t) \quad [8.7(a)]$$

$$B_y = B_0 \sin(kz - \omega t) \quad [8.7(b)]$$

Here k is related to the wave length λ of the wave by the usual equation

$$k = \frac{2\pi}{\lambda} \quad (8.8)$$

and ω is the angular frequency. k is the magnitude of the wave vector (or propagation vector) \mathbf{k} and its direction describes the direction of propagation of the wave. The speed of propagation of the wave is (ω/k) . Using Eqs. [8.7(a) and (b)] for E_x and B_y and Maxwell's equations, one finds that



Heinrich Rudolf Hertz
(1857 – 1894) German physicist who was the first to broadcast and receive radio waves. He produced electromagnetic waves, sent them through space, and measured their wavelength and speed. He showed that the nature of their vibration, reflection and refraction was the same as that of light and heat waves, establishing their identity for the first time. He also pioneered research on discharge of electricity through gases, and discovered the photoelectric effect.

HEINRICH RUDOLF HERTZ (1857–1894)

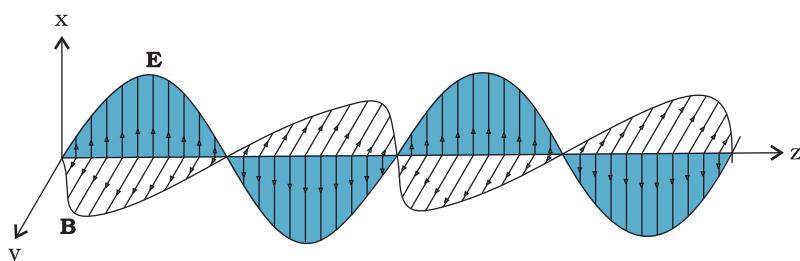


FIGURE 8.4 A linearly polarised electromagnetic wave, propagating in the z -direction with the oscillating electric field \mathbf{E} along the x -direction and the oscillating magnetic field \mathbf{B} along the y -direction.

Physics

Simulate propagation of electromagnetic waves
(i) <http://www.amanogawa.com/waves.html>
(ii) <http://www.phys.hawaii.edu/~teb/java/ntnujava/emWave/emWave.html>


$$\omega = ck, \text{ where, } c = 1/\sqrt{\mu_0 \epsilon_0} \quad [8.9(a)]$$

The relation $\omega = ck$ is the standard one for waves (see for example, Section 15.4 of class XI Physics textbook). This relation is often written in terms of frequency, $v (= \omega/2\pi)$ and wavelength, $\lambda (=2\pi/k)$ as

$$2\pi v = c \left(\frac{2\pi}{\lambda} \right) \quad \text{or} \quad v\lambda = c \quad [8.9(b)]$$

It is also seen from Maxwell's equations that the magnitude of the electric and the magnetic fields in an electromagnetic wave are related as

$$B_0 = (E_0/c) \quad (8.10)$$

We here make remarks on some features of electromagnetic waves. They are self-sustaining oscillations of electric and magnetic fields in free space, or vacuum. They differ from all the other waves we have studied so far, in respect that *no material medium* is involved in the vibrations of the electric and magnetic fields. Sound waves in air are longitudinal waves of compression and rarefaction. Transverse waves on the surface of water consist of water moving up and down as the wave spreads horizontally and radially onwards. Transverse elastic (sound) waves can also propagate in a solid, which is rigid and that resists shear. Scientists in the nineteenth century were so much used to this mechanical picture that they thought that there must be some medium pervading all space and all matter, which responds to electric and magnetic fields just as any elastic medium does. They called this medium *ether*. They were so convinced of the reality of this medium, that there is even a novel called *The Poison Belt* by Sir Arthur Conan Doyle (the creator of the famous detective *Sherlock Holmes*) where the solar system is supposed to pass through a poisonous region of ether! We now accept that no such physical medium is needed. The famous experiment of Michelson and Morley in 1887 demolished conclusively the hypothesis of ether. Electric and magnetic fields, oscillating in space and time, can sustain each other in vacuum.

But what if a material medium is actually there? We know that light, an electromagnetic wave, does propagate through glass, for example. We have seen earlier that the total electric and magnetic fields inside a medium are described in terms of a permittivity ϵ and a magnetic permeability μ (these describe the factors by which the total fields differ from the external fields). These replace ϵ_0 and μ_0 in the description to electric and magnetic fields in Maxwell's equations with the result that in a material medium of permittivity ϵ and magnetic permeability μ , the velocity of light becomes,

$$v = \frac{1}{\sqrt{\mu\epsilon}} \quad (8.11)$$

Thus, the velocity of light depends on electric and magnetic properties of the medium. We shall see in the next chapter that the *refractive index* of one medium with respect to the other is equal to the ratio of velocities of light in the two media.

The velocity of electromagnetic waves in free space or vacuum is an important fundamental constant. It has been shown by experiments on electromagnetic waves of different wavelengths that this velocity is the

same (independent of wavelength) to within a few metres per second, out of a value of 3×10^8 m/s. The constancy of the velocity of em waves in vacuum is so strongly supported by experiments and the actual value is so well known now that this is used to define a standard of *length*. Namely, the metre is now *defined* as the distance travelled by light in vacuum in a time ($1/c$) seconds = $(2.99792458 \times 10^8)^{-1}$ seconds. This has come about for the following reason. The basic unit of time can be defined very accurately in terms of some atomic frequency, i.e., frequency of light emitted by an atom in a particular process. The basic unit of length is harder to define as accurately in a direct way. Earlier measurement of c using earlier units of length (metre rods, etc.) converged to a value of about 2.9979246×10^8 m/s. Since c is such a strongly fixed number, unit of length can be defined in terms of c and the unit of time!

Hertz not only showed the existence of electromagnetic waves, but also demonstrated that the waves, which had wavelength ten million times that of the light waves, could be diffracted, refracted and polarised. Thus, he conclusively established the wave nature of the radiation. Further, he produced stationary electromagnetic waves and determined their wavelength by measuring the distance between two successive nodes. Since the frequency of the wave was known (being equal to the frequency of the oscillator), he obtained the speed of the wave using the formula $v = \nu\lambda$ and found that the waves travelled with the same speed as the speed of light.

The fact that electromagnetic waves are polarised can be easily seen in the response of a portable AM radio to a broadcasting station. If an AM radio has a telescopic antenna, it responds to the electric part of the signal. When the antenna is turned horizontal, the signal will be greatly diminished. Some portable radios have horizontal antenna (usually inside the case of radio), which are sensitive to the magnetic component of the electromagnetic wave. Such a radio must remain horizontal in order to receive the signal. In such cases, response also depends on the orientation of the radio with respect to the station.

Do electromagnetic waves carry energy and momentum like other waves? Yes, they do. We have seen in chapter 2 that in a region of free space with electric field E , there is an energy density ($\epsilon_0 E^2 / 2$). Similarly, as seen in Chapter 6, associated with a magnetic field B is a magnetic energy density ($B^2 / 2\mu_0$). As electromagnetic wave contains both electric and magnetic fields, there is a non-zero energy density associated with it. Now consider a plane perpendicular to the direction of propagation of the electromagnetic wave (Fig. 8.4). If there are, on this plane, electric charges, they will be set and sustained in motion by the electric and magnetic fields of the electromagnetic wave. The charges thus acquire energy and momentum from the waves. This just illustrates the fact that an electromagnetic wave (like other waves) carries energy and momentum. Since it carries momentum, an electromagnetic wave also exerts pressure, called *radiation pressure*.

If the total energy transferred to a surface in time t is U , it can be shown that the magnitude of the total momentum delivered to this surface (*for complete absorption*) is,

$$p = \frac{U}{c} \quad (8.12)$$

When the sun shines on your hand, you feel the energy being absorbed from the electromagnetic waves (your hands get warm). Electromagnetic waves also transfer momentum to your hand but because c is very large, the amount of momentum transferred is extremely small and you do not feel the pressure. In 1903, the American scientists Nicols and Hull succeeded in measuring radiation pressure of visible light and verified Eq. (8.12). It was found to be of the order of $7 \times 10^{-6} \text{ N/m}^2$. Thus, on a surface of area 10 cm^2 , the force due to radiation is only about $7 \times 10^{-9} \text{ N}$.

The great technological importance of electromagnetic waves stems from their capability to carry energy from one place to another. The radio and TV signals from broadcasting stations carry energy. Light carries energy from the sun to the earth, thus making life possible on the earth.

Example 8.2 A plane electromagnetic wave of frequency 25 MHz travels in free space along the x -direction. At a particular point in space and time, $\mathbf{E} = 6.3 \hat{\mathbf{j}} \text{ V/m}$. What is \mathbf{B} at this point?

Solution Using Eq. (8.10), the magnitude of \mathbf{B} is

$$\begin{aligned} B &= \frac{E}{c} \\ &= \frac{6.3 \text{ V/m}}{3 \times 10^8 \text{ m/s}} = 2.1 \times 10^{-8} \text{ T} \end{aligned}$$

To find the direction, we note that \mathbf{E} is along y -direction and the wave propagates along x -axis. Therefore, \mathbf{B} should be in a direction perpendicular to both x - and y -axes. Using vector algebra, $\mathbf{E} \times \mathbf{B}$ should be along x -direction. Since, $(+\hat{\mathbf{j}}) \times (+\hat{\mathbf{k}}) = \hat{\mathbf{i}}$, \mathbf{B} is along the z -direction. Thus, $\mathbf{B} = 2.1 \times 10^{-8} \hat{\mathbf{k}} \text{ T}$

EXAMPLE 8.2

Example 8.3 The magnetic field in a plane electromagnetic wave is given by $B_y = 2 \times 10^{-7} \sin(0.5 \times 10^3 x + 1.5 \times 10^{11} t) \text{ T}$.

(a) What is the wavelength and frequency of the wave?

(b) Write an expression for the electric field.

Solution

(a) Comparing the given equation with

$$B_y = B_0 \sin \left[2\pi \left(\frac{x}{\lambda} + \frac{t}{T} \right) \right]$$

$$\text{We get, } \lambda = \frac{2\pi}{0.5 \times 10^3} \text{ m} = 1.26 \text{ cm},$$

$$\text{and } \frac{1}{T} = \nu = (1.5 \times 10^{11}) / 2\pi = 23.9 \text{ GHz}$$

(b) $E_0 = B_0 c = 2 \times 10^{-7} \text{ T} \times 3 \times 10^8 \text{ m/s} = 6 \times 10^1 \text{ V/m}$

The electric field component is perpendicular to the direction of propagation and the direction of magnetic field. Therefore, the electric field component along the z -axis is obtained as

$$E_z = 60 \sin(0.5 \times 10^3 x + 1.5 \times 10^{11} t) \text{ V/m}$$

EXAMPLE 8.3

Example 8.4 Light with an energy flux of 18 W/cm^2 falls on a non-reflecting surface at normal incidence. If the surface has an area of 20 cm^2 , find the average force exerted on the surface during a 30 minute time span.

Solution

The total energy falling on the surface is

$$U = (18 \text{ W/cm}^2) \times (20 \text{ cm}^2) \times (30 \times 60) \\ = 6.48 \times 10^5 \text{ J}$$

Therefore, the total momentum delivered (for complete absorption) is

$$p = \frac{U}{c} = \frac{6.48 \times 10^5 \text{ J}}{3 \times 10^8 \text{ m/s}} = 2.16 \times 10^{-3} \text{ kg m/s}$$

The average force exerted on the surface is

$$F = \frac{p}{t} = \frac{2.16 \times 10^{-3}}{0.18 \times 10^4} = 1.2 \times 10^{-6} \text{ N}$$

How will your result be modified if the surface is a perfect reflector?

Example 8.5 Calculate the electric and magnetic fields produced by the radiation coming from a 100 W bulb at a distance of 3 m. Assume that the efficiency of the bulb is 2.5% and it is a point source.

Solution The bulb, as a point source, radiates light in all directions uniformly. At a distance of 3 m, the surface area of the surrounding sphere is

$$A = 4\pi r^2 = 4\pi (3)^2 = 113 \text{ m}^2$$

The intensity at this distance is

$$I = \frac{\text{Power}}{\text{Area}} = \frac{100 \text{ W} \times 2.5 \%}{113 \text{ m}^2} \\ = 0.022 \text{ W/m}^2$$

Half of this intensity is provided by the electric field and half by the magnetic field.

$$\frac{1}{2} I = \frac{1}{2} (\epsilon_0 E_{rms}^2 c) \\ = \frac{1}{2} (0.022 \text{ W/m}^2)$$

$$E_{rms} = \sqrt{\frac{0.022}{(8.85 \times 10^{-12})(3 \times 10^8)}} \text{ V/m} \\ = 2.9 \text{ V/m}$$

The value of E found above is the root mean square value of the electric field. Since the electric field in a light beam is sinusoidal, the peak electric field, E_0 is

$$E_0 = \sqrt{2} E_{rms} = \sqrt{2} \times 2.9 \text{ V/m} \\ = 4.07 \text{ V/m}$$

Thus, you see that the electric field strength of the light that you use for reading is fairly large. Compare it with electric field strength of TV or FM waves, which is of the order of a few microvolts per metre.

EXAMPLE 8.4

EXAMPLE 8.5

EXAMPLE 8.5

Now, let us calculate the strength of the magnetic field. It is

$$B_{rms} = \frac{E_{rms}}{c} = \frac{2.9 \text{ V m}^{-1}}{3 \times 10^8 \text{ m s}^{-1}} = 9.6 \times 10^{-9} \text{ T}$$

Again, since the field in the light beam is sinusoidal, the peak magnetic field is $B_0 = \sqrt{2} B_{rms} = 1.4 \times 10^{-8} \text{ T}$. Note that although the energy in the magnetic field is equal to the energy in the electric field, the magnetic field strength is evidently very weak.

8.4 ELECTROMAGNETIC SPECTRUM

At the time Maxwell predicted the existence of electromagnetic waves, the only familiar electromagnetic waves were the visible light waves. The existence of ultraviolet and infrared waves was barely established. By the end of the nineteenth century, X-rays and gamma rays had also been discovered. We now know that, electromagnetic waves include visible light waves, X-rays, gamma rays, radio waves, microwaves, ultraviolet and infrared waves. The classification of em waves according to frequency is the electromagnetic spectrum (Fig. 8.5). There is no sharp division between one kind of wave and the next. The classification is based roughly on how the waves are produced and/or detected.

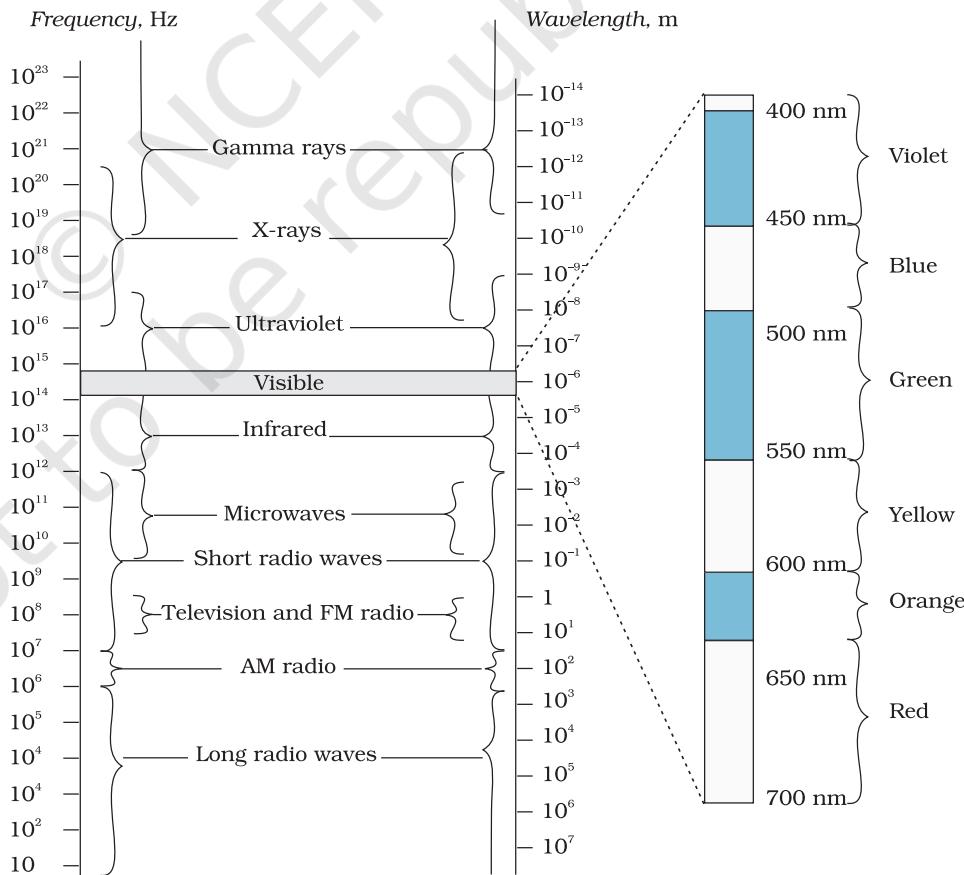


FIGURE 8.5 The electromagnetic spectrum, with common names for various parts of it. The various regions do not have sharply defined boundaries.

Electromagnetic Waves

We briefly describe these different types of electromagnetic waves, in order of decreasing wavelengths.

8.4.1 Radio waves

Radio waves are produced by the accelerated motion of charges in conducting wires. They are used in radio and television communication systems. They are generally in the frequency range from 500 kHz to about 1000 MHz. The AM (amplitude modulated) band is from 530 kHz to 1710 kHz. Higher frequencies upto 54 MHz are used for *short wave* bands. TV waves range from 54 MHz to 890 MHz. The FM (frequency modulated) radio band extends from 88 MHz to 108 MHz. Cellular phones use radio waves to transmit voice communication in the ultrahigh frequency (UHF) band. How these waves are transmitted and received is described in Chapter 15.

8.4.2 Microwaves

Microwaves (short-wavelength radio waves), with frequencies in the gigahertz (GHz) range, are produced by special vacuum tubes (called klystrons, magnetrons and Gunn diodes). Due to their short wavelengths, they are suitable for the radar systems used in aircraft navigation. Radar also provides the basis for the speed guns used to time fast balls, tennis-serves, and automobiles. Microwave ovens are an interesting domestic application of these waves. In such ovens, the frequency of the microwaves is selected to match the resonant frequency of water molecules so that energy from the waves is transferred efficiently to the kinetic energy of the molecules. This raises the temperature of any food containing water.



MICROWAVE OVEN

The spectrum of *electromagnetic radiation* contains a part known as *microwaves*. These waves have frequency and energy smaller than visible light and wavelength larger than it. What is the principle of a microwave oven and how does it work?

Our objective is to cook food or warm it up. All food items such as fruit, vegetables, meat, cereals, etc., contain water as a constituent. Now, what does it mean when we say that a certain object has become warmer? When the temperature of a body rises, the energy of the random motion of atoms and molecules increases and the molecules travel or vibrate or rotate with higher energies. The frequency of rotation of water molecules is about 300 crore hertz, which is 3 gigahertz (GHz). If water receives microwaves of this frequency, its molecules absorb this radiation, which is equivalent to heating up water. These molecules share this energy with neighbouring food molecules, heating up the food.

One should use porcelain vessels and not metal containers in a microwave oven because of the danger of getting a shock from accumulated electric charges. Metals may also melt from heating. The porcelain container remains unaffected and cool, because its large molecules vibrate and rotate with much smaller frequencies, and thus cannot absorb microwaves. Hence, they do not get heated up.

Thus, the basic principle of a microwave oven is to generate microwave radiation of appropriate frequency in the working space of the oven where we keep food. This way energy is not wasted in heating up the vessel. In the conventional heating method, the vessel on the burner gets heated first, and then the food inside gets heated because of transfer of energy from the vessel. In the microwave oven, on the other hand, energy is directly delivered to water molecules which is shared by the entire food.

8.4.3 Infrared waves

Infrared waves are produced by hot bodies and molecules. This band lies adjacent to the low-frequency or long-wave length end of the visible spectrum. Infrared waves are sometimes referred to as *heat waves*. This is because water molecules present in most materials readily absorb infrared waves (many other molecules, for example, CO₂, NH₃, also absorb infrared waves). After absorption, their thermal motion increases, that is, they heat up and heat their surroundings. Infrared lamps are used in physical therapy. Infrared radiation also plays an important role in maintaining the earth's warmth or average temperature through the greenhouse effect. Incoming visible light (which passes relatively easily through the atmosphere) is absorbed by the earth's surface and re-radiated as infrared (longer wavelength) radiations. This radiation is trapped by greenhouse gases such as carbon dioxide and water vapour. Infrared detectors are used in Earth satellites, both for military purposes and to observe growth of crops. Electronic devices (for example semiconductor light emitting diodes) also emit infrared and are widely used in the remote switches of household electronic systems such as TV sets, video recorders and hi-fi systems.

8.4.4 Visible rays

It is the most familiar form of electromagnetic waves. It is the part of the spectrum that is detected by the human eye. It runs from about 4×10^{14} Hz to about 7×10^{14} Hz or a wavelength range of about 700 – 400 nm. Visible light emitted or reflected from objects around us provides us information about the world. Our eyes are sensitive to this range of wavelengths. Different animals are sensitive to different range of wavelengths. For example, snakes can detect infrared waves, and the 'visible' range of many insects extends well into the ultraviolet.

8.4.5 Ultraviolet rays

It covers wavelengths ranging from about 4×10^{-7} m (400 nm) down to 6×10^{-10} m (0.6 nm). Ultraviolet (UV) radiation is produced by special lamps and very hot bodies. The sun is an important source of ultraviolet light. But fortunately, most of it is absorbed in the ozone layer in the atmosphere at an altitude of about 40 – 50 km. UV light in large quantities has harmful effects on humans. Exposure to UV radiation induces the production of more melanin, causing tanning of the skin. UV radiation is absorbed by ordinary glass. Hence, one cannot get tanned or sunburn through glass windows.

Welders wear special glass goggles or face masks with glass windows to protect their eyes from large amount of UV produced by welding arcs. Due to its shorter wavelengths, UV radiations can be focussed into very narrow beams for high precision applications such as LASIK (*Laser-assisted in situ keratomileusis*) eye surgery. UV lamps are used to kill germs in water purifiers.

Ozone layer in the atmosphere plays a protective role, and hence its depletion by chlorofluorocarbons (CFCs) gas (such as freon) is a matter of international concern.

Electromagnetic Waves

8.4.6 X-rays

Beyond the UV region of the electromagnetic spectrum lies the X-ray region. We are familiar with X-rays because of its medical applications. It covers wavelengths from about 10^{-8} m (10 nm) down to 10^{-13} m (10^{-4} nm). One common way to generate X-rays is to bombard a metal target by high energy electrons. X-rays are used as a diagnostic tool in medicine and as a treatment for certain forms of cancer. Because X-rays damage or destroy living tissues and organisms, care must be taken to avoid unnecessary or over exposure.

8.4.7 Gamma rays

They lie in the upper frequency range of the electromagnetic spectrum and have wavelengths of from about 10^{-10} m to less than 10^{-14} m. This high frequency radiation is produced in nuclear reactions and also emitted by radioactive nuclei. They are used in medicine to destroy cancer cells.

Table 8.1 summarises different types of electromagnetic waves, their production and detections. As mentioned earlier, the demarcation between different region is not sharp and there are overlaps.

TABLE 8.1 DIFFERENT TYPES OF ELECTROMAGNETIC WAVES

Type	Wavelength range	Production	Detection
Radio	> 0.1 m	Rapid acceleration and decelerations of electrons in aerials	Receiver's aerials
Microwave	0.1m to 1 mm	Klystron valve or magnetron valve	Point contact diodes
Infra-red	1mm to 700 nm	Vibration of atoms and molecules	Thermopiles Bolometer, Infrared photographic film
Light	700 nm to 400 nm	Electrons in atoms emit light when they move from one energy level to a lower energy level	The eye Photocells Photographic film
Ultraviolet	400 nm to 1nm	Inner shell electrons in atoms moving from one energy level to a lower level	Photocells Photographic film
X-rays	1nm to 10^{-3} nm	X-ray tubes or inner shell electrons	Photographic film Geiger tubes Ionisation chamber
Gamma rays	< 10^{-3} nm	Radioactive decay of the nucleus	-do-

SUMMARY

- Maxwell found an inconsistency in the Ampere's law and suggested the existence of an additional current, called displacement current, to remove this inconsistency. This displacement current is due to time-varying electric field and is given by

$$i_d = \epsilon_0 \frac{d\Phi_E}{dt}$$

and acts as a source of magnetic field in exactly the same way as conduction current.

- An accelerating charge produces electromagnetic waves. An electric charge oscillating harmonically with frequency ν , produces electromagnetic waves of the same frequency ν . An electric dipole is a basic source of electromagnetic waves.
- Electromagnetic waves with wavelength of the order of a few metres were first produced and detected in the laboratory by Hertz in 1887. He thus verified a basic prediction of Maxwell's equations.
- Electric and magnetic fields oscillate sinusoidally in space and time in an electromagnetic wave. The oscillating electric and magnetic fields, **E** and **B** are perpendicular to each other, and to the direction of propagation of the electromagnetic wave. For a wave of frequency ν , wavelength λ , propagating along z-direction, we have

$$\begin{aligned} E &= E_x(t) = E_0 \sin(kz - \omega t) \\ &= E_0 \sin \left[2\pi \left(\frac{z}{\lambda} - \nu t \right) \right] = E_0 \sin \left[2\pi \left(\frac{z}{\lambda} - \frac{t}{T} \right) \right] \\ B &= B_y(t) = B_0 \sin(kz - \omega t) \\ &= B_0 \sin \left[2\pi \left(\frac{z}{\lambda} - \nu t \right) \right] = B_0 \sin \left[2\pi \left(\frac{z}{\lambda} - \frac{t}{T} \right) \right] \end{aligned}$$

They are related by $E_0/B_0 = c$.

- The speed c of electromagnetic wave in vacuum is related to μ_0 and ϵ_0 (the free space permeability and permittivity constants) as follows:

$c = 1/\sqrt{\mu_0 \epsilon_0}$. The value of c equals the speed of light obtained from optical measurements.

Light is an electromagnetic wave; c is, therefore, also the speed of light. Electromagnetic waves other than light also have the same velocity c in free space.

The speed of light, or of electromagnetic waves in a material medium is given by $v = 1/\sqrt{\mu \epsilon}$

where μ is the permeability of the medium and ϵ its permittivity.

- Electromagnetic waves carry energy as they travel through space and this energy is shared equally by the electric and magnetic fields.

Electromagnetic waves transport momentum as well. When these waves strike a surface, a pressure is exerted on the surface. If total energy transferred to a surface in time t is U , total momentum delivered to this surface is $p = U/c$.

- The spectrum of electromagnetic waves stretches, in principle, over an infinite range of wavelengths. Different regions are known by different

names; γ -rays, X-rays, ultraviolet rays, visible rays, infrared rays, microwaves and radio waves in order of increasing wavelength from 10^{-2} Å or 10^{-12} m to 10^6 m.

They interact with matter via their electric and magnetic fields which set in oscillation charges present in all matter. The detailed interaction and so the mechanism of absorption, scattering, etc., depend on the wavelength of the electromagnetic wave, and the nature of the atoms and molecules in the medium.

POINTS TO PONDER

1. The basic difference between various types of electromagnetic waves lies in their wavelengths or frequencies since all of them travel through vacuum with the same speed. Consequently, the waves differ considerably in their mode of interaction with matter.
2. Accelerated charged particles radiate electromagnetic waves. The wavelength of the electromagnetic wave is often correlated with the characteristic size of the system that radiates. Thus, gamma radiation, having wavelength of 10^{-14} m to 10^{-15} m, typically originate from an atomic nucleus. X-rays are emitted from heavy atoms. Radio waves are produced by accelerating electrons in a circuit. A transmitting antenna can most efficiently radiate waves having a wavelength of about the same size as the antenna. Visible radiation emitted by atoms is, however, much longer in wavelength than atomic size.
3. The oscillating fields of an electromagnetic wave can accelerate charges and can produce oscillating currents. Therefore, an apparatus designed to detect electromagnetic waves is based on this fact. Hertz original 'receiver' worked in exactly this way. The same basic principle is utilised in practically all modern receiving devices. High frequency electromagnetic waves are detected by other means based on the physical effects they produce on interacting with matter.
4. Infrared waves, with frequencies lower than those of visible light, vibrate not only the electrons, but entire atoms or molecules of a substance. This vibration increases the internal energy and consequently, the temperature of the substance. This is why infrared waves are often called *heat waves*.
5. The centre of sensitivity of our eyes coincides with the centre of the wavelength distribution of the sun. It is because humans have evolved with visions most sensitive to the strongest wavelengths from the sun.

EXERCISES

- 8.1** Figure 8.6 shows a capacitor made of two circular plates each of radius 12 cm, and separated by 5.0 cm. The capacitor is being charged by an external source (not shown in the figure). The charging current is constant and equal to 0.15A.
- (a) Calculate the capacitance and the rate of change of potential difference between the plates.

Physics

- (b) Obtain the displacement current across the plates.
 (c) Is Kirchhoff's first rule (junction rule) valid at each plate of the capacitor? Explain.

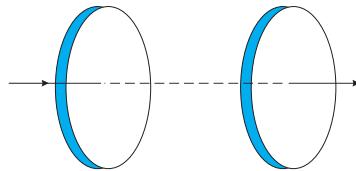


FIGURE 8.6

- 8.2** A parallel plate capacitor (Fig. 8.7) made of circular plates each of radius $R = 6.0\text{ cm}$ has a capacitance $C = 100\text{ pF}$. The capacitor is connected to a 230 V ac supply with a (angular) frequency of 300 rad s^{-1} .
 (a) What is the rms value of the conduction current?
 (b) Is the conduction current equal to the displacement current?
 (c) Determine the amplitude of \mathbf{B} at a point 3.0 cm from the axis between the plates.

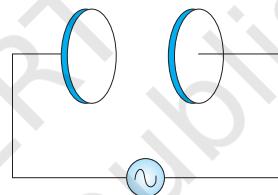


FIGURE 8.7

- 8.3** What physical quantity is the same for X-rays of wavelength 10^{-10} m , red light of wavelength 6800 \AA and radiowaves of wavelength 500 m ?
8.4 A plane electromagnetic wave travels in vacuum along z -direction. What can you say about the directions of its electric and magnetic field vectors? If the frequency of the wave is 30 MHz , what is its wavelength?
8.5 A radio can tune in to any station in the 7.5 MHz to 12 MHz band. What is the corresponding wavelength band?
8.6 A charged particle oscillates about its mean equilibrium position with a frequency of 10^9 Hz . What is the frequency of the electromagnetic waves produced by the oscillator?
8.7 The amplitude of the magnetic field part of a harmonic electromagnetic wave in vacuum is $B_0 = 510\text{ nT}$. What is the amplitude of the electric field part of the wave?
8.8 Suppose that the electric field amplitude of an electromagnetic wave is $E_0 = 120\text{ N/C}$ and that its frequency is $v = 50.0\text{ MHz}$. (a) Determine, B_0, ω , k , and λ . (b) Find expressions for \mathbf{E} and \mathbf{B} .
8.9 The terminology of different parts of the electromagnetic spectrum is given in the text. Use the formula $E = hv$ (for energy of a quantum of radiation: photon) and obtain the photon energy in units of eV for different parts of the electromagnetic spectrum. In what way are the different scales of photon energies that you obtain related to the sources of electromagnetic radiation?
8.10 In a plane electromagnetic wave, the electric field oscillates sinusoidally at a frequency of $2.0 \times 10^{10}\text{ Hz}$ and amplitude 48 V m^{-1} .

- (a) What is the wavelength of the wave?
- (b) What is the amplitude of the oscillating magnetic field?
- (c) Show that the average energy density of the **E** field equals the average energy density of the **B** field. [$c = 3 \times 10^8 \text{ m s}^{-1}$.]

ADDITIONAL EXERCISES

8.11 Suppose that the electric field part of an electromagnetic wave in vacuum is $\mathbf{E} = \{(3.1 \text{ N/C}) \cos [(1.8 \text{ rad/m}) y + (5.4 \times 10^6 \text{ rad/s})t]\}\hat{\mathbf{i}}$.

- (a) What is the direction of propagation?
- (b) What is the wavelength λ ?
- (c) What is the frequency v ?
- (d) What is the amplitude of the magnetic field part of the wave?
- (e) Write an expression for the magnetic field part of the wave.

8.12 About 5% of the power of a 100 W light bulb is converted to visible radiation. What is the average intensity of visible radiation

- (a) at a distance of 1m from the bulb?
- (b) at a distance of 10 m?

Assume that the radiation is emitted isotropically and neglect reflection.

8.13 Use the formula $\lambda_m T = 0.29 \text{ cm K}$ to obtain the characteristic temperature ranges for different parts of the electromagnetic spectrum. What do the numbers that you obtain tell you?

8.14 Given below are some famous numbers associated with electromagnetic radiations in different contexts in physics. State the part of the electromagnetic spectrum to which each belongs.

- (a) 21 cm (wavelength emitted by atomic hydrogen in interstellar space).
- (b) 1057 MHz (frequency of radiation arising from two close energy levels in hydrogen; known as Lamb shift).
- (c) 2.7 K [temperature associated with the isotropic radiation filling all space-thought to be a relic of the 'big-bang' origin of the universe].
- (d) 5890 Å - 5896 Å [double lines of sodium]
- (e) 14.4 keV [energy of a particular transition in ^{57}Fe nucleus associated with a famous high resolution spectroscopic method (Mössbauer spectroscopy)].

8.15 Answer the following questions:

- (a) Long distance radio broadcasts use short-wave bands. Why?
- (b) It is necessary to use satellites for long distance TV transmission. Why?
- (c) Optical and radiotelescopes are built on the ground but X-ray astronomy is possible only from satellites orbiting the earth. Why?
- (d) The small ozone layer on top of the stratosphere is crucial for human survival. Why?
- (e) If the earth did not have an atmosphere, would its average surface temperature be higher or lower than what it is now?
- (f) Some scientists have predicted that a global nuclear war on the earth would be followed by a severe 'nuclear winter' with a devastating effect on life on earth. What might be the basis of this prediction?

Chapter Nine

RAY OPTICS AND OPTICAL INSTRUMENTS



9.1 INTRODUCTION

Nature has endowed the human eye (retina) with the sensitivity to detect electromagnetic waves within a small range of the electromagnetic spectrum. Electromagnetic radiation belonging to this region of the spectrum (wavelength of about 400 nm to 750 nm) is called light. It is mainly through light and the sense of vision that we know and interpret the world around us.

There are two things that we can intuitively mention about light from common experience. First, that it travels with enormous speed and second, that it travels in a straight line. It took some time for people to realise that the speed of light is finite and measurable. Its presently accepted value in vacuum is $c = 2.99792458 \times 10^8 \text{ m s}^{-1}$. For many purposes, it suffices to take $c = 3 \times 10^8 \text{ m s}^{-1}$. The speed of light in vacuum is the highest speed attainable in nature.

The intuitive notion that light travels in a straight line seems to contradict what we have learnt in Chapter 8, that light is an electromagnetic wave of wavelength belonging to the visible part of the spectrum. How to reconcile the two facts? The answer is that the wavelength of light is very small compared to the size of ordinary objects that we encounter commonly (generally of the order of a few cm or larger). In this situation, as you will learn in Chapter 10, a light wave can be considered to travel from one point to another, along a straight line joining

them. The path is called a *ray* of light, and a bundle of such rays constitutes a *beam* of light.

In this chapter, we consider the phenomena of reflection, refraction and dispersion of light, using the ray picture of light. Using the basic laws of reflection and refraction, we shall study the image formation by plane and spherical reflecting and refracting surfaces. We then go on to describe the construction and working of some important optical instruments, including the human eye.

PARTICLE MODEL OF LIGHT

Newton's fundamental contributions to mathematics, mechanics, and gravitation often blind us to his deep experimental and theoretical study of light. He made pioneering contributions in the field of optics. He further developed the corpuscular model of light proposed by Descartes. It presumes that light energy is concentrated in tiny particles called *corpuscles*. He further assumed that corpuscles of light were massless elastic particles. With his understanding of mechanics, he could come up with a simple model of reflection and refraction. It is a common observation that a ball bouncing from a smooth plane surface obeys the laws of reflection. When this is an elastic collision, the magnitude of the velocity remains the same. As the surface is smooth, there is no force acting parallel to the surface, so the component of momentum in this direction also remains the same. Only the component perpendicular to the surface, i.e., the normal component of the momentum, gets reversed in reflection. Newton argued that smooth surfaces like mirrors reflect the corpuscles in a similar manner.

In order to explain the phenomena of refraction, Newton postulated that the speed of the corpuscles was greater in water or glass than in air. However, later on it was discovered that the speed of light is less in water or glass than in air.

In the field of optics, Newton – the experimenter, was greater than Newton – the theorist. He himself observed many phenomena, which were difficult to understand in terms of particle nature of light. For example, the colours observed due to a thin film of oil on water. Property of partial reflection of light is yet another such example. Everyone who has looked into the water in a pond sees image of the face in it, but also sees the bottom of the pond. Newton argued that some of the corpuscles, which fall on the water, get reflected and some get transmitted. But what property could distinguish these two kinds of corpuscles? Newton had to postulate some kind of unpredictable, chance phenomenon, which decided whether an individual corpuscle would be reflected or not. In explaining other phenomena, however, the corpuscles were presumed to behave as if they are identical. Such a dilemma does not occur in the wave picture of light. An incoming wave can be divided into two weaker waves at the boundary between air and water.

9.2 REFLECTION OF LIGHT BY SPHERICAL MIRRORS

We are familiar with the laws of reflection. The angle of reflection (i.e., the angle between reflected ray and the normal to the reflecting surface or the mirror) equals the angle of incidence (angle between incident ray and the normal). Also that the incident ray, reflected ray and the normal to the reflecting surface at the point of incidence lie in the same plane (Fig. 9.1). These laws are valid at each point on any reflecting surface whether plane or curved. However, we shall restrict our discussion to the special case of curved surfaces, that is, spherical surfaces. The normal in

Ray Optics and Optical Instruments

this case is to be taken as normal to the tangent to surface at the point of incidence. That is, the normal is along the radius, the line joining the centre of curvature of the mirror to the point of incidence.

We have already studied that the geometric centre of a spherical mirror is called its pole while that of a spherical lens is called its optical centre. The line joining the pole and the centre of curvature of the spherical mirror is known as the *principal axis*. In the case of spherical lenses, the principal axis is the line joining the optical centre with its principal focus as you will see later.

9.2.1 Sign convention

To derive the relevant formulae for reflection by spherical mirrors and refraction by spherical lenses, we must first adopt a sign convention for measuring distances. In this book, we shall follow the *Cartesian sign convention*. According to this convention, all distances are measured from the pole of the mirror or the optical centre of the lens. The distances measured in the same direction as the incident light are taken as positive and those measured in the direction opposite to the direction of incident light are taken as negative (Fig. 9.2). The heights measured upwards with respect to *x*-axis and normal to the principal axis (*x*-axis) of the mirror/lens are taken as positive (Fig. 9.2). The heights measured downwards are taken as negative.

With a common accepted convention, it turns out that a single formula for spherical mirrors and a single formula for spherical lenses can handle all different cases.

9.2.2 Focal length of spherical mirrors

Figure 9.3 shows what happens when a parallel beam of light is incident on (a) a concave mirror, and (b) a convex mirror. We assume that the rays are *paraxial*, i.e., they are incident at points close to the pole P of the mirror and make small angles with the principal axis. The reflected rays converge at a point F on the principal axis of a concave mirror [Fig. 9.3(a)]. For a convex mirror, the reflected rays appear to diverge from a point F on its principal axis [Fig. 9.3(b)]. The point F is called the *principal focus* of the mirror. If the parallel paraxial beam of light were incident, making some angle with the principal axis, the reflected rays would converge (or appear to diverge) from a point in a plane through F normal to the principal axis. This is called the *focal plane* of the mirror [Fig. 9.3(c)].

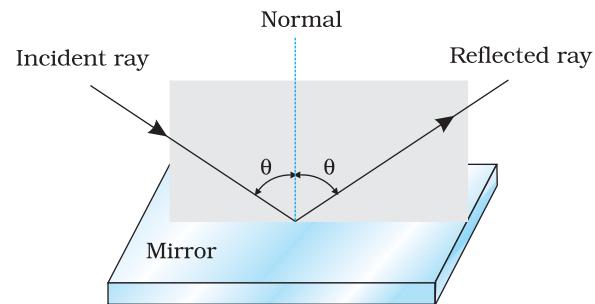


FIGURE 9.1 The incident ray, reflected ray and the normal to the reflecting surface lie in the same plane.

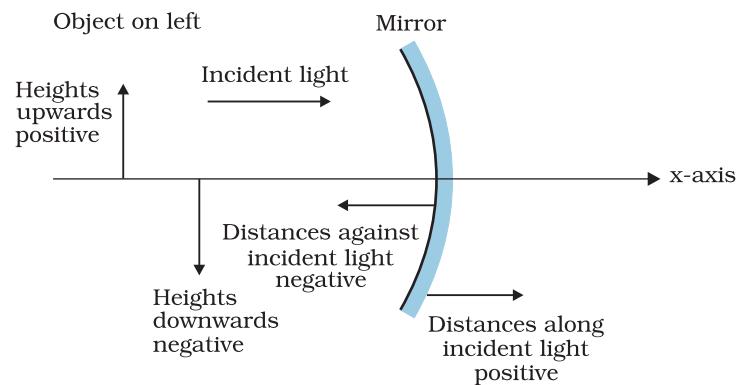


FIGURE 9.2 The Cartesian Sign Convention.

Physics

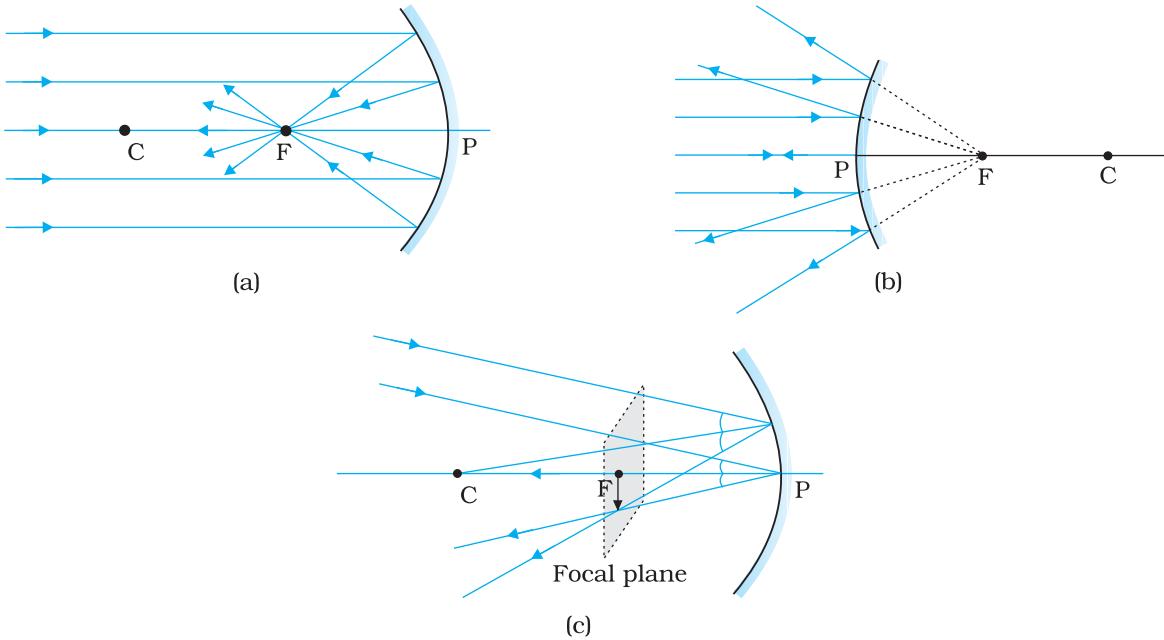


FIGURE 9.3 Focus of a concave and convex mirror.

The distance between the focus F and the pole P of the mirror is called the *focal length* of the mirror, denoted by f . We now show that $f = R/2$, where R is the radius of curvature of the mirror. The geometry of reflection of an incident ray is shown in Fig. 9.4.

Let C be the centre of curvature of the mirror. Consider a ray parallel to the principal axis striking the mirror at M . Then CM will be perpendicular to the mirror at M . Let θ be the angle of incidence, and MD be the perpendicular from M on the principal axis. Then,

$$\angle MCP = \theta \text{ and } \angle MFP = 2\theta$$

Now,

$$\tan \theta = \frac{MD}{CD} \text{ and } \tan 2\theta = \frac{MD}{FD} \quad (9.1)$$

For small θ , which is true for paraxial rays, $\tan \theta \approx \theta$, $\tan 2\theta \approx 2\theta$. Therefore, Eq. (9.1) gives

$$\frac{MD}{FD} = 2 \frac{MD}{CD}$$

$$\text{or, } FD = \frac{CD}{2} \quad (9.2)$$

Now, for small θ , the point D is very close to the point P . Therefore, $FD = f$ and $CD = R$. Equation (9.2) then gives

$$f = R/2 \quad (9.3)$$

9.2.3 The mirror equation

If rays emanating from a point actually meet at another point after reflection and/or refraction, that point is called the *image* of the first point. The image is *real* if the rays actually converge to the point; it is

Ray Optics and Optical Instruments

virtual if the rays do not actually meet but appear to diverge from the point when produced backwards. An image is thus a point-to-point correspondence with the object established through reflection and/or refraction.

In principle, we can take any two rays emanating from a point on an object, trace their paths, find their point of intersection and thus, obtain the image of the point due to reflection at a spherical mirror. In practice, however, it is convenient to choose any two of the following rays:

- The ray from the point which is parallel to the principal axis. The reflected ray goes through the focus of the mirror.
- The ray passing through the centre of curvature of a concave mirror or appearing to pass through it for a convex mirror. The reflected ray simply retraces the path.
- The ray passing through (or directed towards) the focus of the concave mirror or appearing to pass through (or directed towards) the focus of a convex mirror. The reflected ray is parallel to the principal axis.
- The ray incident at any angle at the pole. The reflected ray follows laws of reflection.

Figure 9.5 shows the ray diagram considering three rays. It shows the image $A'B'$ (in this case, real) of an object AB formed by a concave mirror. It does not mean that only three rays emanate from the point A . An infinite number of rays emanate from any source, in all directions. Thus, point A' is image point of A if every ray originating at point A and falling on the concave mirror after reflection passes through the point A' .

We now derive the mirror equation or the relation between the object distance (u), image distance (v) and the focal length (f).

From Fig. 9.5, the two right-angled triangles $A'B'F$ and MPF are similar. (For paraxial rays, MP can be considered to be a straight line perpendicular to CP .) Therefore,

$$\frac{BA}{PM} = \frac{BF}{FP}$$

$$\text{or } \frac{BA}{BA} = \frac{BF}{FP} \quad (\because PM = AB) \quad (9.4)$$

Since $\angle APB = \angle A'PB'$, the right angled triangles $A'B'P$ and ABP are also similar. Therefore,

$$\frac{BA}{BA} = \frac{BP}{BP} \quad (9.5)$$

Comparing Eqs. (9.4) and (9.5), we get

$$\frac{BF}{FP} = \frac{BP - FP}{FP} = \frac{BP}{FP} \quad (9.6)$$

Equation (9.6) is a relation involving magnitude of distances. We now apply the sign convention. We note that light travels from the object to the mirror MPN . Hence this is taken as the positive direction. To reach

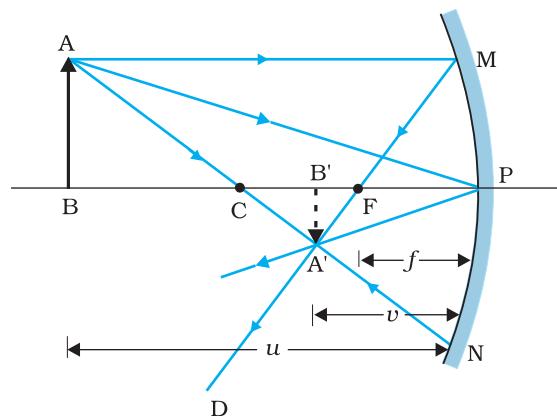


FIGURE 9.5 Ray diagram for image formation by a concave mirror.

Physics

the object AB, image A'B' as well as the focus F from the pole P, we have to travel opposite to the direction of incident light. Hence, all the three will have negative signs. Thus,

$$B'P = -v, FP = -f, BP = -u$$

Using these in Eq. (9.6), we get

$$\begin{aligned} & \frac{-v}{-f} = \frac{-v}{-u} \\ \text{or } & \frac{v-f}{f} = \frac{v}{u} \\ & \frac{1}{v} - \frac{1}{u} = \frac{1}{f} \end{aligned} \quad (9.7)$$

This relation is known as the *mirror equation*.

The size of the image relative to the size of the object is another important quantity to consider. We define linear *magnification* (m) as the ratio of the height of the image (h') to the height of the object (h):

$$m = \frac{h'}{h} \quad (9.8)$$

h and h' will be taken positive or negative in accordance with the accepted sign convention. In triangles A'B'P and ABP, we have,

$$\frac{BA}{BA} = \frac{B'P}{BP}$$

With the sign convention, this becomes

$$\frac{-h}{h} = \frac{-v}{-u}$$

so that

$$m = \frac{h}{h} = -\frac{v}{u} \quad (9.9)$$

We have derived here the mirror equation, Eq. (9.7), and the magnification formula, Eq. (9.9), for the case of real, inverted image formed by a concave mirror. With the proper use of sign convention, these are, in fact, valid for all the cases of reflection by a spherical mirror (concave or convex) whether the image formed is real or virtual. Figure 9.6 shows the ray diagrams for virtual image formed by a concave and convex mirror. You should verify that Eqs. (9.7) and (9.9) are valid for these cases as well.

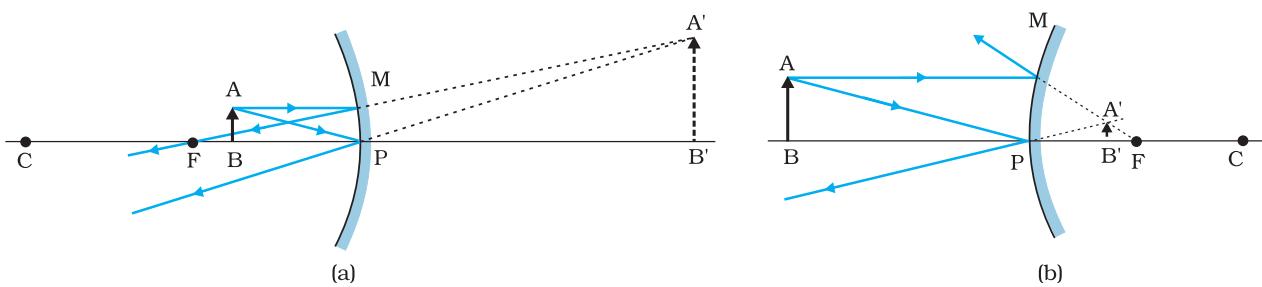


FIGURE 9.6 Image formation by (a) a concave mirror with object between P and F, and (b) a convex mirror.

Example 9.1 Suppose that the lower half of the concave mirror's reflecting surface in Fig. 9.5 is covered with an opaque (non-reflective) material. What effect will this have on the image of an object placed in front of the mirror?

Solution You may think that the image will now show only half of the object, but taking the laws of reflection to be true for all points of the remaining part of the mirror, the image will be that of the whole object. However, as the area of the reflecting surface has been reduced, the intensity of the image will be low (in this case, half).

EXAMPLE 9.1

Example 9.2 A mobile phone lies along the principal axis of a concave mirror, as shown in Fig. 9.7. Show by suitable diagram, the formation of its image. Explain why the magnification is not uniform. Will the distortion of image depend on the location of the phone with respect to the mirror?

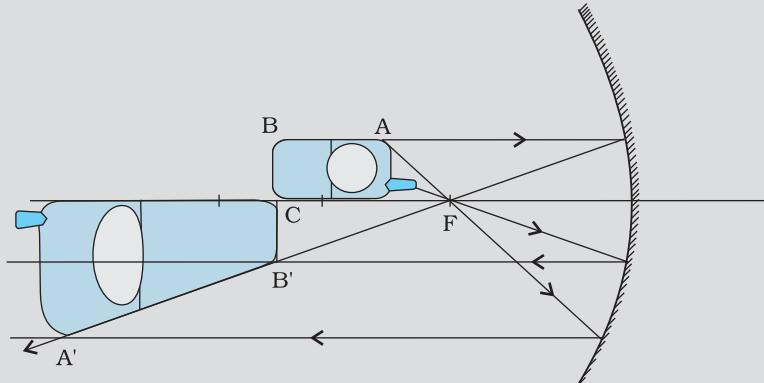


FIGURE 9.7

Solution

The ray diagram for the formation of the image of the phone is shown in Fig. 9.7. The image of the part which is on the plane perpendicular to principal axis will be on the same plane. It will be of the same size, i.e., $B'C = BC$. You can yourself realise why the image is distorted.

EXAMPLE 9.2

Example 9.3 An object is placed at (i) 10 cm, (ii) 5 cm in front of a concave mirror of radius of curvature 15 cm. Find the position, nature, and magnification of the image in each case.

Solution

The focal length $f = -15/2 \text{ cm} = -7.5 \text{ cm}$

(i) The object distance $u = -10 \text{ cm}$. Then Eq. (9.7) gives

$$\frac{1}{v} - \frac{1}{-10} = \frac{1}{-7.5}$$

$$\text{or } v = \frac{10 \cdot 7.5}{2.5} = -30 \text{ cm}$$

The image is 30 cm from the mirror on the same side as the object.

$$\text{Also, magnification } m = -\frac{v}{u} = -\frac{(30)}{(10)} = 3$$

The image is magnified, real and inverted.

EXAMPLE 9.3

EXAMPLE 9.3

- (ii) The object distance $u = -5 \text{ cm}$. Then from Eq. (9.7),

$$\frac{1}{v} = \frac{1}{5} + \frac{1}{7.5}$$

$$\text{or } v = \frac{5 \times 7.5}{7.5 - 5} = 15 \text{ cm}$$

This image is formed at 15 cm behind the mirror. It is a virtual image.

$$\text{Magnification } m = -\frac{v}{u} = -\frac{15}{(-5)} = 3$$

The image is magnified, virtual and erect.

EXAMPLE 9.4

Example 9.4 Suppose while sitting in a parked car, you notice a jogger approaching towards you in the side view mirror of $R = 2 \text{ m}$. If the jogger is running at a speed of 5 m s^{-1} , how fast the image of the jogger appear to move when the jogger is (a) 39 m, (b) 29 m, (c) 19 m, and (d) 9 m away.

Solution

From the mirror equation, Eq. (9.7), we get

$$v = \frac{fu}{u - f}$$

For convex mirror, since $R = 2 \text{ m}$, $f = 1 \text{ m}$. Then

$$\text{for } u = -39 \text{ m}, v = \frac{(39)}{39 - 1} = \frac{39}{40} \text{ m}$$

Since the jogger moves at a constant speed of 5 m s^{-1} , after 1 s the position of the image v (for $u = -39 + 5 = -34$) is $(34/35) \text{ m}$.

The shift in the position of image in 1 s is

$$\frac{39}{40} - \frac{34}{35} = \frac{1365 - 1360}{1400} = \frac{5}{1400} = \frac{1}{280} \text{ m}$$

Therefore, the average speed of the image when the jogger is between 39 m and 34 m from the mirror, is $(1/280) \text{ m s}^{-1}$

Similarly, it can be seen that for $u = -29 \text{ m}$, -19 m and -9 m , the speed with which the image appears to move is

$$\frac{1}{150} \text{ m s}^{-1}, \frac{1}{60} \text{ m s}^{-1} \text{ and } \frac{1}{10} \text{ m s}^{-1}, \text{ respectively.}$$

Although the jogger has been moving with a constant speed, the speed of his/her image appears to increase substantially as he/she moves closer to the mirror. This phenomenon can be noticed by any person sitting in a stationary car or a bus. In case of moving vehicles, a similar phenomenon could be observed if the vehicle in the rear is moving closer with a constant speed.

9.3 REFRACTION

When a beam of light encounters another transparent medium, a part of light gets reflected back into the first medium while the rest enters the other. A ray of light represents a beam. The direction of propagation of an obliquely incident ray of light that enters the other medium, changes

Ray Optics and Optical Instruments

at the interface of the two media. This phenomenon is called *refraction of light*. Snell experimentally obtained the following laws of refraction:

- The incident ray, the refracted ray and the normal to the interface at the point of incidence, all lie in the same plane.
- The ratio of the sine of the angle of incidence to the sine of angle of refraction is constant. Remember that the angles of incidence (i) and refraction (r) are the angles that the incident and its refracted ray make with the normal, respectively. We have

$$\frac{\sin i}{\sin r} = n_{21}$$

(9.10)

where n_{21} is a constant, called the *refractive index* of the second medium with respect to the first medium. Equation (9.10) is the well-known Snell's law of refraction. We note that n_{21} is a characteristic of the pair of media (and also depends on the wavelength of light), but is independent of the angle of incidence.

From Eq. (9.10), if $n_{21} > 1$, $r < i$, i.e., the refracted ray bends towards the normal. In such a case medium 2 is said to be *optically denser* (or *denser*, in short) than medium 1. On the other hand, if $n_{21} < 1$, $r > i$, the refracted ray bends away from the normal. This is the case when incident ray in a denser medium refracts into a rarer medium.

Note: Optical density should not be confused with mass density, which is mass per unit volume. It is possible that mass density of an optically denser medium may be less than that of an optically rarer medium (optical density is the ratio of the speed of light in two media). For example, turpentine and water. Mass density of turpentine is less than that of water but its optical density is higher.

If n_{21} is the refractive index of medium 2 with respect to medium 1 and n_{12} the refractive index of medium 1 with respect to medium 2, then it should be clear that

$$n_{12} = \frac{1}{n_{21}} \quad (9.11)$$

It also follows that if n_{32} is the refractive index of medium 3 with respect to medium 2 then $n_{32} = n_{31} \times n_{12}$, where n_{31} is the refractive index of medium 3 with respect to medium 1.

Some elementary results based on the laws of refraction follow immediately. For a rectangular slab, refraction takes place at two interfaces (air-glass and glass-air). It is easily seen from Fig. 9.9 that $r_2 = i_1$, i.e., the emergent ray is parallel to the incident ray—there is no

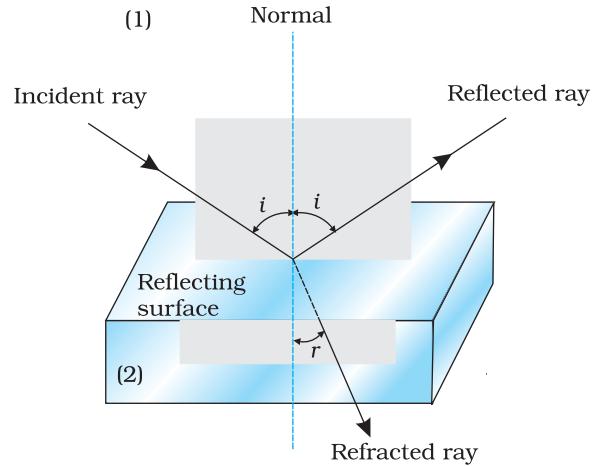


FIGURE 9.8 Refraction and reflection of light.

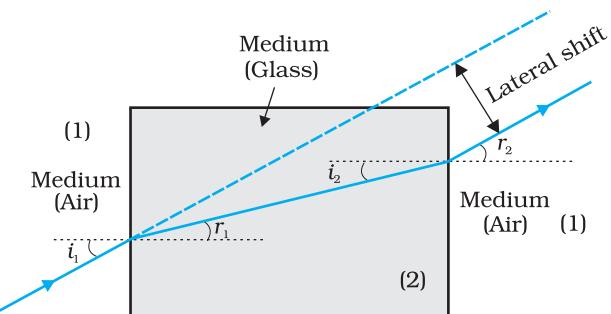


FIGURE 9.9 Lateral shift of a ray refracted through a parallel-sided slab.

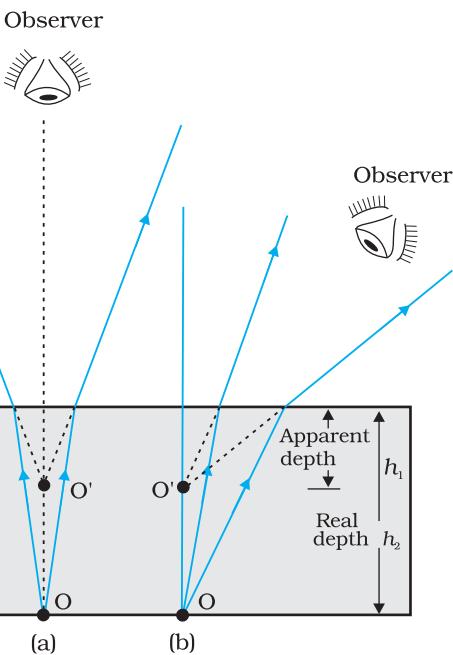


FIGURE 9.10 Apparent depth for (a) normal, and (b) oblique viewing.

deviation, but it does suffer lateral displacement/shift with respect to the incident ray. Another familiar observation is that the bottom of a tank filled with water appears to be raised (Fig. 9.10). For viewing near the normal direction, it can be shown that the apparent depth, (h_1) is real depth (h_2) divided by the refractive index of the medium (water).

The refraction of light through the atmosphere is responsible for many interesting phenomena. For example, the sun is visible a little before the actual sunrise and until a little after the actual sunset due to refraction of light through the atmosphere (Fig. 9.11). By actual sunrise we mean the actual crossing of the horizon by the sun. Figure 9.11 shows the actual and apparent positions of the sun with respect to the horizon. The figure is highly exaggerated to show the effect. The refractive index of air with respect to vacuum is 1.00029. Due to this, the apparent shift in the direction of the sun is by about half a degree and the corresponding time difference between actual sunset and apparent sunset is about 2 minutes (see Example 9.5). The apparent flattening (oval shape) of the sun at sunset and sunrise is also due to the same phenomenon.

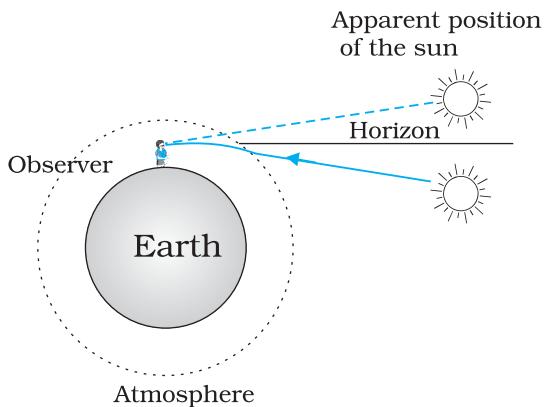


FIGURE 9.11 Advance sunrise and delayed sunset due to atmospheric refraction.

EXAMPLE 9.5

Example 9.5 The earth takes 24 h to rotate once about its axis. How much time does the sun take to shift by 1° when viewed from the earth?

Solution

Time taken for 360° shift = 24 h

Time taken for 1° shift = $24/360$ h = 4 min.

THE DROWNING CHILD, LIFEGUARD AND SNELL'S LAW

Consider a rectangular swimming pool PQSR; see figure here. A lifeguard sitting at G outside the pool notices a child drowning at a point C. The guard wants to reach the child in the shortest possible time. Let SR be the side of the pool between G and C. Should he/she take a straight line path GAC between G and C or GBC in which the path BC in water would be the shortest, or some other path GXC? The guard knows that his/her running speed v_1 on ground is higher than his/her swimming speed v_2 .

Suppose the guard enters water at X. Let $GX = l_1$ and $XC = l_2$. Then the time taken to reach from G to C would be

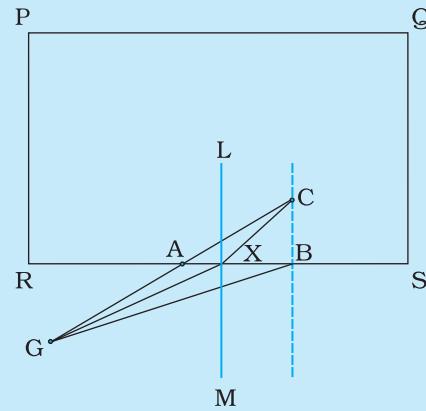
$$t = \frac{l_1}{v_1} + \frac{l_2}{v_2}$$

To make this time minimum, one has to differentiate it (with respect to the coordinate of X) and find the point X when t is a minimum. On doing all this algebra (which we skip here), we find that the guard should enter water at a point where Snell's law is satisfied. To understand this, draw a perpendicular LM to side SR at X. Let $\angle GXM = i$ and $\angle CXL = r$. Then it can be seen that t is minimum when

$$\frac{\sin i}{\sin r} = \frac{v_1}{v_2}$$

In the case of light v_1/v_2 , the ratio of the velocity of light in vacuum to that in the medium, is the refractive index n of the medium.

In short, whether it is a wave or a particle or a human being, whenever two mediums and two velocities are involved, one must follow Snell's law if one wants to take the shortest time.



9.4 TOTAL INTERNAL REFLECTION

When light travels from an optically denser medium to a rarer medium at the interface, it is partly reflected back into the same medium and partly refracted to the second medium. This reflection is called the *internal reflection*.

When a ray of light enters from a denser medium to a rarer medium, it bends away from the normal, for example, the ray AO₁B in Fig. 9.12. The incident ray AO₁ is partially reflected (O₁C) and partially transmitted (O₁B) or refracted, the angle of refraction (r) being larger than the angle of incidence (i). As the angle of incidence increases, so does the angle of refraction, till for the ray AO₃, the angle of refraction is $\pi/2$. The refracted ray is bent so much away from the normal that it grazes the surface at the interface between the two media. This is shown by the ray AO₃D in Fig. 9.12. If the angle of incidence is increased still further (e.g., the ray AO₄), refraction is not possible, and the incident ray is totally reflected.

Physics

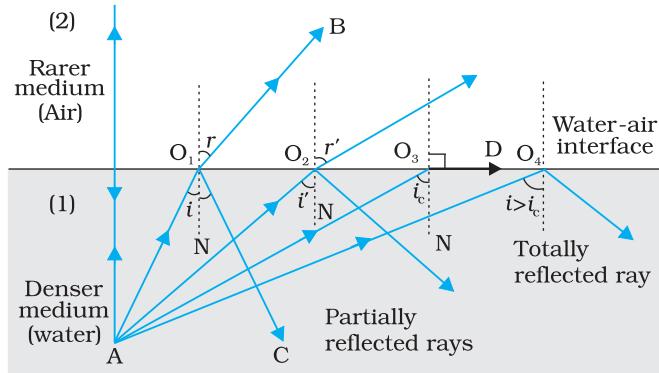


FIGURE 9.12 Refraction and internal reflection of rays from a point A in the denser medium (water) incident at different angles at the interface with a rarer medium (air).

to the value of $\sin i$ for which the law can be satisfied, that is, $i = i_c$ such that

$$\sin i_c = n_{21} \quad (9.12)$$

For values of i larger than i_c , Snell's law of refraction cannot be satisfied, and hence no refraction is possible.

The refractive index of denser medium 1 with respect to rarer medium 2 will be $n_{12} = 1/\sin i_c$. Some typical critical angles are listed in Table 9.1.

TABLE 9.1 CRITICAL ANGLE OF SOME TRANSPARENT MEDIA WITH RESPECT TO AIR

Substance medium	Refractive index	Critical angle
Water	1.33	48.75°
Crown glass	1.52	41.14°
Dense flint glass	1.62	37.31°
Diamond	2.42	24.41°

A demonstration for total internal reflection

All optical phenomena can be demonstrated very easily with the use of a laser torch or pointer, which is easily available nowadays. Take a glass beaker with clear water in it. Stir the water a few times with a piece of soap, so that it becomes a little turbid. Take a laser pointer and shine its beam through the turbid water. You will find that the path of the beam inside the water shines brightly.

Shine the beam from below the beaker such that it strikes at the upper water surface at the other end. Do you find that it undergoes partial reflection (which is seen as a spot on the table below) and partial refraction [which comes out in the air and is seen as a spot on the roof; Fig. 9.13(a)]? Now direct the laser beam from one side of the beaker such that it strikes the upper surface of water more obliquely [Fig. 9.13(b)]. Adjust the direction of laser beam until you find the angle for which the refraction

This is called *total internal reflection*. When light gets reflected by a surface, normally some fraction of it gets transmitted. The reflected ray, therefore, is always less intense than the incident ray, howsoever smooth the reflecting surface may be. In total internal reflection, on the other hand, no transmission of light takes place.

The angle of incidence corresponding to an angle of refraction 90° , say $\angle AO_3N$, is called the *critical angle* (i_c) for the given pair of media. We see from Snell's law [Eq. (9.10)] that if the relative refractive index is less than one then, since the maximum value of $\sin r$ is unity, there is an upper limit

above the water surface is totally absent and the beam is totally reflected back to water. This is total internal reflection at its simplest.

Pour this water in a long test tube and shine the laser light from top, as shown in Fig. 9.13(c). Adjust the direction of the laser beam such that it is totally internally reflected every time it strikes the walls of the tube. This is similar to what happens in optical fibres.

Take care not to look into the laser beam directly and not to point it at anybody's face.

9.4.1 Total internal reflection in nature and its technological applications

- (i) **Mirage:** On hot summer days, the air near the ground becomes hotter than the air at higher levels. The refractive index of air increases with its density. Hotter air is less dense, and has smaller refractive index than the cooler air. If the air currents are small, that is, the air is still, the optical density at different layers of air increases with height. As a result, light from a tall object such as a tree, passes through a medium whose refractive index decreases towards the ground. Thus, a ray of light from such an object successively bends away from the normal and undergoes total internal reflection, if the angle of incidence for the air near the ground exceeds the critical angle. This is shown in Fig. 9.14(b). To a distant observer, the light appears to be coming from somewhere below the ground. The observer naturally assumes that light is being reflected from the ground, say, by a pool of water near the tall object. Such inverted images of distant tall objects cause an optical illusion to the observer. This phenomenon is called *mirage*. This type of mirage is especially common in hot deserts. Some of you might have noticed that while moving in a bus or a car during a hot summer day, a distant patch of road, especially on a highway, appears to be wet. But, you do not find any evidence of wetness when you reach that spot. This is also due to mirage.

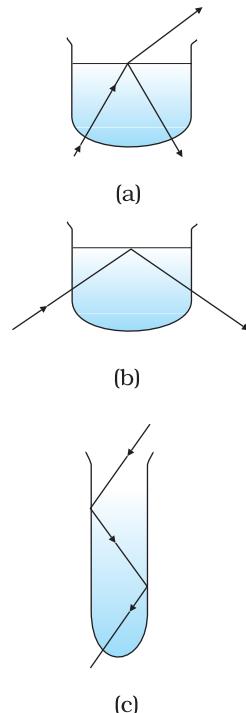


FIGURE 9.13
Observing total internal reflection in water with a laser beam (refraction due to glass of beaker neglected being very thin).

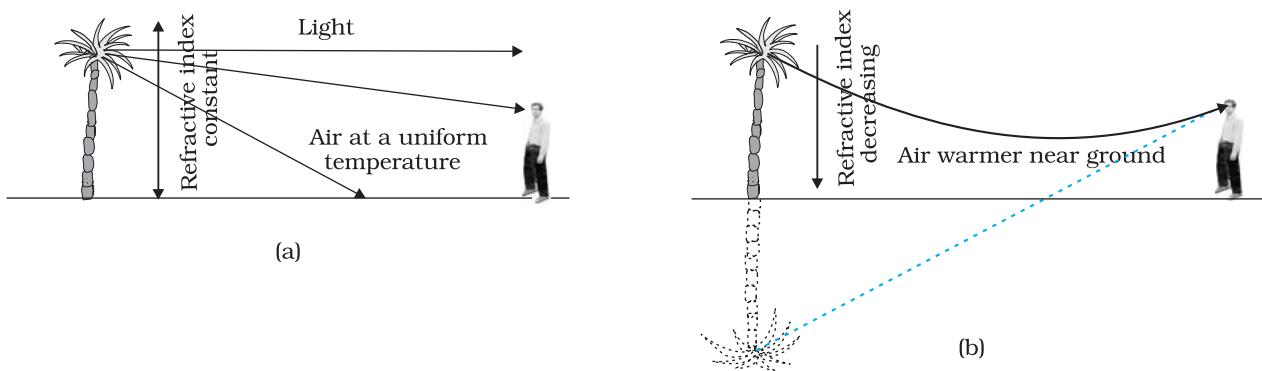


FIGURE 9.14 (a) A tree is seen by an observer at its place when the air above the ground is at uniform temperature. (b) When the layers of air close to the ground have varying temperature with hottest layers near the ground, light from a distant tree may undergo total internal reflection, and the apparent image of the tree may create an illusion to the observer that the tree is near a pool of water.

Physics

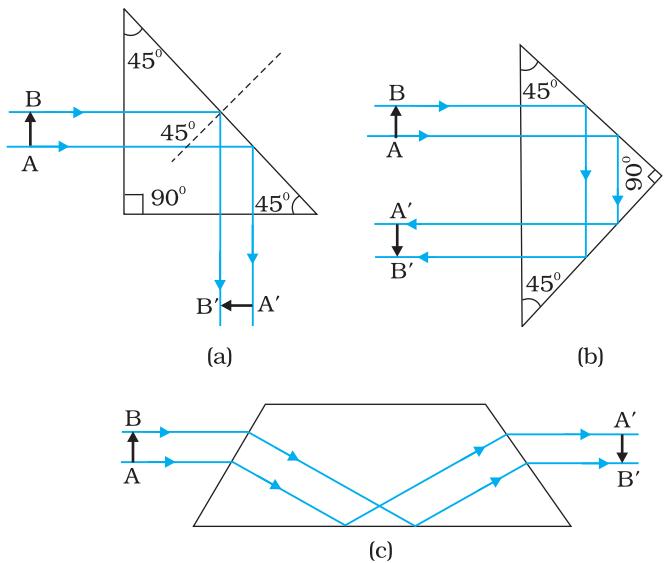


FIGURE 9.15 Prisms designed to bend rays by 90° and 180° or to invert image without changing its size make use of total internal reflection.

In the first two cases, the critical angle i_c for the material of the prism must be less than 45° . We see from Table 9.1 that this is true for both crown glass and dense flint glass.

(iv) *Optical fibres:* Now-a-days optical fibres are extensively used for transmitting audio and video signals through long distances. Optical fibres too make use of the phenomenon of total internal reflection. Optical fibres are fabricated with high quality composite glass/quartz fibres. Each fibre consists of a core and cladding. The refractive index of the material of the core is higher than that of the cladding.

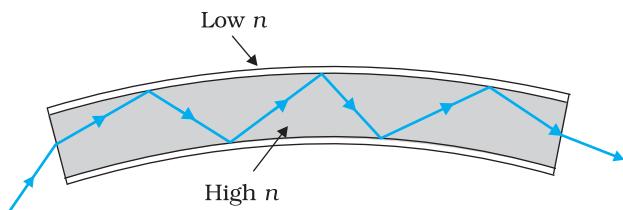


FIGURE 9.16 Light undergoes successive total internal reflections as it moves through an optical fibre.

length. Thus, an optical fibre can be used to act as an optical pipe.

A bundle of optical fibres can be put to several uses. Optical fibres are extensively used for transmitting and receiving electrical signals which are converted to light by suitable transducers. Obviously, optical fibres can also be used for transmission of optical signals. For example, these are used as a 'light pipe' to facilitate visual examination of internal organs like esophagus, stomach and intestines. You might have seen a commonly

(ii) *Diamond:* Diamonds are known for their spectacular brilliance. Their brilliance is mainly due to the total internal reflection of light inside them. The critical angle for diamond-air interface ($\approx 24.4^\circ$) is very small, therefore once light enters a diamond, it is very likely to undergo total internal reflection inside it. Diamonds found in nature rarely exhibit the brilliance for which they are known. It is the technical skill of a diamond cutter which makes diamonds to sparkle so brilliantly. By cutting the diamond suitably, multiple total internal reflections can be made to occur.

(iii) *Prism:* Prisms designed to bend light by 90° or by 180° make use of total internal reflection [Fig. 9.15(a) and (b)]. Such a prism is also used to invert images without changing their size [Fig. 9.15(c)].

When a signal in the form of light is directed at one end of the fibre at a suitable angle, it undergoes repeated total internal reflections along the length of the fibre and finally comes out at the other end (Fig. 9.16). Since light undergoes total internal reflection at each stage, there is no appreciable loss in the intensity of the light signal. Optical fibres are fabricated such that light reflected at one side of inner surface strikes the other at an angle larger than the critical angle. Even if the fibre is bent, light can easily travel along its

available decorative lamp with fine plastic fibres with their free ends forming a fountain like structure. The other end of the fibres is fixed over an electric lamp. When the lamp is switched on, the light travels from the bottom of each fibre and appears at the tip of its free end as a dot of light. The fibres in such decorative lamps are optical fibres.

The main requirement in fabricating optical fibres is that there should be very little absorption of light as it travels for long distances inside them. This has been achieved by purification and special preparation of materials such as quartz. In silica glass fibres, it is possible to transmit more than 95% of the light over a fibre length of 1 km. (Compare with what you expect for a block of ordinary window glass 1 km thick.)

9.5 REFRACTION AT SPHERICAL SURFACES AND BY LENSES

We have so far considered refraction at a plane interface. We shall now consider refraction at a spherical interface between two transparent media. An infinitesimal part of a spherical surface can be regarded as planar and the same laws of refraction can be applied at every point on the surface. Just as for reflection by a spherical mirror, the normal at the point of incidence is perpendicular to the tangent plane to the spherical surface at that point and, therefore, passes through its centre of curvature. We first consider refraction by a single spherical surface and follow it by thin lenses. A thin lens is a transparent optical medium bounded by two surfaces; at least one of which should be spherical. Applying the formula for image formation by a single spherical surface successively at the two surfaces of a lens, we shall obtain the lens maker's formula and then the lens formula.

9.5.1 Refraction at a spherical surface

Figure 9.17 shows the geometry of formation of image I of an object O on the principal axis of a spherical surface with centre of curvature C , and radius of curvature R . The rays are incident from a medium of refractive index n_1 , to another of refractive index n_2 . As before, we take the aperture (or the lateral size) of the surface to be small compared to other distances involved, so that small angle approximation can be made. In particular, NM will be taken to be nearly equal to the length of the perpendicular from the point N on the principal axis. We have, for small angles,

$$\tan \angle NOM = \frac{MN}{OM}$$

$$\tan \angle NCM = \frac{MN}{MC}$$

$$\tan \angle NIM = \frac{MN}{MI}$$

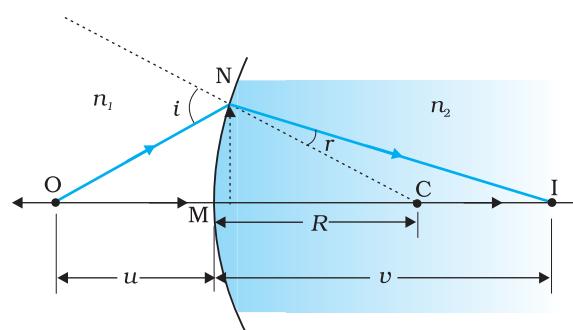


FIGURE 9.17 Refraction at a spherical surface separating two media.

LIGHT SOURCES AND PHOTOMETRY

It is known that a body above absolute zero temperature emits electromagnetic radiation. The wavelength region in which the body emits the radiation depends on its absolute temperature. Radiation emitted by a hot body, for example, a tungsten filament lamp having temperature 2850 K are partly invisible and mostly in infrared (or heat) region. As the temperature of the body increases radiation emitted by it is in visible region. The sun with temperature of about 5500 K emits radiation whose energy versus wavelength graph peaks approximately at 550 nm corresponding to green light and is almost in the middle of the visible region. The energy versus wavelength distribution graph for a given body peaks at some wavelength, which is inversely proportional to the absolute temperature of that body.

The measurement of light as perceived by human eye is called *photometry*. Photometry is measurement of a physiological phenomenon, being the stimulus of light as received by the human eye, transmitted by the optic nerves and analysed by the brain. The main physical quantities in photometry are (i) the *luminous intensity* of the source, (ii) the *luminous flux* or flow of light from the source, and (iii) *illuminance* of the surface. The SI unit of *luminous intensity* (I) is candela (cd). The candela is the luminous intensity, in a given direction, of a source that emits monochromatic radiation of frequency 540×10^{12} Hz and that has a radiant intensity in that direction of 1/683 watt per steradian. If a light source emits one candela of luminous intensity into a solid angle of one steradian, the total luminous flux emitted into that solid angle is one *lumen* (lm). A standard 100 watt incandescent light bulb emits approximately 1700 lumens.

In photometry, the only parameter, which can be measured directly is *illuminance*. It is defined as luminous flux incident per unit area on a surface (lm/m² or lux). Most light meters measure this quantity. The illuminance E , produced by a source of luminous intensity I , is given by $E = I/r^2$, where r is the normal distance of the surface from the source. A quantity named *luminance* (L), is used to characterise the brightness of emitting or reflecting flat surfaces. Its unit is cd/m² (sometimes called 'nit' in industry). A good LCD computer monitor has a brightness of about 250 nits.

Now, for Δ NOC, i is the exterior angle. Therefore, $i = \angle NOM + \angle NCM$

$$i = \frac{MN}{OM} \quad \frac{MN}{MC} \quad (9.13)$$

Similarly,

$$r = \angle NCM - \angle NIM$$

$$\text{i.e., } r = \frac{MN}{MC} \quad \frac{MN}{MI} \quad (9.14)$$

Now, by Snell's law

$$n_1 \sin i = n_2 \sin r$$

or for small angles

$$n_1 i = n_2 r$$

Substituting i and r from Eqs. (9.13) and (9.14), we get

$$\frac{n_1}{OM} \frac{n_2}{MI} = \frac{n_2}{MC} \quad (9.15)$$

Here, OM, MI and MC represent magnitudes of distances. Applying the Cartesian sign convention,

$$OM = -u, MI = +v, MC = +R$$

Substituting these in Eq. (9.15), we get

$$\frac{n_2}{v} \frac{n_1}{u} = \frac{n_2}{R} \quad (9.16)$$

Equation (9.16) gives us a relation between object and image distance in terms of refractive index of the medium and the radius of curvature of the curved spherical surface. It holds for any curved spherical surface.

Example 9.6 Light from a point source in air falls on a spherical glass surface ($n = 1.5$ and radius of curvature = 20 cm). The distance of the light source from the glass surface is 100 cm. At what position the image is formed?

Solution

We use the relation given by Eq. (9.16). Here $u = -100$ cm, $v = ?$, $R = +20$ cm, $n_1 = 1$, and $n_2 = 1.5$. We then have

$$\frac{1.5}{v} \frac{1}{100} \frac{0.5}{20}$$

or $v = +100$ cm

The image is formed at a distance of 100 cm from the glass surface, in the direction of incident light.

EXAMPLE 9.6

9.5.2 Refraction by a lens

Figure 9.18(a) shows the geometry of image formation by a double convex lens. The image formation can be seen in terms of two steps: (i) The first refracting surface forms the image I_1 of the object O [Fig. 9.18(b)]. The image I_1 acts as a virtual object for the second surface that forms the image at I [Fig. 9.18(c)]. Applying Eq. (9.15) to the first interface ABC, we get

$$\frac{n_1}{OB} \frac{n_2}{BI_1} = \frac{n_2}{BC_1} \quad (9.17)$$

A similar procedure applied to the second interface* ADC gives,

$$\frac{n_2}{DI_1} \frac{n_1}{DI} = \frac{n_2}{DC_2} \quad (9.18)$$

* Note that now the refractive index of the medium on the right side of ADC is n_1 while on its left it is n_2 . Further DI_1 is negative as the distance is measured against the direction of incident light.

Physics

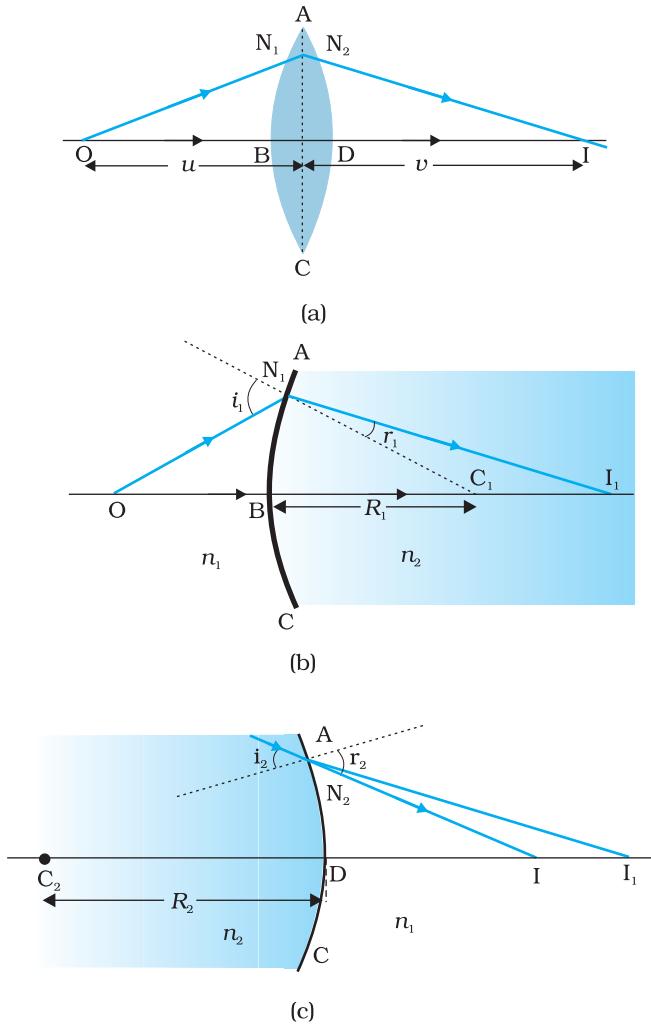


FIGURE 9.18 (a) The position of object, and the image formed by a double convex lens,
 (b) Refraction at the first spherical surface and
 (c) Refraction at the second spherical surface.

For a thin lens, $BI_1 = DI_1$. Adding Eqs. (9.17) and (9.18), we get

$$\frac{n_1}{OB} \frac{n_1}{DI} (n_2 - n_1) \frac{1}{BC_1} \frac{1}{DC_2} \quad (9.19)$$

Suppose the object is at infinity, i.e., $OB \rightarrow \infty$ and $DI = f$, Eq. (9.19) gives

$$\frac{n_1}{f} (n_2 - n_1) \frac{1}{BC_1} \frac{1}{DC_2} \quad (9.20)$$

The point where image of an object placed at infinity is formed is called the *focus F*, of the lens and the distance *f* gives its *focal length*. A lens has two foci, F and F', on either side of it (Fig. 9.19). By the sign convention,

$$BC_1 = +R_1,$$

$$DC_2 = -R_2$$

So Eq. (9.20) can be written as

$$\frac{1}{f} n_{21} 1 \frac{1}{R_1} \frac{1}{R_2} \quad \because n_{21} \frac{n_2}{n_1} \quad (9.21)$$

Equation (9.21) is known as the *lens maker's formula*. It is useful to design lenses of desired focal length using surfaces of suitable radii of curvature. Note that the formula is true for a concave lens also. In that case R_1 is negative, R_2 positive and therefore, f is negative.

From Eqs. (9.19) and (9.20), we get

$$\frac{n_1}{OB} \frac{n_1}{DI} \frac{n_1}{f} \quad (9.22)$$

Again, in the thin lens approximation, B and D are both close to the optical centre of the lens. Applying the sign convention,

$$BO = -u, DI = +v, \text{ we get}$$

$$\frac{1}{v} \frac{1}{u} \frac{1}{f} \quad (9.23)$$

Equation (9.23) is the familiar *thin lens formula*. Though we derived it for a real image formed by a convex lens, the formula is valid for both convex as well as concave lenses and for both real and virtual images.

It is worth mentioning that the two foci, F and F', of a double convex or concave lens are equidistant from the optical centre. The focus on the side of the (original) source of light is called the *first focal point*, whereas the other is called the *second focal point*.

To find the image of an object by a lens, we can, in principle, take any two rays emanating from a point on an object; trace their paths using

Ray Optics and Optical Instruments

the laws of refraction and find the point where the refracted rays meet (or appear to meet). In practice, however, it is convenient to choose any two of the following rays:

- A ray emanating from the object parallel to the principal axis of the lens after refraction passes through the second principal focus F' (in a convex lens) or appears to diverge (in a concave lens) from the first principal focus F .
- A ray of light, passing through the optical centre of the lens, emerges without any deviation after refraction.
- A ray of light passing through the first principal focus (for a convex lens) or appearing to meet at it (for a concave lens) emerges parallel to the principal axis after refraction.

Figures 9.19(a) and (b) illustrate these rules for a convex and a concave lens, respectively. You should practice drawing similar ray diagrams for different positions of the object with respect to the lens and also verify that the lens formula, Eq. (9.23), holds good for all cases.

Here again it must be remembered that each point on an object gives out infinite number of rays. All these rays will pass through the same image point after refraction at the lens.

Magnification (m) produced by a lens is defined, like that for a mirror, as the ratio of the size of the image to that of the object. Proceeding in the same way as for spherical mirrors, it is easily seen that for a lens

$$m = \frac{h'}{h} = \frac{v}{u} \quad (9.24)$$

When we apply the sign convention, we see that, for erect (and virtual) image formed by a convex or concave lens, m is positive, while for an inverted (and real) image, m is negative.

Example 9.7 A magician during a show makes a glass lens with $n = 1.47$ disappear in a trough of liquid. What is the refractive index of the liquid? Could the liquid be water?

Solution

The refractive index of the liquid must be equal to 1.47 in order to make the lens disappear. This means $n_1 = n_2$. This gives $1/f = 0$ or $f \rightarrow \infty$. The lens in the liquid will act like a plane sheet of glass. No, the liquid is not water. It could be glycerine.

EXAMPLE 9.7

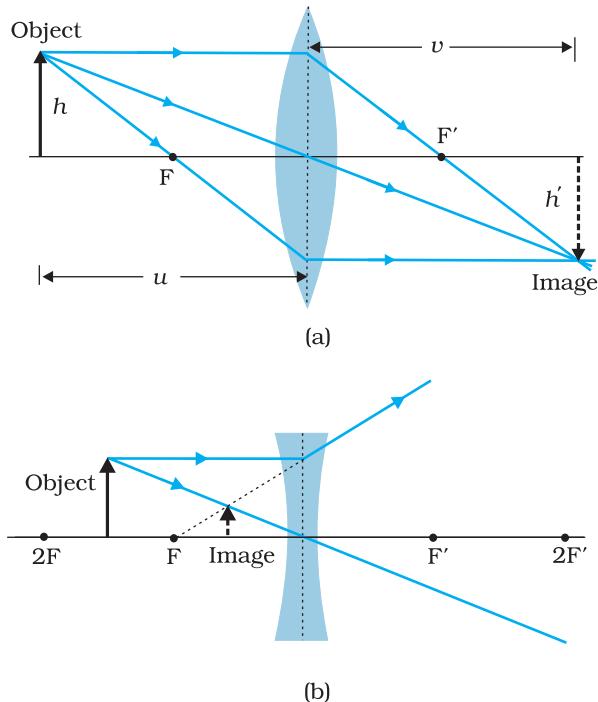
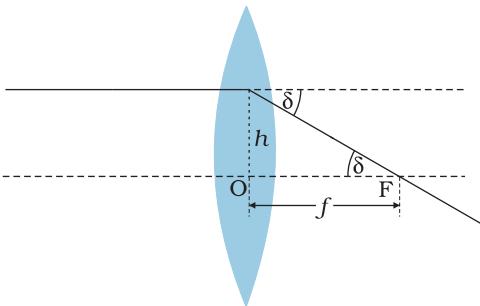


FIGURE 9.19 Tracing rays through (a) convex lens (b) concave lens.

9.5.3 Power of a lens

Power of a lens is a measure of the convergence or divergence, which a lens introduces in the light falling on it. Clearly, a lens of shorter focal



length bends the incident light more, while converging it in case of a convex lens and diverging it in case of a concave lens. The *power P* of a lens is defined as the tangent of the angle by which it converges or diverges a beam of light falling at unit distant from the optical centre (Fig. 9.20).

$$\tan \frac{h}{f}; \text{ if } h > 1 \quad \tan \frac{1}{f} \quad \text{or} \quad \frac{1}{f} \quad \text{for small value of } \delta. \text{ Thus,}$$

FIGURE 9.20 Power of a lens.

$$P = \frac{1}{f} \quad (9.25)$$

The SI unit for power of a lens is dioptre (D): $1\text{D} = 1\text{m}^{-1}$. The power of a lens of focal length of 1 metre is one dioptre. Power of a lens is positive for a converging lens and negative for a diverging lens. Thus, when an optician prescribes a corrective lens of power + 2.5 D, the required lens is a convex lens of focal length + 40 cm. A lens of power of - 4.0 D means a concave lens of focal length - 25 cm.

Example 9.8 (i) If $f = 0.5\text{ m}$ for a glass lens, what is the power of the lens? (ii) The radii of curvature of the faces of a double convex lens are 10 cm and 15 cm. Its focal length is 12 cm. What is the refractive index of glass? (iii) A convex lens has 20 cm focal length in air. What is focal length in water? (Refractive index of air-water = 1.33, refractive index for air-glass = 1.5.)

Solution

- (i) Power = +2 dioptre.
- (ii) Here, we have $f = +12\text{ cm}$, $R_1 = +10\text{ cm}$, $R_2 = -15\text{ cm}$. Refractive index of air is taken as unity. We use the lens formula of Eq. (9.22). The sign convention has to be applied for f , R_1 and R_2 . Substituting the values, we have

$$\frac{1}{12} = (n - 1) \left(\frac{1}{10} - \frac{1}{15} \right)$$

This gives $n = 1.5$.

- (iii) For a glass lens in air, $n_2 = 1.5$, $n_1 = 1$, $f = +20\text{ cm}$. Hence, the lens formula gives

$$\frac{1}{20} = 0.5 \left(\frac{1}{R_1} - \frac{1}{R_2} \right)$$

For the same glass lens in water, $n_2 = 1.5$, $n_1 = 1.33$. Therefore,

$$\frac{1.33}{f} = (1.5 - 1.33) \left(\frac{1}{R_1} - \frac{1}{R_2} \right) \quad (9.26)$$

Combining these two equations, we find $f = +78.2\text{ cm}$.

EXAMPLE 9.8

9.5.4 Combination of thin lenses in contact

Consider two lenses A and B of focal length f_1 and f_2 placed in contact with each other. Let the object be placed at a point O beyond the focus of

Ray Optics and Optical Instruments

the first lens A (Fig. 9.21). The first lens produces an image at I_1 . Since image I_1 is real, it serves as a virtual object for the second lens B, producing the final image at I. It must, however, be borne in mind that formation of image by the first lens is presumed only to facilitate determination of the position of the final image. In fact, the direction of rays emerging from the first lens gets modified in accordance with the angle at which they strike the second lens. Since the lenses are thin, we assume the optical centres of the lenses to be coincident. Let this central point be denoted by P.

For the image formed by the first lens A, we get

$$\frac{1}{v_1} - \frac{1}{u} = \frac{1}{f_1} \quad (9.27)$$

For the image formed by the second lens B, we get

$$\frac{1}{v} - \frac{1}{v_1} = \frac{1}{f_2} \quad (9.28)$$

Adding Eqs. (9.27) and (9.28), we get

$$\frac{1}{v} - \frac{1}{u} - \frac{1}{f_1} - \frac{1}{f_2} \quad (9.29)$$

If the two lens-system is regarded as equivalent to a single lens of focal length f , we have

$$\frac{1}{v} - \frac{1}{u} = \frac{1}{f}$$

so that we get

$$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2} \quad (9.30)$$

The derivation is valid for any number of thin lenses in contact. If several thin lenses of focal length f_1, f_2, f_3, \dots are in contact, the effective focal length of their combination is given by

$$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2} + \frac{1}{f_3} + \dots \quad (9.31)$$

In terms of power, Eq. (9.31) can be written as

$$P = P_1 + P_2 + P_3 + \dots \quad (9.32)$$

where P is the net power of the lens combination. Note that the sum in Eq. (9.32) is an algebraic sum of individual powers, so some of the terms on the right side may be positive (for convex lenses) and some negative (for concave lenses). Combination of lenses helps to obtain diverging or converging lenses of desired magnification. It also enhances sharpness of the image. Since the image formed by the first lens becomes the object for the second, Eq. (9.25) implies that the total magnification m of the combination is a product of magnification (m_1, m_2, m_3, \dots) of individual lenses

$$m = m_1 m_2 m_3 \dots \quad (9.33)$$

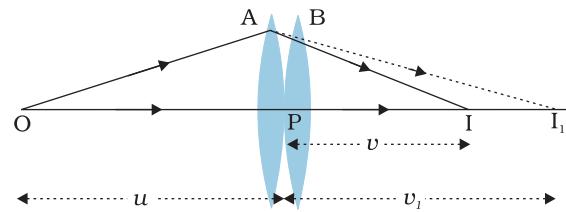


FIGURE 9.21 Image formation by a combination of two thin lenses in contact.

Such a system of combination of lenses is commonly used in designing lenses for cameras, microscopes, telescopes and other optical instruments.

Example 9.9 Find the position of the image formed by the lens combination given in the Fig. 9.22.

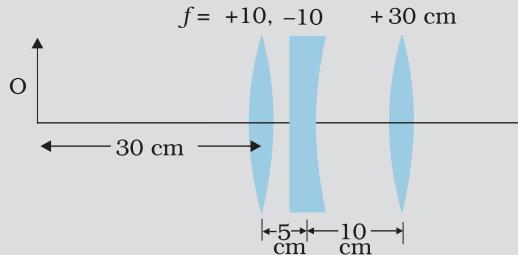


FIGURE 9.22

Solution Image formed by the first lens

$$\frac{1}{v_1} - \frac{1}{u_1} - \frac{1}{f_1}$$

$$\frac{1}{v_1} - \frac{1}{30} - \frac{1}{10}$$

$$\text{or } v_1 = 15 \text{ cm}$$

The image formed by the first lens serves as the object for the second. This is at a distance of $(15 - 5) \text{ cm} = 10 \text{ cm}$ to the right of the second lens. Though the image is real, it serves as a virtual object for the second lens, which means that the rays appear to come from it for the second lens.

$$\frac{1}{v_2} - \frac{1}{10} - \frac{1}{10}$$

$$\text{or } v_2 = \infty$$

The virtual image is formed at an infinite distance to the left of the second lens. This acts as an object for the third lens.

$$\frac{1}{v_3} - \frac{1}{u_3} - \frac{1}{f_3}$$

$$\text{or } \frac{1}{v_3} - \frac{1}{v_2} - \frac{1}{30}$$

$$\text{or } v_3 = 30 \text{ cm}$$

The final image is formed 30 cm to the right of the third lens.

EXAMPLE 9.9

9.6 REFRACTION THROUGH A PRISM

Figure 9.23 shows the passage of light through a triangular prism ABC. The angles of incidence and refraction at the first face AB are i and r_1 , while the angle of incidence (from glass to air) at the second face AC is r_2 and the angle of refraction or emergence e . The angle between the emergent ray RS and the direction of the incident ray PQ is called the *angle of deviation*, δ .

Ray Optics and Optical Instruments

In the quadrilateral AQNR, two of the angles (at the vertices Q and R) are right angles. Therefore, the sum of the other angles of the quadrilateral is 180° .

$$\angle A + \angle QNR = 180^\circ$$

From the triangle QNR,

$$r_1 + r_2 + \angle QNR = 180^\circ$$

Comparing these two equations, we get

$$r_1 + r_2 = A \quad (9.34)$$

The total deviation δ is the sum of deviations at the two faces,

$$\delta = (i - r_1) + (e - r_2)$$

that is,

$$\delta = i + e - A \quad (9.35)$$

Thus, the angle of deviation depends on the angle of incidence. A plot between the angle of deviation and angle of incidence is shown in Fig. 9.24. You can see that, in general, any given value of δ , except for $i = e$, corresponds to two values i and hence of e . This, in fact, is expected from the symmetry of i and e in Eq. (9.35), i.e., δ remains the same if i and e are interchanged. Physically, this is related to the fact that the path of ray in Fig. 9.23 can be traced back, resulting in the same angle of deviation. At the minimum deviation D_m , the refracted ray inside the prism becomes parallel to its base. We have

$$\delta = D_m, \quad i = e \text{ which implies } r_1 = r_2.$$

Equation (9.34) gives

$$2r = A \text{ or } r = \frac{A}{2} \quad (9.36)$$

In the same way, Eq. (9.35) gives

$$D_m = 2i - A, \text{ or } i = (A + D_m)/2 \quad (9.37)$$

The refractive index of the prism is

$$n_{21} = \frac{n_2}{n_1} \frac{\sin[(A - D_m)/2]}{\sin[A/2]} \quad (9.38)$$

The angles A and D_m can be measured experimentally. Equation (9.38) thus provides a method of determining refractive index of the material of the prism.

For a small angle prism, i.e., a thin prism, D_m is also very small, and we get

$$n_{21} \frac{\sin[(A - D_m)/2]}{\sin[A/2]} \approx \frac{A - D_m / 2}{A/2}$$

$$D_m = (n_{21} - 1)A$$

It implies that, thin prisms do not deviate light much.

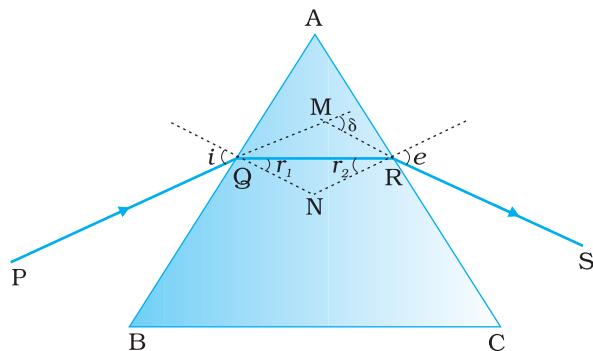


FIGURE 9.23 A ray of light passing through a triangular glass prism.

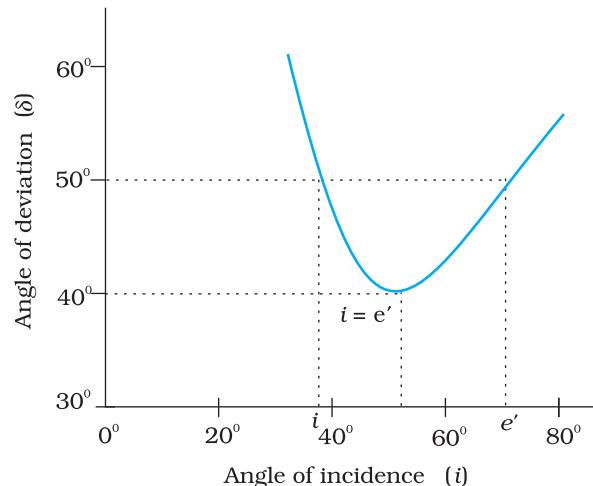


FIGURE 9.24 Plot of angle of deviation (δ) versus angle of incidence (i) for a triangular prism.

9.7 DISPERSION BY A PRISM

It has been known for a long time that when a narrow beam of sunlight, usually called white light, is incident on a glass prism, the emergent light is seen to be consisting of several colours. There is actually a continuous variation of colour, but broadly, the different component

colours that appear in sequence are: **violet, indigo, blue, green, yellow, orange** and **red** (given by the acronym VIBGYOR). The red light bends the least, while the violet light bends the most (Fig. 9.25).

The phenomenon of splitting of light into its component colours is known as *dispersion*. The pattern of colour components of light is called the spectrum of light. The word *spectrum* is now used in a much more general sense: we discussed in Chapter 8 the electromagnetic spectrum over the large range of wavelengths, from γ -rays to radio waves, of which the spectrum of light (visible spectrum) is only a small part.

Though the reason for appearance of spectrum is now common knowledge, it was a matter of much debate in the history of physics. Does the prism itself create colour in some way or does it only separate the colours already present in white light?

In a classic experiment known for its simplicity but great significance, Isaac Newton settled the issue once for all. He put another similar prism, but in an inverted position, and let the emergent beam from the first prism fall on the second prism (Fig. 9.26). The resulting emergent beam was found to be white light. The explanation was clear—the first prism splits the white light into its component colours, while the inverted prism recombines them to give white light. Thus, white light itself consists of light of different colours, which are separated by the prism.

It must be understood here that a ray of light, as defined mathematically, does not exist. An actual ray is really a beam of many rays of light. Each ray splits into component colours when it enters the glass prism. When those coloured rays come out on the other side, they again produce a white beam.

We now know that colour is associated with wavelength of light. In the visible spectrum, red light is at the long wavelength end (~ 700 nm) while the violet light is at the short wavelength end (~ 400 nm). Dispersion takes place because the refractive index of medium for different wavelengths (colours) is different. For example, the bending

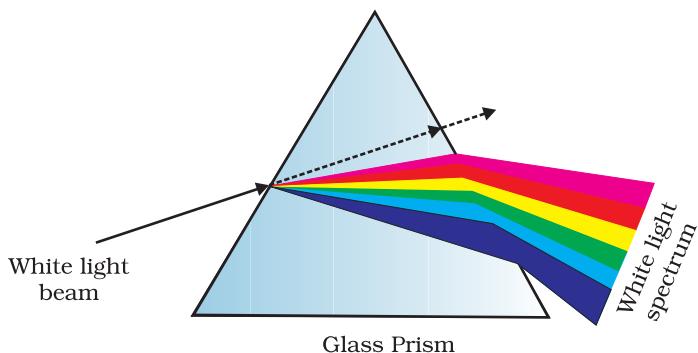


FIGURE 9.25 Dispersion of sunlight or white light on passing through a glass prism. The relative deviation of different colours shown is highly exaggerated.

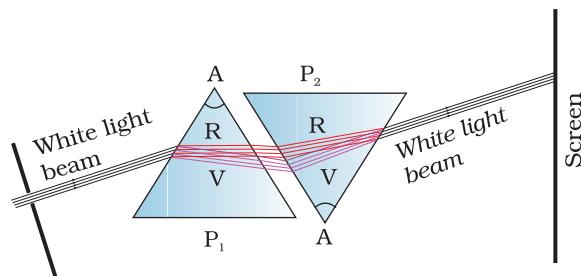


FIGURE 9.26 Schematic diagram of Newton's classic experiment on dispersion of white light.

of red component of white light is least while it is most for the violet. Equivalently, red light travels faster than violet light in a glass prism. Table 9.2 gives the refractive indices for different wavelength for crown glass and flint glass. Thick lenses could be assumed as made of many prisms, therefore, thick lenses show *chromatic aberration* due to dispersion of light.

TABLE 9.2 REFRACTIVE INDICES FOR DIFFERENT WAVELENGTHS

Colour	Wavelength (nm)	Crown glass	Flint glass
Violet	396.9	1.533	1.663
Blue	486.1	1.523	1.639
Yellow	589.3	1.517	1.627
Red	656.3	1.515	1.622

The variation of refractive index with wavelength may be more pronounced in some media than the other. In vacuum, of course, the speed of light is independent of wavelength. Thus, vacuum (or air approximately) is a non-dispersive medium in which all colours travel with the same speed. This also follows from the fact that sunlight reaches us in the form of white light and not as its components. On the other hand, glass is a dispersive medium.

9.8 SOME NATURAL PHENOMENA DUE TO SUNLIGHT

The interplay of light with things around us gives rise to several beautiful phenomena. The spectacle of colour that we see around us all the time is possible only due to sunlight. The blue of the sky, white clouds, the red-hue at sunrise and sunset, the rainbow, the brilliant colours of some pearls, shells, and wings of birds, are just a few of the natural wonders we are used to. We describe some of them here from the point of view of physics.

9.8.1 The rainbow

The rainbow is an example of the dispersion of sunlight by the water drops in the atmosphere. This is a phenomenon due to combined effect of dispersion, refraction and reflection of sunlight by spherical water droplets of rain. The conditions for observing a rainbow are that the sun should be shining in one part of the sky (say near western horizon) while it is raining in the opposite part of the sky (say eastern horizon). An observer can therefore see a rainbow only when his back is towards the sun.

In order to understand the formation of rainbows, consider Fig. (9.27(a)). Sunlight is first refracted as it enters a raindrop, which causes the different wavelengths (colours) of white light to separate. Longer wavelength of light (red) are bent the least while the shorter wavelength (violet) are bent the most. Next, these component rays strike



Formation of rainbows
<http://www.eo.ucar.edu/rainbows>
<http://www.atoptics.co.uk/bows.htm>

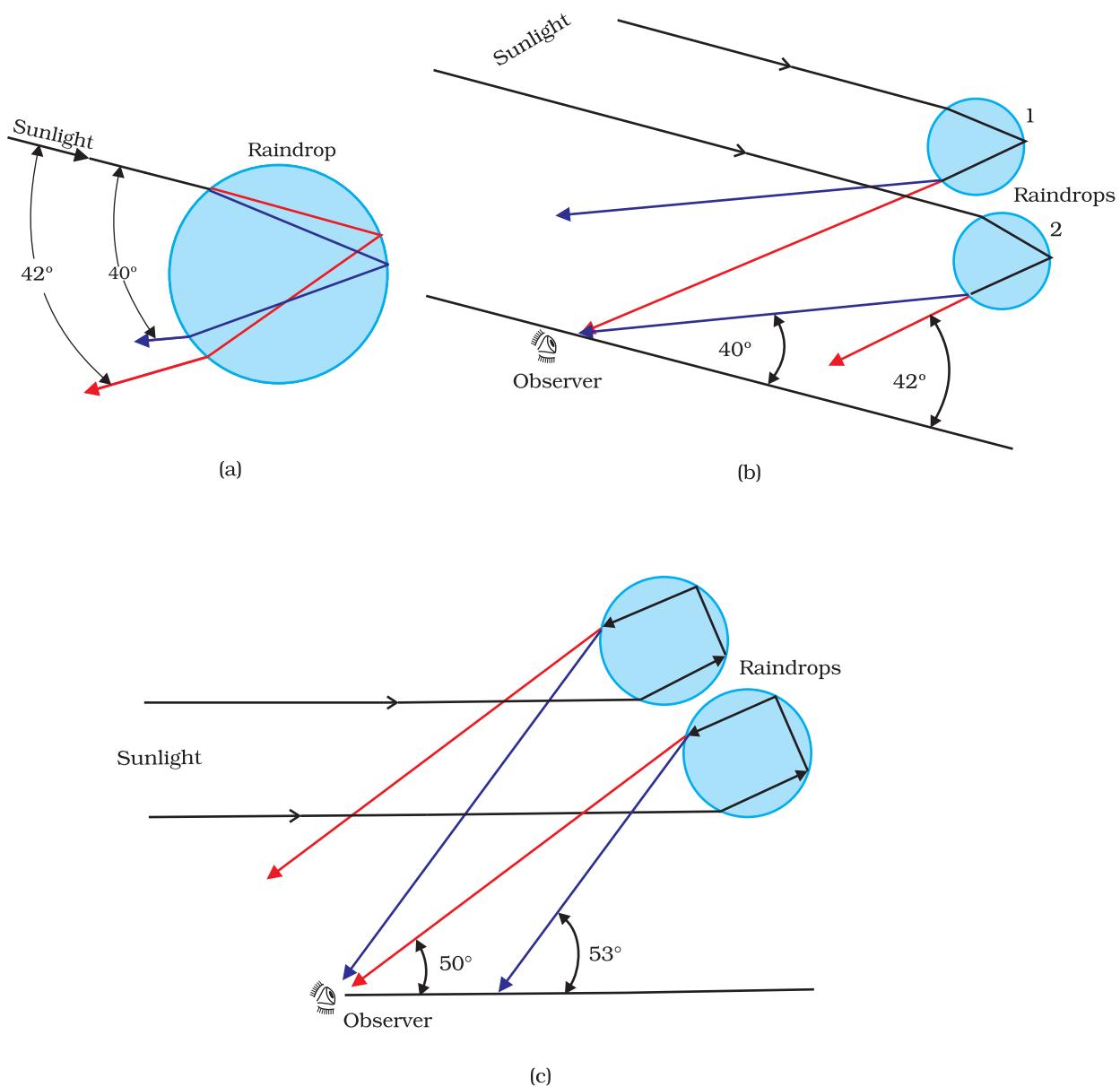


FIGURE 9.27 Rainbow: (a) The sun rays incident on a water drop get refracted twice and reflected internally by a drop; (b) Enlarge view of internal reflection and refraction of a ray of light inside a drop form primary rainbow; and (c) secondary rainbow is formed by rays undergoing internal reflection twice inside the drop.

the inner surface of the water drop and get internally reflected if the angle between the refracted ray and normal to the drop surface is greater than the critical angle (48° , in this case). The reflected light is refracted again as it comes out of the drop as shown in the figure. It is found that the violet light emerges at an angle of 40° related to the incoming sunlight and red light emerges at an angle of 42° . For other colours, angles lie in between these two values.

Figure 9.27(b) explains the formation of primary rainbow. We see that red light from drop 1 and violet light from drop 2 reach the observers eye. The violet from drop 1 and red light from drop 2 are directed at level above or below the observer. Thus the observer sees a rainbow with red colour on the top and violet on the bottom. Thus, the primary rainbow is a result of three-step process, that is, refraction, reflection and refraction.

When light rays undergoes *two* internal reflections inside a raindrop, instead of *one* as in the primary rainbow, a secondary rainbow is formed as shown in Fig. 9.27(c). It is due to four-step process. The intensity of light is reduced at the second reflection and hence the secondary rainbow is fainter than the primary rainbow. Further, the order of the colours is reversed in it as is clear from Fig. 9.27(c).

9.8.2 Scattering of light

As sunlight travels through the earth's atmosphere, it gets *scattered* (changes its direction) by the atmospheric particles. Light of shorter wavelengths is scattered much more than light of longer wavelengths. (The amount of scattering is inversely proportional to the fourth power of the wavelength. This is known as Rayleigh scattering). Hence, the bluish colour predominates in a clear sky, since blue has a shorter wavelength than red and is scattered much more strongly. In fact, violet gets scattered even more than blue, having a shorter wavelength. But since our eyes are more sensitive to blue than violet, we see the sky blue.

Large particles like dust and water droplets present in the atmosphere behave differently. The relevant quantity here is the relative size of the wavelength of light λ , and the scatterer (of typical size, say, a). For $a \ll \lambda$, one has Rayleigh scattering which is proportional to $1/\lambda^4$. For $a \gg \lambda$, i.e., large scattering objects (for example, raindrops, large dust or ice particles) this is not true; all wavelengths are scattered nearly equally. Thus, clouds which have droplets of water with $a \gg \lambda$ are generally white.

At sunset or sunrise, the sun's rays have to pass through a larger distance in the atmosphere (Fig. 9.28). Most of the blue and other shorter wavelengths are removed by scattering. The least scattered light reaching our eyes, therefore, the sun looks reddish. This explains the reddish appearance of the sun and full moon near the horizon.

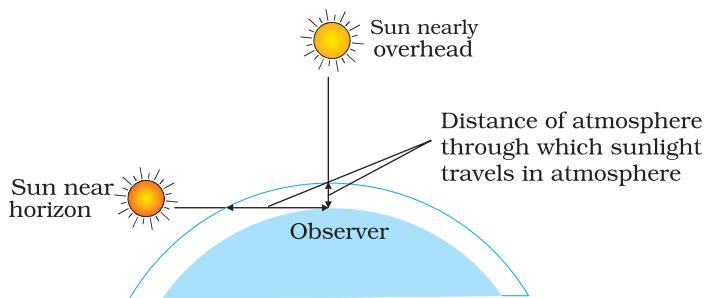


FIGURE 9.28 Sunlight travels through a longer distance in the atmosphere at sunset and sunrise.

9.9 OPTICAL INSTRUMENTS

A number of optical devices and instruments have been designed utilising reflecting and refracting properties of mirrors, lenses and prisms. Periscope, kaleidoscope, binoculars, telescopes, microscopes are some

examples of optical devices and instruments that are in common use. Our eye is, of course, one of the most important optical device the nature has endowed us with. Starting with the eye, we then go on to describe the principles of working of the microscope and the telescope.

9.9.1 The eye

Figure 9.29 (a) shows the eye. Light enters the eye through a curved front surface, the cornea. It passes through the pupil which is the central hole in the iris. The size of the pupil can change under control of muscles. The light is further focussed by the eye lens on the retina. The retina is a film of nerve fibres covering the curved back surface of the eye. The retina contains rods and cones which sense light intensity and colour, respectively, and transmit electrical signals via the optic nerve to the brain which finally processes this information. The shape (curvature) and therefore the focal length of the lens can be modified somewhat by the ciliary muscles. For example, when the muscle is relaxed, the focal length is about 2.5 cm and objects at infinity are in sharp focus on the retina. When the object is brought closer to the eye, in order to maintain the same image-lens distance (≈ 2.5 cm), the focal length of the eye lens becomes shorter by the action of the ciliary muscles. This property of the eye is called *accommodation*. If the object is too close to the eye, the lens cannot curve enough to focus the image on to the retina, and the image is blurred. The closest distance for which the lens can focus light on the retina is called the *least distance of distinct vision*, or the *near point*. The standard value for normal vision is taken as 25 cm. (Often the near point is given the symbol D.) This distance increases with age, because of the decreasing effectiveness of the ciliary muscle and the loss of flexibility of the lens. The near point may be as close as about 7 to 8 cm in a child ten years of age, and may increase to as much as 200 cm at 60 years of age. Thus, if an elderly person tries to read a book at about 25 cm from the eye, the image appears blurred. This condition (defect of the eye) is called *presbyopia*. It is corrected by using a converging lens for reading.

Thus, our eyes are marvellous organs that have the capability to interpret incoming electromagnetic waves as images through a complex process. These are our greatest assets and we must take proper care to protect them. Imagine the world without a pair of functional eyes. Yet many amongst us bravely face this challenge by effectively overcoming their limitations to lead a normal life. They deserve our appreciation for their courage and conviction.

In spite of all precautions and proactive action, our eyes may develop some defects due to various reasons. We shall restrict our discussion to some common optical defects of the eye. For example, the light from a distant object arriving at the eye-lens may get converged at a point in front of the retina. This type of defect is called *nearsightedness* or *myopia*. This means that the eye is producing too much convergence in the incident beam. To compensate this, we interpose a concave lens between the eye and the object, with the diverging effect desired to get the image focussed on the retina [Fig. 9.29(b)].

Ray Optics and Optical Instruments

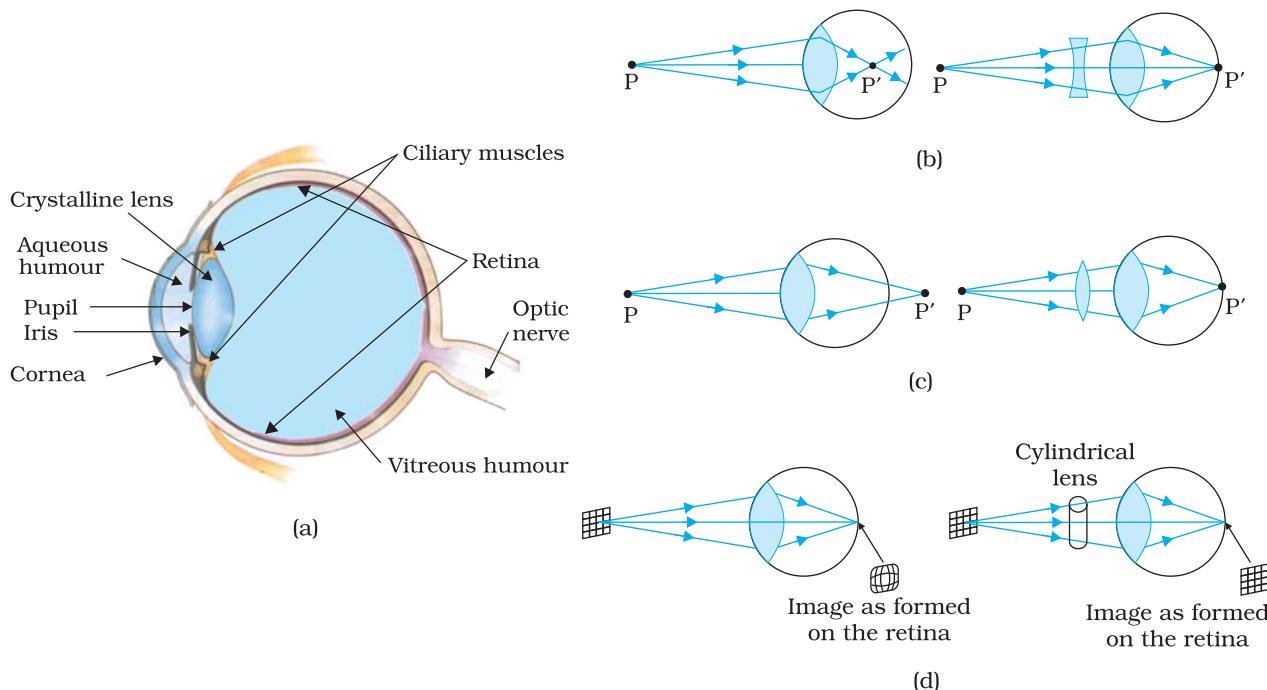


FIGURE 9.29 (a) The structure of the eye; (b) shortsighted or myopic eye and its correction; (c) farsighted or hypermetropic eye and its correction; and (d) astigmatic eye and its correction.

Similarly, if the eye-lens focusses the incoming light at a point behind the retina, a convergent lens is needed to compensate for the defect in vision. This defect is called *farsightedness* or *hypermetropia* [Fig. 9.29(c)].

Another common defect of vision is called *astigmatism*. This occurs when the cornea is not spherical in shape. For example, the cornea could have a larger curvature in the vertical plane than in the horizontal plane or vice-versa. If a person with such a defect in eye-lens looks at a wire mesh or a grid of lines, focussing in either the vertical or the horizontal plane may not be as sharp as in the other plane. Astigmatism results in lines in one direction being well focussed while those in a perpendicular direction may appear distorted [Fig. 9.29(d)]. Astigmatism can be corrected by using a cylindrical lens of desired radius of curvature with an appropriately directed axis. This defect can occur along with myopia or hypermetropia.

Example 9.10 What focal length should the reading spectacles have for a person for whom the least distance of distinct vision is 50 cm?

Solution The distance of normal vision is 25 cm. So if a book is at $u = -25$ cm, its image should be formed at $v = -50$ cm. Therefore, the desired focal length is given by

$$\frac{1}{f} = \frac{1}{v} - \frac{1}{u}$$

$$\text{or } \frac{1}{f} = \frac{1}{-50} - \frac{1}{-25} = \frac{1}{50}$$

or $f = +50$ cm (convex lens).

EXAMPLE 9.10

EXAMPLE 9.11
Example 9.11

- The far point of a myopic person is 80 cm in front of the eye. What is the power of the lens required to enable him to see very distant objects clearly?
- In what way does the corrective lens help the above person? Does the lens magnify very distant objects? Explain carefully.
- The above person prefers to remove his spectacles while reading a book. Explain why?

Solution

- Solving as in the previous example, we find that the person should use a concave lens of focal length = - 80 cm, i.e., of power = - 1.25 dioptries.
- No. The concave lens, in fact, reduces the size of the object, but the angle subtended by the distant object at the eye is the same as the angle subtended by the image (at the far point) at the eye. The eye is able to see distant objects not because the corrective lens magnifies the object, but because it brings the object (i.e., it produces virtual image of the object) at the far point of the eye which then can be focussed by the eye-lens on the retina.
- The myopic person may have a normal near point, i.e., about 25 cm (or even less). In order to read a book with the spectacles, such a person must keep the book at a distance greater than 25 cm so that the image of the book by the concave lens is produced not closer than 25 cm. The angular size of the book (or its image) at the greater distance is evidently less than the angular size when the book is placed at 25 cm and no spectacles are needed. Hence, the person prefers to remove the spectacles while reading.

EXAMPLE 9.12
Example 9.12 (a) The near point of a hypermetropic person is 75 cm from the eye. What is the power of the lens required to enable the person to read clearly a book held at 25 cm from the eye? (b) In what way does the corrective lens help the above person? Does the lens magnify objects held near the eye? (c) The above person prefers to remove the spectacles while looking at the sky. Explain why?

Solution

- $u = -25 \text{ cm}$, $v = -75 \text{ cm}$
 $1/f = 1/25 - 1/75$, i.e., $f = 37.5 \text{ cm}$.
The corrective lens needs to have a converging power of +2.67 dioptries.
- The corrective lens produces a virtual image (at 75 cm) of an object at 25 cm. The angular size of this image is the same as that of the object. In this sense the lens does not magnify the object but merely brings the object to the near point of the hypermetropic eye, which then gets focussed on the retina. However, the angular size is greater than that of the same object at the near point (75 cm) viewed without the spectacles.
- A hypermetropic eye may have normal far point i.e., it may have enough converging power to focus parallel rays from infinity on the retina of the shortened eyeball. Wearing spectacles of converging lenses (used for near vision) will amount to more converging power than needed for parallel rays. Hence the person prefers not to use the spectacles for far objects.

9.9.2 The microscope

A simple magnifier or microscope is a converging lens of small focal length (Fig. 9.30). In order to use such a lens as a microscope, the lens is held near the object, one focal length away or less, and the eye is positioned close to the lens on the other side. The idea is to get an erect, magnified and virtual image of the object at a distance so that it can be viewed comfortably, i.e., at 25 cm or more. If the object is at a distance f , the image is at infinity. However, if the object is at a distance slightly less than the focal length of the lens, the image is virtual and closer than infinity. Although the closest comfortable distance for viewing the image is when it is at the near point (distance $D \approx 25$ cm), it causes some strain on the eye. Therefore, the image formed at infinity is often considered most suitable for viewing by the relaxed eye. We show both cases, the first in Fig. 9.30(a), and the second in Fig. 9.30(b) and (c).

The linear magnification m , for the image formed at the near point D , by a simple microscope can be obtained by using the relation

$$m = \frac{v}{u} = v \cdot \frac{1}{v} - \frac{1}{f} = 1 - \frac{v}{f}$$

Now according to our sign convention, v is negative, and is equal in magnitude to D . Thus, the magnification is

$$m = 1 - \frac{D}{f} \quad (9.39)$$

Since D is about 25 cm, to have a magnification of six, one needs a convex lens of focal length, $f = 5$ cm.

Note that $m = h'/h$ where h is the size of the object and h' the size of the image. This is also the ratio of the angle subtended by the image to that subtended by the object, if placed at D for comfortable viewing. (Note that this is not the angle actually subtended by the object at the eye, which is h/u .) What a single-lens simple magnifier achieves is that it allows the object to be brought closer to the eye than D .

We will now find the magnification when the image is at infinity. In this case we will have to obtain the *angular* magnification. Suppose the object has a height h . The maximum angle it can subtend, and be clearly visible (without a lens), is when it is at the near point, i.e., a distance D . The angle subtended is then given by

$$\tan \theta_o \approx \frac{h}{D} \quad (9.40)$$

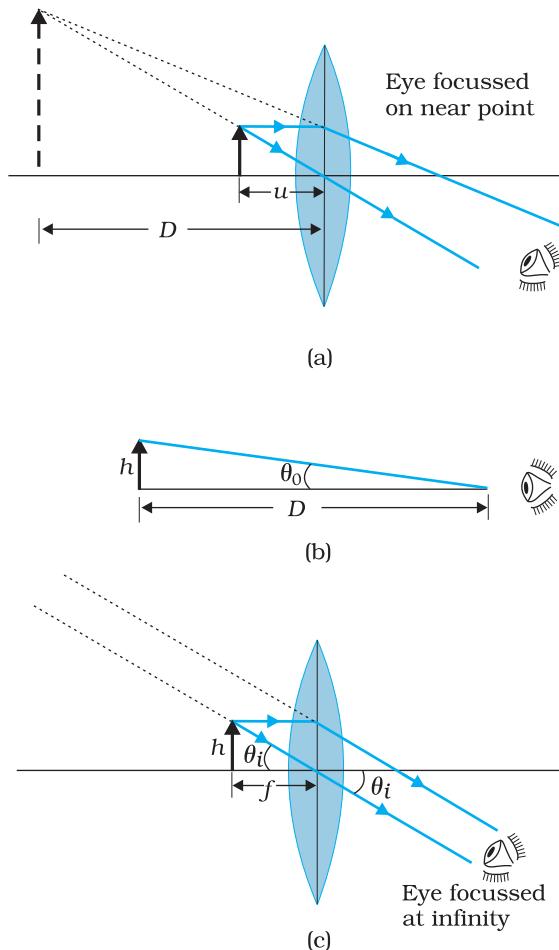


FIGURE 9.30 A simple microscope; (a) the magnifying lens is located such that the image is at the near point, (b) the angle subtended by the object, is the same as that at the near point, and (c) the object near the focal point of the lens; the image is far off but closer than infinity.

Physics

We now find the angle subtended at the eye by the image when the object is at u . From the relations

$$\frac{h}{h} = m \quad \frac{v}{u}$$

we have the angle subtended by the image

$\tan i = \frac{h}{v} = \frac{h}{v} \frac{v}{u} = \frac{h}{u} \approx \theta$. The angle subtended by the object, when it is at $u = -f$.

$$\tan i = \frac{h}{f} \quad (9.41)$$

as is clear from Fig. 9.29(c). The angular magnification is, therefore

$$m = \frac{-i}{o} = \frac{D}{f} \quad (9.42)$$

This is one less than the magnification when the image is at the near point, Eq. (9.39), but the viewing is more comfortable and the difference in magnification is usually small. In subsequent discussions of optical instruments (microscope and telescope) we shall assume the image to be at infinity.

A simple microscope has a limited maximum magnification (≤ 9) for realistic focal lengths. For much larger magnifications, one uses two lenses, one compounding the effect of the other. This is known as a

compound microscope. A schematic diagram of a compound microscope is shown in Fig. 9.31. The lens nearest the object, called the *objective*, forms a real, inverted, magnified image of the object. This serves as the object for the second lens, the *eyepiece*, which functions essentially like a simple microscope or magnifier, produces the final image, which is enlarged and virtual. The first inverted image is thus near (at or within) the focal plane of the eyepiece, at a distance appropriate for final image formation at infinity, or a little closer for image formation at the near point. Clearly, the final image is inverted with respect to the original object.

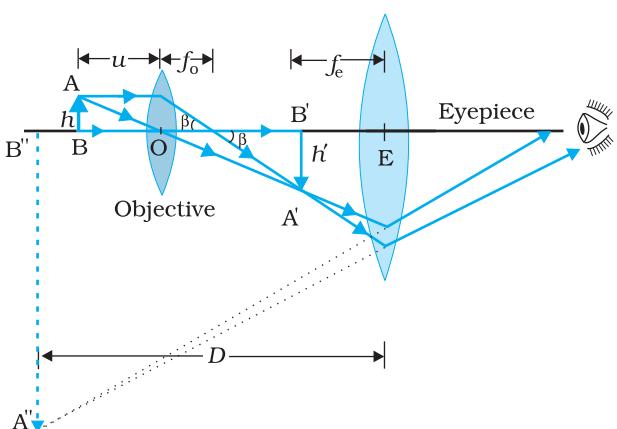


FIGURE 9.31 Ray diagram for the formation of image by a compound microscope.

We now obtain the magnification due to a compound microscope. The ray diagram of Fig. 9.31 shows that the (linear) magnification due to the objective, namely h'/h , equals

$$m_o = \frac{h'}{h} = \frac{L}{f_o} \quad (9.43)$$

where we have used the result

$$\tan i = \frac{h}{f_o} = \frac{h}{L}$$

Here h' is the size of the first image, the object size being h and f_o being the focal length of the objective. The first image is formed near the focal point of the eyepiece. The distance L , i.e., the distance between the second focal point of the objective and the first focal point of the eyepiece (focal length f_e) is called the tube length of the compound microscope.

As the first inverted image is near the focal point of the eyepiece, we use the result from the discussion above for the simple microscope to obtain the (angular) magnification m_e due to it [Eq. (9.39)], when the final image is formed at the near point, is

$$m_e = 1 - \frac{D}{f_e} \quad [9.44(a)]$$

When the final image is formed at infinity, the angular magnification due to the eyepiece [Eq. (9.42)] is

$$m_e = (D/f_e) \quad [9.44(b)]$$

Thus, the total magnification [(according to Eq. (9.33)], when the image is formed at infinity, is

$$m = m_o m_e = \frac{L}{f_o} \cdot \frac{D}{f_e} \quad (9.45)$$

Clearly, to achieve a large magnification of a *small* object (hence the name microscope), the objective and eyepiece should have small focal lengths. In practice, it is difficult to make the focal length much smaller than 1 cm. Also large lenses are required to make L large.

For example, with an objective with $f_o = 1.0$ cm, and an eyepiece with focal length $f_e = 2.0$ cm, and a tube length of 20 cm, the magnification is

$$m = m_o m_e = \frac{L}{f_o} \cdot \frac{D}{f_e}$$

$$= \frac{20}{1} \cdot \frac{25}{2} = 250$$

Various other factors such as illumination of the object, contribute to the quality and visibility of the image. In modern microscopes, multi-component lenses are used for both the objective and the eyepiece to improve image quality by minimising various optical aberrations (defects) in lenses.

9.9.3 Telescope

The telescope is used to provide angular magnification of distant objects (Fig. 9.32). It also has an objective and an eyepiece. But here, the objective has a large focal length and a much larger aperture than the eyepiece. Light from a distant object enters the objective and a real image is formed in the tube at its second focal point. The eyepiece magnifies this image producing a final inverted image. The magnifying power m is the ratio of the angle β subtended at the eye by the final image to the angle α which the object subtends at the lens or the eye. Hence

$$m = \frac{h}{f_e} \cdot \frac{f_o}{h} = \frac{f_o}{f_e} \quad (9.46)$$

In this case, the length of the telescope tube is $f_o + f_e$.



The world's largest optical telescopes
<http://astro.nineplanets.org/bigeyes.html>

Physics

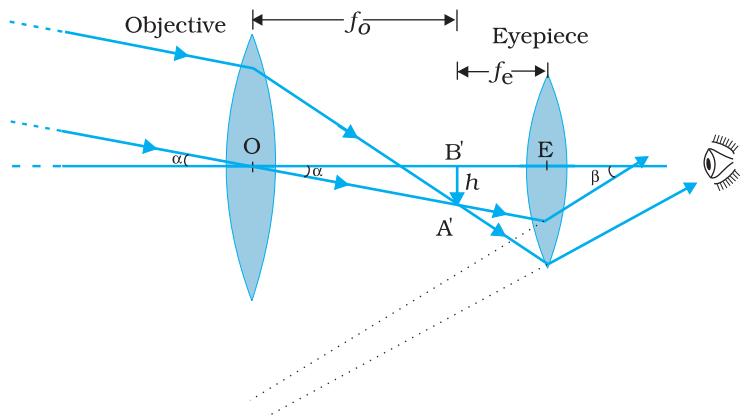


FIGURE 9.32 A refracting telescope.

The main considerations with an astronomical telescope are its light gathering power and its resolution or resolving power. The former clearly depends on the area of the objective. With larger diameters, fainter objects can be observed. The resolving power, or the ability to observe two objects distinctly, which are in very nearly the same direction, also depends on the diameter of the objective. So, the desirable aim in optical telescopes is to make them with objective of large diameter. The largest lens objective in use has a diameter of 40 inch (~ 1.02 m). It is at the Yerkes Observatory in Wisconsin, USA. Such big lenses tend to be very heavy and therefore, difficult to make and support by their edges. Further, it is rather difficult and expensive to make such large sized lenses which form images that are free from any kind of chromatic aberration and distortions.

For these reasons, modern telescopes use a concave mirror rather than a lens for the objective. Telescopes with mirror objectives are called *reflecting telescopes*. They have several advantages. First, there is no chromatic aberration in a mirror. Second, if a parabolic reflecting surface is chosen, spherical aberration is also removed. Mechanical support is much less of a problem since a mirror weighs much less than a lens of equivalent optical quality, and can be supported over its entire back surface, not just over its rim. One obvious problem with a reflecting telescope is that the objective mirror focusses light inside the telescope tube. One must have an eyepiece and the observer right there, obstructing some light (depending on the size of the observer cage). This is what is done in the very large 200 inch (~ 5.08 m) diameters, Mt. Palomar telescope, California. The viewer sits near the focal point of the mirror, in a small cage. Another solution to the problem is to deflect the light being focussed by another mirror. One such arrangement using a convex secondary mirror to focus the incident light, which now passes through a hole in the objective primary mirror, is shown

Terrestrial telescopes have, in addition, a pair of inverting lenses to make the final image erect. Refracting telescopes can be used both for terrestrial and astronomical observations. For example, consider a telescope whose objective has a focal length of 100 cm and the eyepiece a focal length of 1 cm. The magnifying power of this telescope is $m = 100/1 = 100$.

Let us consider a pair of stars of actual separation $1'$ (one minute of arc). The stars appear as though they are separated by an angle of $100 \times 1' = 100' = 1.67^\circ$.

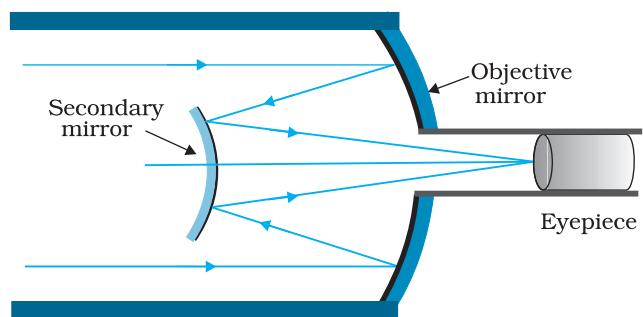


FIGURE 9.33 Schematic diagram of a reflecting telescope (Cassegrain).

in Fig. 9.33. This is known as a *Cassegrain* telescope, after its inventor. It has the advantages of a large focal length in a short telescope. The largest telescope in India is in Kavalur, Tamil Nadu. It is a 2.34 m diameter reflecting telescope (Cassegrain). It was ground, polished, set up, and is being used by the Indian Institute of Astrophysics, Bangalore. The largest reflecting telescopes in the world are the pair of Keck telescopes in Hawaii, USA, with a reflector of 10 metre in diameter.

SUMMARY

1. Reflection is governed by the equation $\angle i = \angle r'$ and refraction by the Snell's law, $\sin i / \sin r = n$, where the incident ray, reflected ray, refracted ray and normal lie in the same plane. Angles of incidence, reflection and refraction are i , r' and r , respectively.
2. The *critical angle of incidence* i_c for a ray incident from a denser to rarer medium, is that angle for which the angle of refraction is 90° . For $i > i_c$, total internal reflection occurs. Multiple internal reflections in diamond ($i_c \approx 24.4^\circ$), totally reflecting prisms and mirage, are some examples of total internal reflection. Optical fibres consist of glass fibres coated with a thin layer of material of *lower* refractive index. Light incident at an angle at one end comes out at the other, after multiple internal reflections, even if the fibre is bent.
3. *Cartesian sign convention*: Distances measured in the same direction as the incident light are positive; those measured in the opposite direction are negative. All distances are measured from the pole/optic centre of the mirror/lens on the principal axis. The heights measured upwards above x -axis and normal to the principal axis of the mirror/lens are taken as positive. The heights measured downwards are taken as negative.
4. *Mirror equation*:

$$\frac{1}{v} - \frac{1}{u} = \frac{1}{f}$$

where u and v are object and image distances, respectively and f is the focal length of the mirror. f is (approximately) half the radius of curvature R . f is negative for concave mirror; f is positive for a convex mirror.

5. For a prism of the angle A , of refractive index n_2 placed in a medium of refractive index n_1 ,

$$n_{21} = \frac{n_2}{n_1} = \frac{\sin A / D_m / 2}{\sin A / 2}$$

where D_m is the angle of minimum deviation.

6. For refraction through a spherical interface (from medium 1 to 2 of refractive index n_1 and n_2 , respectively)

$$\frac{n_2}{v} - \frac{n_1}{u} = \frac{n_2 - n_1}{R}$$

Thin lens formula

$$\frac{1}{v} - \frac{1}{u} = \frac{1}{f}$$

Lens maker's formula

$$\frac{1}{f} = \frac{n_2 - n_1}{n_1} \cdot \frac{1}{R_1} - \frac{1}{R_2}$$

R_1 and R_2 are the radii of curvature of the lens surfaces. f is positive for a converging lens; f is negative for a diverging lens. The power of a lens $P = 1/f$.

The SI unit for power of a lens is dioptre (D): $1 \text{ D} = 1 \text{ m}^{-1}$.

If several thin lenses of focal length f_1, f_2, f_3, \dots are in contact, the effective focal length of their combination, is given by

$$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2} + \frac{1}{f_3} + \dots$$

The total power of a combination of several lenses is

$$P = P_1 + P_2 + P_3 + \dots$$

7. *Dispersion* is the splitting of light into its constituent colours.
8. *The Eye:* The eye has a convex lens of focal length about 2.5 cm. This focal length can be varied somewhat so that the image is always formed on the retina. This ability of the eye is called *accommodation*. In a defective eye, if the image is focussed before the retina (myopia), a diverging corrective lens is needed; if the image is focussed beyond the retina (hypermetropia), a converging corrective lens is needed. Astigmatism is corrected by using cylindrical lenses.
9. *Magnifying power m of a simple microscope* is given by $m = 1 + (D/f)$, where $D = 25 \text{ cm}$ is the least distance of distinct vision and f is the focal length of the convex lens. If the image is at infinity, $m = D/f$. For a compound microscope, the magnifying power is given by $m = m_e \times m_o$ where $m_e = 1 + (D/f_e)$, is the magnification due to the eyepiece and m_o is the magnification produced by the objective. Approximately,

$$m = \frac{L}{f_o} \cdot \frac{D}{f_e}$$

where f_o and f_e are the focal lengths of the objective and eyepiece, respectively, and L is the distance between their focal points.

10. *Magnifying power m of a telescope* is the ratio of the angle β subtended at the eye by the image to the angle α subtended at the eye by the object.

$$m = \frac{f_o}{f_e}$$

where f_o and f_e are the focal lengths of the objective and eyepiece, respectively.

POINTS TO PONDER

1. The laws of reflection and refraction are true for all surfaces and pairs of media at the point of the incidence.
2. The real image of an object placed between f and $2f$ from a convex lens can be seen on a screen placed at the image location. If the screen is removed, is the image still there? This question puzzles many, because it is difficult to reconcile ourselves with an image suspended in air.

without a screen. But the image does exist. Rays from a given point on the object are converging to an image point in space and diverging away. The screen simply diffuses these rays, some of which reach our eye and we see the image. This can be seen by the images formed in air during a laser show.

3. Image formation needs regular reflection/refraction. In principle, all rays from a given point should reach the same image point. This is why you do not see your image by an irregular reflecting object, say the page of a book.
4. Thick lenses give coloured images due to dispersion. The variety in colour of objects we see around us is due to the constituent colours of the light incident on them. A monochromatic light may produce an entirely different perception about the colours on an object as seen in white light.
5. For a simple microscope, the angular size of the object equals the angular size of the image. Yet it offers magnification because we can keep the small object much closer to the eye than 25 cm and hence have it subtend a large angle. The image is at 25 cm which we can see. Without the microscope, you would need to keep the small object at 25 cm which would subtend a very small angle.

EXERCISES

- 9.1** A small candle, 2.5 cm in size is placed at 27 cm in front of a concave mirror of radius of curvature 36 cm. At what distance from the mirror should a screen be placed in order to obtain a sharp image? Describe the nature and size of the image. If the candle is moved closer to the mirror, how would the screen have to be moved?
- 9.2** A 4.5 cm needle is placed 12 cm away from a convex mirror of focal length 15 cm. Give the location of the image and the magnification. Describe what happens as the needle is moved farther from the mirror.
- 9.3** A tank is filled with water to a height of 12.5 cm. The apparent depth of a needle lying at the bottom of the tank is measured by a microscope to be 9.4 cm. What is the refractive index of water? If water is replaced by a liquid of refractive index 1.63 up to the same height, by what distance would the microscope have to be moved to focus on the needle again?
- 9.4** Figures 9.34(a) and (b) show refraction of a ray in air incident at 60° with the normal to a glass-air and water-air interface, respectively. Predict the angle of refraction in glass when the angle of incidence in water is 45° with the normal to a water-glass interface [Fig. 9.34(c)].

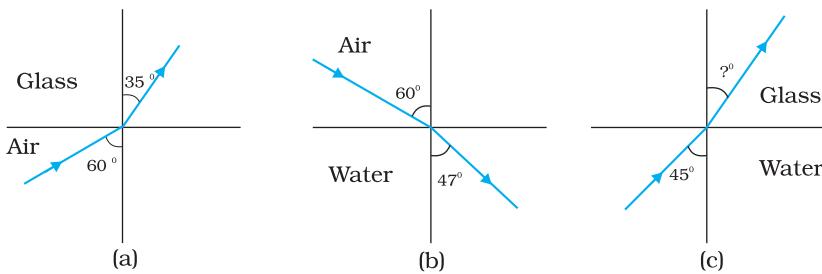


FIGURE 9.34

Physics

- 9.5** A small bulb is placed at the bottom of a tank containing water to a depth of 80cm. What is the area of the surface of water through which light from the bulb can emerge out? Refractive index of water is 1.33. (Consider the bulb to be a point source.)
- 9.6** A prism is made of glass of unknown refractive index. A parallel beam of light is incident on a face of the prism. The angle of minimum deviation is measured to be 40° . What is the refractive index of the material of the prism? The refracting angle of the prism is 60° . If the prism is placed in water (refractive index 1.33), predict the new angle of minimum deviation of a parallel beam of light.
- 9.7** Double-convex lenses are to be manufactured from a glass of refractive index 1.55, with both faces of the same radius of curvature. What is the radius of curvature required if the focal length is to be 20cm?
- 9.8** A beam of light converges at a point P. Now a lens is placed in the path of the convergent beam 12cm from P. At what point does the beam converge if the lens is (a) a convex lens of focal length 20cm, and (b) a concave lens of focal length 16cm?
- 9.9** An object of size 3.0cm is placed 14cm in front of a concave lens of focal length 21cm. Describe the image produced by the lens. What happens if the object is moved further away from the lens?
- 9.10** What is the focal length of a convex lens of focal length 30cm in contact with a concave lens of focal length 20cm? Is the system a converging or a diverging lens? Ignore thickness of the lenses.
- 9.11** A compound microscope consists of an objective lens of focal length 2.0cm and an eyepiece of focal length 6.25cm separated by a distance of 15cm. How far from the objective should an object be placed in order to obtain the final image at (a) the least distance of distinct vision (25cm), and (b) at infinity? What is the magnifying power of the microscope in each case?
- 9.12** A person with a normal near point (25cm) using a compound microscope with objective of focal length 8.0 mm and an eyepiece of focal length 2.5cm can bring an object placed at 9.0mm from the objective in sharp focus. What is the separation between the two lenses? Calculate the magnifying power of the microscope,
- 9.13** A small telescope has an objective lens of focal length 144cm and an eyepiece of focal length 6.0cm. What is the magnifying power of the telescope? What is the separation between the objective and the eyepiece?
- 9.14** (a) A giant refracting telescope at an observatory has an objective lens of focal length 15m. If an eyepiece of focal length 1.0cm is used, what is the angular magnification of the telescope?
(b) If this telescope is used to view the moon, what is the diameter of the image of the moon formed by the objective lens? The diameter of the moon is 3.48×10^6 m, and the radius of lunar orbit is 3.8×10^8 m.
- 9.15** Use the mirror equation to deduce that:
(a) an object placed between f and $2f$ of a concave mirror produces a real image beyond $2f$.
(b) a convex mirror always produces a virtual image independent of the location of the object.
(c) the virtual image produced by a convex mirror is always diminished in size and is located between the focus and the pole.

Ray Optics and Optical Instruments

- (d) an object placed between the pole and focus of a concave mirror produces a virtual and enlarged image.

[Note: This exercise helps you deduce algebraically properties of images that one obtains from explicit ray diagrams.]

- 9.16** A small pin fixed on a table top is viewed from above from a distance of 50 cm. By what distance would the pin appear to be raised if it is viewed from the same point through a 15 cm thick glass slab held parallel to the table? Refractive index of glass = 1.5. Does the answer depend on the location of the slab?

- 9.17** (a) Figure 9.35 shows a cross-section of a 'light pipe' made of a glass fibre of refractive index 1.68. The outer covering of the pipe is made of a material of refractive index 1.44. What is the range of the angles of the incident rays with the axis of the pipe for which total reflections inside the pipe take place, as shown in the figure.
(b) What is the answer if there is no outer covering of the pipe?

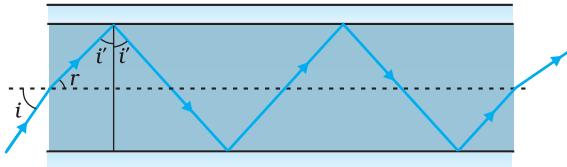


FIGURE 9.35

- 9.18** Answer the following questions:

- (a) You have learnt that plane and convex mirrors produce virtual images of objects. Can they produce real images under some circumstances? Explain.
(b) A virtual image, we always say, cannot be caught on a screen. Yet when we 'see' a virtual image, we are obviously bringing it on to the 'screen' (i.e., the retina) of our eye. Is there a contradiction?
(c) A diver under water, looks obliquely at a fisherman standing on the bank of a lake. Would the fisherman look taller or shorter to the diver than what he actually is?
(d) Does the apparent depth of a tank of water change if viewed obliquely? If so, does the apparent depth increase or decrease?
(e) The refractive index of diamond is much greater than that of ordinary glass. Is this fact of some use to a diamond cutter?

- 9.19** The image of a small electric bulb fixed on the wall of a room is to be obtained on the opposite wall 3 m away by means of a large convex lens. What is the maximum possible focal length of the lens required for the purpose?

- 9.20** A screen is placed 90 cm from an object. The image of the object on the screen is formed by a convex lens at two different locations separated by 20 cm. Determine the focal length of the lens.

- 9.21** (a) Determine the 'effective focal length' of the combination of the two lenses in Exercise 9.10, if they are placed 8.0 cm apart with their principal axes coincident. Does the answer depend on which side of the combination a beam of parallel light is incident? Is the notion of effective focal length of this system useful at all?
(b) An object 1.5 cm in size is placed on the side of the convex lens in the arrangement (a) above. The distance between the object

Physics

and the convex lens is 40 cm. Determine the magnification produced by the two-lens system, and the size of the image.

- 9.22** At what angle should a ray of light be incident on the face of a prism of refracting angle 60° so that it just suffers total internal reflection at the other face? The refractive index of the material of the prism is 1.524.
- 9.23** You are given prisms made of crown glass and flint glass with a wide variety of angles. Suggest a combination of prisms which will
(a) deviate a pencil of white light without much dispersion,
(b) disperse (and displace) a pencil of white light without much deviation.
- 9.24** For a normal eye, the far point is at infinity and the near point of distinct vision is about 25 cm in front of the eye. The cornea of the eye provides a converging power of about 40 dioptres, and the least converging power of the eye-lens behind the cornea is about 20 dioptres. From this rough data estimate the range of accommodation (i.e., the range of converging power of the eye-lens) of a normal eye.
- 9.25** Does short-sightedness (myopia) or long-sightedness (hypermetropia) imply necessarily that the eye has partially lost its ability of accommodation? If not, what might cause these defects of vision?
- 9.26** A myopic person has been using spectacles of power -1.0 dioptre for distant vision. During old age he also needs to use separate reading glass of power + 2.0 dioptres. Explain what may have happened.
- 9.27** A person looking at a person wearing a shirt with a pattern comprising vertical and horizontal lines is able to see the vertical lines more distinctly than the horizontal ones. What is this defect due to? How is such a defect of vision corrected?
- 9.28** A man with normal near point (25 cm) reads a book with small print using a magnifying glass: a thin convex lens of focal length 5 cm.
(a) What is the closest and the farthest distance at which he should keep the lens from the page so that he can read the book when viewing through the magnifying glass?
(b) What is the maximum and the minimum angular magnification (magnifying power) possible using the above simple microscope?
- 9.29** A card sheet divided into squares each of size 1 mm^2 is being viewed at a distance of 9 cm through a magnifying glass (a converging lens of focal length 9 cm) held close to the eye.
(a) What is the magnification produced by the lens? How much is the area of each square in the virtual image?
(b) What is the angular magnification (magnifying power) of the lens?
(c) Is the magnification in (a) equal to the magnifying power in (b)? Explain.
- 9.30** (a) At what distance should the lens be held from the figure in Exercise 9.29 in order to view the squares distinctly with the maximum possible magnifying power?
(b) What is the magnification in this case?
(c) Is the magnification equal to the magnifying power in this case? Explain.
- 9.31** What should be the distance between the object in Exercise 9.30 and the magnifying glass if the virtual image of each square in the figure is to have an area of 6.25 mm^2 . Would you be able to see the squares distinctly with your eyes very close to the magnifier?

Ray Optics and Optical Instruments

[Note: Exercises 9.29 to 9.31 will help you clearly understand the difference between magnification in absolute size and the angular magnification (or magnifying power) of an instrument.]

9.32 Answer the following questions:

- The angle subtended at the eye by an object is equal to the angle subtended at the eye by the virtual image produced by a magnifying glass. In what sense then does a magnifying glass provide angular magnification?
- In viewing through a magnifying glass, one usually positions one's eyes very close to the lens. Does angular magnification change if the eye is moved back?
- Magnifying power of a simple microscope is inversely proportional to the focal length of the lens. What then stops us from using a convex lens of smaller and smaller focal length and achieving greater and greater magnifying power?
- Why must both the objective and the eyepiece of a compound microscope have short focal lengths?
- When viewing through a compound microscope, our eyes should be positioned not on the eyepiece but a short distance away from it for best viewing. Why? How much should be that short distance between the eye and eyepiece?

9.33 An angular magnification (magnifying power) of 30X is desired using an objective of focal length 1.25 cm and an eyepiece of focal length 5 cm. How will you set up the compound microscope?

9.34 A small telescope has an objective lens of focal length 140 cm and an eyepiece of focal length 5.0 cm. What is the magnifying power of the telescope for viewing distant objects when

- the telescope is in normal adjustment (i.e., when the final image is at infinity)?
- the final image is formed at the least distance of distinct vision (25 cm)?

9.35 (a) For the telescope described in Exercise 9.34 (a), what is the separation between the objective lens and the eyepiece?
 (b) If this telescope is used to view a 100 m tall tower 3 km away, what is the height of the image of the tower formed by the objective lens?
 (c) What is the height of the final image of the tower if it is formed at 25 cm?

9.36 A Cassegrain telescope uses two mirrors as shown in Fig. 9.33. Such a telescope is built with the mirrors 20 mm apart. If the radius of curvature of the large mirror is 220 mm and the small mirror is 140 mm, where will the final image of an object at infinity be?

9.37 Light incident normally on a plane mirror attached to a galvanometer coil retraces backwards as shown in Fig. 9.36. A current in the coil produces a deflection of 3.5° of the mirror. What is the displacement of the reflected spot of light on a screen placed 1.5 m away?

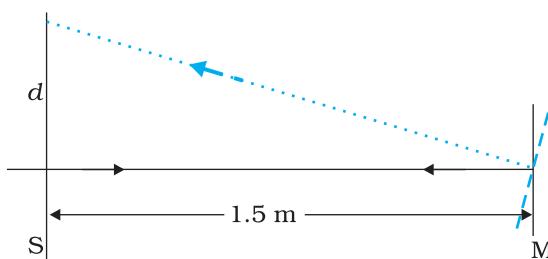


FIGURE 9.36

Physics

- 9.38** Figure 9.37 shows an equiconvex lens (of refractive index 1.50) in contact with a liquid layer on top of a plane mirror. A small needle with its tip on the principal axis is moved along the axis until its inverted image is found at the position of the needle. The distance of the needle from the lens is measured to be 45.0cm. The liquid is removed and the experiment is repeated. The new distance is measured to be 30.0cm. What is the refractive index of the liquid?

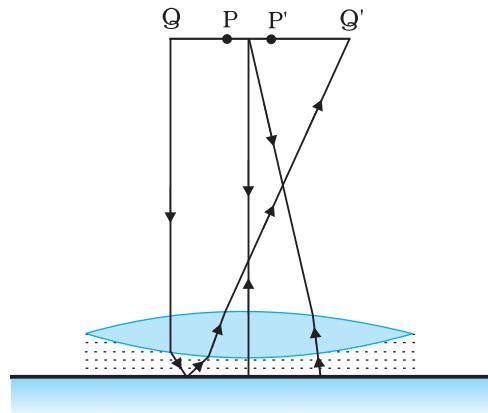


FIGURE 9.37

Chapter Ten

WAVE OPTICS

10.1 INTRODUCTION

In 1637 Descartes gave the corpuscular model of light and derived Snell's law. It explained the laws of reflection and refraction of light at an interface. The corpuscular model predicted that if the ray of light (on refraction) bends towards the normal then the speed of light would be greater in the second medium. This corpuscular model of light was further developed by Isaac Newton in his famous book entitled *OPTICKS* and because of the tremendous popularity of this book, the corpuscular model is very often attributed to Newton.

In 1678, the Dutch physicist Christiaan Huygens put forward the wave theory of light – it is this wave model of light that we will discuss in this chapter. As we will see, the wave model could satisfactorily explain the phenomena of reflection and refraction; however, it predicted that on refraction if the wave bends towards the normal then the speed of light would be less in the second medium. This is in contradiction to the prediction made by using the corpuscular model of light. It was much later confirmed by experiments where it was shown that the speed of light in water is less than the speed in air confirming the prediction of the wave model; Foucault carried out this experiment in 1850.

The wave theory was not readily accepted primarily because of Newton's authority and also because light could travel through vacuum

■ Physics

and it was felt that a wave would always require a medium to propagate from one point to the other. However, when Thomas Young performed his famous interference experiment in 1801, it was firmly established that light is indeed a wave phenomenon. The wavelength of visible light was measured and found to be extremely small; for example, the wavelength of yellow light is about $0.5\text{ }\mu\text{m}$. Because of the smallness of the wavelength of visible light (in comparison to the dimensions of typical mirrors and lenses), light can be assumed to approximately travel in straight lines. This is the field of geometrical optics, which we had discussed in the previous chapter. Indeed, the branch of optics in which one completely neglects the finiteness of the wavelength is called geometrical optics and a ray is defined as the path of energy propagation in the limit of wavelength tending to zero.

After the interference experiment of Young in 1801, for the next 40 years or so, many experiments were carried out involving the interference and diffraction of lightwaves; these experiments could only be satisfactorily explained by assuming a wave model of light. Thus, around the middle of the nineteenth century, the wave theory seemed to be very well established. The only major difficulty was that since it was thought that a wave required a medium for its propagation, how could light waves propagate through vacuum. This was explained when Maxwell put forward his famous electromagnetic theory of light. Maxwell had developed a set of equations describing the laws of electricity and magnetism and using these equations he derived what is known as the wave equation from which he *predicted* the existence of electromagnetic waves*. From the wave equation, Maxwell could calculate the speed of electromagnetic waves in free space and he found that the theoretical value was very close to the measured value of speed of light. From this, he propounded that *light must be an electromagnetic wave*. Thus, according to Maxwell, light waves are associated with changing electric and magnetic fields; changing electric field produces a time and space varying magnetic field and a changing magnetic field produces a time and space varying electric field. The changing electric and magnetic fields result in the propagation of electromagnetic waves (or light waves) even in vacuum.

In this chapter we will first discuss the original formulation of the *Huygens principle* and derive the laws of reflection and refraction. In Sections 10.4 and 10.5, we will discuss the phenomenon of interference which is based on the principle of superposition. In Section 10.6 we will discuss the phenomenon of diffraction which is based on Huygens-Fresnel principle. Finally in Section 10.7 we will discuss the phenomenon of polarisation which is based on the fact that the light waves are *transverse electromagnetic waves*.

* Maxwell had predicted the existence of electromagnetic waves around 1855; it was much later (around 1890) that Heinrich Hertz produced radiowaves in the laboratory. J.C. Bose and G. Marconi made practical applications of the *Hertzian waves*.

DOES LIGHT TRAVEL IN A STRAIGHT LINE?

Light travels in a straight line in Class VI; it does not do so in Class XII and beyond! Surprised, aren't you?

In school, you are shown an experiment in which you take three cardboards with pinholes in them, place a candle on one side and look from the other side. If the flame of the candle and the three pinholes are in a straight line, you can see the candle. Even if one of them is displaced a little, you cannot see the candle. *This proves, so your teacher says, that light travels in a straight line.*

In the present book, there are two consecutive chapters, one on ray optics and the other on wave optics. Ray optics is based on rectilinear propagation of light, and deals with mirrors, lenses, reflection, refraction, etc. Then you come to the chapter on wave optics, and you are told that light travels as a wave, that it can bend around objects, it can diffract and interfere, etc.

In optical region, light has a wavelength of about half a micrometre. If it encounters an obstacle of about this size, it can bend around it and can be seen on the other side. Thus a micrometre size obstacle will not be able to stop a light ray. If the obstacle is much larger, however, light will not be able to bend to that extent, and will not be seen on the other side.

This is a property of a wave in general, and can be seen in sound waves too. The sound wave of our speech has a wavelength of about 50 cm to 1 m. If it meets an obstacle of the size of a few metres, it bends around it and reaches points behind the obstacle. But when it comes across a larger obstacle of a few hundred metres, such as a hillock, most of it is reflected and is heard as an echo.

Then what about the primary school experiment? What happens there is that when we move any cardboard, the displacement is of the order of a few millimetres, which is much larger than the wavelength of light. Hence the candle cannot be seen. If we are able to move one of the cardboards by a micrometer or less, light will be able to diffract, and the candle will still be seen.

One could add to the first sentence in this box: *It learns how to bend as it grows up!*

10.2 HUYGENS PRINCIPLE

We would first define a wavefront: when we drop a small stone on a calm pool of water, waves spread out from the point of impact. Every point on the surface starts oscillating with time. At any instant, a photograph of the surface would show circular rings on which the disturbance is maximum. Clearly, all points on such a circle are oscillating in phase because they are at the same distance from the source. Such a locus of points, which oscillate in phase is called a *wavefront*; thus *a wavefront is defined as a surface of constant phase*. The speed with which the wavefront moves outwards from the source is called the speed of the wave. The energy of the wave travels in a direction perpendicular to the wavefront.

If we have a point source emitting waves uniformly in all directions, then the locus of points which have the same amplitude and vibrate in the same phase are spheres and we have what is known as a *spherical wave* as shown in Fig. 10.1(a). At a large distance from the source, a

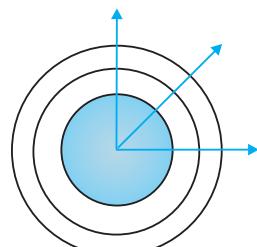


FIGURE 10.1 (a) A diverging spherical wave emanating from a point source. The wavefronts are spherical.

Physics

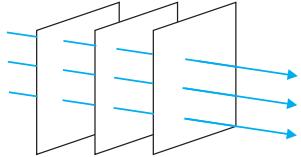


FIGURE 10.1 (b) At a large distance from the source, a small portion of the spherical wave can be approximated by a plane wave.

small portion of the sphere can be considered as a plane and we have what is known as a *plane wave* [Fig. 10.1(b)].

Now, if we know the shape of the wavefront at $t = 0$, then Huygens principle allows us to determine the shape of the wavefront at a later time τ . Thus, Huygens principle is essentially a geometrical construction, which given the shape of the wafefront at any time allows us to determine the shape of the wavefront at a later time. Let us consider a diverging wave and let $F_1 F_2$ represent a portion of the spherical wavefront at $t = 0$ (Fig. 10.2). Now, according to Huygens principle, *each point of the wavefront is the source of a secondary disturbance and the wavelets emanating from these points spread out in all directions with the speed of the wave. These wavelets emanating from the wavefront are usually referred to as secondary wavelets and if we draw a common tangent to all these spheres, we obtain the new position of the wavefront at a later time*.

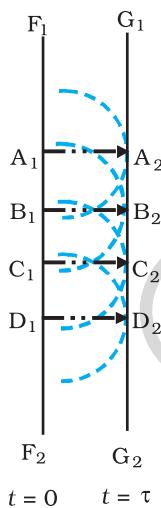


FIGURE 10.3
Huygens geometrical construction for a plane wave propagating to the right. $F_1 F_2$ is the plane wavefront at $t = 0$ and $G_1 G_2$ is the wavefront at a later time τ . The lines $A_1 A_2$, $B_1 B_2$... etc, are normal to both $F_1 F_2$ and $G_1 G_2$ and represent rays.

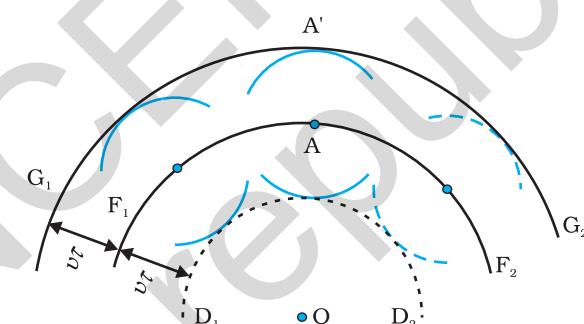


FIGURE 10.2 $F_1 F_2$ represents the spherical wavefront (with O as centre) at $t = 0$. The envelope of the secondary wavelets emanating from $F_1 F_2$ produces the forward moving wavefront $G_1 G_2$. The backwave $D_1 D_2$ does not exist.

Thus, if we wish to determine the shape of the wavefront at $t = \tau$, we draw spheres of radius $v\tau$ from each point on the spherical wavefront where v represents the speed of the waves in the medium. If we now draw a common tangent to all these spheres, we obtain the new position of the wavefront at $t = \tau$. The new wavefront shown as $G_1 G_2$ in Fig. 10.2 is again spherical with point O as the centre.

The above model has one shortcoming: we also have a backwave which is shown as $D_1 D_2$ in Fig. 10.2. Huygens argued that the amplitude of the secondary wavelets is maximum in the forward direction and zero in the backward direction; by making this adhoc assumption, Huygens could explain the absence of the backwave. However, this adhoc assumption is not satisfactory and the absence of the backwave is really justified from more rigorous wave theory.

In a similar manner, we can use Huygens principle to determine the shape of the wavefront for a plane wave propagating through a medium (Fig. 10.3).

10.3 REFRACTION AND REFLECTION OF PLANE WAVES USING HUYGENS PRINCIPLE

10.3.1 Refraction of a plane wave

We will now use Huygens principle to derive the laws of refraction. Let PP' represent the surface separating medium 1 and medium 2, as shown in Fig. 10.4. Let v_1 and v_2 represent the speed of light in medium 1 and medium 2, respectively. We assume a plane wavefront AB propagating in the direction $A'A$ incident on the interface at an angle i as shown in the figure. Let τ be the time taken by the wavefront to travel the distance BC . Thus,

$$BC = v_1 \tau$$

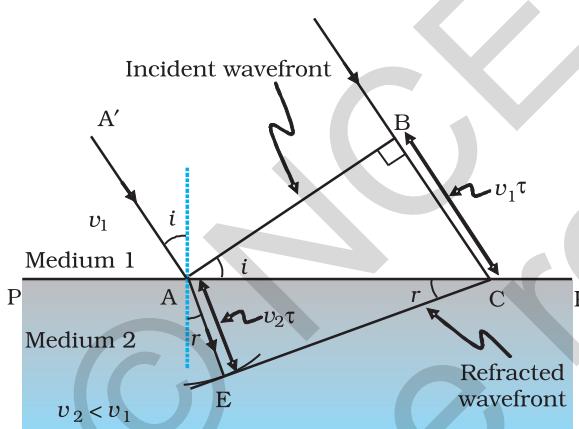


FIGURE 10.4 A plane wave AB is incident at an angle i on the surface PP' separating medium 1 and medium 2. The plane wave undergoes refraction and CE represents the refracted wavefront. The figure corresponds to $v_2 < v_1$ so that the refracted waves bends towards the normal.

In order to determine the shape of the refracted wavefront, we draw a sphere of radius $v_2\tau$ from the point A in the second medium (the speed of the wave in the second medium is v_2). Let CE represent a tangent plane drawn from the point C on to the sphere. Then, $AE = v_2\tau$ and CE would represent the refracted wavefront. If we now consider the triangles ABC and AEC , we readily obtain

$$\sin i = \frac{BC}{AC} = \frac{v_1\tau}{AC} \quad (10.1)$$

and

$$\sin r = \frac{AE}{AC} = \frac{v_2\tau}{AC} \quad (10.2)$$

where i and r are the angles of incidence and refraction, respectively.



Christiaan Huygens (1629 – 1695) Dutch physicist, astronomer, mathematician and the founder of the wave theory of light. His book, *Treatise on light*, makes fascinating reading even today. He brilliantly explained the double refraction shown by the mineral calcite in this work in addition to reflection and refraction. He was the first to analyse circular and simple harmonic motion and designed and built improved clocks and telescopes. He discovered the true geometry of Saturn's rings.

CHRISTIAAN HUYGENS (1629 – 1695)

Physics

Thus we obtain

$$\frac{\sin i}{\sin r} = \frac{v_1}{v_2} \quad (10.3)$$

From the above equation, we get the important result that if $r < i$ (i.e., if the ray bends toward the normal), the speed of the light wave in the second medium (v_2) will be less than the speed of the light wave in the first medium (v_1). This prediction is opposite to the prediction from the corpuscular model of light and as later experiments showed, the prediction of the wave theory is correct. Now, if c represents the speed of light in vacuum, then,

$$n_1 = \frac{c}{v_1} \quad (10.4)$$

and

$$n_2 = \frac{c}{v_2} \quad (10.5)$$

are known as the refractive indices of medium 1 and medium 2, respectively. In terms of the refractive indices, Eq. (10.3) can be written as

$$n_1 \sin i = n_2 \sin r \quad (10.6)$$

This is the *Snell's law of refraction*. Further, if λ_1 and λ_2 denote the wavelengths of light in medium 1 and medium 2, respectively and if the distance BC is equal to λ_1 , then the distance AE will be equal to λ_2 (because if the crest from B has reached C in time τ , then the crest from A should have also reached E in time τ); thus,

$$\frac{\lambda_1}{\lambda_2} = \frac{BC}{AE} = \frac{v_1}{v_2}$$

or

$$\frac{v_1}{\lambda_1} = \frac{v_2}{\lambda_2} \quad (10.7)$$

The above equation implies that when a wave gets refracted into a denser medium ($v_1 > v_2$) the wavelength and the speed of propagation decrease but the frequency $v (= v/\lambda)$ remains the same.

10.3.2 Refraction at a rarer medium

We now consider refraction of a plane wave at a rarer medium, i.e., $v_2 > v_1$. Proceeding in an exactly similar manner we can construct a refracted wavefront as shown in Fig. 10.5. The angle of refraction will now be greater than angle of incidence; however, we will still have $n_1 \sin i = n_2 \sin r$. We define an angle i_c by the following equation

$$\sin i_c = \frac{n_2}{n_1} \quad (10.8)$$

Thus, if $i = i_c$ then $\sin r = 1$ and $r = 90^\circ$. Obviously, for $i > i_c$ there can not be any refracted wave. The angle i_c is known as the *critical angle* and for all angles of incidence greater than the critical angle, we will not have

any refracted wave and the wave will undergo what is known as *total internal reflection*. The phenomenon of total internal reflection and its applications was discussed in Section 9.4.

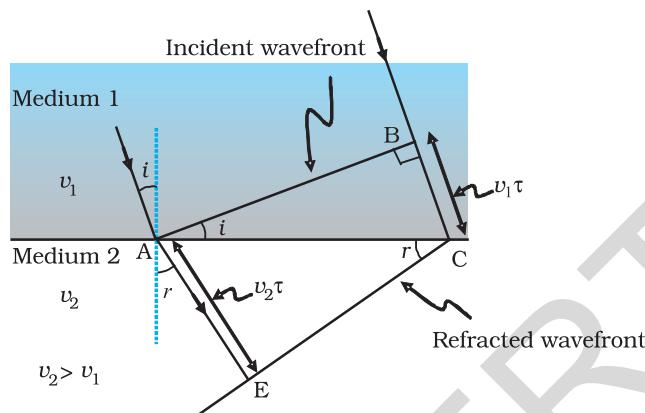


FIGURE 10.5 Refraction of a plane wave incident on a rarer medium for which $v_2 > v_1$. The plane wave bends away from the normal.

10.3.3 Reflection of a plane wave by a plane surface

We next consider a plane wave AB incident at an angle i on a reflecting surface MN. If v represents the speed of the wave in the medium and if τ represents the time taken by the wavefront to advance from the point B to C then the distance

$$BC = v\tau$$

In order to construct the reflected wavefront we draw a sphere of radius $v\tau$ from the point A as shown in Fig. 10.6. Let CE represent the tangent plane drawn from the point C to this sphere. Obviously

$$AE = BC = v\tau$$

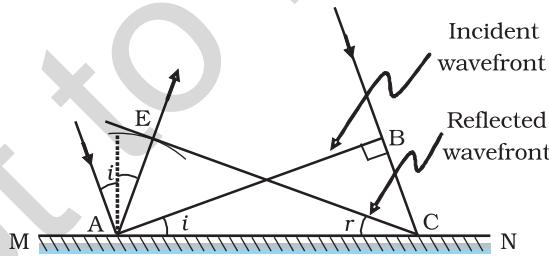


FIGURE 10.6 Reflection of a plane wave AB by the reflecting surface MN. AB and CE represent incident and reflected wavefronts.

If we now consider the triangles EAC and BAC we will find that they are congruent and therefore, the angles i and r (as shown in Fig. 10.6) would be equal. This is the *law of reflection*.

Once we have the laws of reflection and refraction, the behaviour of prisms, lenses, and mirrors can be understood. These phenomena were

discussed in detail in Chapter 9 on the basis of rectilinear propagation of light. Here we just describe the behaviour of the wavefronts as they undergo reflection or refraction. In Fig. 10.7(a) we consider a plane wave passing through a thin prism. Clearly, since the speed of light waves is less in glass, the lower portion of the incoming wavefront (which travels through the greatest thickness of glass) will get delayed resulting in a tilt in the emerging wavefront as shown in the figure. In Fig. 10.7(b) we consider a plane wave incident on a thin convex lens; the central part of the incident plane wave traverses the thinnest portion of the lens and is delayed the most. The emerging wavefront has a depression at the centre and therefore the wavefront becomes spherical and converges to the point F which is known as the *focus*. In Fig. 10.7(c) a plane wave is incident on a concave mirror and on reflection we have a spherical wave converging to the focal point F. In a similar manner, we can understand refraction and reflection by concave lenses and convex mirrors.

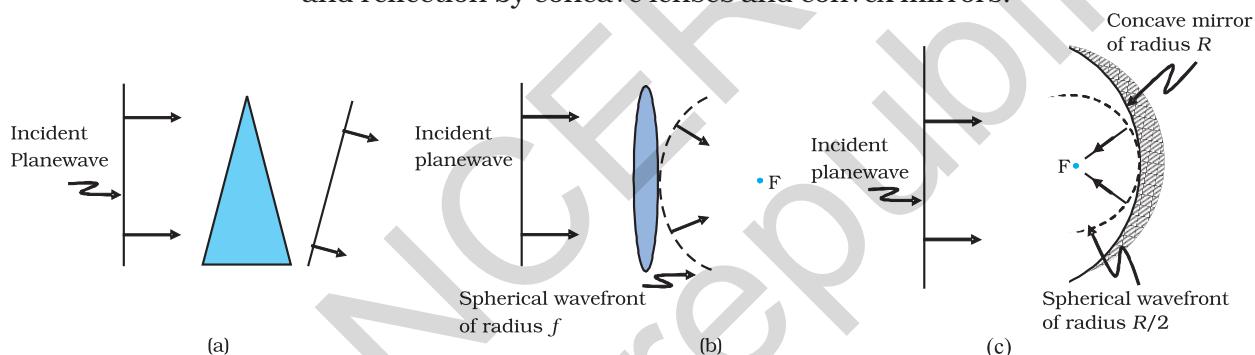


FIGURE 10.7 Refraction of a plane wave by (a) a thin prism, (b) a convex lens. (c) Reflection of a plane wave by a concave mirror.

From the above discussion it follows that the total time taken from a point on the object to the corresponding point on the image is the same measured along any ray. For example, when a convex lens focusses light to form a real image, although the ray going through the centre traverses a shorter path, but because of the slower speed in glass, the time taken is the same as for rays travelling near the edge of the lens.

10.3.4 The doppler effect

We should mention here that one should be careful in constructing the wavefronts if the source (or the observer) is moving. For example, if there is no medium and the source moves away from the observer, then later wavefronts have to travel a greater distance to reach the observer and hence take a longer time. The time taken between the arrival of two successive wavefronts is hence longer at the observer than it is at the source. Thus, when the source moves away from the observer the frequency as measured by the source will be smaller. This is known as the *Doppler effect*. Astronomers call the increase in wavelength due to doppler effect as *red shift* since a wavelength in the middle of the visible region of the spectrum moves towards the red end of the spectrum. When waves are received from a source moving towards the observer, there is an apparent decrease in wavelength, this is referred to as *blue shift*.

Wave Optics

You have already encountered Doppler effect for sound waves in Chapter 15 of Class XI textbook. For velocities small compared to the speed of light, we can use the same formulae which we use for sound waves. The fractional change in frequency $\Delta\nu/\nu$ is given by $-\nu_{\text{radial}}/c$, where ν_{radial} is the component of the source velocity along the line joining the observer to the source relative to the observer; ν_{radial} is considered positive when the source moves away from the observer. Thus, the Doppler shift can be expressed as:

$$\frac{\Delta\nu}{\nu} = -\frac{\nu_{\text{radial}}}{c} \quad (10.9)$$

The formula given above is valid only when the speed of the source is small compared to that of light. A more accurate formula for the Doppler effect which is valid even when the speeds are close to that of light, requires the use of Einstein's special theory of relativity. The Doppler effect for light is very important in astronomy. It is the basis for the measurements of the radial velocities of distant galaxies.

Example 10.1 What speed should a galaxy move with respect to us so that the sodium line at 589.0 nm is observed at 589.6 nm?

Solution Since $\nu\lambda = c$, $\frac{\Delta\nu}{\nu} = -\frac{\Delta\lambda}{\lambda}$ (for small changes in ν and λ). For

$$\Delta\lambda = 589.6 - 589.0 = +0.6 \text{ nm}$$

we get [using Eq. (10.9)]

$$\frac{\Delta\nu}{\nu} = -\frac{\Delta\lambda}{\lambda} = -\frac{\nu_{\text{radial}}}{c}$$

$$\text{or, } \nu_{\text{radial}} \approx +c \frac{0.6}{589.0} = +3.06 \times 10^5 \text{ ms}^{-1}$$
$$= 306 \text{ km/s}$$

Therefore, the galaxy is moving away from us.

EXAMPLE 10.1

Example 10.2

- When monochromatic light is incident on a surface separating two media, the reflected and refracted light both have the same frequency as the incident frequency. Explain why?
- When light travels from a rarer to a denser medium, the speed decreases. Does the reduction in speed imply a reduction in the energy carried by the light wave?
- In the wave picture of light, intensity of light is determined by the square of the amplitude of the wave. What determines the intensity of light in the photon picture of light?

Solution

- Reflection and refraction arise through interaction of incident light with the atomic constituents of matter. Atoms may be viewed as

EXAMPLE 10.2

EXAMPLE 10.2

oscillators, which take up the frequency of the external agency (light) causing forced oscillations. The frequency of light emitted by a charged oscillator equals its frequency of oscillation. Thus, the frequency of scattered light equals the frequency of incident light.

- No. Energy carried by a wave depends on the amplitude of the wave, not on the speed of wave propagation.
- For a given frequency, intensity of light in the photon picture is determined by the number of photons crossing an unit area per unit time.

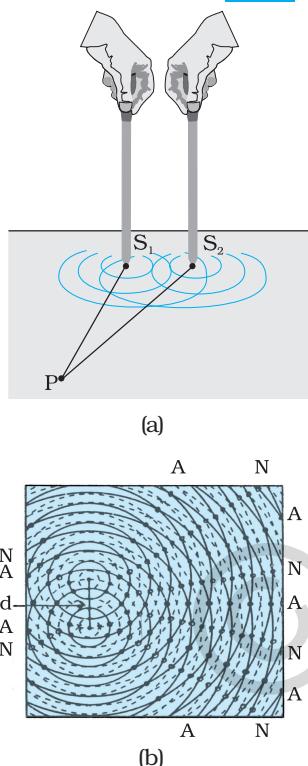


FIGURE 10.8 (a) Two needles oscillating in phase in water represent two coherent sources.

(b) The pattern of displacement of water molecules at an instant on the surface of water showing nodal N (no displacement) and antinodal A (maximum displacement) lines.

10.4 COHERENT AND INCOHERENT ADDITION OF WAVES

In this section we will discuss the interference pattern produced by the superposition of two waves. You may recall that we had discussed the superposition principle in Chapter 15 of your Class XI textbook. Indeed the entire field of interference is based on the *superposition principle* according to which *at a particular point in the medium, the resultant displacement produced by a number of waves is the vector sum of the displacements produced by each of the waves*.

Consider two needles S_1 and S_2 moving periodically up and down in an identical fashion in a trough of water [Fig. 10.8(a)]. They produce two water waves, and at a particular point, the phase difference between the displacements produced by each of the waves does not change with time; when this happens the two sources are said to be *coherent*. Figure 10.8(b) shows the position of crests (solid circles) and troughs (dashed circles) at a given instant of time. Consider a point P for which

$$S_1 P = S_2 P$$

Since the distances $S_1 P$ and $S_2 P$ are equal, waves from S_1 and S_2 will take the same time to travel to the point P and waves that emanate from S_1 and S_2 in phase will also arrive, at the point P , in phase.

Thus, if the displacement produced by the source S_1 at the point P is given by

$$y_1 = a \cos \omega t$$

then, the displacement produced by the source S_2 (at the point P) will also be given by

$$y_2 = a \cos \omega t$$

Thus, the resultant of displacement at P would be given by

$$y = y_1 + y_2 = 2 a \cos \omega t$$

Since the intensity is proportional to the square of the amplitude, the resultant intensity will be given by

$$I = 4 I_0$$

where I_0 represents the intensity produced by each one of the individual sources; I_0 is proportional to a^2 . In fact at any point on the perpendicular bisector of $S_1 S_2$, the intensity will be $4I_0$. The two sources are said to

Wave Optics

interfere constructively and we have what is referred to as *constructive interference*. We next consider a point Q [Fig. 10.9(a)] for which

$$S_2 Q - S_1 Q = 2\lambda$$

The waves emanating from S_1 will arrive exactly two cycles earlier than the waves from S_2 and will again be in phase [Fig. 10.9(a)]. Thus, if the displacement produced by S_1 is given by

$$y_1 = a \cos \omega t$$

then the displacement produced by S_2 will be given by

$$y_2 = a \cos (\omega t - 4\pi) = a \cos \omega t$$

where we have used the fact that a path difference of 2λ corresponds to a phase difference of 4π . The two displacements are once again in phase and the intensity will again be $4I_0$ giving rise to constructive interference. In the above analysis we have assumed that the distances $S_1 Q$ and $S_2 Q$ are much greater than d (which represents the distance between S_1 and S_2) so that although $S_1 Q$ and $S_2 Q$ are not equal, the amplitudes of the displacement produced by each wave are very nearly the same.

We next consider a point R [Fig. 10.9(b)] for which

$$S_2 R - S_1 R = -2.5\lambda$$

The waves emanating from S_1 will arrive exactly two and a half cycles later than the waves from S_2 [Fig. 10.10(b)]. Thus if the displacement produced by S_1 is given by

$$y_1 = a \cos \omega t$$

then the displacement produced by S_2 will be given by

$$y_2 = a \cos (\omega t + 5\pi) = -a \cos \omega t$$

where we have used the fact that a path difference of 2.5λ corresponds to a phase difference of 5π . The two displacements are now out of phase and the two displacements will cancel out to give zero intensity. This is referred to as *destructive interference*.

To summarise: If we have two coherent sources S_1 and S_2 vibrating in phase, then for an arbitrary point P whenever the path difference,

$$S_1 P \sim S_2 P = n\lambda \quad (n = 0, 1, 2, 3, \dots) \quad (10.10)$$

we will have constructive interference and the resultant intensity will be $4I_0$; the sign \sim between $S_1 P$ and $S_2 P$ represents the difference between $S_1 P$ and $S_2 P$. On the other hand, if the point P is such that the path difference,

$$S_1 P \sim S_2 P = (n + \frac{1}{2})\lambda \quad (n = 0, 1, 2, 3, \dots) \quad (10.11)$$

we will have *destructive interference* and the resultant intensity will be zero. Now, for any other arbitrary point G (Fig. 10.10) let the phase difference between the two displacements be ϕ . Thus, if the displacement produced by S_1 is given by

$$y_1 = a \cos \omega t$$

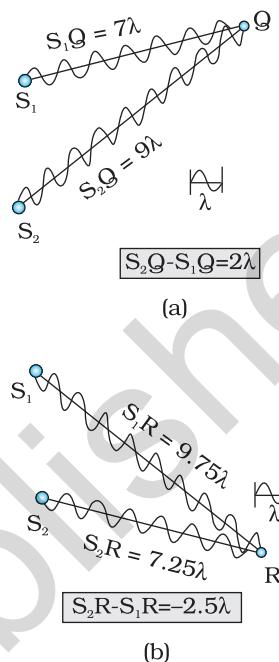


FIGURE 10.9

- (a) Constructive interference at a point Q for which the path difference is 2λ .
 (b) Destructive interference at a point R for which the path difference is 2.5λ .

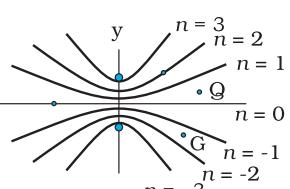


FIGURE 10.10 Locus of points for which $S_1 P - S_2 P$ is equal to zero, $\pm\lambda, \pm 2\lambda, \pm 3\lambda$.

Physics

then, the displacement produced by S_2 would be

$$y_2 = a \cos(\omega t + \phi)$$

and the resultant displacement will be given by

$$\begin{aligned}y &= y_1 + y_2 \\&= a [\cos \omega t + \cos (\omega t + \phi)] \\&= 2 a \cos(\phi/2) \cos(\omega t + \phi/2)\end{aligned}$$

The amplitude of the resultant displacement is $2a \cos(\phi/2)$ and therefore the intensity at that point will be

$$I = 4 I_0 \cos^2(\phi/2) \quad (10.12)$$

If $\phi = 0, \pm 2\pi, \pm 4\pi, \dots$ which corresponds to the condition given by Eq. (10.10) we will have constructive interference leading to maximum intensity. On the other hand, if $\phi = \pm\pi, \pm 3\pi, \pm 5\pi, \dots$ [which corresponds to the condition given by Eq. (10.11)] we will have destructive interference leading to zero intensity.

Now if the two sources are coherent (i.e., if the two needles are going up and down regularly) then the phase difference ϕ at any point will not change with time and we will have a stable interference pattern; i.e., the positions of maxima and minima will not change with time. However, if the two needles do not maintain a constant phase difference, then the interference pattern will also change with time and, if the phase difference changes very rapidly with time, the positions of maxima and minima will also vary rapidly with time and we will see a “time-averaged” intensity distribution. When this happens, we will observe an average intensity that will be given by

$$\langle I \rangle = 4 I_0 \langle \cos^2(\phi/2) \rangle \quad (10.13)$$

where angular brackets represent time averaging. Indeed it is shown in Section 7.2 that if $\phi(t)$ varies randomly with time, the time-averaged quantity $\langle \cos^2(\phi/2) \rangle$ will be $1/2$. This is also intuitively obvious because the function $\cos^2(\phi/2)$ will randomly vary between 0 and 1 and the average value will be $1/2$. The resultant intensity will be given by

$$I = 2 I_0 \quad (10.14)$$

at all points.

When the phase difference between the two vibrating sources changes rapidly with time, we say that the two sources are incoherent and when this happens the intensities just add up. This is indeed what happens when two separate light sources illuminate a wall.

10.5 INTERFERENCE OF LIGHT WAVES AND YOUNG'S EXPERIMENT

We will now discuss interference using light waves. If we use two sodium lamps illuminating two pinholes (Fig. 10.11) we will not observe any interference fringes. This is because of the fact that the light wave emitted from an ordinary source (like a sodium lamp) undergoes abrupt phase

Wave Optics

changes in times of the order of 10^{-10} seconds. Thus the light waves coming out from two independent sources of light will not have any fixed phase relationship and would be incoherent, when this happens, as discussed in the previous section, the intensities on the screen will add up.

The British physicist Thomas Young used an ingenious technique to “lock” the phases of the waves emanating from S_1 and S_2 . He made two pinholes S_1 and S_2 (very close to each other) on an opaque screen [Fig. 10.12(a)]. These were illuminated by another pinhole that was in turn, lit by a bright source. Light waves spread out from S and fall on both S_1 and S_2 . S_1 and S_2 then behave like two coherent sources because light waves coming out from S_1 and S_2 are derived from the same original source and any abrupt phase change in S will manifest in exactly similar phase changes in the light coming out from S_1 and S_2 . Thus, the two sources S_1 and S_2 will be *locked* in phase; i.e., they will be coherent like the two vibrating needle in our water wave example [Fig. 10.8(a)].

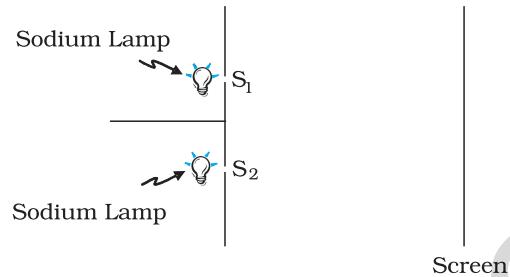


FIGURE 10.11 If two sodium lamps illuminate two pinholes S_1 and S_2 , the intensities will add up and no interference fringes will be observed on the screen.

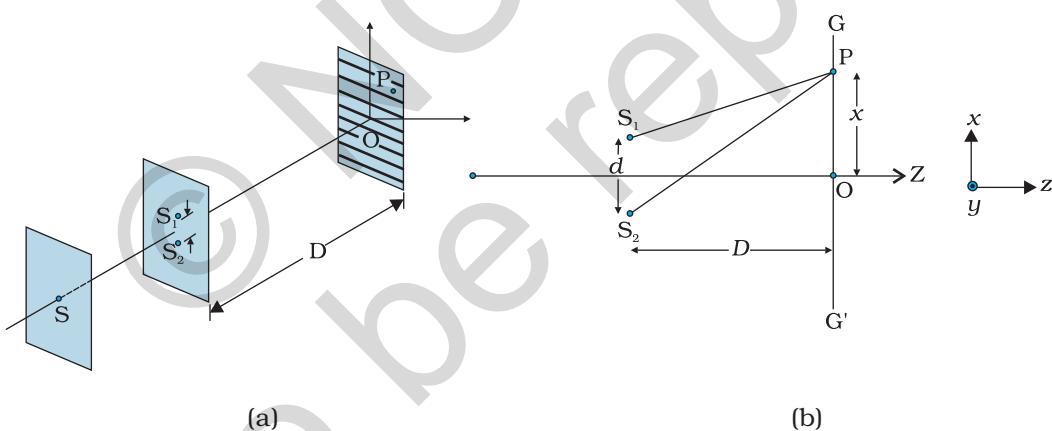


FIGURE 10.12 Young's arrangement to produce interference pattern.

Thus spherical waves emanating from S_1 and S_2 will produce interference fringes on the screen GG' , as shown in Fig. 10.12(b). The positions of maximum and minimum intensities can be calculated by using the analysis given in Section 10.4 where we had shown that for an arbitrary point P on the line GG' [Fig. 10.12(b)] to correspond to a maximum, we must have

$$S_2P - S_1P = n\lambda; \quad n = 0, 1, 2 \dots \quad (10.15)$$

Now,

$$(S_2P)^2 - (S_1P)^2 = D^2 + x + \frac{d}{2}^2 - D^2 + x - \frac{d}{2}^2 = 2xd$$

Physics

THOMAS YOUNG (1773 – 1829)



Thomas Young (1773 – 1829) English physicist, physician and Egyptologist. Young worked on a wide variety of scientific problems, ranging from the structure of the eye and the mechanism of vision to the decipherment of the Rosetta stone. He revived the wave theory of light and recognised that interference phenomena provide proof of the wave properties of light.

where $S_1S_2 = d$ and $OP = x$. Thus

$$S_2P - S_1P = \frac{2xd}{S_2P + S_1P} \quad (10.16)$$

If $x, d \ll D$ then negligible error will be introduced if $S_2P + S_1P$ (in the denominator) is replaced by $2D$. For example, for $d = 0.1$ cm, $D = 100$ cm, $OP = 1$ cm (which correspond to typical values for an interference experiment using light waves), we have

$$S_2P + S_1P = [(100)^2 + (1.05)^2]^{1/2} + [(100)^2 + (0.95)^2]^{1/2} \approx 200.01 \text{ cm}$$

Thus if we replace $S_2P + S_1P$ by $2D$, the error involved is about 0.005%. In this approximation, Eq. (10.16) becomes

$$S_2P - S_1P \approx \quad (10.17)$$

Hence we will have constructive interference resulting in a bright region when

$$x = x_n = \frac{n\lambda D}{d}; n = 0, \pm 1, \pm 2, \dots \quad (10.18)$$

On the other hand, we will have a dark region near

$$x = x_{n+1/2} = (n + \frac{1}{2}) \frac{\lambda D}{d}; n = 0, \pm 1, \pm 2 \quad (10.19)$$

Thus dark and bright bands appear on the screen, as shown in Fig. 10.13. Such bands are called *fringes*. Equations (10.18) and (10.19) show that dark and bright fringes are equally spaced and the distance between two consecutive bright and dark fringes is given by

$$\beta = x_{n+1} - x_n$$

$$\text{or } \beta = \frac{\lambda D}{d} \quad (10.20)$$

which is the expression for the *fringe width*. Obviously, the central point O (in Fig. 10.12) will be bright because $S_1O = S_2O$ and it will correspond to $n = 0$. If we consider the line perpendicular to the plane of the paper and passing through O [i.e., along the y -axis] then all points on this line will be equidistant from S_1 and S_2 and we will have a bright central fringe which is a straight line as shown in Fig. 10.13. In order to determine the shape of the interference pattern on the screen we note that a particular fringe would correspond to the locus of points with a constant value of $S_2P - S_1P$. Whenever this constant is an integral multiple of λ , the fringe will be bright and whenever it is an odd integral multiple of $\lambda/2$ it will be a dark fringe. Now, the locus of the point P lying in the $x-y$ plane such that $S_2P - S_1P (= \Delta)$ is a constant, is a hyperbola. Thus the fringe pattern will strictly be a hyperbola; however, if the distance D is very large compared to the fringe width, the fringes will be very nearly straight lines as shown in Fig. 10.13.

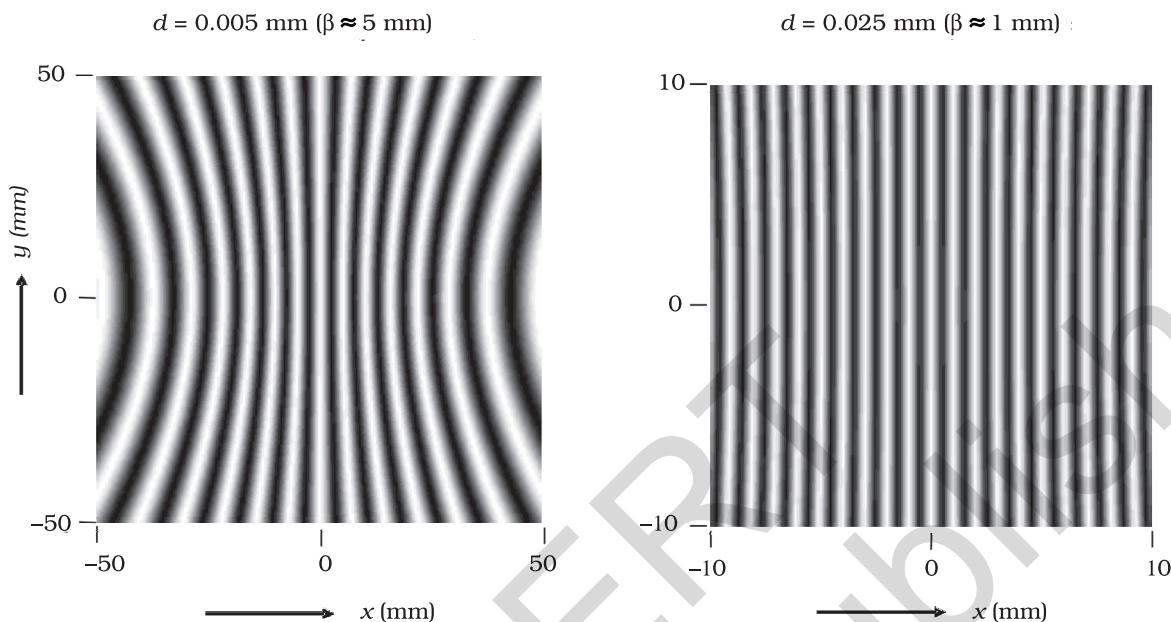


FIGURE 10.13 Computer generated fringe pattern produced by two point source S_1 and S_2 on the screen GG' (Fig. 10.12); (a) and (b) correspond to $d = 0.005 \text{ mm}$ and 0.025 mm , respectively (both figures correspond to $D = 5 \text{ cm}$ and $\lambda = 5 \times 10^{-5} \text{ cm}$) (Adopted from OPTICS by A. Ghatak, Tata McGraw Hill Publishing Co. Ltd., New Delhi, 2000.)

In the double-slit experiment shown in Fig. 10.12, we have taken the source hole S on the perpendicular bisector of the two slits, which is shown as the line SO . What happens if the source S is slightly away from the perpendicular bisector. Consider that the source is moved to some new point S' and suppose that Q is the mid-point of S_1 and S_2 . If the angle $S'QS$ is ϕ , then the central bright fringe occurs at an angle $-\phi$, on the other side. Thus, if the source S is on the perpendicular bisector, then the central fringe occurs at O , also on the perpendicular bisector. If S is shifted by an angle ϕ to point S' , then the central fringe appears at a point O' at an angle $-\phi$, which means that it is shifted by the same angle on the other side of the bisector. This also means that the source S' , the mid-point Q and the point O' of the central fringe are in a straight line.

We end this section by quoting from the Nobel lecture of Dennis Gabor*

The wave nature of light was demonstrated convincingly for the first time in 1801 by Thomas Young by a wonderfully simple experiment. He let a ray of sunlight into a dark room, placed a dark screen in front of it, pierced with two small pinholes, and beyond this, at some distance, a white screen. He then saw two darkish lines at both sides of a bright line, which gave him sufficient encouragement to repeat the experiment, this time with spirit flame as light source, with a little salt in it to produce the bright yellow sodium light. This time he saw a number of dark lines, regularly spaced; the first clear proof that light added to light can produce darkness. This phenomenon is called

* Dennis Gabor received the 1971 Nobel Prize in Physics for discovering the principles of holography.

Physics

Interactive animation of Young's experiment
<http://vsg.quasihome.com/interferer.html>



EXAMPLE 10.3

Example 10.3 Two slits are made one millimetre apart and the screen is placed one metre away. What is the fringe separation when blue-green light of wavelength 500 nm is used?

Solution Fringe spacing $= \frac{D\lambda}{d} = \frac{1 \times 5 \times 10^{-7}}{1 \times 10^{-3}} \text{ m}$
 $= 5 \times 10^{-4} \text{ m} = 0.5 \text{ mm}$

EXAMPLE 10.4

Example 10.4 What is the effect on the interference fringes in a Young's double-slit experiment due to each of the following operations:

- the screen is moved away from the plane of the slits;
- the (monochromatic) source is replaced by another (monochromatic) source of shorter wavelength;
- the separation between the two slits is increased;
- the source slit is moved closer to the double-slit plane;
- the width of the source slit is increased;
- the monochromatic source is replaced by a source of white light?

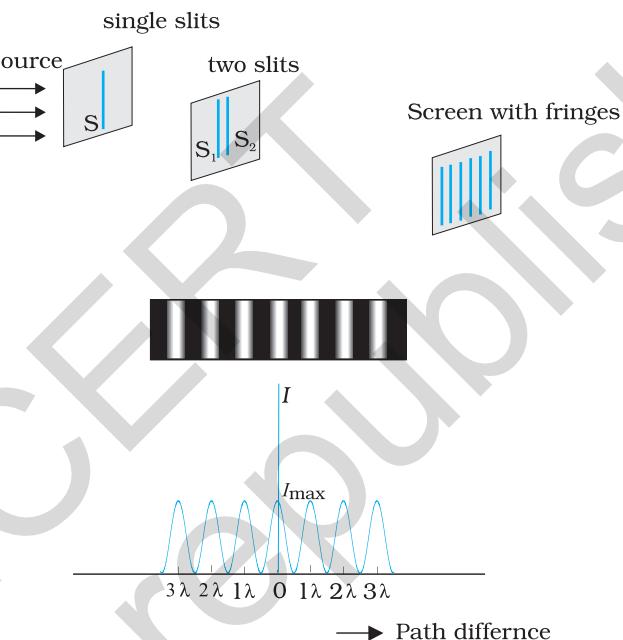


FIGURE 10.14 Photograph and the graph of the intensity distribution in Young's double-slit experiment.

(In each operation, take all parameters, other than the one specified, to remain unchanged.)

Solution

- Angular separation of the fringes remains constant ($= \lambda/d$). The actual separation of the fringes increases in proportion to the distance of the screen from the plane of the two slits.
- The separation of the fringes (and also angular separation) decreases. See, however, the condition mentioned in (d) below.
- The separation of the fringes (and also angular separation) decreases. See, however, the condition mentioned in (d) below.
- Let s be the size of the source and S its distance from the plane of the two slits. For interference fringes to be seen, the condition $s/S < \lambda/d$ should be satisfied; otherwise, interference patterns produced by different parts of the source overlap and no fringes are seen. Thus, as S decreases (i.e., the source slit is brought closer), the interference pattern gets less and less sharp, and when the source is brought too close for this condition to be valid, the fringes disappear. Till this happens, the fringe separation remains fixed.
- Same as in (d). As the source slit width increases, fringe pattern gets less and less sharp. When the source slit is so wide that the condition $s/S \leq \lambda/d$ is not satisfied, the interference pattern disappears.
- The interference patterns due to different component colours of white light overlap (incoherently). The central bright fringes for different colours are at the same position. Therefore, the central fringe is white. For a point P for which $S_2P - S_1P = \lambda_b/2$, where λ_b ($\approx 4000 \text{ \AA}$) represents the wavelength for the blue colour, the blue component will be absent and the fringe will appear red in colour. Slightly farther away where $S_2Q - S_1Q = \lambda_b = \lambda_r/2$ where λ_r ($\approx 8000 \text{ \AA}$) is the wavelength for the red colour, the fringe will be predominantly blue.

Thus, the fringe closest on either side of the central white fringe is red and the farthest will appear blue. After a few fringes, no clear fringe pattern is seen.

EXAMPLE 10.4

10.6 DIFFRACTION

If we look clearly at the shadow cast by an opaque object, close to the region of geometrical shadow, there are alternate dark and bright regions just like in interference. This happens due to the phenomenon of diffraction. Diffraction is a general characteristic exhibited by all types of waves, be it sound waves, light waves, water waves or matter waves. Since the wavelength of light is much smaller than the dimensions of most obstacles; we do not encounter diffraction effects of light in everyday observations. However, the finite resolution of our eye or of optical

■ Physics

instruments such as telescopes or microscopes is limited due to the phenomenon of diffraction. Indeed the colours that you see when a CD is viewed is due to diffraction effects. We will now discuss the phenomenon of diffraction.

10.6.1 The single slit

In the discussion of Young's experiment, we stated that a single narrow slit acts as a new source from which light spreads out. Even before Young, early experimenters – including Newton – had noticed that light spreads out from narrow holes and slits. It seems to turn around corners and enter regions where we would expect a shadow. These effects, known as *diffraction*, can only be properly understood using wave ideas. After all, you are hardly surprised to hear sound waves from someone talking around a corner!

When the double slit in Young's experiment is replaced by a single narrow slit (illuminated by a monochromatic source), a broad pattern with a central bright region is seen. On both sides, there are alternate dark and bright regions, the intensity becoming weaker away from the centre (Fig. 10.16). To understand this, go to Fig. 10.15, which shows a parallel beam of light falling normally on a single slit LN of width a . The diffracted light goes on to meet a screen. The midpoint of the slit is M.

A straight line through M perpendicular to the slit plane meets the screen at C. We want the intensity at any point P on the screen. As before, straight lines joining P to the different points L, M, N, etc., can be treated as parallel, making an angle θ with the normal MC.

The basic idea is to divide the slit into much smaller parts, and add their contributions at P with the proper phase differences. We are treating different parts of the wavefront at the slit as secondary sources. Because the incoming wavefront is parallel to the plane of the slit, these sources are in phase.

The path difference $NP - LP$ between the two edges of the slit can be calculated exactly as for Young's experiment. From Fig. 10.15,

$$\begin{aligned} NP - LP &= NQ \\ &= a \sin \theta \\ &\approx a\theta \end{aligned} \tag{10.21}$$

Similarly, if two points M_1 and M_2 in the slit plane are separated by y , the path difference $M_2P - M_1P \approx y\theta$. We now have to sum up equal, coherent contributions from a large number of sources, each with a different phase. This calculation was made by Fresnel using integral calculus, so we omit it here. The main features of the diffraction pattern can be understood by simple arguments.

At the central point C on the screen, the angle θ is zero. All path differences are zero and hence all the parts of the slit contribute in phase. This gives maximum intensity at C. Experimental observation shown in

Wave Optics

Fig. 10.15 indicates that the intensity has a central maximum at $\theta = 0$ and other secondary maxima at $\theta \approx (n+1/2) \lambda/a$, and has minima (zero intensity) at $\theta \approx n\lambda/a$, $n = \pm 1, \pm 2, \pm 3, \dots$. It is easy to see why it has minima at these values of angle. Consider first the angle θ where the path difference $a\theta$ is λ . Then,

$$\theta \approx \lambda/a. \quad (10.22)$$

Now, divide the slit into two equal halves LM and MN each of size $a/2$. For every point M_1 in LM, there is a point M_2 in MN such that $M_1 M_2 = a/2$. The path difference between M_1 and M_2 at P = $M_2 P - M_1 P = \theta a/2 = \lambda/2$ for the angle chosen. This means that the contributions from M_1 and M_2 are 180° out of phase and cancel in the direction $\theta = \lambda/a$. Contributions from the two halves of the slit LM and MN, therefore, cancel each other. Equation (10.22) gives the angle at which the intensity falls to zero. One can similarly show that the intensity is zero for $\theta = n\lambda/a$, with n being any integer (except zero!). Notice that the angular size of the central maximum increases when the slit width a decreases.

It is also easy to see why there are maxima at $\theta = (n + 1/2)\lambda/a$ and why they go on becoming weaker and weaker with increasing n . Consider an angle $\theta = 3\lambda/2a$ which is midway between two of the dark fringes. Divide the slit into three equal parts. If we take the first two thirds of the slit, the path difference between the two ends would be

$$\frac{2}{3}a \times \theta = \frac{2a}{3} \times \frac{3\lambda}{2a} = \lambda \quad (10.23)$$

The first two-thirds of the slit can therefore be divided into two halves which have a $\lambda/2$ path difference. The contributions of these two halves cancel in the same manner as described earlier. Only the remaining one-third of the slit contributes to the intensity at a point between the two minima. Clearly, this will be much weaker than the central maximum (where the entire slit contributes in phase). One can similarly show that there are maxima at $(n + 1/2)\lambda/a$ with $n = 2, 3, \dots$. These become weaker with increasing n , since only one-fifth, one-seventh, etc., of the slit contributes in these cases. The photograph and intensity pattern corresponding to it is shown in Fig. 10.16.

There has been prolonged discussion about difference between interference and diffraction among scientists since the discovery of these phenomena. In this context, it is

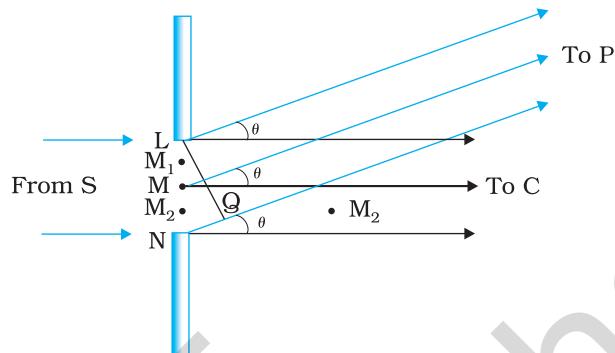


FIGURE 10.15 The geometry of path differences for diffraction by a single slit.

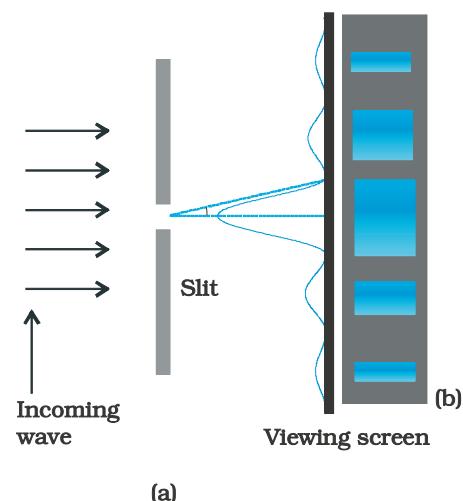


FIGURE 10.16 Intensity distribution and photograph of fringes due to diffraction at single slit.

Physics

interesting to note what Richard Feynman* has said in his famous Feynman Lectures on Physics:

No one has ever been able to define the difference between interference and diffraction satisfactorily. It is just a question of usage, and there is no specific, important physical difference between them. The best we can do is, roughly speaking, is to say that when there are only a few sources, say two interfering sources, then the result is usually called interference, but if there is a large number of them, it seems that the word diffraction is more often used.

In the double-slit experiment, we must note that the pattern on the screen is actually a superposition of single-slit diffraction from each slit or hole, and the double-slit interference pattern. This is shown in Fig. 10.17. It shows a broader diffraction peak in which there appear several fringes of smaller width due to double-slit interference. The number of interference fringes occurring in the broad diffraction peak depends on the ratio d/a , that is the ratio of the distance between the two slits to the width of a slit. In the limit of a becoming very small, the diffraction pattern will become very flat and we will observe the two-slit interference pattern [see Fig. 10.13(b)].

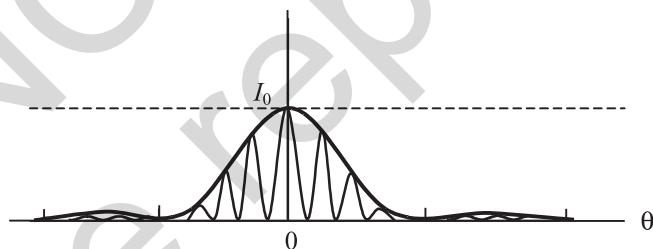


FIGURE 10.17 The actual double-slit interference pattern. The envelope shows the single slit diffraction.

Example 10.5 In Example 10.3, what should the width of each slit be to obtain 10 maxima of the double slit pattern within the central maximum of the single slit pattern?

Solution We want $a\theta = \lambda, \theta = \frac{\lambda}{a}$

$$10 \frac{\lambda}{d} = 2 \frac{\lambda}{a} \quad a = \frac{d}{5} = 0.2 \text{ mm}$$

Notice that the wavelength of light and distance of the screen do not enter in the calculation of a .

In the double-slit interference experiment of Fig. 10.12, what happens if we close one slit? You will see that it now amounts to a single slit. But you will have to take care of some shift in the pattern. We now have a source at S, and only one hole (or slit) S₁ or S₂. This will produce a single-

Interactive animation on single slit diffraction pattern
<http://www.phys.hawaii.edu/~teb/optics/java/sliddiffr/>

PHYSICS

EXAMPLE 10.5

* Richard Feynman was one of the recipients of the 1965 Nobel Prize in Physics for his fundamental work in quantum electrodynamics.

slit diffraction pattern on the screen. The centre of the central bright fringe will appear at a point which lies on the straight line SS_1 or SS_2 , as the case may be.

We now compare and contrast the interference pattern with that seen for a coherently illuminated single slit (usually called the single slit diffraction pattern).

- (i) The interference pattern has a number of equally spaced bright and dark bands. The diffraction pattern has a central bright maximum which is twice as wide as the other maxima. The intensity falls as we go to successive maxima away from the centre, on either side.
- (ii) We calculate the interference pattern by superposing two waves originating from the two narrow slits. The diffraction pattern is a superposition of a continuous family of waves originating from each point on a single slit.
- (iii) For a single slit of width a , the first null of the interference pattern occurs at an angle of λ/a . At the same angle of λ/a , we get a maximum (not a null) for two narrow slits separated by a distance a .

One must understand that both d and a have to be quite small, to be able to observe good interference and diffraction patterns. For example, the separation d between the two slits must be of the order of a milimetre or so. The width a of each slit must be even smaller, of the order of 0.1 or 0.2 mm.

In our discussion of Young's experiment and the single-slit diffraction, we have assumed that the screen on which the fringes are formed is at a large distance. The two or more paths from the slits to the screen were treated as parallel. This situation also occurs when we place a converging lens after the slits and place the screen at the focus. Parallel paths from the slit are combined at a single point on the screen. *Note that the lens does not introduce any extra path differences in a parallel beam.* This arrangement is often used since it gives more intensity than placing the screen far away. If f is the focal length of the lens, then we can easily work out the size of the central bright maximum. In terms of angles, the separation of the central maximum from the first null of the diffraction pattern is λ/a . Hence, the size on the screen will be $f\lambda/a$.

10.6.2 Seeing the single slit diffraction pattern

It is surprisingly easy to see the single-slit diffraction pattern for oneself. The equipment needed can be found in most homes — two razor blades and one clear glass electric bulb preferably with a straight filament. One has to hold the two blades so that the edges are parallel and have a narrow slit in between. This is easily done with the thumb and forefingers (Fig. 10.18).

Keep the slit parallel to the filament, right in front of the eye. Use spectacles if you normally do. With slight adjustment of the width of the slit and the parallelism of the edges, the pattern should be seen with its bright and dark bands. Since the position of all the bands (except the central one) depends on wavelength, they will show some colours. Using a filter for red or blue will make the fringes clearer. With both filters available, the wider fringes for red compared to blue can be seen.

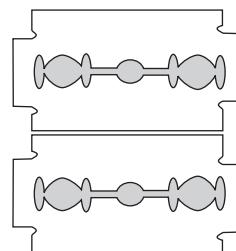


FIGURE 10.18
Holding two blades to form a single slit. A bulb filament viewed through this shows clear diffraction bands.

Physics

In this experiment, the filament plays the role of the first slit S in Fig. 10.16. The lens of the eye focuses the pattern on the screen (the retina of the eye).

With some effort, one can cut a double slit in an aluminium foil with a blade. The bulb filament can be viewed as before to repeat Young's experiment. In daytime, there is another suitable bright source subtending a small angle at the eye. This is the reflection of the Sun in any shiny convex surface (e.g., a cycle bell). Do not try direct sunlight – it can damage the eye and will not give fringes anyway as the Sun subtends an angle of $(1/2)^\circ$.

In interference and diffraction, light energy is redistributed. If it reduces in one region, producing a dark fringe, it increases in another region, producing a bright fringe. There is no gain or loss of energy, which is consistent with the principle of conservation of energy.

10.6.3 Resolving power of optical instruments

In Chapter 9 we had discussed about telescopes. The angular resolution of the telescope is determined by the objective of the telescope. The stars which are not resolved in the image produced by the objective cannot be resolved by any further magnification produced by the eyepiece. The primary purpose of the eyepiece is to provide magnification of the image produced by the objective.

Consider a parallel beam of light falling on a convex lens. If the lens is well corrected for aberrations, then geometrical optics tells us that the beam will get focused to a point. However, because of diffraction, the beam instead of getting focused to a point gets focused to a spot of finite area. In this case the effects due to diffraction can be taken into account by considering a plane wave incident on a circular aperture followed by a convex lens (Fig. 10.19). The analysis of the corresponding diffraction pattern is quite involved; however, in principle, it is similar to the analysis carried out to obtain the single-slit diffraction pattern. Taking into account the effects due to diffraction, the pattern on the focal plane would consist of a central bright region surrounded by concentric dark and bright rings (Fig. 10.19). A detailed analysis shows that the radius of the central bright region is approximately given by

$$r_0 \approx \frac{1.22 \lambda f}{2a} = \frac{0.61 \lambda f}{a} \quad (10.24)$$

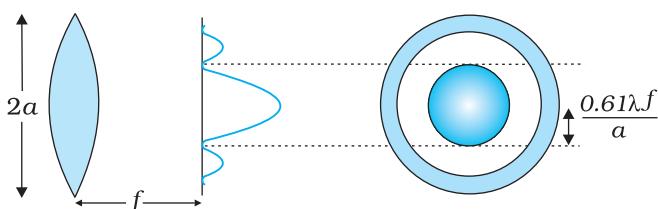


FIGURE 10.19 A parallel beam of light is incident on a convex lens. Because of diffraction effects, the beam gets focused to a spot of radius $\approx 0.61 \lambda f/a$.

where f is the focal length of the lens and $2a$ is the diameter of the circular aperture or the diameter of the lens, whichever is smaller. Typically if

$$\lambda \approx 0.5 \text{ } \mu\text{m}, f \approx 20 \text{ cm} \text{ and } a \approx 5 \text{ cm}$$

we have

$$r_0 \approx 1.2 \text{ } \mu\text{m}$$

Although the size of the spot is very small, it plays an important role in determining the limit of resolution of optical instruments like a telescope or a microscope. For the two stars to be just resolved

$$f\Delta\theta \approx r_0 \approx \frac{0.61\lambda f}{a}$$

implying

$$\Delta\theta \approx \frac{0.61\lambda}{a} \quad (10.25)$$

Thus $\Delta\theta$ will be small if the diameter of the objective is large. This implies that the telescope will have better resolving power if a is large. It is for this reason that for better resolution, a telescope must have a large diameter objective.

Example 10.6 Assume that light of wavelength 6000\AA is coming from a star. What is the limit of resolution of a telescope whose objective has a diameter of 100 inch?

Solution A 100 inch telescope implies that $2a = 100$ inch = 254 cm. Thus if,

$$\lambda \approx 6000\text{\AA} = 6 \times 10^{-5} \text{ cm}$$

then

$$\Delta\theta \approx \frac{0.61 \times 6 \times 10^{-5}}{127} \approx 2.9 \times 10^{-7} \text{ radians}$$

EXAMPLE 10.6

We can apply a similar argument to the objective lens of a microscope. In this case, the object is placed slightly beyond f , so that a real image is formed at a distance v [Fig. 10.20]. The magnification – ratio of image size to object size – is given by $m \approx v/f$. It can be seen from Fig. 10.20 that

$$D/f \approx 2 \tan \beta \quad (10.26)$$

where 2β is the angle subtended by the diameter of the objective lens at the focus of the microscope.

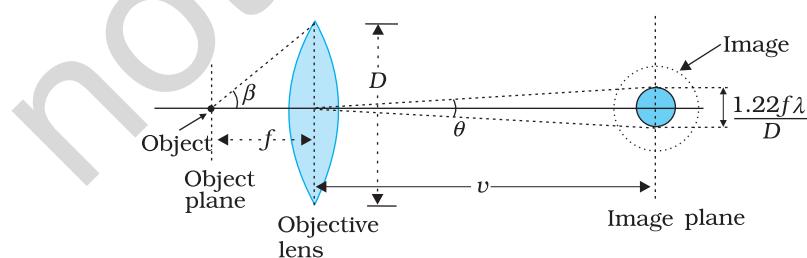


FIGURE 10.20 Real image formed by the objective lens of the microscope.

Physics

DETERMINE THE RESOLVING POWER OF YOUR EYE

You can estimate the resolving power of your eye with a simple experiment. Make black stripes of equal width separated by white stripes; see figure here. All the black stripes should be of equal width, while the width of the intermediate white stripes should increase as you go from the left to the right. For example, let all black stripes have a width of 5 mm. Let the width of the first two white stripes be 0.5 mm each, the next two white stripes be 1 mm each, the next two 1.5 mm each, etc. Paste this pattern on a wall in a room or laboratory, at the height of your eye.



Now watch the pattern, preferably with one eye. By moving away or closer to the wall, find the position where you can just see some two black stripes as separate stripes. All the black stripes to the left of this stripe would merge into one another and would not be distinguishable. On the other hand, the black stripes to the right of this would be more and more clearly visible. Note the width d of the white stripe which separates the two regions, and measure the distance D of the wall from your eye. Then d/D is the resolution of your eye.

You have watched specks of dust floating in air in a sunbeam entering through your window. Find the distance (of a speck) which you can clearly see and distinguish from a neighbouring speck. Knowing the resolution of your eye and the distance of the speck, estimate the size of the speck of dust.

When the separation between two points in a microscopic specimen is comparable to the wavelength λ of the light, the diffraction effects become important. The image of a point object will again be a diffraction pattern whose size in the image plane will be

$$v\theta = v \frac{1.22 \lambda}{D} \quad (10.27)$$

Two objects whose images are closer than this distance will not be resolved, they will be seen as one. The corresponding minimum separation, d_{\min} , in the object plane is given by

$$\begin{aligned} d_{\min} &= v \frac{1.22 \lambda}{D} / m \\ &= \frac{1.22 \lambda}{D} \cdot \frac{v}{m} \\ &= \frac{1.22 f \lambda}{D} \end{aligned} \quad (10.28)$$

Now, combining Eqs. (10.26) and (10.28), we get

$$d_{\min} = \frac{1.22 \lambda}{2 \tan \beta}$$

$$\square \frac{1.22\lambda}{2 \sin \beta} \quad (10.29)$$

If the medium between the object and the objective lens is not air but a medium of refractive index n , Eq. (10.29) gets modified to

$$d_{\min} = \frac{1.22\lambda}{2n \sin \beta} \quad (10.30)$$

The product $n \sin \beta$ is called the *numerical aperture* and is sometimes marked on the objective.

The resolving power of the microscope is given by the reciprocal of the minimum separation of two points seen as distinct. It can be seen from Eq. (10.30) that the resolving power can be increased by choosing a medium of higher refractive index. Usually an oil having a refractive index close to that of the objective glass is used. Such an arrangement is called an '*oil immersion objective*'. Notice that it is not possible to make $\sin \beta$ larger than unity. Thus, we see that the resolving power of a microscope is basically determined by the wavelength of the light used.

There is a likelihood of confusion between resolution and magnification, and similarly between the role of a telescope and a microscope to deal with these parameters. A telescope produces images of far objects nearer to our eye. Therefore objects which are not resolved at far distance, can be resolved by looking at them through a telescope. A microscope, on the other hand, magnifies objects (which are near to us) and produces their larger image. We may be looking at two stars or two satellites of a far-away planet, or we may be looking at different regions of a living cell. In this context, it is good to remember that a telescope resolves whereas a microscope magnifies.

10.6.4 The validity of ray optics

An aperture (i.e., slit or hole) of size a illuminated by a parallel beam sends diffracted light into an angle of approximately $\approx \lambda/a$. This is the angular size of the bright central maximum. In travelling a distance z , the diffracted beam therefore acquires a width $z\lambda/a$ due to diffraction. It is interesting to ask at what value of z the spreading due to diffraction becomes comparable to the size a of the aperture. We thus approximately equate $z\lambda/a$ with a . This gives the distance beyond which divergence of the beam of width a becomes significant. Therefore,

$$z \square \frac{a^2}{\lambda} \quad (10.31)$$

We define a quantity z_F called the *Fresnel distance* by the following equation

$$z_F \square a^2 / \lambda$$

Equation (10.31) shows that for distances much smaller than z_F , the spreading due to diffraction is smaller compared to the size of the beam. It becomes comparable when the distance is approximately z_F . For distances much greater than z_F , the spreading due to diffraction

EXAMPLE 10.7

dominates over that due to ray optics (i.e., the size a of the aperture). Equation (10.31) also shows that ray optics is valid in the limit of wavelength tending to zero.

Example 10.7 For what distance is ray optics a good approximation when the aperture is 3 mm wide and the wavelength is 500 nm?

$$\text{Solution} \quad z_F = \frac{a^2}{\lambda} = \frac{(3 \times 10^{-3})^2}{5 \times 10^{-7}} = 18 \text{ m}$$

This example shows that even with a small aperture, diffraction spreading can be neglected for rays many metres in length. Thus, ray optics is valid in many common situations.

10.7 POLARISATION

Consider holding a long string that is held horizontally, the other end of which is assumed to be fixed. If we move the end of the string up and down in a periodic manner, we will generate a wave propagating in the $+x$ direction (Fig. 10.21). Such a wave could be described by the following equation

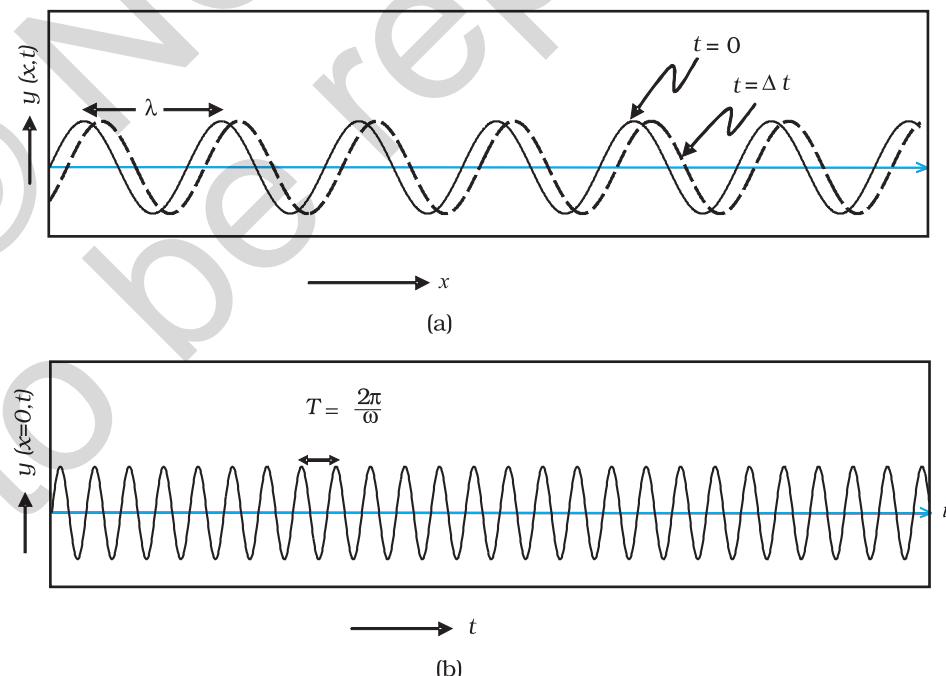


FIGURE 10.21 (a) The curves represent the displacement of a string at $t = 0$ and at $t = \Delta t$, respectively when a sinusoidal wave is propagating in the $+x$ -direction. (b) The curve represents the time variation of the displacement at $x = 0$ when a sinusoidal wave is propagating in the $+x$ -direction. At $x = \Delta x$, the time variation of the displacement will be slightly displaced to the right.

Wave Optics

$$y(x,t) = a \sin(kx - \omega t) \quad (10.32)$$

where a and $\omega (= 2\pi\nu)$ represent the amplitude and the angular frequency of the wave, respectively; further,

$$\lambda = \frac{2\pi}{k} \quad (10.33)$$

represents the wavelength associated with the wave. We had discussed propagation of such waves in Chapter 15 of Class XI textbook. Since the displacement (which is along the y direction) is at right angles to the direction of propagation of the wave, we have what is known as a *transverse wave*. Also, since the displacement is in the y direction, it is often referred to as a y -polarised wave. Since each point on the string moves on a straight line, the wave is also referred to as a linearly polarised wave. Further, the string always remains confined to the x - y plane and therefore it is also referred to as a *plane polarised wave*.

In a similar manner we can consider the vibration of the string in the x - z plane generating a z -polarised wave whose displacement will be given by

$$z(x,t) = a \sin(kx - \omega t) \quad (10.34)$$

It should be mentioned that the linearly polarised waves [described by Eqs. (10.33) and (10.34)] are all transverse waves; i.e., the displacement of each point of the string is always at right angles to the direction of propagation of the wave. Finally, if the plane of vibration of the string is changed randomly in very short intervals of time, then we have what is known as an *unpolarised wave*. Thus, for an unpolarised wave the displacement will be randomly changing with time though it will always be perpendicular to the direction of propagation.

Light waves are transverse in nature; i.e., the electric field associated with a propagating light wave is always at right angles to the direction of propagation of the wave. This can be easily demonstrated using a simple polaroid. You must have seen thin plastic like sheets, which are called *polaroids*. A polaroid consists of long chain molecules aligned in a particular direction. The electric vectors (associated with the propagating light wave) along the direction of the aligned molecules get absorbed. Thus, if an unpolarised light wave is incident on such a polaroid then the light wave will get linearly polarised with the electric vector oscillating along a direction perpendicular to the aligned molecules; this direction is known as the *pass-axis* of the polaroid.

Thus, if the light from an ordinary source (like a sodium lamp) passes through a polaroid sheet P_1 , it is observed that its intensity is reduced by half. Rotating P_1 has no effect on the transmitted beam and transmitted intensity remains constant. Now, let an identical piece of polaroid P_2 be placed before P_1 . As expected, the light from the lamp is reduced in intensity on passing through P_2 alone. But now rotating P_1 has a dramatic effect on the light coming from P_2 . In one position, the intensity transmitted

by P_2 followed by P_1 is nearly zero. When turned by 90° from this position, P_1 transmits nearly the full intensity emerging from P_2 (Fig. 10.22).

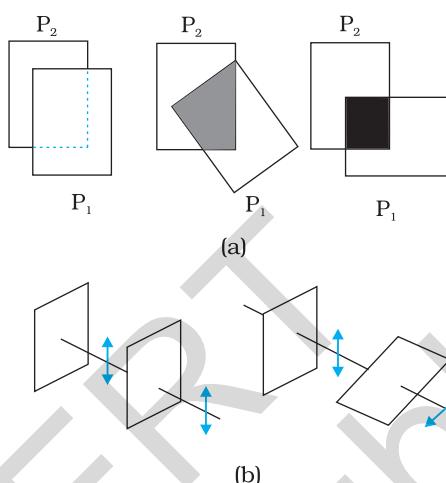


FIGURE 10.22 (a) Passage of light through two polaroids P_2 and P_1 . The transmitted fraction falls from 1 to 0 as the angle between them varies from 0° to 90° . Notice that the light seen through a single polaroid P_1 does not vary with angle. (b) Behaviour of the electric vector when light passes through two polaroids. The transmitted polarisation is the component parallel to the polaroid axis. The double arrows show the oscillations of the electric vector.

The above experiment can be easily understood by assuming that light passing through the polaroid P_2 gets polarised along the pass-axis of P_2 . If the pass-axis of P_2 makes an angle θ with the pass-axis of P_1 , then when the polarised beam passes through the polaroid P_2 , the component $E \cos \theta$ (along the pass-axis of P_2) will pass through P_2 . Thus, as we rotate the polaroid P_1 (or P_2), the intensity will vary as:

$$I = I_0 \cos^2 \theta \quad (10.35)$$

where I_0 is the intensity of the polarized light after passing through P_1 . This is known as *Malus' law*. The above discussion shows that the intensity coming out of a single polaroid is half of the incident intensity. By putting a second polaroid, the intensity can be further controlled from 50% to zero of the incident intensity by adjusting the angle between the pass-axes of two polaroids.

Polaroids can be used to control the intensity, in sunglasses, windowpanes, etc. Polaroids are also used in photographic cameras and 3D movie cameras.

Example 10.8 Discuss the intensity of transmitted light when a polaroid sheet is rotated between two crossed polaroids?

Solution Let I_0 be the intensity of polarised light after passing through the first polariser P_1 . Then the intensity of light after passing through second polariser P_2 will be

$$I = I_0 \cos^2 \theta,$$

where θ is the angle between pass axes of P_1 and P_2 . Since P_1 and P_3 are crossed the angle between the pass axes of P_2 and P_3 will be $(\pi/2 - \theta)$. Hence the intensity of light emerging from P_3 will be

$$\begin{aligned} I &= I_0 \cos^2 \theta \cos^2 \frac{\pi}{2} - \theta \\ &= I_0 \cos^2 \theta \sin^2 \theta = (I_0/4) \sin^2 2\theta \end{aligned}$$

Therefore, the transmitted intensity will be maximum when $\theta = \pi/4$.

10.7.1 Polarisation by scattering

The light from a clear blue portion of the sky shows a rise and fall of intensity when viewed through a polaroid which is rotated. This is nothing but sunlight, which has changed its direction (having been scattered) on encountering the molecules of the earth's atmosphere. As Fig. 10.23(a) shows, the incident sunlight is unpolarised. The dots stand for polarisation perpendicular to the plane of the figure. The double arrows show polarisation in the plane of the figure. (There is no phase relation between these two in unpolarised light). Under the influence of the electric field of the incident wave the electrons in the molecules acquire components of motion in both these directions. We have drawn an observer looking at 90° to the direction of the sun. Clearly, charges accelerating parallel to the double arrows do not radiate energy towards this observer since their acceleration has no transverse component. The radiation scattered by the molecule is therefore represented by dots. It is polarised perpendicular to the plane of the figure. This explains the polarisation of scattered light from the sky.

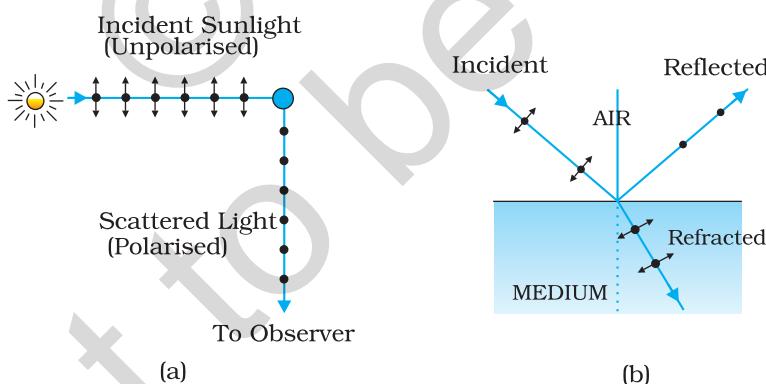


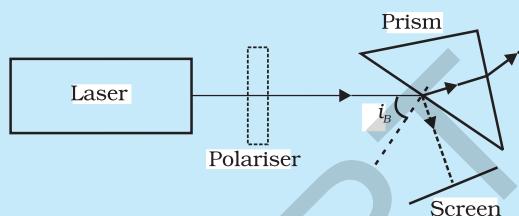
FIGURE 10.23 (a) Polarisation of the blue scattered light from the sky.

The incident sunlight is unpolarised (dots and arrows). A typical molecule is shown. It scatters light by 90° polarised normal to the plane of the paper (dots only). (b) Polarisation of light reflected from a transparent medium at the Brewster angle (reflected ray perpendicular to refracted ray).

The scattering of light by molecules was intensively investigated by C.V. Raman and his collaborators in Kolkata in the 1920s. Raman was awarded the Nobel Prize for Physics in 1930 for this work.

A SPECIAL CASE OF TOTAL TRANSMISSION

When light is incident on an interface of two media, it is observed that some part of it gets reflected and some part gets transmitted. Consider a related question: *Is it possible that under some conditions a monochromatic beam of light incident on a surface (which is normally reflective) gets completely transmitted with no reflection?* To your surprise, the answer is *yes*.



Let us try a simple experiment and check what happens. Arrange a laser, a good polariser, a prism and screen as shown in the figure here.

Let the light emitted by the laser source pass through the polariser and be incident on the surface of the prism at the Brewster's angle of incidence i_B . Now rotate the polariser carefully and you will observe that for a specific alignment of the polariser, the light incident on the prism is completely transmitted and no light is reflected from the surface of the prism. The reflected spot will completely vanish.

10.7.2 Polarisation by reflection

Figure 10.23(b) shows light reflected from a transparent medium, say, water. As before, the dots and arrows indicate that both polarisations are present in the incident and refracted waves. We have drawn a situation in which the reflected wave travels at right angles to the refracted wave. The oscillating electrons in the water produce the reflected wave. These move in the two directions transverse to the radiation from wave in the medium, i.e., the *refracted wave*. The arrows are parallel to the direction of the *reflected wave*. Motion in this direction does not contribute to the reflected wave. As the figure shows, the reflected light is therefore linearly polarised perpendicular to the plane of the figure (represented by dots). This can be checked by looking at the reflected light through an analyser. The transmitted intensity will be zero when the axis of the analyser is in the plane of the figure, i.e., the plane of incidence.

When unpolarised light is incident on the boundary between two transparent media, the reflected light is polarised with its electric vector perpendicular to the plane of incidence when the refracted and reflected rays make a right angle with each other. Thus we have seen that when reflected wave is perpendicular to the refracted wave, the reflected wave is a totally polarised wave. The angle of incidence in this case is called *Brewster's angle* and is denoted by i_B . We can see that i_B is related to the refractive index of the denser medium. Since we have $i_B + r = \pi/2$, we get from Snell's law

$$\mu = \frac{\sin i_B}{\sin r} = \frac{\sin i_B}{\sin(\pi/2 - i_B)}$$

$$= \frac{\sin i_B}{\cos i_B} = \tan i_B \quad (10.36)$$

This is known as *Brewster's law*.

Example 10.9 Unpolarised light is incident on a plane glass surface. What should be the angle of incidence so that the reflected and refracted rays are perpendicular to each other?

Solution For $i + r$ to be equal to $\pi/2$, we should have $\tan i_B = \mu = 1.5$. This gives $i_B = 57^\circ$. This is the Brewster's angle for air to glass interface.

EXAMPLE 10.9

For simplicity, we have discussed scattering of light by 90° , and reflection at the Brewster angle. In this special situation, one of the two perpendicular components of the electric field is zero. At other angles, both components are present but one is stronger than the other. There is no stable phase relationship between the two perpendicular components since these are derived from two perpendicular components of an unpolarised beam. When such light is viewed through a rotating analyser, one sees a maximum and a minimum of intensity but not complete darkness. This kind of light is called *partially polarised*.

Let us try to understand the situation. When an unpolarised beam of light is incident at the Brewster's angle on an interface of two media, only part of light with electric field vector perpendicular to the plane of incidence will be reflected. Now by using a good polariser, if we completely remove all the light with its electric vector perpendicular to the plane of incidence and let this light be incident on the surface of the prism at Brewster's angle, you will then observe no reflection and there will be total transmission of light.

We began this chapter by pointing out that there are some phenomena which can be explained only by the wave theory. In order to develop a proper understanding, we first described how some phenomena like reflection and refraction, which were studied on this basis of Ray Optics in Chapter 9, can also be understood on the basis of Wave Optics. Then we described Young's double slit experiment which was a turning point in the study of optics. Finally, we described some associated points such as diffraction, resolution, polarisation, and validity of ray optics. In the next chapter, you will see how new experiments led to new theories at the turn of the century around 1900 A.D.

SUMMARY

1. Huygens' principle tells us that each point on a wavefront is a source of secondary waves, which add up to give the wavefront at a later time.
2. Huygens' construction tells us that the new wavefront is the forward envelope of the secondary waves. When the speed of light is independent of direction, the secondary waves are spherical. The rays are then perpendicular to both the wavefronts and the time of travel

is the same measured along any ray. This principle leads to the well known laws of reflection and refraction.

3. The principle of superposition of waves applies whenever two or more sources of light illuminate the same point. When we consider the intensity of light due to these sources at the given point, there is an interference term in addition to the sum of the individual intensities. But this term is important only if it has a non-zero average, which occurs only if the sources have the same frequency and a stable phase difference.
4. Young's double slit of separation d gives equally spaced fringes of angular separation λ/d . The source, mid-point of the slits, and central bright fringe lie in a straight line. An extended source will destroy the fringes if it subtends angle more than λ/d at the slits.
5. A single slit of width a gives a diffraction pattern with a central maximum. The intensity falls to zero at angles of $\pm \frac{\lambda}{a}, \pm \frac{2\lambda}{a}$, etc., with successively weaker secondary maxima in between. Diffraction limits the angular resolution of a telescope to λ/D where D is the diameter. Two stars closer than this give strongly overlapping images. Similarly, a microscope objective subtending angle 2β at the focus, in a medium of refractive index n , will just separate two objects spaced at a distance $\lambda/(2n \sin \beta)$, which is the resolution limit of a microscope. Diffraction determines the limitations of the concept of light rays. A beam of width a travels a distance a^2/λ , called the Fresnel distance, before it starts to spread out due to diffraction.
6. Natural light, e.g., from the sun is unpolarised. This means the electric vector takes all possible directions in the transverse plane, rapidly and randomly, during a measurement. A polaroid transmits only one component (parallel to a special axis). The resulting light is called linearly polarised or plane polarised. When this kind of light is viewed through a second polaroid whose axis turns through 2π , two maxima and minima of intensity are seen. Polarised light can also be produced by reflection at a special angle (called the Brewster angle) and by scattering through $\pi/2$ in the earth's atmosphere.

POINTS TO PONDER

1. Waves from a point source spread out in all directions, while light was seen to travel along narrow rays. It required the insight and experiment of Huygens, Young and Fresnel to understand how a wave theory could explain all aspects of the behaviour of light.
2. The crucial new feature of waves is interference of amplitudes from different sources which can be both constructive and destructive, as shown in Young's experiment.
3. Even a wave falling on single slit should be regarded as a large number of sources which interfere constructively in the forward direction ($\theta = 0$), and destructively in other directions.
4. Diffraction phenomena define the limits of ray optics. The limit of the ability of microscopes and telescopes to distinguish very close objects is set by the wavelength of light.
5. Most interference and diffraction effects exist even for longitudinal waves like sound in air. But polarisation phenomena are special to transverse waves like light waves.

EXERCISES

- 10.1** Monochromatic light of wavelength 589 nm is incident from air on a water surface. What are the wavelength, frequency and speed of (a) reflected, and (b) refracted light? Refractive index of water is 1.33.
- 10.2** What is the shape of the wavefront in each of the following cases:
(a) Light diverging from a point source.
(b) Light emerging out of a convex lens when a point source is placed at its focus.
(c) The portion of the wavefront of light from a distant star intercepted by the Earth.
- 10.3** (a) The refractive index of glass is 1.5. What is the speed of light in glass? (Speed of light in vacuum is $3.0 \times 10^8 \text{ m s}^{-1}$)
(b) Is the speed of light in glass independent of the colour of light? If not, which of the two colours red and violet travels slower in a glass prism?
- 10.4** In a Young's double-slit experiment, the slits are separated by 0.28 mm and the screen is placed 1.4 m away. The distance between the central bright fringe and the fourth bright fringe is measured to be 1.2 cm. Determine the wavelength of light used in the experiment.
- 10.5** In Young's double-slit experiment using monochromatic light of wavelength λ , the intensity of light at a point on the screen where path difference is λ , is K units. What is the intensity of light at a point where path difference is $\lambda/3$?
- 10.6** A beam of light consisting of two wavelengths, 650 nm and 520 nm, is used to obtain interference fringes in a Young's double-slit experiment.
(a) Find the distance of the third bright fringe on the screen from the central maximum for wavelength 650 nm.
(b) What is the least distance from the central maximum where the bright fringes due to both the wavelengths coincide?
- 10.7** In a double-slit experiment the angular width of a fringe is found to be 0.2° on a screen placed 1 m away. The wavelength of light used is 600 nm. What will be the angular width of the fringe if the entire experimental apparatus is immersed in water? Take refractive index of water to be $4/3$.
- 10.8** What is the Brewster angle for air to glass transition? (Refractive index of glass = 1.5.)
- 10.9** Light of wavelength 5000 Å falls on a plane reflecting surface. What are the wavelength and frequency of the reflected light? For what angle of incidence is the reflected ray normal to the incident ray?
- 10.10** Estimate the distance for which ray optics is good approximation for an aperture of 4 mm and wavelength 400 nm.

ADDITIONAL EXERCISES

- 10.11** The 6563 \AA $\text{H}\alpha$ line emitted by hydrogen in a star is found to be red-shifted by 15 \AA . Estimate the speed with which the star is receding from the Earth.
- 10.12** Explain how Corpuscular theory predicts the speed of light in a medium, say, water, to be greater than the speed of light in vacuum. Is the prediction confirmed by experimental determination of the speed of light in water? If not, which alternative picture of light is consistent with experiment?
- 10.13** You have learnt in the text how Huygens' principle leads to the laws of reflection and refraction. Use the same principle to deduce directly that a point object placed in front of a plane mirror produces a virtual image whose distance from the mirror is equal to the object distance from the mirror.
- 10.14** Let us list some of the factors, which could possibly influence the speed of wave propagation:
- (i) nature of the source.
 - (ii) direction of propagation.
 - (iii) motion of the source and/or observer.
 - (iv) wavelength.
 - (v) intensity of the wave.
- On which of these factors, if any, does
- (a) the speed of light in vacuum,
 - (b) the speed of light in a medium (say, glass or water),
- depend?
- 10.15** For sound waves, the Doppler formula for frequency shift differs slightly between the two situations: (i) source at rest; observer moving, and (ii) source moving; observer at rest. The exact Doppler formulas for the case of light waves in vacuum are, however, strictly identical for these situations. Explain why this should be so. Would you expect the formulas to be strictly identical for the two situations in case of light travelling in a medium?
- 10.16** In double-slit experiment using light of wavelength 600 nm , the angular width of a fringe formed on a distant screen is 0.1° . What is the spacing between the two slits?
- 10.17** Answer the following questions:
- (a) In a single slit diffraction experiment, the width of the slit is made double the original width. How does this affect the size and intensity of the central diffraction band?
 - (b) In what way is diffraction from each slit related to the interference pattern in a double-slit experiment?
 - (c) When a tiny circular obstacle is placed in the path of light from a distant source, a bright spot is seen at the centre of the shadow of the obstacle. Explain why?
 - (d) Two students are separated by a 7 m partition wall in a room 10 m high. If both light and sound waves can bend around

Wave Optics

obstacles, how is it that the students are unable to see each other even though they can converse easily.

- (e) Ray optics is based on the assumption that light travels in a straight line. Diffraction effects (observed when light propagates through small apertures/slits or around small obstacles) disprove this assumption. Yet the ray optics assumption is so commonly used in understanding location and several other properties of images in optical instruments. What is the justification?

10.18 Two towers on top of two hills are 40 km apart. The line joining them passes 50 m above a hill halfway between the towers. What is the longest wavelength of radio waves, which can be sent between the towers without appreciable diffraction effects?

10.19 A parallel beam of light of wavelength 500 nm falls on a narrow slit and the resulting diffraction pattern is observed on a screen 1 m away. It is observed that the first minimum is at a distance of 2.5 mm from the centre of the screen. Find the width of the slit.

10.20 Answer the following questions:

- (a) When a low flying aircraft passes overhead, we sometimes notice a slight shaking of the picture on our TV screen. Suggest a possible explanation.
- (b) As you have learnt in the text, the principle of linear superposition of wave displacement is basic to understanding intensity distributions in diffraction and interference patterns. What is the justification of this principle?

10.21 In deriving the single slit diffraction pattern, it was stated that the intensity is zero at angles of $n\lambda/a$. Justify this by suitably dividing the slit to bring out the cancellation.

Chapter Eleven

DUAL NATURE OF RADIATION AND MATTER

11.1 INTRODUCTION

The Maxwell's equations of electromagnetism and Hertz experiments on the generation and detection of electromagnetic waves in 1887 strongly established the wave nature of light. Towards the same period at the end of 19th century, experimental investigations on conduction of electricity (electric discharge) through gases at low pressure in a discharge tube led to many historic discoveries. The discovery of X-rays by Roentgen in 1895, and of electron by J. J. Thomson in 1897, were important milestones in the understanding of atomic structure. It was found that at sufficiently low pressure of about 0.001 mm of mercury column, a discharge took place between the two electrodes on applying the electric field to the gas in the discharge tube. A fluorescent glow appeared on the glass opposite to cathode. The colour of glow of the glass depended on the type of glass, it being yellowish-green for soda glass. The cause of this fluorescence was attributed to the radiation which appeared to be coming from the cathode. These *cathode rays* were discovered, in 1870, by William Crookes who later, in 1879, suggested that these rays consisted of streams of fast moving negatively charged particles. The British physicist J. J. Thomson (1856 -1940) confirmed this hypothesis. By applying mutually perpendicular electric and magnetic fields across the discharge tube, J. J. Thomson was the first to determine experimentally the speed and the specific charge [charge to mass ratio (e/m)] of the cathode ray.

Dual Nature of Radiation and Matter

particles. They were found to travel with speeds ranging from about 0.1 to 0.2 times the speed of light (3×10^8 m/s). The presently accepted value of e/m is 1.76×10^{11} C/kg. Further, the value of e/m was found to be independent of the nature of the material/metal used as the cathode (emitter), or the gas introduced in the discharge tube. This observation suggested the universality of the cathode ray particles.

Around the same time, in 1887, it was found that certain metals, when irradiated by ultraviolet light, emitted negatively charged particles having small speeds. Also, certain metals when heated to a high temperature were found to emit negatively charged particles. The value of e/m of these particles was found to be the same as that for cathode ray particles. These observations thus established that all these particles, although produced under different conditions, were identical in nature. J. J. Thomson, in 1897, named these particles as *electrons*, and suggested that they were fundamental, universal constituents of matter. For his epoch-making discovery of electron, through his theoretical and experimental investigations on conduction of electricity by gasses, he was awarded the Nobel Prize in Physics in 1906. In 1913, the American physicist R. A. Millikan (1868–1953) performed the pioneering oil-drop experiment for the precise measurement of the charge on an electron. He found that the charge on an oil-droplet was always an integral multiple of an elementary charge, 1.602×10^{-19} C. Millikan's experiment established that *electric charge is quantised*. From the values of charge (e) and specific charge (e/m), the mass (m) of the electron could be determined.

11.2 ELECTRON EMISSION

We know that metals have free electrons (negatively charged particles) that are responsible for their conductivity. However, the free electrons cannot normally escape out of the metal surface. If an electron attempts to come out of the metal, the metal surface acquires a positive charge and pulls the electron back to the metal. The free electron is thus held inside the metal surface by the attractive forces of the ions. Consequently, the electron can come out of the metal surface only if it has got sufficient energy to overcome the attractive pull. A certain minimum amount of energy is required to be given to an electron to pull it out from the surface of the metal. This minimum energy required by an electron to escape from the metal surface is called the *work function* of the metal. It is generally denoted by ϕ_0 and measured in eV (electron volt). One electron volt is the energy gained by an electron when it has been accelerated by a potential difference of 1 volt, so that $1 \text{ eV} = 1.602 \times 10^{-19} \text{ J}$.

This unit of energy is commonly used in atomic and nuclear physics. The work function (ϕ_0) depends on the properties of the metal and the nature of its surface. The values of work function of some metals are given in Table 11.1. These values are approximate as they are very sensitive to surface impurities.

Note from Table 11.1 that the work function of platinum is the highest ($\phi_0 = 5.65$ eV) while it is the lowest ($\phi_0 = 2.14$ eV) for caesium.

The minimum energy required for the electron emission from the metal surface can be supplied to the free electrons by any one of the following physical processes:

TABLE 11.1 WORK FUNCTIONS OF SOME METALS

Metal	Work function ϕ_0 (eV)	Metal	Work function ϕ_0 (eV)
Cs	2.14	Al	4.28
K	2.30	Hg	4.49
Na	2.75	Cu	4.65
Ca	3.20	Ag	4.70
Mo	4.17	Ni	5.15
Pb	4.25	Pt	5.65

- (i) *Thermionic emission*: By suitably heating, sufficient thermal energy can be imparted to the free electrons to enable them to come out of the metal.
- (ii) *Field emission*: By applying a very strong electric field (of the order of 10^8 V m^{-1}) to a metal, electrons can be pulled out of the metal, as in a spark plug.
- (iii) *Photo-electric emission*: When light of suitable frequency illuminates a metal surface, electrons are emitted from the metal surface. These photo(light)-generated electrons are called *photoelectrons*.

11.3 PHOTOELECTRIC EFFECT

11.3.1 Hertz's observations

The phenomenon of photoelectric emission was discovered in 1887 by Heinrich Hertz (1857-1894), during his electromagnetic wave experiments. In his experimental investigation on the production of electromagnetic waves by means of a spark discharge, Hertz observed that high voltage sparks across the detector loop were enhanced when the emitter plate was illuminated by ultraviolet light from an arc lamp.

Light shining on the metal surface somehow facilitated the escape of free, charged particles which we now know as electrons. When light falls on a metal surface, some electrons near the surface absorb enough energy from the incident radiation to overcome the attraction of the positive ions in the material of the surface. After gaining sufficient energy from the incident light, the electrons escape from the surface of the metal into the surrounding space.

11.3.2 Hallwachs' and Lenard's observations

Wilhelm Hallwachs and Philipp Lenard investigated the phenomenon of photoelectric emission in detail during 1886-1902.

Lenard (1862-1947) observed that when ultraviolet radiations were allowed to fall on the emitter plate of an evacuated glass tube enclosing two electrodes (metal plates), current flows in the circuit (Fig. 11.1). As soon as the ultraviolet radiations were stopped, the current flow also

Dual Nature of Radiation and Matter

stopped. These observations indicate that when ultraviolet radiations fall on the emitter plate C, electrons are ejected from it which are attracted towards the positive, collector plate A by the electric field. The electrons flow through the evacuated glass tube, resulting in the current flow. Thus, light falling on the surface of the emitter causes current in the external circuit. Hallwachs and Lenard studied how this photo current varied with collector plate potential, and with frequency and intensity of incident light.

Hallwachs, in 1888, undertook the study further and connected a negatively charged zinc plate to an electroscope. He observed that the zinc plate lost its charge when it was illuminated by ultraviolet light. Further, the uncharged zinc plate became positively charged when it was irradiated by ultraviolet light. Positive charge on a positively charged zinc plate was found to be further enhanced when it was illuminated by ultraviolet light. From these observations he concluded that negatively charged particles were emitted from the zinc plate under the action of ultraviolet light.

After the discovery of the electron in 1897, it became evident that the incident light causes electrons to be emitted from the emitter plate. Due to negative charge, the emitted electrons are pushed towards the collector plate by the electric field. Hallwachs and Lenard also observed that when ultraviolet light fell on the emitter plate, no electrons were emitted at all when the frequency of the incident light was smaller than a certain minimum value, called the *threshold frequency*. This minimum frequency depends on the nature of the material of the emitter plate.

It was found that certain metals like zinc, cadmium, magnesium, etc., responded only to ultraviolet light, having short wavelength, to cause electron emission from the surface. However, some alkali metals such as lithium, sodium, potassium, caesium and rubidium were sensitive even to visible light. All these *photosensitive substances* emit electrons when they are illuminated by light. After the discovery of electrons, these electrons were termed as *photoelectrons*. The phenomenon is called *photoelectric effect*.

11.4 EXPERIMENTAL STUDY OF PHOTOELECTRIC EFFECT

Figure 11.1 depicts a schematic view of the arrangement used for the experimental study of the photoelectric effect. It consists of an evacuated glass/quartz tube having a photosensitive plate C and another metal plate A. Monochromatic light from the source S of sufficiently short wavelength passes through the window W and falls on the photosensitive plate C (emitter). A transparent quartz window is sealed on to the glass tube, which permits ultraviolet radiation to pass through it and irradiate the photosensitive plate C. The electrons are emitted by the plate C and are collected by the plate A (collector), by the electric field created by the battery. The battery maintains the potential difference between the plates C and A, that can be varied. The polarity of the plates C and A can be reversed by a commutator. Thus, the plate A can be maintained at a desired positive or negative potential with respect to emitter C. When the collector plate A is positive with respect to the emitter plate C, the electrons are



Simulate experiments on photoelectric effect
<http://www.kcvs.ca/site/projects/physics.html>

Physics

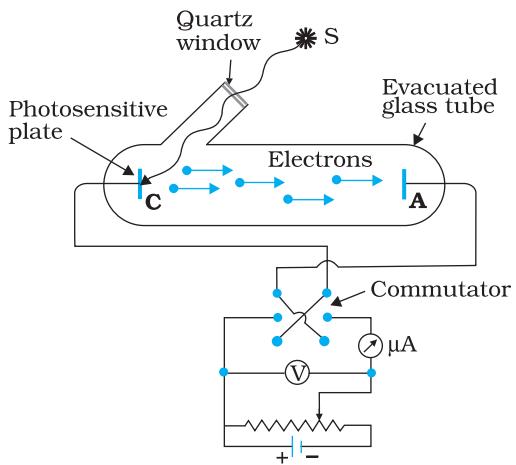


FIGURE 11.1 Experimental arrangement for study of photoelectric effect.

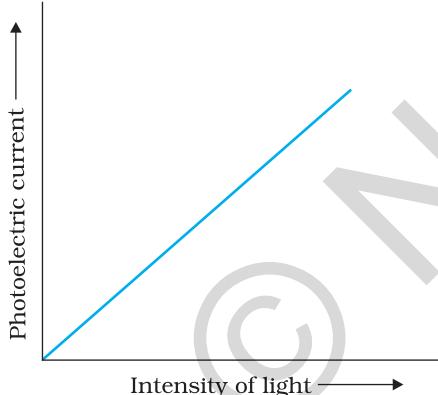


FIGURE 11.2 Variation of Photoelectric current with intensity of light.

attracted to it. The emission of electrons causes flow of electric current in the circuit. The potential difference between the emitter and collector plates is measured by a voltmeter (V) whereas the resulting photo current flowing in the circuit is measured by a microammeter (μA). The photoelectric current can be increased or decreased by varying the potential of collector plate A with respect to the emitter plate C. The intensity and frequency of the incident light can be varied, as can the potential difference V between the emitter C and the collector A.

We can use the experimental arrangement of Fig. 11.1 to study the variation of photocurrent with (a) intensity of radiation, (b) frequency of incident radiation, (c) the potential difference between the plates A and C, and (d) the nature of the material of plate C. Light of different frequencies can be used by putting appropriate coloured filter or coloured glass in the path of light falling on the emitter C. The intensity of light is varied by changing the distance of the light source from the emitter.

11.4.1 Effect of intensity of light on photocurrent

The collector A is maintained at a positive potential with respect to emitter C so that electrons ejected from C are attracted towards collector A. Keeping the frequency of the incident radiation and the accelerating potential fixed, the intensity of light is varied and the resulting photoelectric current is measured each time. It is found that the photocurrent increases linearly with intensity of incident light as shown graphically in Fig. 11.2. The photocurrent is directly proportional to the number of photoelectrons emitted per second. This implies that *the number of photoelectrons emitted per second is directly proportional to the intensity of incident radiation*.

11.4.2 Effect of potential on photoelectric current

We first keep the plate A at some positive accelerating potential with respect to the plate C and illuminate the plate C with light of fixed frequency v and fixed intensity I_1 . We next vary the positive potential of plate A gradually and measure the resulting photocurrent each time. It is found that the photoelectric current increases with increase in accelerating (positive) potential. At some stage, for a certain positive potential of plate A, all the emitted electrons are collected by the plate A and the photoelectric current becomes maximum or saturates. If we increase the accelerating potential of plate A further, the photocurrent does not increase. This maximum value of the photoelectric current is called *saturation current*. Saturation current corresponds to the case when all the photoelectrons emitted by the emitter plate C reach the collector plate A.

We now apply a negative (retarding) potential to the plate A with respect to the plate C and make it increasingly negative gradually. When the

Dual Nature of Radiation and Matter

polarity is reversed, the electrons are repelled and only the most energetic electrons are able to reach the collector A. The photocurrent is found to decrease rapidly until it drops to zero at a certain sharply defined, critical value of the negative potential V_0 on the plate A. For a particular frequency of incident radiation, *the minimum negative (retarding) potential V_0 given to the plate A for which the photocurrent stops or becomes zero is called the cut-off or stopping potential.*

The interpretation of the observation in terms of photoelectrons is straightforward. All the photoelectrons emitted from the metal do not have the same energy. Photoelectric current is zero when the stopping potential is sufficient to repel even the most energetic photoelectrons, with the maximum kinetic energy (K_{\max}), so that

$$K_{\max} = e V_0 \quad (11.1)$$

We can now repeat this experiment with incident radiation of the same frequency but of higher intensity I_2 and I_3 ($I_3 > I_2 > I_1$). We note that the saturation currents are now found to be at higher values. This shows that more electrons are being emitted per second, proportional to the intensity of incident radiation. But the stopping potential remains the same as that for the incident radiation of intensity I_1 , as shown graphically in Fig. 11.3. Thus, *for a given frequency of the incident radiation, the stopping potential is independent of its intensity.* In other words, the maximum kinetic energy of photoelectrons depends on the light source and the emitter plate material, but is independent of intensity of incident radiation.

11.4.3 Effect of frequency of incident radiation on stopping potential

We now study the relation between the frequency ν of the incident radiation and the stopping potential V_0 . We suitably adjust the same intensity of light radiation at various frequencies and study the variation of photocurrent with collector plate potential. The resulting variation is shown in Fig. 11.4. We obtain different values of stopping potential but the same value of the saturation current for incident radiation of different frequencies. The energy of the emitted electrons depends on the frequency of the incident radiations. The stopping potential is more negative for higher frequencies of incident radiation. Note from

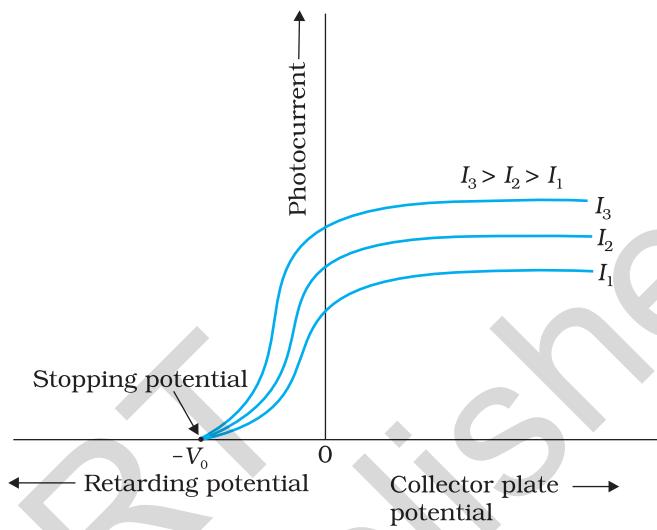


FIGURE 11.3 Variation of photocurrent with collector plate potential for different intensity of incident radiation.

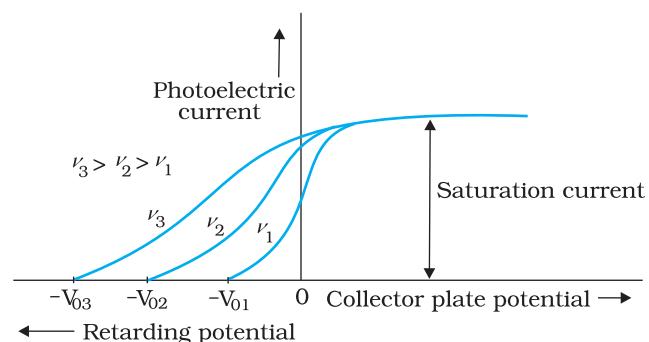


FIGURE 11.4 Variation of photoelectric current with collector plate potential for different frequencies of incident radiation.

Physics

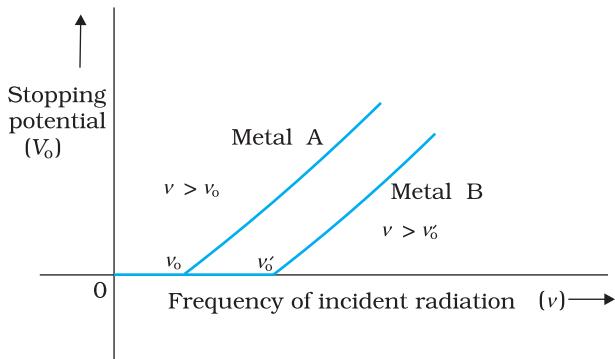


FIGURE 11.5 Variation of stopping potential V_0 with frequency ν of incident radiation for a given photosensitive material.

- (ii) there exists a certain minimum cut-off frequency ν_0 for which the stopping potential is zero.

These observations have two implications:

- (i) *The maximum kinetic energy of the photoelectrons varies linearly with the frequency of incident radiation, but is independent of its intensity.*
- (ii) *For a frequency ν of incident radiation, lower than the cut-off frequency ν_0 , no photoelectric emission is possible even if the intensity is large.*

This minimum, cut-off frequency ν_0 , is called the *threshold frequency*.

It is different for different metals.

Different photosensitive materials respond differently to light. Selenium is more sensitive than zinc or copper. The same photosensitive substance gives different response to light of different wavelengths. For example, ultraviolet light gives rise to photoelectric effect in copper while green or red light does not.

Note that in all the above experiments, it is found that, if frequency of the incident radiation exceeds the threshold frequency, the photoelectric emission starts instantaneously without any apparent time lag, even if the incident radiation is very dim. It is now known that emission starts in a time of the order of 10^{-9} s or less.

We now summarise the experimental features and observations described in this section.

- (i) For a given photosensitive material and frequency of incident radiation (above the threshold frequency), the photoelectric current is directly proportional to the intensity of incident light (Fig. 11.2).
- (ii) For a given photosensitive material and frequency of incident radiation, saturation current is found to be proportional to the intensity of incident radiation whereas the stopping potential is independent of its intensity (Fig. 11.3).
- (iii) For a given photosensitive material, there exists a certain minimum cut-off frequency of the incident radiation, called the *threshold frequency*, below which no emission of photoelectrons takes place, no matter how intense the incident light is. Above the threshold frequency, the stopping potential or equivalently the maximum kinetic

Fig. 11.4 that the stopping potentials are in the order $V_{03} > V_{02} > V_{01}$ if the frequencies are in the order $\nu_3 > \nu_2 > \nu_1$. This implies that greater the frequency of incident light, greater is the maximum kinetic energy of the photoelectrons. Consequently, we need greater retarding potential to stop them completely. If we plot a graph between the frequency of incident radiation and the corresponding stopping potential for different metals we get a straight line, as shown in Fig. 11.5.

The graph shows that

- (i) the stopping potential V_0 varies linearly with the frequency of incident radiation for a given photosensitive material.

energy of the emitted photoelectrons increases linearly with the frequency of the incident radiation, but is independent of its intensity (Fig. 11.5).

- (iv) The photoelectric emission is an instantaneous process without any apparent time lag ($\sim 10^{-9}$ s or less), even when the incident radiation is made exceedingly dim.

11.5 PHOTOELECTRIC EFFECT AND WAVE THEORY OF LIGHT

The wave nature of light was well established by the end of the nineteenth century. The phenomena of interference, diffraction and polarisation were explained in a natural and satisfactory way by the wave picture of light. According to this picture, light is an electromagnetic wave consisting of electric and magnetic fields with continuous distribution of energy over the region of space over which the wave is extended. Let us now see if this wave picture of light can explain the observations on photoelectric emission given in the previous section.

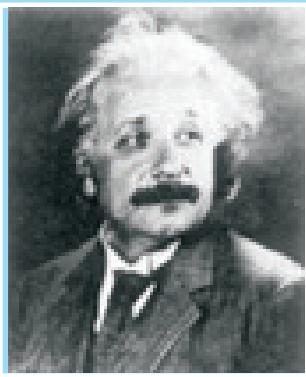
According to the wave picture of light, the free electrons at the surface of the metal (over which the beam of radiation falls) absorb the radiant energy continuously. The greater the intensity of radiation, the greater are the amplitude of electric and magnetic fields. Consequently, the greater the intensity, the greater should be the energy absorbed by each electron. In this picture, the maximum kinetic energy of the photoelectrons on the surface is then expected to increase with increase in intensity. Also, no matter what the frequency of radiation is, a sufficiently intense beam of radiation (over sufficient time) should be able to impart enough energy to the electrons, so that they exceed the minimum energy needed to escape from the metal surface. A threshold frequency, therefore, should not exist. These expectations of the wave theory directly contradict observations (i), (ii) and (iii) given at the end of sub-section 11.4.3.

Further, we should note that in the wave picture, the absorption of energy by electron takes place continuously over the entire wavefront of the radiation. Since a large number of electrons absorb energy, the energy absorbed per electron per unit time turns out to be small. Explicit calculations estimate that it can take hours or more for a single electron to pick up sufficient energy to overcome the work function and come out of the metal. This conclusion is again in striking contrast to observation (iv) that the photoelectric emission is instantaneous. In short, the wave picture is unable to explain the most basic features of photoelectric emission.

11.6 EINSTEIN'S PHOTOELECTRIC EQUATION: ENERGY QUANTUM OF RADIATION

In 1905, Albert Einstein (1879-1955) proposed a radically new picture of electromagnetic radiation to explain photoelectric effect. In this picture, photoelectric emission does not take place by continuous absorption of energy from radiation. Radiation energy is built up of discrete units – the so called *quanta of energy of radiation*. Each quantum of radiant energy

Physics



Albert Einstein (1879 – 1955) Einstein, one of the greatest physicists of all time, was born in Ulm, Germany. In 1905, he published three path-breaking papers. In the first paper, he introduced the notion of light quanta (now called photons) and used it to explain the features of photoelectric effect. In the second paper, he developed a theory of Brownian motion, confirmed experimentally a few years later and provided a convincing evidence of the atomic picture of matter. The third paper gave birth to the special theory of relativity. In 1916, he published the general theory of relativity. Some of Einstein's most significant later contributions are: the notion of stimulated emission introduced in an alternative derivation of Planck's blackbody radiation law, static model of the universe which started modern cosmology, quantum statistics of a gas of massive bosons, and a critical analysis of the foundations of quantum mechanics. In 1921, he was awarded the Nobel Prize in physics for his contribution to theoretical physics and the photoelectric effect.

ALBERT EINSTEIN (1879 – 1955)

has energy hv , where h is Planck's constant and v the frequency of light. In photoelectric effect, an electron absorbs a quantum of energy (hv) of radiation. If this quantum of energy absorbed exceeds the minimum energy needed for the electron to escape from the metal surface (work function ϕ_0), the electron is emitted with maximum kinetic energy

$$K_{\max} = hv - \phi_0 \quad (11.2)$$

More tightly bound electrons will emerge with kinetic energies less than the maximum value. Note that the intensity of light of a given frequency is determined by the number of photons incident per second. Increasing the intensity will increase the number of emitted electrons per second. However, the maximum kinetic energy of the emitted photoelectrons is determined by the energy of each photon.

Equation (11.2) is known as *Einstein's photoelectric equation*. We now see how this equation accounts in a simple and elegant manner all the observations on photoelectric effect given at the end of sub-section 11.4.3.

- According to Eq. (11.2), K_{\max} depends linearly on v , and is independent of intensity of radiation, in agreement with observation. This has happened because in Einstein's picture, photoelectric effect arises from the absorption of a single quantum of radiation by a single electron. The intensity of radiation (that is proportional to the number of energy quanta per unit area per unit time) is irrelevant to this basic process.
- Since K_{\max} must be non-negative, Eq. (11.2) implies that photoelectric emission is possible only if $hv > \phi_0$
or $v > v_0$, where

$$v_0 = \frac{\phi_0}{h} \quad (11.3)$$

Equation (11.3) shows that the greater the work function ϕ_0 , the higher the minimum or threshold frequency v_0 needed to emit photoelectrons. Thus, there exists a threshold frequency $v_0 (= \phi_0/h)$ for the metal surface, below which no photoelectric emission is possible, no matter how intense the incident radiation may be or how long it falls on the surface.

- In this picture, intensity of radiation as noted above, is proportional to the number of energy quanta per unit area per unit time. The greater the number of energy quanta available, the greater is the number of electrons absorbing the energy quanta and greater, therefore, is the number of electrons coming out of the metal (for $v > v_0$). This explains why, for $v > v_0$, photoelectric current is proportional to intensity.

- In Einstein's picture, the basic elementary process involved in photoelectric effect is the absorption of a light quantum by an electron. This process is instantaneous. Thus, whatever may be the intensity i.e., the number of quanta of radiation per unit area per unit time, photoelectric emission is instantaneous. Low intensity does not mean delay in emission, since the basic elementary process is the same. Intensity only determines how many electrons are able to participate in the elementary process (absorption of a light quantum by a single electron) and, therefore, the photoelectric current.

Using Eq. (11.1), the photoelectric equation, Eq. (11.2), can be written as

$$eV_0 = h\nu - \phi_0; \text{ for } \nu \geq \nu_0$$

$$\text{or } V_0 = \frac{h}{e} \nu - \frac{\phi_0}{e} \quad (11.4)$$

This is an important result. It predicts that the V_0 versus ν curve is a straight line with slope $= (h/e)$, independent of the nature of the material. During 1906-1916, Millikan performed a series of experiments on photoelectric effect, aimed at disproving Einstein's photoelectric equation. He measured the slope of the straight line obtained for sodium, similar to that shown in Fig. 11.5. Using the known value of e , he determined the value of Planck's constant h . This value was close to the value of Planck's constant ($= 6.626 \times 10^{-34}$ J s) determined in an entirely different context. In this way, in 1916, Millikan proved the validity of Einstein's photoelectric equation, instead of disproving it.

The successful explanation of photoelectric effect using the hypothesis of light quanta and the experimental determination of values of h and ϕ_0 , in agreement with values obtained from other experiments, led to the acceptance of Einstein's picture of photoelectric effect. Millikan verified photoelectric equation with great precision, for a number of alkali metals over a wide range of radiation frequencies.

11.7 PARTICLE NATURE OF LIGHT: THE PHOTON

Photoelectric effect thus gave evidence to the strange fact that light in interaction with matter behaved as if it was made of quanta or packets of energy, each of energy $h\nu$.

Is the light quantum of energy to be associated with a particle? Einstein arrived at the important result, that the light quantum can also be associated with momentum ($h\nu/c$). A definite value of energy as well as momentum is a strong sign that the light quantum can be associated with a particle. This particle was later named *photon*. The particle-like behaviour of light was further confirmed, in 1924, by the experiment of A.H. Compton (1892-1962) on scattering of X-rays from electrons. In 1921, Einstein was awarded the Nobel Prize in Physics for his contribution to theoretical physics and the photoelectric effect. In 1923, Millikan was awarded the Nobel Prize in physics for his work on the elementary charge of electricity and on the photoelectric effect.

We can summarise the photon picture of electromagnetic radiation as follows:

Physics

- (i) In interaction of radiation with matter, radiation behaves as if it is made up of particles called photons.
- (ii) Each photon has energy $E (=hv)$ and momentum $p (= h v/c)$, and speed c , the speed of light.
- (iii) All photons of light of a particular frequency v , or wavelength λ , have the same energy $E (=hv = hc/\lambda)$ and momentum $p (= hv/c = h/\lambda)$, whatever the intensity of radiation may be. By increasing the intensity of light of given wavelength, there is only an increase in the number of photons per second crossing a given area, with each photon having the same energy. Thus, photon energy is independent of intensity of radiation.
- (iv) Photons are electrically neutral and are not deflected by electric and magnetic fields.
- (v) In a photon-particle collision (such as photon-electron collision), the total energy and total momentum are conserved. However, the number of photons may not be conserved in a collision. The photon may be absorbed or a new photon may be created.

Example 11.1 Monochromatic light of frequency 6.0×10^{14} Hz is produced by a laser. The power emitted is 2.0×10^{-3} W. (a) What is the energy of a photon in the light beam? (b) How many photons per second, on an average, are emitted by the source?

Solution

(a) Each photon has an energy
$$E = h v = (6.63 \times 10^{-34} \text{ J s}) (6.0 \times 10^{14} \text{ Hz})$$
$$= 3.98 \times 10^{-19} \text{ J}$$

(b) If N is the number of photons emitted by the source per second, the power P transmitted in the beam equals N times the energy per photon E , so that $P = N E$. Then

$$N = \frac{P}{E} = \frac{2.0 \times 10^{-3} \text{ W}}{3.98 \times 10^{-19} \text{ J}}$$
$$= 5.0 \times 10^{15} \text{ photons per second.}$$

Example 11.2 The work function of caesium is 2.14 eV. Find (a) the threshold frequency for caesium, and (b) the wavelength of the incident light if the photocurrent is brought to zero by a stopping potential of 0.60 V.

Solution

(a) For the cut-off or threshold frequency, the energy $h v_0$ of the incident radiation must be equal to work function ϕ_0 , so that

$$v_0 = \frac{\phi_0}{h} = \frac{2.14 \text{ eV}}{6.63 \times 10^{-34} \text{ J s}}$$
$$= \frac{2.14 \times 1.6 \times 10^{-19} \text{ J}}{6.63 \times 10^{-34} \text{ J s}} = 5.16 \times 10^{14} \text{ Hz}$$

Thus, for frequencies less than this threshold frequency, no photoelectrons are ejected.

(b) Photocurrent reduces to zero, when maximum kinetic energy of the emitted photoelectrons equals the potential energy $e V_0$ by the retarding potential V_0 . Einstein's Photoelectric equation is

Dual Nature of Radiation and Matter

$$eV_0 = h\nu - \phi_0 = \frac{hc}{\lambda} - \phi_0$$

or, $\lambda = hc/(eV_0 + \phi_0)$

$$= \frac{(6.63 \times 10^{-34} \text{ J s}) \times (3 \times 10^8 \text{ m/s})}{(0.60 \text{ eV} + 2.14 \text{ eV})}$$

$$= \frac{19.89 \times 10^{-26} \text{ J m}}{(2.74 \text{ eV})}$$

$$\lambda = \frac{19.89 \times 10^{-26} \text{ J m}}{2.74 \times 1.6 \times 10^{-19} \text{ J}} = 454 \text{ nm}$$

EXAMPLE 11.2

Example 11.3 The wavelength of light in the visible region is about 390 nm for violet colour, about 550 nm (average wavelength) for yellow-green colour and about 760 nm for red colour.

- What are the energies of photons in (eV) at the (i) violet end, (ii) average wavelength, yellow-green colour, and (iii) red end of the visible spectrum? (Take $h = 6.63 \times 10^{-34} \text{ J s}$ and $1 \text{ eV} = 1.6 \times 10^{-19} \text{ J}$.)
- From which of the photosensitive materials with work functions listed in Table 11.1 and using the results of (i), (ii) and (iii) of (a), can you build a photoelectric device that operates with visible light?

Solution

- (a) Energy of the incident photon, $E = h\nu = hc/\lambda$
 $E = (6.63 \times 10^{-34} \text{ J s}) (3 \times 10^8 \text{ m/s})/\lambda$

$$= \frac{1.989 \times 10^{-25} \text{ J m}}{\lambda}$$

- (i) For violet light, $\lambda_1 = 390 \text{ nm}$ (lower wavelength end)

$$\text{Incident photon energy, } E_1 = \frac{1.989 \times 10^{-25} \text{ J m}}{390 \times 10^{-9} \text{ m}}$$

$$= 5.10 \times 10^{-19} \text{ J}$$

$$= \frac{5.10 \times 10^{-19} \text{ J}}{1.6 \times 10^{-19} \text{ J/eV}}$$

$$= 3.19 \text{ eV}$$

- (ii) For yellow-green light, $\lambda_2 = 550 \text{ nm}$ (average wavelength)

$$\text{Incident photon energy, } E_2 = \frac{1.989 \times 10^{-25} \text{ J m}}{550 \times 10^{-9} \text{ m}}$$

$$= 3.62 \times 10^{-19} \text{ J} = 2.26 \text{ eV}$$

- (iii) For red light, $\lambda_3 = 760 \text{ nm}$ (higher wavelength end)

$$\text{Incident photon energy, } E_3 = \frac{1.989 \times 10^{-25} \text{ J m}}{760 \times 10^{-9} \text{ m}}$$

$$= 2.62 \times 10^{-19} \text{ J} = 1.64 \text{ eV}$$

- (b) For a photoelectric device to operate, we require incident light energy E to be equal to or greater than the work function ϕ_0 of the material. Thus, the photoelectric device will operate with violet light (with $E = 3.19 \text{ eV}$) photosensitive material Na (with $\phi_0 = 2.75 \text{ eV}$), K (with $\phi_0 = 2.30 \text{ eV}$) and Cs (with $\phi_0 = 2.14 \text{ eV}$). It will also operate with yellow-green light (with $E = 2.26 \text{ eV}$) for Cs (with $\phi_0 = 2.14 \text{ eV}$) only. However, it will not operate with red light (with $E = 1.64 \text{ eV}$) for any of these photosensitive materials.

EXAMPLE 11.3

11.8 WAVE NATURE OF MATTER

The dual (wave-particle) nature of light (electromagnetic radiation, in general) comes out clearly from what we have learnt in this and the preceding chapters. The wave nature of light shows up in the phenomena of interference, diffraction and polarisation. On the other hand, in photoelectric effect and Compton effect which involve energy and momentum transfer, radiation behaves as if it is made up of a bunch of particles – the photons. Whether a particle or wave description is best suited for understanding an experiment depends on the nature of the experiment. For example, in the familiar phenomenon of seeing an object by our eye, both descriptions are important. The gathering and focussing mechanism of light by the eye-lens is well described in the wave picture. But its absorption by the rods and cones (of the retina) requires the photon picture of light.

A natural question arises: If radiation has a dual (wave-particle) nature, might not the particles of nature (the electrons, protons, etc.) also exhibit wave-like character? In 1924, the French physicist Louis Victor de Broglie (pronounced as de Broglie) (1892–1987) put forward the bold hypothesis that moving particles of matter should display wave-like properties under suitable conditions. He reasoned that nature was symmetrical and that the two basic physical entities – matter and energy, must have symmetrical character. If radiation shows dual aspects, so should matter. De Broglie proposed that the wave length λ associated with a particle of momentum p is given as

$$\lambda = \frac{h}{p} = \frac{h}{mv} \quad (11.5)$$

where m is the mass of the particle and v its speed. Equation (11.5) is known as the *de Broglie relation* and the wavelength λ of the *matter wave* is called *de Broglie wavelength*. The dual aspect of matter is evident in the de Broglie relation. On the left hand side of Eq. (11.5), λ is the attribute of a wave while on the right hand side the momentum p is a typical attribute of a particle. Planck's constant h relates the two attributes.

Equation (11.5) for a material particle is basically a hypothesis whose validity can be tested only by experiment. However, it is interesting to see that it is satisfied also by a photon. For a photon, as we have seen,

$$p = hv/c \quad (11.6)$$

Therefore,

$$\frac{h}{p} = \frac{c}{v} = \lambda \quad (11.7)$$

That is, the de Broglie wavelength of a photon given by Eq. (11.5) equals the wavelength of electromagnetic radiation of which the photon is a quantum of energy and momentum.

Clearly, from Eq. (11.5), λ is smaller for a heavier particle (large m) or more energetic particle (large v). For example, the de Broglie wavelength of a ball of mass 0.12 kg moving with a speed of 20 m s^{-1} is easily calculated:

Dual Nature of Radiation and Matter

PHOTOCELL

A photocell is a technological application of the photoelectric effect. It is a device whose electrical properties are affected by light. It is also sometimes called an electric eye. A photocell consists of a semi-cylindrical photo-sensitive metal plate C (emitter) and a wire loop A (collector) supported in an evacuated glass or quartz bulb. It is connected to the external circuit having a high-tension battery B and microammeter (μA) as shown in the Figure. Sometimes, instead of the plate C, a thin layer of photosensitive material is pasted on the inside of the bulb. A part of the bulb is left clean for the light to enter it.

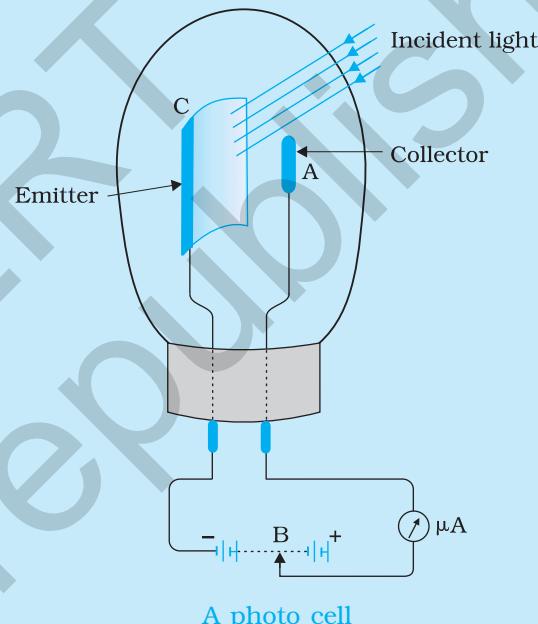
When light of suitable wavelength falls on the emitter C, photoelectrons are emitted. These photoelectrons are drawn to the collector A. Photocurrent of the order of a few microampere can be normally obtained from a photo cell.

A photocell converts a change in intensity of illumination into a change in photocurrent. This current can be used to operate control systems and in light measuring devices. A photocell of lead sulphide sensitive to infrared radiation is used in electronic ignition circuits.

In scientific work, photo cells are used whenever it is necessary to measure the intensity of light. Light meters in photographic cameras make use of photo cells to measure the intensity of incident light. The photocells, inserted in the door light electric circuit, are used as automatic door opener. A person approaching a doorway may interrupt a light beam which is incident on a photocell. The abrupt change in photocurrent may be used to start a motor which opens the door or rings an alarm. They are used in the control of a counting device which records every interruption of the light beam caused by a person or object passing across the beam. So photocells help count the persons entering an auditorium, provided they enter the hall one by one. They are used for detection of traffic law defaulters: an alarm may be sounded whenever a beam of (*invisible*) radiation is intercepted.

In burglar alarm, (*invisible*) ultraviolet light is continuously made to fall on a photocell installed at the doorway. A person entering the door interrupts the beam falling on the photocell. The abrupt change in photocurrent is used to start an electric bell ringing. In fire alarm, a number of photocells are installed at suitable places in a building. In the event of breaking out of fire, light radiations fall upon the photocell. This completes the electric circuit through an electric bell or a siren which starts operating as a warning signal.

Photocells are used in the reproduction of sound in motion pictures and in the television camera for scanning and telecasting scenes. They are used in industries for detecting minor flaws or holes in metal sheets.



$$p = m v = 0.12 \text{ kg} \times 20 \text{ m s}^{-1} = 2.40 \text{ kg m s}^{-1}$$

$$\lambda = \frac{h}{p} = \frac{6.63 \times 10^{-34} \text{ J s}}{2.40 \text{ kg m s}^{-1}} = 2.76 \times 10^{-34} \text{ m}$$

Physics

LOUIS VICTOR DE BROGLIE (1892 – 1987)



Louis Victor de Broglie (1892 – 1987) French physicist who put forth revolutionary idea of wave nature of matter. This idea was developed by Erwin Schrödinger into a full-fledged theory of quantum mechanics commonly known as wave mechanics. In 1929, he was awarded the Nobel Prize in Physics for his discovery of the wave nature of electrons.

This wavelength is so small that it is beyond any measurement. This is the reason why macroscopic objects in our daily life do not show wave-like properties. On the other hand, in the sub-atomic domain, the wave character of particles is significant and measurable.

Consider an electron (mass m , charge e) accelerated from rest through a potential V . The kinetic energy K of the electron equals the work done (eV) on it by the electric field:

$$K = eV \quad (11.8)$$

$$\text{Now, } K = \frac{1}{2} m v^2 = \frac{p^2}{2m}, \text{ so that}$$

$$p = \sqrt{2mK} = \sqrt{2meV} \quad (11.9)$$

The de Broglie wavelength λ of the electron is then

$$\lambda = \frac{h}{p} = \frac{h}{\sqrt{2mK}} = \frac{h}{\sqrt{2meV}} \quad (11.10)$$

Substituting the numerical values of h , m , e , we get

$$\lambda = \frac{1.227}{\sqrt{V}} \text{ nm} \quad (11.11)$$

where V is the magnitude of accelerating potential in volts. For a 120 V accelerating potential, Eq. (11.11) gives $\lambda = 0.112$ nm. This wavelength is of the same order as the spacing between the atomic planes in crystals. This

suggests that matter waves associated with an electron could be verified by crystal diffraction experiments analogous to X-ray diffraction. We describe the experimental verification of the de Broglie hypothesis in the next section. In 1929, de Broglie was awarded the Nobel Prize in Physics for his discovery of the wave nature of electrons.

The matter-wave picture elegantly incorporated the Heisenberg's *uncertainty principle*. According to the principle, it is not possible to measure *both* the position and momentum of an electron (or any other particle) *at the same time* exactly. There is always some uncertainty (Δx) in the specification of position and some uncertainty (Δp) in the specification of momentum. The product of Δx and Δp is of the order of \hbar^* (with $\hbar = h/2\pi$), i.e.,

$$\Delta x \Delta p \approx \hbar \quad (11.12)$$

Equation (11.12) allows the possibility that Δx is zero; but then Δp must be infinite in order that the product is non-zero. Similarly, if Δp is zero, Δx must be infinite. Ordinarily, both Δx and Δp are non-zero such that their product is of the order of \hbar .

Now, if an electron has a definite momentum p , (i.e. $\Delta p = 0$), by the de Broglie relation, it has a definite wavelength λ . A wave of definite (single)

Dual Nature of Radiation and Matter

wavelength extends all over space. By Born's probability interpretation this means that the electron is not localised in any finite region of space. That is, its position uncertainty is infinite ($\Delta x \rightarrow \infty$), which is consistent with the uncertainty principle.

In general, the matter wave associated with the electron is not extended all over space. It is a wave packet extending over some finite region of space. In that case Δx is not infinite but has some finite value depending on the extension of the wave packet. Also, you must appreciate that a wave packet of finite extension does not have a single wavelength. It is built up of wavelengths spread around some central wavelength.

By de Broglie's relation, then, the momentum of the electron will also have a spread – an uncertainty Δp . This is as expected from the uncertainty principle. It can be shown that the wave packet description together with de Broglie relation and Born's probability interpretation reproduce the Heisenberg's uncertainty principle exactly.

In Chapter 12, the de Broglie relation will be seen to justify Bohr's postulate on quantisation of angular momentum of electron in an atom.

Figure 11.6 shows a schematic diagram of (a) a localised wave packet, and (b) an extended wave with fixed wavelength.

Example 11.4 What is the de Broglie wavelength associated with (a) an electron moving with a speed of 5.4×10^6 m/s, and (b) a ball of mass 150 g travelling at 30.0 m/s?

Solution

(a) For the electron:

Mass $m = 9.11 \times 10^{-31}$ kg, speed $v = 5.4 \times 10^6$ m/s. Then, momentum

$$p = m v = 9.11 \times 10^{-31} \text{ (kg)} \times 5.4 \times 10^6 \text{ (m/s)}$$

$$p = 4.92 \times 10^{-24} \text{ kg m/s}$$

de Broglie wavelength, $\lambda = h/p$

$$= \frac{6.63 \times 10^{-34} \text{ Js}}{4.92 \times 10^{-24} \text{ kg m/s}}$$

$$\lambda = 0.135 \text{ nm}$$

(b) For the ball:

Mass $m' = 0.150$ kg, speed $v' = 30.0$ m/s.

Then momentum $p' = m' v' = 0.150 \text{ (kg)} \times 30.0 \text{ (m/s)}$

$$p' = 4.50 \text{ kg m/s}$$

de Broglie wavelength $\lambda' = h/p'$.

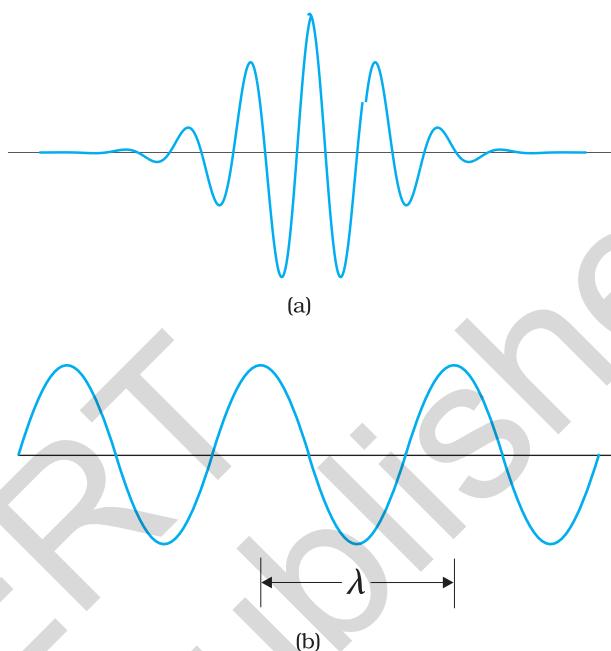


FIGURE 11.6 (a) The wave packet description of an electron. The wave packet corresponds to a spread of wavelength around some central wavelength (and hence by de Broglie relation, a spread in momentum). Consequently, it is associated with an uncertainty in position (Δx) and an uncertainty in momentum (Δp).

(b) The matter wave corresponding to a definite momentum of an electron extends all over space. In this case, $\Delta p = 0$ and $\Delta x \rightarrow \infty$.

Physics

EXAMPLE 11.4

$$= \frac{6.63 \times 10^{-34} \text{ Js}}{4.50 \times \text{kg m/s}}$$

$$\lambda' = 1.47 \times 10^{-34} \text{ m}$$

The de Broglie wavelength of electron is comparable with X-ray wavelengths. However, for the ball it is about 10^{-19} times the size of the proton, quite beyond experimental measurement.

EXAMPLE 11.5

Example 11.5 An electron, an α -particle, and a proton have the same kinetic energy. Which of these particles has the shortest de Broglie wavelength?

Solution

For a particle, de Broglie wavelength, $\lambda = h/p$

$$\text{Kinetic energy, } K = p^2/2m$$

$$\text{Then, } \lambda = h/\sqrt{2mK}$$

For the same kinetic energy K , the de Broglie wavelength associated with the particle is inversely proportional to the square root of their masses. A proton (${}^1_1\text{H}$) is 1836 times massive than an electron and an α -particle (${}^4_2\text{He}$) four times that of a proton.

Hence, α - particle has the shortest de Broglie wavelength.

PROBABILITY INTERPRETATION TO MATTER WAVES

It is worth pausing here to reflect on just what a matter wave associated with a particle, say, an electron, means. Actually, a truly satisfactory physical understanding of the dual nature of matter and radiation has not emerged so far. The great founders of quantum mechanics (Niels Bohr, Albert Einstein, and many others) struggled with this and related concepts for long. Still the deep physical interpretation of quantum mechanics continues to be an area of active research. Despite this, the concept of matter wave has been mathematically introduced in modern quantum mechanics with great success. An important milestone in this connection was when Max Born (1882-1970) suggested a probability interpretation to the matter wave amplitude. According to this, the intensity (square of the amplitude) of the matter wave at a point determines the probability density of the particle at that point. Probability density means probability per unit volume. Thus, if A is the amplitude of the wave at a point, $|A|^2 \Delta V$ is the probability of the particle being found in a small volume ΔV around that point. Thus, if the intensity of matter wave is large in a certain region, there is a greater probability of the particle being found there than where the intensity is small.

EXAMPLE 11.6

Example 11.6 A particle is moving three times as fast as an electron. The ratio of the de Broglie wavelength of the particle to that of the electron is 1.813×10^{-4} . Calculate the particle's mass and identify the particle.

Solution

de Broglie wavelength of a moving particle, having mass m and velocity v :

$$\lambda = \frac{h}{p} = \frac{h}{mv}$$

Mass, $m = h/\lambda v$

For an electron, mass $m_e = h/\lambda_e v_e$

Now, we have $v/v_e = 3$ and

$$\lambda/\lambda_e = 1.813 \times 10^{-4}$$

Then, mass of the particle, $m = m_e \frac{\lambda_e}{\lambda} \frac{v_e}{v}$

$$m = (9.11 \times 10^{-31} \text{ kg}) \times (1/3) \times (1/1.813 \times 10^{-4})$$

$$m = 1.675 \times 10^{-27} \text{ kg.}$$

Thus, the particle, with this mass could be a proton or a neutron.

EXAMPLE 11.6

Example 11.7 What is the de Broglie wavelength associated with an electron, accelerated through a potential difference of 100 volts?

Solution Accelerating potential $V = 100 \text{ V}$. The de Broglie wavelength λ is

$$\lambda = h/p = \frac{1.227}{\sqrt{V}} \text{ nm}$$

$$\lambda = \frac{1.227}{\sqrt{100}} \text{ nm} = 0.123 \text{ nm}$$

The de Broglie wavelength associated with an electron in this case is of the order of X-ray wavelengths.

EXAMPLE 11.7

11.9 DAVISSON AND GERMER EXPERIMENT

The wave nature of electrons was first experimentally verified by C.J. Davisson and L.H. Germer in 1927 and independently by G.P. Thomson, in 1928, who observed diffraction effects with beams of electrons scattered by crystals. Davisson and Thomson shared the Nobel Prize in 1937 for their experimental discovery of diffraction of electrons by crystals.

The experimental arrangement used by Davisson and Germer is schematically shown in Fig. 11.7. It consists of an electron gun which comprises of a tungsten filament F, coated with barium oxide and heated by a low voltage power supply (L.T. or battery). Electrons emitted by the filament are accelerated to a desired velocity

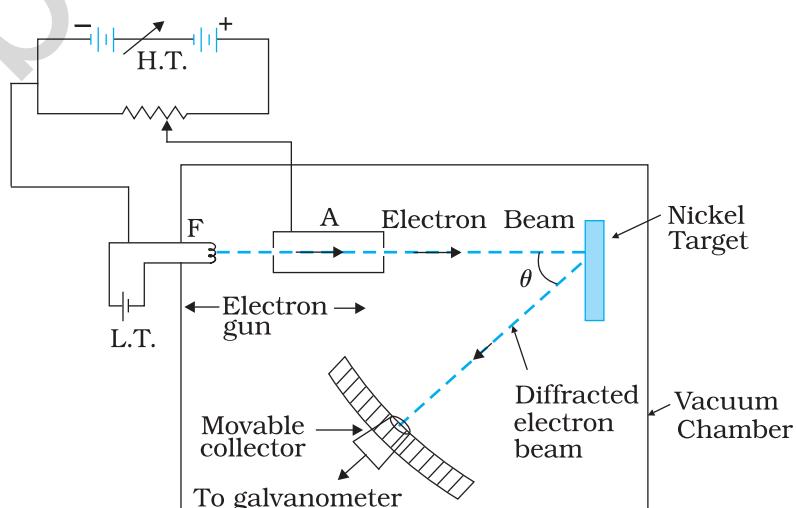


FIGURE 11.7 Davisson-Germer electron diffraction arrangement.

Physics

by applying suitable potential/voltage from a high voltage power supply (H.T. or battery). They are made to pass through a cylinder with fine holes along its axis, producing a fine collimated beam. The beam is made to fall on the surface of a nickel crystal. The electrons are scattered in all directions by the atoms of the crystal. The intensity of the electron beam, scattered in a given direction, is measured by the electron detector (collector). The detector can be moved on a circular scale and is connected to a sensitive galvanometer, which records the current. The deflection of the galvanometer is proportional to the intensity of the electron beam entering the collector. The apparatus is enclosed in an evacuated chamber. By moving the detector on the circular scale at different positions, the intensity of the scattered electron beam is measured for different values of angle of scattering θ which is the angle between the incident and the scattered electron beams. The variation of the intensity (I) of the scattered electrons with the angle of scattering θ is obtained for different accelerating voltages.

The experiment was performed by varying the accelerating voltage from 44 V to 68 V. It was noticed that a strong peak appeared in the intensity (I) of the scattered electron for an accelerating voltage of 54V at a scattering angle $\theta = 50^\circ$

The appearance of the peak in a particular direction is due to the constructive interference of electrons scattered from different layers of the regularly spaced atoms of the crystals. From the electron diffraction measurements, the wavelength of matter waves was found to be 0.165 nm.

The de Broglie wavelength λ associated with electrons, using Eq. (11.11), for $V = 54$ V is given by

$$\lambda = h/p = \frac{1.227}{\sqrt{V}} \text{ nm}$$

$$\lambda = \frac{1.227}{\sqrt{54}} \text{ nm} = 0.167 \text{ nm}$$

Thus, there is an excellent agreement between the theoretical value and the experimentally obtained value of de Broglie wavelength. Davisson-Germer experiment thus strikingly confirms the wave nature of electrons and the de Broglie relation. More recently, in 1989, the wave nature of a beam of electrons was experimentally demonstrated in a double-slit experiment, similar to that used for the wave nature of light. Also, in an experiment in 1994, interference fringes were obtained with the beams of iodine molecules, which are about a million times more massive than electrons.

The de Broglie hypothesis has been basic to the development of modern quantum mechanics. It has also led to the field of electron optics. The wave properties of electrons have been utilised in the design of electron microscope which is a great improvement, with higher resolution, over the optical microscope.

Dual Nature of Radiation and Matter

SUMMARY

1. The minimum energy needed by an electron to come out from a metal surface is called the work function of the metal. Energy (greater than the work function (ϕ_0) required for electron emission from the metal surface can be supplied by suitably heating or applying strong electric field or irradiating it by light of suitable frequency.
2. Photoelectric effect is the phenomenon of emission of electrons by metals when illuminated by light of suitable frequency. Certain metals respond to ultraviolet light while others are sensitive even to the visible light. Photoelectric effect involves conversion of light energy into electrical energy. It follows the law of conservation of energy. The photoelectric emission is an instantaneous process and possesses certain special features.
3. Photoelectric current depends on (i) the intensity of incident light, (ii) the potential difference applied between the two electrodes, and (iii) the nature of the emitter material.
4. The stopping potential (V_0) depends on (i) the frequency of incident light, and (ii) the nature of the emitter material. For a given frequency of incident light, it is independent of its intensity. The stopping potential is directly related to the maximum kinetic energy of electrons emitted: $e V_0 = (1/2) m v_{max}^2 = K_{max}$.
5. Below a certain frequency (threshold frequency) v_0 , characteristic of the metal, no photoelectric emission takes place, no matter how large the intensity may be.
6. The classical wave theory could not explain the main features of photoelectric effect. Its picture of continuous absorption of energy from radiation could not explain the independence of K_{max} on intensity, the existence of v_0 and the instantaneous nature of the process. Einstein explained these features on the basis of photon picture of light. According to this, light is composed of discrete packets of energy called quanta or photons. Each photon carries an energy $E (= h\nu)$ and momentum $p (= h/\lambda)$, which depend on the frequency (ν) of incident light and not on its intensity. Photoelectric emission from the metal surface occurs due to absorption of a photon by an electron.
7. Einstein's photoelectric equation is in accordance with the energy conservation law as applied to the photon absorption by an electron in the metal. The maximum kinetic energy $(1/2)m v_{max}^2$ is equal to the photon energy ($h\nu$) minus the work function $\phi_0 (= h\nu_0)$ of the target metal:

$$\frac{1}{2} m v_{max}^2 = V_0 e = h\nu - \phi_0 = h (\nu - \nu_0)$$

This photoelectric equation explains all the features of the photoelectric effect. Millikan's first precise measurements confirmed the Einstein's photoelectric equation and obtained an accurate value of Planck's constant h . This led to the acceptance of particle or photon description (nature) of electromagnetic radiation, introduced by Einstein.

8. Radiation has dual nature: wave and particle. The nature of experiment determines whether a wave or particle description is best suited for understanding the experimental result. Reasoning that radiation and matter should be symmetrical in nature, Louis Victor de Broglie

Physics

attributed a wave-like character to matter (material particles). The waves associated with the moving material particles are called matter waves or de Broglie waves.

9. The de Broglie wavelength (λ) associated with a moving particle is related to its momentum p as: $\lambda = h/p$. The dualism of matter is inherent in the de Broglie relation which contains a wave concept (λ) and a particle concept (p). The de Broglie wavelength is independent of the charge and nature of the material particle. It is significantly measurable (of the order of the atomic-planes spacing in crystals) only in case of sub-atomic particles like electrons, protons, etc. (due to smallness of their masses and hence, momenta). However, it is indeed very small, quite beyond measurement, in case of macroscopic objects, commonly encountered in everyday life.
10. Electron diffraction experiments by Davisson and Germer, and by G. P. Thomson, as well as many later experiments, have verified and confirmed the wave-nature of electrons. The de Broglie hypothesis of matter waves supports the Bohr's concept of stationary orbits.

Physical Quantity	Symbol	Dimensions	Unit	Remarks
Planck's constant	h	$[ML^2T^{-1}]$	J s	$E = hv$
Stopping potential	V_0	$[ML^2T^{-3}A^{-1}]$	V	$eV_0 = K_{\max}$
Work function	ϕ_0	$[ML^2T^{-2}]$	J; eV	$K_{\max} = E - \phi_0$
Threshold frequency	v_0	$[T^{-1}]$	Hz	$v_0 = \phi_0/h$
de Broglie wavelength	λ	[L]	m	$\lambda = h/p$

POINTS TO PONDER

1. Free electrons in a metal are free in the sense that they move inside the metal in a constant potential (This is only an approximation). They are not free to move out of the metal. They need additional energy to get out of the metal.
2. Free electrons in a metal do not all have the same energy. Like molecules in a gas jar, the electrons have a certain energy distribution at a given temperature. This distribution is different from the usual Maxwell's distribution that you have learnt in the study of kinetic theory of gases. You will learn about it in later courses, but the difference has to do with the fact that electrons obey Pauli's exclusion principle.
3. Because of the energy distribution of free electrons in a metal, the energy required by an electron to come out of the metal is different for different electrons. Electrons with higher energy require less additional energy to come out of the metal than those with lower energies. Work function is the least energy required by an electron to come out of the metal.

4. Observations on photoelectric effect imply that in the event of matter-light interaction, *absorption of energy takes place in discrete units of $h\nu$.* This is not quite the same as saying that light consists of particles, each of energy $h\nu$.
5. Observations on the stopping potential (its independence of intensity and dependence on frequency) are the crucial discriminator between the wave-picture and photon-picture of photoelectric effect.
6. The wavelength of a matter wave given by $\lambda = \frac{h}{p}$ has physical significance; its phase velocity v_p has no physical significance. However, the group velocity of the matter wave is physically meaningful and equals the velocity of the particle.

EXERCISES

- 11.1** Find the
(a) maximum frequency, and
(b) minimum wavelength of X-rays produced by 30 kV electrons.
- 11.2** The work function of caesium metal is 2.14 eV. When light of frequency $6 \times 10^{14} \text{ Hz}$ is incident on the metal surface, photoemission of electrons occurs. What is the
(a) maximum kinetic energy of the emitted electrons,
(b) Stopping potential, and
(c) maximum speed of the emitted photoelectrons?
- 11.3** The photoelectric cut-off voltage in a certain experiment is 1.5 V. What is the maximum kinetic energy of photoelectrons emitted?
- 11.4** Monochromatic light of wavelength 632.8 nm is produced by a helium-neon laser. The power emitted is 9.42 mW.
(a) Find the energy and momentum of each photon in the light beam,
(b) How many photons per second, on the average, arrive at a target irradiated by this beam? (Assume the beam to have uniform cross-section which is less than the target area), and
(c) How fast does a hydrogen atom have to travel in order to have the same momentum as that of the photon?
- 11.5** The energy flux of sunlight reaching the surface of the earth is $1.388 \times 10^3 \text{ W/m}^2$. How many photons (nearly) per square metre are incident on the Earth per second? Assume that the photons in the sunlight have an average wavelength of 550 nm.
- 11.6** In an experiment on photoelectric effect, the slope of the cut-off voltage versus frequency of incident light is found to be $4.12 \times 10^{-15} \text{ V s}$. Calculate the value of Planck's constant.
- 11.7** A 100W sodium lamp radiates energy uniformly in all directions. The lamp is located at the centre of a large sphere that absorbs all the sodium light which is incident on it. The wavelength of the sodium light is 589 nm. (a) What is the energy per photon associated

Physics

with the sodium light? (b) At what rate are the photons delivered to the sphere?

- 11.8** The threshold frequency for a certain metal is 3.3×10^{14} Hz. If light of frequency 8.2×10^{14} Hz is incident on the metal, predict the cut-off voltage for the photoelectric emission.
- 11.9** The work function for a certain metal is 4.2 eV. Will this metal give photoelectric emission for incident radiation of wavelength 330 nm?
- 11.10** Light of frequency 7.21×10^{14} Hz is incident on a metal surface. Electrons with a maximum speed of 6.0×10^5 m/s are ejected from the surface. What is the threshold frequency for photoemission of electrons?
- 11.11** Light of wavelength 488 nm is produced by an argon laser which is used in the photoelectric effect. When light from this spectral line is incident on the emitter, the stopping (cut-off) potential of photoelectrons is 0.38 V. Find the work function of the material from which the emitter is made.
- 11.12** Calculate the
(a) momentum, and
(b) de Broglie wavelength of the electrons accelerated through a potential difference of 56 V.
- 11.13** What is the
(a) momentum,
(b) speed, and
(c) de Broglie wavelength of an electron with kinetic energy of 120 eV.
- 11.14** The wavelength of light from the spectral emission line of sodium is 589 nm. Find the kinetic energy at which
(a) an electron, and
(b) a neutron, would have the same de Broglie wavelength.
- 11.15** What is the de Broglie wavelength of
(a) a bullet of mass 0.040 kg travelling at the speed of 1.0 km/s,
(b) a ball of mass 0.060 kg moving at a speed of 1.0 m/s, and
(c) a dust particle of mass 1.0×10^{-9} kg drifting with a speed of 2.2 m/s?
- 11.16** An electron and a photon each have a wavelength of 1.00 nm. Find
(a) their momenta,
(b) the energy of the photon, and
(c) the kinetic energy of electron.
- 11.17** (a) For what kinetic energy of a neutron will the associated de Broglie wavelength be 1.40×10^{-10} m?
(b) Also find the de Broglie wavelength of a neutron, in thermal equilibrium with matter, having an average kinetic energy of $(3/2) kT$ at 300 K.
- 11.18** Show that the wavelength of electromagnetic radiation is equal to the de Broglie wavelength of its quantum (photon).
- 11.19** What is the de Broglie wavelength of a nitrogen molecule in air at 300 K? Assume that the molecule is moving with the root-mean-square speed of molecules at this temperature. (Atomic mass of nitrogen = 14.0076 u)

ADDITIONAL EXERCISES

- 11.20** (a) Estimate the speed with which electrons emitted from a heated emitter of an evacuated tube impinge on the collector maintained at a potential difference of 500 V with respect to the emitter. Ignore the small initial speeds of the electrons. The *specific charge* of the electron, i.e., its e/m is given to be $1.76 \times 10^{11} \text{ C kg}^{-1}$.
- (b) Use the same formula you employ in (a) to obtain electron speed for an collector potential of 10 MV. Do you see what is wrong? In what way is the formula to be modified?
- 11.21** (a) A monoenergetic electron beam with electron speed of $5.20 \times 10^6 \text{ m s}^{-1}$ is subject to a magnetic field of $1.30 \times 10^{-4} \text{ T}$ normal to the beam velocity. What is the radius of the circle traced by the beam, given e/m for electron equals $1.76 \times 10^{11} \text{ C kg}^{-1}$.
- (b) Is the formula you employ in (a) valid for calculating radius of the path of a 20 MeV electron beam? If not, in what way is it modified?
- [**Note:** Exercises 11.20(b) and 11.21(b) take you to relativistic mechanics which is beyond the scope of this book. They have been inserted here simply to emphasise the point that the formulas you use in part (a) of the exercises are not valid at very high speeds or energies. See answers at the end to know what ‘very high speed or energy’ means.]
- 11.22** An electron gun with its collector at a potential of 100 V fires out electrons in a spherical bulb containing hydrogen gas at low pressure ($\sim 10^{-2} \text{ mm of Hg}$). A magnetic field of $2.83 \times 10^{-4} \text{ T}$ curves the path of the electrons in a circular orbit of radius 12.0 cm. (The path can be viewed because the gas ions in the path focus the beam by attracting electrons, and emitting light by electron capture; this method is known as the ‘fine beam tube’ method.) Determine e/m from the data.
- 11.23** (a) An X-ray tube produces a continuous spectrum of radiation with its short wavelength end at 0.45 \AA . What is the maximum energy of a photon in the radiation?
- (b) From your answer to (a), guess what order of accelerating voltage (for electrons) is required in such a tube?
- 11.24** In an accelerator experiment on high-energy collisions of electrons with positrons, a certain event is interpreted as annihilation of an electron-positron pair of total energy 10.2 BeV into two γ -rays of equal energy. What is the wavelength associated with each γ -ray? (1BeV = 10^9 eV)
- 11.25** Estimating the following two numbers should be interesting. The first number will tell you why radio engineers do not need to worry much about photons! The second number tells you why our eye can never ‘count photons’, even in barely detectable light.
- (a) The number of photons emitted per second by a Medium wave transmitter of 10 kW power, emitting radiowaves of wavelength 500 m.
- (b) The number of photons entering the pupil of our eye per second corresponding to the minimum intensity of white light that we

Physics

humans can perceive ($\sim 10^{-10} \text{ W m}^{-2}$). Take the area of the pupil to be about 0.4 cm^2 , and the average frequency of white light to be about $6 \times 10^{14} \text{ Hz}$.

- 11.26** Ultraviolet light of wavelength 2271 \AA from a 100 W mercury source irradiates a photo-cell made of molybdenum metal. If the stopping potential is -1.3 V , estimate the work function of the metal. How would the photo-cell respond to a high intensity ($\sim 10^5 \text{ W m}^{-2}$) red light of wavelength 6328 \AA produced by a He-Ne laser?

- 11.27** Monochromatic radiation of wavelength 640.2 nm ($1\text{nm} = 10^{-9} \text{ m}$) from a neon lamp irradiates photosensitive material made of caesium on tungsten. The stopping voltage is measured to be 0.54 V . The source is replaced by an iron source and its 427.2 nm line irradiates the same photo-cell. Predict the new stopping voltage.

- 11.28** A mercury lamp is a convenient source for studying frequency dependence of photoelectric emission, since it gives a number of spectral lines ranging from the UV to the red end of the visible spectrum. In our experiment with rubidium photo-cell, the following lines from a mercury source were used:

$$\lambda_1 = 3650 \text{ \AA}, \lambda_2 = 4047 \text{ \AA}, \lambda_3 = 4358 \text{ \AA}, \lambda_4 = 5461 \text{ \AA}, \lambda_5 = 6907 \text{ \AA},$$

The stopping voltages, respectively, were measured to be:

$$V_{01} = 1.28 \text{ V}, V_{02} = 0.95 \text{ V}, V_{03} = 0.74 \text{ V}, V_{04} = 0.16 \text{ V}, V_{05} = 0 \text{ V}$$

Determine the value of Planck's constant h , the threshold frequency and work function for the material.

[Note: You will notice that to get h from the data, you will need to know e (which you can take to be $1.6 \times 10^{-19} \text{ C}$). Experiments of this kind on Na, Li, K, etc. were performed by Millikan, who, using his own value of e (from the oil-drop experiment) confirmed Einstein's photoelectric equation and at the same time gave an independent estimate of the value of h .]

- 11.29** The work function for the following metals is given:

Na: 2.75 eV ; K: 2.30 eV ; Mo: 4.17 eV ; Ni: 5.15 eV . Which of these metals will not give photoelectric emission for a radiation of wavelength 3300 \AA from a He-Cd laser placed 1 m away from the photocell? What happens if the laser is brought nearer and placed 50 cm away?

- 11.30** Light of intensity 10^{-5} W m^{-2} falls on a sodium photo-cell of surface area 2 cm^2 . Assuming that the top 5 layers of sodium absorb the incident energy, estimate time required for photoelectric emission in the wave-picture of radiation. The work function for the metal is given to be about 2 eV . What is the implication of your answer?

- 11.31** Crystal diffraction experiments can be performed using X-rays, or electrons accelerated through appropriate voltage. Which probe has greater energy? (For quantitative comparison, take the wavelength of the probe equal to 1 \AA , which is of the order of inter-atomic spacing in the lattice) ($m_e = 9.11 \times 10^{-31} \text{ kg}$).

- 11.32** (a) Obtain the de Broglie wavelength of a neutron of kinetic energy 150 eV . As you have seen in Exercise 11.31, an electron beam of this energy is suitable for crystal diffraction experiments. Would a neutron beam of the same energy be equally suitable? Explain. ($m_n = 1.675 \times 10^{-27} \text{ kg}$)

Dual Nature of Radiation and Matter

- (b) Obtain the de Broglie wavelength associated with thermal neutrons at room temperature (27°C). Hence explain why a fast neutron beam needs to be thermalised with the environment before it can be used for neutron diffraction experiments.
- 11.33** An electron microscope uses electrons accelerated by a voltage of 50 kV. Determine the de Broglie wavelength associated with the electrons. If other factors (such as numerical aperture, etc.) are taken to be roughly the same, how does the resolving power of an electron microscope compare with that of an optical microscope which uses yellow light?
- 11.34** The wavelength of a probe is roughly a measure of the size of a structure that it can probe in some detail. The quark structure of protons and neutrons appears at the minute length-scale of 10^{-15} m or less. This structure was first probed in early 1970's using high energy electron beams produced by a linear accelerator at Stanford, USA. Guess what might have been the order of energy of these electron beams. (Rest mass energy of electron = 0.511 MeV.)
- 11.35** Find the typical de Broglie wavelength associated with a He atom in helium gas at room temperature (27°C) and 1 atm pressure; and compare it with the mean separation between two atoms under these conditions.
- 11.36** Compute the typical de Broglie wavelength of an electron in a metal at 27°C and compare it with the mean separation between two electrons in a metal which is given to be about $2 \times 10^{-10}\text{ m}$.
- [Note: Exercises 11.35 and 11.36 reveal that while the wave-packets associated with gaseous molecules under ordinary conditions are non-overlapping, the electron wave-packets in a metal strongly overlap with one another. This suggests that whereas molecules in an ordinary gas can be distinguished apart, electrons in a metal cannot be distinguished apart from one another. This indistinguishability has many fundamental implications which you will explore in more advanced Physics courses.]
- 11.37** Answer the following questions:
- Quarks inside protons and neutrons are thought to carry fractional charges $\{(+2/3)e ; (-1/3)e\}$. Why do they not show up in Millikan's oil-drop experiment?
 - What is so special about the combination e/m ? Why do we not simply talk of e and m separately?
 - Why should gases be insulators at ordinary pressures and start conducting at very low pressures?
 - Every metal has a definite work function. Why do all photoelectrons not come out with the same energy if incident radiation is monochromatic? Why is there an energy distribution of photoelectrons?
 - The energy and momentum of an electron are related to the frequency and wavelength of the associated matter wave by the relations:

$$E = h\nu, p = \frac{h}{\lambda}$$

But while the value of λ is physically significant, the value of ν (and therefore, the value of the phase speed $v\lambda$) has no physical significance. Why?

APPENDIX

11.1 The history of wave-particle flip-flop

What is light? This question has haunted mankind for a long time. But systematic experiments were done by scientists since the dawn of the scientific and industrial era, about four centuries ago. Around the same time, theoretical models about what light is made of were developed. While building a model in any branch of science, it is essential to see that it is able to explain all the experimental observations existing at that time. It is therefore appropriate to summarize some observations about light that were known in the seventeenth century.

The properties of light known at that time included (a) rectilinear propagation of light, (b) reflection from plane and curved surfaces, (c) refraction at the boundary of two media, (d) dispersion into various colours, (e) high speed. Appropriate laws were formulated for the first four phenomena. For example, Snell formulated his laws of refraction in 1621. Several scientists right from the days of Galileo had tried to measure the speed of light. But they had not been able to do so. They had only concluded that it was higher than the limit of their measurement.

Two models of light were also proposed in the seventeenth century. Descartes, in early decades of seventeenth century, proposed that light consists of particles, while Huygens, around 1650-60, proposed that light consists of waves. Descartes' proposal was merely a philosophical model, devoid of any experiments or scientific arguments. Newton soon after, around 1660-70, extended Descartes' particle model, known as *corpuscular theory*, built it up as a scientific theory, and explained various known properties with it. These models, light as waves and as particles, in a sense, are quite opposite of each other. But both models could explain all the known properties of light. There was nothing to choose between them.

The history of the development of these models over the next few centuries is interesting. Bartholinus, in 1669, discovered double refraction of light in some crystals, and Huygens, in 1678, was quick to explain it on the basis of his wave theory of light. In spite of this, for over one hundred years, Newton's particle model was firmly believed and preferred over the wave model. This was partly because of its simplicity and partly because of Newton's influence on contemporary physics.

Then in 1801, Young performed his double-slit experiment and observed interference fringes. This phenomenon could be explained only by wave theory. It was realized that diffraction was also another phenomenon which could be explained only by wave theory. In fact, it was a natural consequence of Huygens idea of secondary wavelets emanating from every point in the path of light. These experiments could not be explained by assuming that light consists of particles. Another phenomenon of polarisation was discovered around 1810, and this too could be naturally explained by the wave theory. Thus wave theory of Huygens came to the forefront and Newton's particle theory went into the background. This situation again continued for almost a century.

Better experiments were performed in the nineteenth century to determine the speed of light. With more accurate experiments, a value of 3×10^8 m/s for speed of light in vacuum was arrived at. Around 1860, Maxwell proposed his equations of electromagnetism and it was realized that *all* electromagnetic phenomena known at that time could be explained by Maxwell's four equations. Soon Maxwell showed that electric and magnetic fields could propagate through empty space (vacuum) in the form of electromagnetic waves. He calculated the speed of these waves and arrived at a theoretical value of 2.998×10^8 m/s. The close agreement of this value with the experimental value suggested that light consists of electromagnetic waves. In 1887 Hertz demonstrated the generation and detection of such waves. This established the wave theory of light on a firm footing. We might say that while eighteenth century belonged to the particle model, the nineteenth century belonged to the wave model of light.

Vast amounts of experiments were done during the period 1850-1900 on heat and related phenomena, an altogether different area of physics. Theories and models like kinetic theory and thermodynamics were developed which quite successfully explained the various phenomena, except one.

Dual Nature of Radiation and Matter

Every body at any temperature emits radiation of all wavelengths. It also absorbs radiation falling on it. A body which absorbs all the radiation falling on it is called a *black body*. It is an ideal concept in physics, like concepts of a point mass or uniform motion. A graph of the intensity of radiation emitted by a body versus wavelength is called the *black body spectrum*. No theory in those days could explain the complete black body spectrum!

In 1900, Planck hit upon a novel idea. If we assume, he said, that radiation is emitted in packets of energy instead of continuously as in a wave, then we can explain the black body spectrum. Planck himself regarded these quanta, or packets, as a property of emission and absorption, rather than that of light. He derived a formula which agreed with the entire spectrum. This was a confusing mixture of wave and particle pictures – radiation is emitted as a particle, it travels as a wave, and is again absorbed as a particle! Moreover, this put physicists in a dilemma. Should we again accept the particle picture of light just to explain one phenomenon? Then what happens to the phenomena of interference and diffraction which cannot be explained by the particle model?

But soon in 1905, Einstein explained the photoelectric effect by assuming the particle picture of light. In 1907, Debye explained the low temperature specific heats of solids by using the particle picture for lattice vibrations in a crystalline solid. Both these phenomena belonging to widely diverse areas of physics could be explained only by the particle model and not by the wave model. In 1923, Compton's x-ray scattering experiments from atoms also went in favour of the particle picture. This increased the dilemma further.

Thus by 1923, physicists faced with the following situation. (a) There were some phenomena like rectilinear propagation, reflection, refraction, which could be explained by either particle model or by wave model. (b) There were some phenomena such as diffraction and interference which could be explained only by the wave model but *not* by the particle model. (c) There were some phenomena such as black body radiation, photoelectric effect, and Compton scattering which could be explained only by the particle model but *not* by the wave model. Somebody in those days aptly remarked that light behaves as a particle on Mondays, Wednesdays and Fridays, and as a wave on Tuesdays, Thursdays and Saturdays, and we don't talk of light on Sundays!

In 1924, de Broglie proposed his theory of wave-particle duality in which he said that not only photons of light but also 'particles' of matter such as electrons and atoms possess a dual character, sometimes behaving like a particle and sometimes as a wave. He gave a formula connecting their mass, velocity, momentum (particle characteristics), with their wavelength and frequency (wave characteristics)! In 1927 Thomson, and Davisson and Germer, in separate experiments, showed that electrons did behave like waves with a wavelength which agreed with that given by de Broglie's formula. Their experiment was on diffraction of electrons through crystalline solids, in which the regular arrangement of atoms acted like a grating. Very soon, diffraction experiments with other 'particles' such as neutrons and protons were performed and these too confirmed with de Broglie's formula. This confirmed wave-particle duality as an established principle of physics. Here was a principle, physicists thought, which explained all the phenomena mentioned above not only for light but also for the so-called particles.

But there was no basic theoretical foundation for wave-particle duality. De Broglie's proposal was merely a qualitative argument based on symmetry of nature. Wave-particle duality was at best a principle, not an outcome of a sound fundamental theory. It is true that all experiments whatever agreed with de Broglie formula. But physics does not work that way. On the one hand, it needs experimental confirmation, while on the other hand, it also needs sound theoretical basis for the models proposed. This was developed over the next two decades. Dirac developed his theory of radiation in about 1928, and Heisenberg and Pauli gave it a firm footing by 1930. Tomonaga, Schwinger, and Feynman, in late 1940s, produced further refinements and cleared the theory of inconsistencies which were noticed. All these theories mainly put wave-particle duality on a theoretical footing.

Although the story continues, it grows more and more complex and beyond the scope of this note. But we have here the essential structure of what happened, and let us be satisfied with it at the moment. Now it is regarded as a natural consequence of present theories of physics that electromagnetic radiation as well as particles of matter exhibit both wave and particle properties in different experiments, and sometimes even in the different parts of the same experiment.

Chapter Twelve

ATOMS



12.1 INTRODUCTION

By the nineteenth century, enough evidence had accumulated in favour of atomic hypothesis of matter. In 1897, the experiments on electric discharge through gases carried out by the English physicist J. J. Thomson (1856 – 1940) revealed that atoms of different elements contain negatively charged constituents (electrons) that are identical for all atoms. However, atoms on a whole are electrically neutral. Therefore, an atom must also contain some positive charge to neutralise the negative charge of the electrons. But what is the arrangement of the positive charge and the electrons inside the atom? In other words, what is the structure of an atom?

The first model of atom was proposed by J. J. Thomson in 1898. According to this model, the positive charge of the atom is uniformly distributed throughout the volume of the atom and the negatively charged electrons are embedded in it like seeds in a watermelon. This model was picturesquely called *plum pudding model* of the atom. However subsequent studies on atoms, as described in this chapter, showed that the distribution of the electrons and positive charges are very different from that proposed in this model.

We know that condensed matter (solids and liquids) and dense gases at all temperatures emit electromagnetic radiation in which a continuous distribution of several wavelengths is present, though with different intensities. This radiation is considered to be due to oscillations of atoms

and molecules, governed by the interaction of each atom or molecule with its neighbours. *In contrast*, light emitted from rarefied gases heated in a flame, or excited electrically in a glow tube such as the familiar neon sign or mercury vapour light has only certain discrete wavelengths. The spectrum appears as a series of bright lines. In such gases, the average spacing between atoms is large. Hence, the radiation emitted can be considered due to individual atoms rather than because of interactions between atoms or molecules.

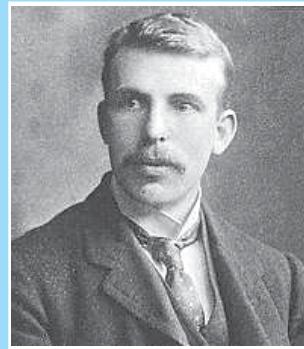
In the early nineteenth century it was also established that each element is associated with a characteristic spectrum of radiation, for example, hydrogen always gives a set of lines with fixed relative position between the lines. This fact suggested an intimate relationship between the internal structure of an atom and the spectrum of radiation emitted by it. In 1885, Johann Jakob Balmer (1825 – 1898) obtained a simple empirical formula which gave the wavelengths of a group of lines emitted by atomic hydrogen. Since hydrogen is simplest of the elements known, we shall consider its spectrum in detail in this chapter.

Ernst Rutherford (1871–1937), a former research student of J. J. Thomson, was engaged in experiments on α -particles emitted by some radioactive elements. In 1906, he proposed a classic experiment of scattering of these α -particles by atoms to investigate the atomic structure. This experiment was later performed around 1911 by Hans Geiger (1882–1945) and Ernst Marsden (1889–1970, who was 20 year-old student and had not yet earned his bachelor's degree). The details are discussed in Section 12.2. The explanation of the results led to the birth of Rutherford's planetary model of atom (also called the *nuclear model of the atom*). According to this the entire positive charge and most of the mass of the atom is concentrated in a small volume called the nucleus with electrons revolving around the nucleus just as planets revolve around the sun.

Rutherford's nuclear model was a major step towards how we see the atom today. However, it could not explain why atoms emit light of only discrete wavelengths. How could an atom as simple as hydrogen, consisting of a single electron and a single proton, emit a complex spectrum of specific wavelengths? In the classical picture of an atom, the electron revolves round the nucleus much like the way a planet revolves round the sun. However, we shall see that there are some serious difficulties in accepting such a model.

12.2 ALPHA-PARTICLE SCATTERING AND RUTHERFORD'S NUCLEAR MODEL OF ATOM

At the suggestion of Ernst Rutherford, in 1911, H. Geiger and E. Marsden performed some experiments. In one of their experiments, as shown in



ERNST RUTHERFORD (1871 – 1937)

Ernst Rutherford (1871 – 1937) British physicist who did pioneering work on radioactive radiation. He discovered alpha-rays and beta-rays. Along with Frederick Soddy, he created the modern theory of radioactivity. He studied the 'emanation' of thorium and discovered a new noble gas, an isotope of radon, now known as thoron. By scattering alpha-rays from the metal foils, he discovered the atomic nucleus and proposed the planetary model of the atom. He also estimated the approximate size of the nucleus.

Physics

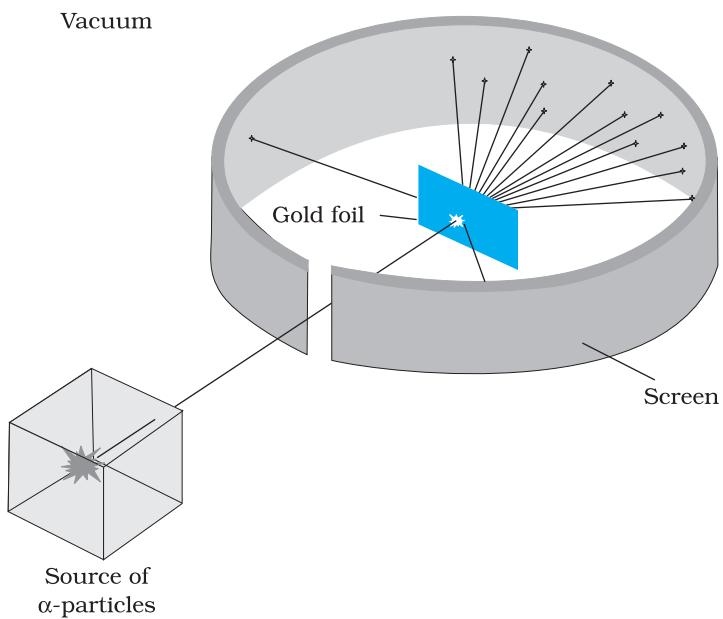


FIGURE 12.1 Geiger-Marsden scattering experiment. The entire apparatus is placed in a vacuum chamber (not shown in this figure).

Fig. 12.1, they directed a beam of 5.5 MeV α -particles emitted from a $^{214}_{83}\text{Bi}$ radioactive source at a thin metal foil made of gold. Figure 12.2 shows a schematic diagram of this experiment. Alpha-particles emitted by a $^{214}_{83}\text{Bi}$ radioactive source were collimated into a narrow beam by their passage through lead bricks. The beam was allowed to fall on a thin foil of gold of thickness 2.1×10^{-7} m. The scattered alpha-particles were observed through a rotatable detector consisting of zinc sulphide screen and a microscope. The scattered alpha-particles on striking the screen produced brief light flashes or scintillations. These flashes may be viewed through a microscope and the distribution of the number of scattered particles may be studied as a function of angle of scattering.

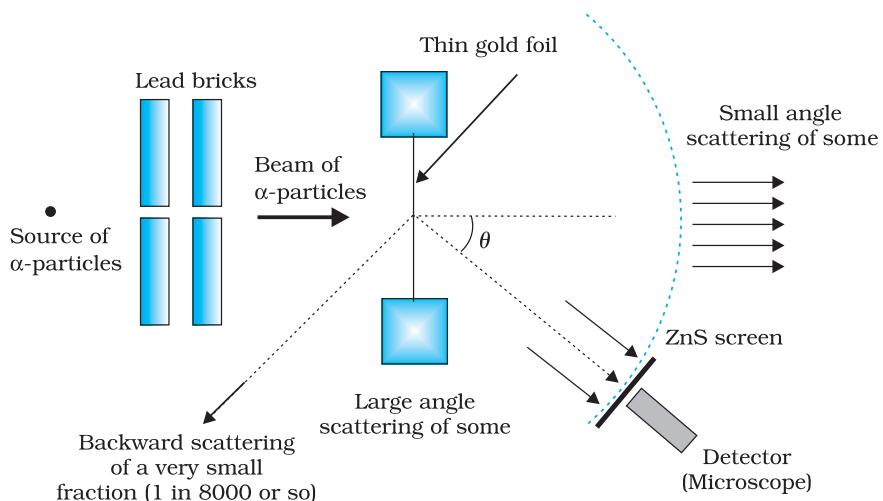


FIGURE 12.2 Schematic arrangement of the Geiger-Marsden experiment.

A typical graph of the total number of α -particles scattered at different angles, in a given interval of time, is shown in Fig. 12.3. The dots in this figure represent the data points and the solid curve is the theoretical prediction based on the assumption that the target atom has a small, dense, positively charged nucleus. Many of the α -particles pass through the foil. It means that they do not suffer any collisions. Only about 0.14% of the incident α -particles scatter by more than 1° ; and about 1 in 8000 deflect by more than 90° . Rutherford argued that, to deflect the α -particle backwards, it must experience a large repulsive force. This force could

Atoms

be provided if the greater part of the mass of the atom and its positive charge were concentrated tightly at its centre. Then the incoming α -particle could get very close to the positive charge without penetrating it, and such a close encounter would result in a large deflection. This agreement supported the hypothesis of the nuclear atom. This is why Rutherford is credited with the *discovery* of the nucleus.

In Rutherford's nuclear model of the atom, the entire positive charge and most of the mass of the atom are concentrated in the nucleus with the electrons some distance away. The electrons would be moving in orbits about the nucleus just as the planets do around the sun. Rutherford's experiments suggested the size of the nucleus to be about 10^{-15} m to 10^{-14} m. From kinetic theory, the size of an atom was known to be 10^{-10} m, about 10,000 to 100,000 times larger

than the size of the nucleus (see Chapter 11, Section 11.6 in Class XI Physics textbook). Thus, the electrons would seem to be at a distance from the nucleus of about 10,000 to 100,000 times the size of the nucleus itself. Thus, most of an atom is empty space. With the atom being largely empty space, it is easy to see why most α -particles go right through a thin metal foil. However, when α -particle happens to come near a nucleus, the intense electric field there scatters it through a large angle. The atomic electrons, being so light, do not appreciably affect the α -particles.

The scattering data shown in Fig. 12.3 can be analysed by employing Rutherford's nuclear model of the atom. As the gold foil is very thin, it can be assumed that α -particles will suffer not more than one scattering during their passage through it. Therefore, computation of the trajectory of an alpha-particle scattered by a single nucleus is enough. Alpha-particles are nuclei of helium atoms and, therefore, carry two units, $2e$, of positive charge and have the mass of the helium atom. The charge of the gold nucleus is Ze , where Z is the atomic number of the atom; for gold $Z=79$. Since the nucleus of gold is about 50 times heavier than an α -particle, it is reasonable to assume that it remains stationary throughout the scattering process. Under these assumptions, the trajectory of an alpha-particle can be computed employing Newton's second law of motion and the Coulomb's law for electrostatic force of repulsion between the alpha-particle and the positively charged nucleus.

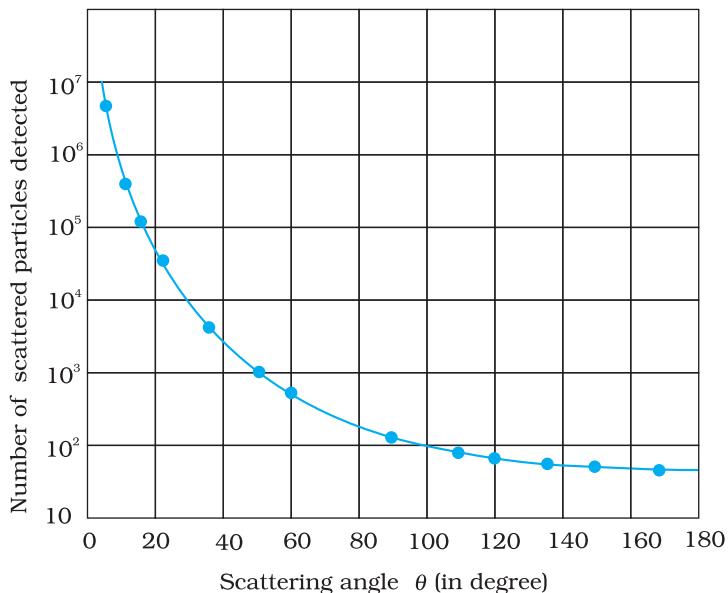


FIGURE 12.3 Experimental data points (shown by dots) on scattering of α -particles by a thin foil at different angles obtained by Geiger and Marsden using the setup shown in Figs. 12.1 and 12.2. Rutherford's nuclear model predicts the solid curve which is seen to be in good agreement with experiment.

The magnitude of this force is

$$F = \frac{1}{4\pi_0} \frac{(2e)(Ze)}{r^2} \quad (12.1)$$

where r is the distance between the α -particle and the nucleus. The force is directed along the line joining the α -particle and the nucleus. The magnitude and direction of the force on an α -particle continuously changes as it approaches the nucleus and recedes away from it.

12.2.1 Alpha-particle trajectory

The trajectory traced by an α -particle depends on the impact parameter, b of collision. The *impact parameter* is the perpendicular distance of the initial velocity vector of the α -particle from the centre of the nucleus (Fig. 12.4).

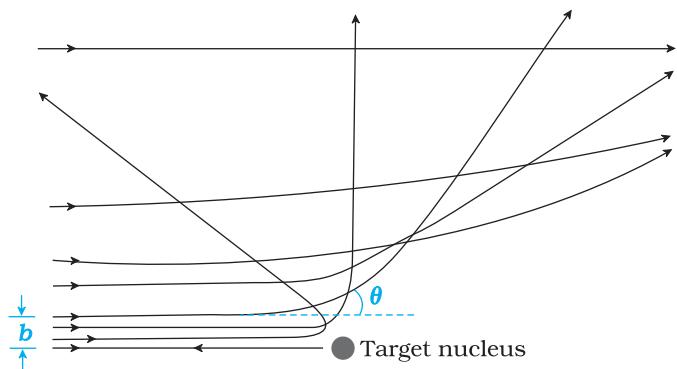


FIGURE 12.4 Trajectory of α -particles in the coulomb field of a target nucleus. The impact parameter, b and scattering angle θ are also depicted.

12.4). A given beam of α -particles has a distribution of impact parameters b , so that the beam is scattered in various directions with different probabilities (Fig. 12.4). (In a beam, all particles have nearly same kinetic energy.) It is seen that an α -particle close to the nucleus (small impact parameter) suffers large scattering. In case of head-on collision, the impact parameter is minimum and the α -particle rebounds back ($\theta \approx \pi$). For a large impact parameter, the α -particle goes nearly undeviated and has a small deflection ($\theta \approx 0$).

The fact that only a small fraction of the number of incident particles rebound back indicates that the number of α -particles undergoing head on collision is small. This,

in turn, implies that the mass of the atom is concentrated in a small volume. Rutherford scattering therefore, is a powerful way to determine an upper limit to the size of the nucleus.

Example 12.1 In the Rutherford's nuclear model of the atom, the nucleus (radius about 10^{-15} m) is analogous to the sun about which the electron move in orbit (radius $\approx 10^{-10}$ m) like the earth orbits around the sun. If the dimensions of the solar system had the same proportions as those of the atom, would the earth be closer to or farther away from the sun than actually it is? The radius of earth's orbit is about 1.5×10^{11} m. The radius of sun is taken as 7×10^8 m.

Solution The ratio of the radius of electron's orbit to the radius of nucleus is $(10^{-10}\text{ m})/(10^{-15}\text{ m}) = 10^5$, that is, the radius of the electron's orbit is 10^5 times larger than the radius of nucleus. If the radius of the earth's orbit around the sun were 10^5 times larger than the radius of the sun, the radius of the earth's orbit would be $10^5 \times 7 \times 10^8\text{ m} = 7 \times 10^{13}\text{ m}$. This is more than 100 times greater than the actual orbital radius of earth. Thus, the earth would be much farther away from the sun.

It implies that an atom contains a much greater fraction of empty space than our solar system does.

Example 12.2 In a Geiger-Marsden experiment, what is the distance of closest approach to the nucleus of a 7.7 MeV α -particle before it comes momentarily to rest and reverses its direction?

Solution The key idea here is that throughout the scattering process, the total mechanical energy of the system consisting of an α -particle and a gold nucleus is conserved. The system's initial mechanical energy is E_i before the particle and nucleus interact, and it is equal to its mechanical energy E_f when the α -particle momentarily stops. The initial energy E_i is just the kinetic energy K of the incoming α -particle. The final energy E_f is just the electric potential energy U of the system. The potential energy U can be calculated from Eq. (12.1).

Let d be the centre-to-centre distance between the α -particle and the gold nucleus when the α -particle is at its stopping point. Then we can write the conservation of energy $E_i = E_f$ as

$$K = \frac{1}{4\epsilon_0} \frac{(2e)(Ze)}{d} = \frac{2Ze^2}{4\epsilon_0 d}$$

Thus the distance of closest approach d is given by

$$d = \frac{2Ze^2}{4\epsilon_0 K}$$

The maximum kinetic energy found in α -particles of natural origin is 7.7 MeV or 1.2×10^{-12} J. Since $1/4\pi\epsilon_0 = 9.0 \times 10^9$ N m 2 /C 2 . Therefore with $e = 1.6 \times 10^{-19}$ C, we have,

$$d = \frac{(2)(9.0 \times 10^9 \text{ Nm}^2/\text{C}^2)(1.6 \times 10^{-19} \text{ C})^2 Z}{1.2 \times 10^{-12} \text{ J}}$$

$$= 3.84 \times 10^{-16} \text{ Z m}$$

The atomic number of foil material gold is $Z = 79$, so that

$$d (\text{Au}) = 3.0 \times 10^{-14} \text{ m} = 30 \text{ fm. (1 fm (i.e. fermi) = } 10^{-15} \text{ m.)}$$

The radius of gold nucleus is, therefore, less than 3.0×10^{-14} m. This is not in very good agreement with the observed result as the actual radius of gold nucleus is 6 fm. The cause of discrepancy is that the distance of closest approach is considerably larger than the sum of the radii of the gold nucleus and the α -particle. Thus, the α -particle reverses its motion without ever actually touching the gold nucleus.



Simulate Rutherford scattering experiment
http://www-outreach.phy.cam.ac.uk/camphy/nucleus/nucleus6_1.htm

EXAMPLE 12.2

12.2.2 Electron orbits

The Rutherford nuclear model of the atom which involves classical concepts, pictures the atom as an electrically neutral sphere consisting of a very small, massive and positively charged nucleus at the centre surrounded by the revolving electrons in their respective dynamically stable orbits. The electrostatic force of attraction, F_e between the revolving electrons and the nucleus provides the requisite centripetal force (F_c) to keep them in their orbits. Thus, for a dynamically stable orbit in a hydrogen atom

$$\frac{F_e}{r} = F_c$$

$$\frac{mv^2}{r} = \frac{1}{4\epsilon_0} \frac{e^2}{r^2} \quad (12.2)$$

Physics

Thus the relation between the orbit radius and the electron velocity is

$$r = \frac{e^2}{4\pi\epsilon_0 mv^2} \quad (12.3)$$

The kinetic energy (K) and electrostatic potential energy (U) of the electron in hydrogen atom are

$$K = \frac{1}{2}mv^2 \quad U = -\frac{e^2}{8\pi\epsilon_0 r}$$

(The negative sign in U signifies that the electrostatic force is in the $-r$ direction.) Thus the total energy E of the electron in a hydrogen atom is

$$E = K + U = \frac{e^2}{8\pi\epsilon_0 r} - \frac{e^2}{4\pi\epsilon_0 r}$$
$$= -\frac{e^2}{8\pi\epsilon_0 r} \quad (12.4)$$

The total energy of the electron is negative. This implies the fact that the electron is bound to the nucleus. If E were positive, an electron will not follow a closed orbit around the nucleus.

Example 12.3 It is found experimentally that 13.6 eV energy is required to separate a hydrogen atom into a proton and an electron. Compute the orbital radius and the velocity of the electron in a hydrogen atom.

Solution Total energy of the electron in hydrogen atom is $-13.6 \text{ eV} = -13.6 \times 1.6 \times 10^{-19} \text{ J} = -2.2 \times 10^{-18} \text{ J}$. Thus from Eq. (12.4), we have

$$\frac{e^2}{8\pi\epsilon_0 r} = 2.2 \times 10^{-18} \text{ J}$$

This gives the orbital radius

$$r = \frac{e^2}{8\pi\epsilon_0 E} = \frac{(9 \times 10^9 \text{ N m}^2/\text{C}^2)(1.6 \times 10^{-19} \text{ C})^2}{(2)(-2.2 \times 10^{-18} \text{ J})}$$
$$= 5.3 \times 10^{-11} \text{ m.}$$

The velocity of the revolving electron can be computed from Eq. (12.3) with $m = 9.1 \times 10^{-31} \text{ kg}$,

$$v = \frac{e}{\sqrt{4\pi\epsilon_0 mr}} = 2.2 \times 10^6 \text{ m/s.}$$

EXAMPLE 12.3

12.3 ATOMIC SPECTRA

As mentioned in Section 12.1, each element has a characteristic spectrum of radiation, which it emits. When an atomic gas or vapour is excited at low pressure, usually by passing an electric current through it, the emitted radiation has a spectrum which contains certain specific wavelengths only. A spectrum of this kind is termed as emission line spectrum and it

Atoms

consists of bright lines on a dark background. The spectrum emitted by atomic hydrogen is shown in Fig. 12.5. Study of emission line spectra of a material can therefore serve as a type of “fingerprint” for identification of the gas. When white light passes through a gas and we analyse the transmitted light using a spectrometer we find some dark lines in the spectrum. These dark lines correspond precisely to those wavelengths which were found in the emission line spectrum of the gas. This is called the *absorption spectrum* of the material of the gas.

12.3.1 Spectral series

We might expect that the frequencies of the light emitted by a particular element would exhibit some regular pattern. Hydrogen is the simplest atom and therefore, has the simplest spectrum. In the observed spectrum, however, at first sight, there does not seem to be any resemblance of order or regularity in spectral lines. But the spacing between lines within certain sets of the hydrogen spectrum decreases in a regular way (Fig. 12.5). Each of these sets is called a *spectral series*. In 1885, the first such series was observed by a Swedish school teacher Johann Jakob Balmer (1825–1898) in the visible region of the hydrogen spectrum. This series is called *Balmer series* (Fig. 12.6). The line with the longest wavelength, 656.3 nm in the red is called H_{α} ; the next line with wavelength 486.1 nm in the blue-green is called H_{β} , the third line 434.1 nm in the violet is called H_{γ} ; and so on. As the wavelength decreases, the lines appear closer together and are weaker in intensity. Balmer found a simple empirical formula for the observed wavelengths

$$\frac{1}{\lambda} = R \frac{1}{2^2} - \frac{1}{n^2} \quad (12.5)$$

where λ is the wavelength, R is a constant called the *Rydberg constant*, and n may have integral values 3, 4, 5, etc. The value of R is $1.097 \times 10^7 \text{ m}^{-1}$. This equation is also called Balmer formula.

Taking $n = 3$ in Eq. (12.5), one obtains the wavelength of the H_{α} line:

$$\begin{aligned} \frac{1}{\lambda} &= 1.097 \times 10^7 \left(\frac{1}{2^2} - \frac{1}{3^2} \right) \text{ m}^{-1} \\ &= 1.522 \times 10^6 \text{ m}^{-1} \end{aligned}$$

i.e., $\lambda = 656.3 \text{ nm}$

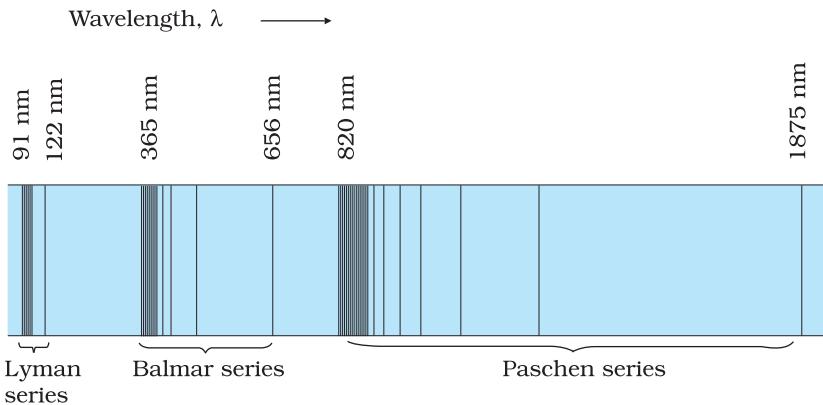


FIGURE 12.5 Emission lines in the spectrum of hydrogen.

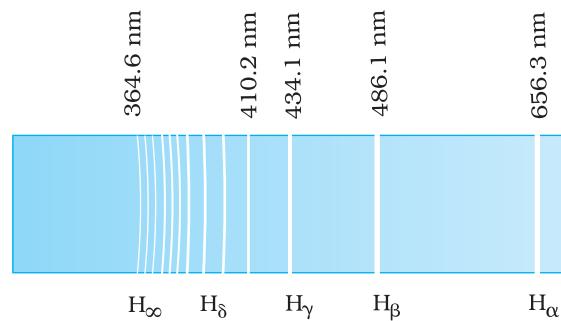


FIGURE 12.6 Balmer series in the emission spectrum of hydrogen.

Physics

For $n = 4$, one obtains the wavelength of H_β line, etc. For $n = \infty$, one obtains the limit of the series, at $\lambda = 364.6$ nm. This is the shortest wavelength in the Balmer series. Beyond this limit, no further distinct lines appear, instead only a faint continuous spectrum is seen.

Other series of spectra for hydrogen were subsequently discovered. These are known, after their discoverers, as Lyman, Paschen, Brackett, and Pfund series. These are represented by the formulae:

Lyman series:

$$\frac{1}{\lambda} = R \frac{1}{1^2} \frac{1}{n^2} \quad n = 2, 3, 4, \dots \quad (12.6)$$

Paschen series:

$$\frac{1}{\lambda} = R \frac{1}{3^2} \frac{1}{n^2} \quad n = 4, 5, 6, \dots \quad (12.7)$$

Brackett series:

$$\frac{1}{\lambda} = R \frac{1}{4^2} \frac{1}{n^2} \quad n = 5, 6, 7, \dots \quad (12.8)$$

Pfund series:

$$\frac{1}{\lambda} = R \frac{1}{5^2} \frac{1}{n^2} \quad n = 6, 7, 8, \dots \quad (12.9)$$

The Lyman series is in the ultraviolet, and the Paschen and Brackett series are in the infrared region.

The Balmer formula Eq. (12.5) may be written in terms of frequency of the light, recalling that

$$c = \nu \lambda$$

$$\text{or } \frac{1}{\lambda} = \frac{c}{\nu}$$

Thus, Eq. (12.5) becomes

$$= R c \frac{1}{2^2} \frac{1}{n^2} \quad (12.10)$$

There are only a few elements (hydrogen, singly ionised helium, and doubly ionised lithium) whose spectra can be represented by simple formula like Eqs. (12.5) – (12.9).

Equations (12.5) – (12.9) are useful as they give the wavelengths that hydrogen atoms radiate or absorb. However, these results are empirical and do not give any reasoning why only certain frequencies are observed in the hydrogen spectrum.

12.4 BOHR MODEL OF THE HYDROGEN ATOM

The model of the atom proposed by Rutherford assumes that the atom, consisting of a central nucleus and revolving electron is stable much like sun-planet system which the model imitates. However, there are some fundamental differences between the two situations. While the planetary system is held by gravitational force, the nucleus-electron system being charged objects, interact by Coulomb's Law of force. We know that an

Atoms

object which moves in a circle is being constantly accelerated – the acceleration being centripetal in nature. According to classical electromagnetic theory, an accelerating charged particle emits radiation in the form of electromagnetic waves. The energy of an accelerating electron should therefore, continuously decrease. The electron would spiral inward and eventually fall into the nucleus (Fig. 12.7). Thus, such an atom can not be stable. Further, according to the classical electromagnetic theory, the frequency of the electromagnetic waves emitted by the revolving electrons is equal to the frequency of revolution. As the electrons spiral inwards, their angular velocities and hence their frequencies would change continuously, and so will the frequency of the light emitted. Thus, they would emit a continuous spectrum, in contradiction to the line spectrum actually observed. Clearly Rutherford model tells only a part of the story implying that the classical ideas are not sufficient to explain the atomic structure.

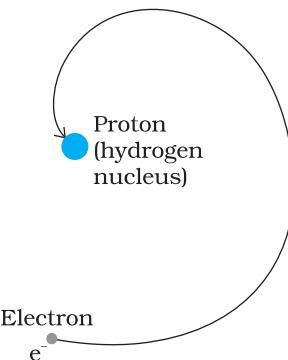


FIGURE 12.7 An accelerated atomic electron must spiral into the nucleus as it loses energy.

Example 12.4 According to the classical electromagnetic theory, calculate the initial frequency of the light emitted by the electron revolving around a proton in hydrogen atom.

Solution From Example 12.3 we know that velocity of electron moving around a proton in hydrogen atom in an orbit of radius 5.3×10^{-11} m is 2.2×10^{-6} m/s. Thus, the frequency of the electron moving around the proton is

$$\frac{v}{2\pi r} = \frac{2.2 \times 10^6 \text{ m s}^{-1}}{2 \times 5.3 \times 10^{-11} \text{ m}}$$
$$\approx 6.6 \times 10^{15} \text{ Hz.}$$

According to the classical electromagnetic theory we know that the frequency of the electromagnetic waves emitted by the revolving electrons is equal to the frequency of its revolution around the nucleus. Thus the initial frequency of the light emitted is 6.6×10^{15} Hz.



Niels Henrik David Bohr (1885 – 1962) Danish physicist who explained the spectrum of hydrogen atom based on quantum ideas. He gave a theory of nuclear fission based on the liquid-drop model of nucleus. Bohr contributed to the clarification of conceptual problems in quantum mechanics, in particular by proposing the complementary principle.

NIELS HENRIK DAVID BOHR (1885 – 1962)

EXAMPLE 12.4

■ Physics

It was Niels Bohr (1885 – 1962) who made certain modifications in this model by adding the ideas of the newly developing quantum hypothesis. Niels Bohr studied in Rutherford's laboratory for several months in 1912 and he was convinced about the validity of Rutherford nuclear model. Faced with the dilemma as discussed above, Bohr, in 1913, concluded that in spite of the success of electromagnetic theory in explaining large-scale phenomena, it could not be applied to the processes at the atomic scale. It became clear that a fairly radical departure from the established principles of classical mechanics and electromagnetism would be needed to understand the structure of atoms and the relation of atomic structure to atomic spectra. Bohr combined classical and early quantum concepts and gave his theory in the form of three postulates. These are :

- (i) Bohr's first postulate was that *an electron in an atom could revolve in certain stable orbits without the emission of radiant energy*, contrary to the predictions of electromagnetic theory. According to this postulate, each atom has certain definite stable states in which it can exist, and each possible state has definite total energy. These are called the stationary states of the atom.
- (ii) Bohr's second postulate defines these stable orbits. This postulate states that the *electron revolves around the nucleus only in those orbits for which the angular momentum is some integral multiple of $h/2\pi$* where h is the Planck's constant ($= 6.6 \times 10^{-34} \text{ J s}$). Thus the angular momentum (L) of the orbiting electron is quantised. That is

$$L = nh/2\pi \quad (12.11)$$

- (iii) Bohr's third postulate incorporated into atomic theory the early quantum concepts that had been developed by Planck and Einstein. It states that *an electron might make a transition from one of its specified non-radiating orbits to another of lower energy. When it does so, a photon is emitted having energy equal to the energy difference between the initial and final states. The frequency of the emitted photon is then given by*

$$hv = E_i - E_f \quad (12.12)$$

where E_i and E_f are the energies of the initial and final states and $E_i > E_f$.

For a hydrogen atom, Eq. (12.4) gives the expression to determine the energies of different energy states. But then this equation requires the radius r of the electron orbit. To calculate r , Bohr's second postulate about the angular momentum of the electron—the quantisation condition – is used. The angular momentum L is given by

$$L = mvr$$

Bohr's second postulate of quantisation [Eq. (12.11)] says that the allowed values of angular momentum are integral multiples of $h/2\pi$.

$$L_n = mv_n r_n = \frac{nh}{2} \quad (12.13)$$

where n is an integer, r_n is the radius of n^{th} possible orbit and v_n is the speed of moving electron in the n^{th} orbit. The allowed orbits are numbered

Atoms

1, 2, 3 ..., according to the values of n , which is called the *principal quantum number* of the orbit.

From Eq. (12.3), the relation between v_n and r_n is

$$v_n = \frac{e}{\sqrt{4\pi\epsilon_0 m r_n}}$$

Combining it with Eq. (12.13), we get the following expressions for v_n and r_n ,

$$v_n = \frac{1}{n} \frac{e^2}{4\pi\epsilon_0 h/2} \quad (12.14)$$

and

$$r_n = \frac{n^2}{m} \frac{\hbar^2}{2} \frac{4\pi\epsilon_0}{e^2} \quad (12.15)$$

Eq. (12.14) depicts that the orbital speed in the n^{th} orbit falls by a factor of n . Using Eq. (12.15), the size of the innermost orbit ($n = 1$) can be obtained as

$$r_1 = \frac{\hbar^2}{me^2}$$

This is called the *Bohr radius*, represented by the symbol a_0 . Thus,

$$a_0 = \frac{\hbar^2}{me^2} \quad (12.16)$$

Substitution of values of \hbar , m , ϵ_0 and e gives $a_0 = 5.29 \times 10^{-11}$ m. From Eq. (12.15), it can also be seen that the radii of the orbits increase as n^2 .

The total energy of the electron in the stationary states of the hydrogen atom can be obtained by substituting the value of orbital radius in Eq. (12.4) as

$$E_n = -\frac{e^2}{8\pi\epsilon_0 n^2} \frac{m}{h^2} \frac{2}{4\pi\epsilon_0} \frac{e^2}{2}$$

$$\text{or } E_n = -\frac{me^4}{8n^2\pi^2\epsilon_0^2 h^2} \quad (12.17)$$

Substituting values, Eq. (12.17) yields

$$E_n = \frac{2.18 \times 10^{-18}}{n^2} \text{ J} \quad (12.18)$$

Atomic energies are often expressed in electron volts (eV) rather than joules. Since $1 \text{ eV} = 1.6 \times 10^{-19} \text{ J}$, Eq. (12.18) can be rewritten as

$$E_n = \frac{13.6}{n^2} \text{ eV} \quad (12.19)$$

The negative sign of the total energy of an electron moving in an orbit means that the electron is bound with the nucleus. Energy will thus be required to remove the electron from the hydrogen atom to a distance infinitely far away from its nucleus (or proton in hydrogen atom).

Physics

The derivation of Eqs. (12.17) – (12.19) involves the assumption that the electronic orbits are circular, though orbits under inverse square force are, in general elliptical. (Planets move in elliptical orbits under the inverse square gravitational force of the sun.) However, it was shown by the German physicist Arnold Sommerfeld (1868 – 1951) that, when the restriction of circular orbit is relaxed, these equations continue to hold even for elliptic orbits.

ORBIT VS STATE (ORBITAL PICTURE) OF ELECTRON IN ATOM

We are introduced to the Bohr Model of atom one time or the other in the course of physics. This model has its place in the history of quantum mechanics and particularly in explaining the structure of an atom. It has become a milestone since Bohr introduced the revolutionary idea of definite energy orbits for the electrons, contrary to the classical picture requiring an accelerating particle to radiate. Bohr also introduced the idea of quantisation of angular momentum of electrons moving in definite orbits. Thus it was a semi-classical picture of the structure of atom.

Now with the development of quantum mechanics, we have a better understanding of the structure of atom. Solutions of the Schrödinger wave equation assign a wave-like description to the electrons bound in an atom due to attractive forces of the protons.

An orbit of the electron in the Bohr model is the circular path of motion of an electron around the nucleus. But according to quantum mechanics, we cannot associate a definite path with the motion of the electrons in an atom. We can only talk about the probability of finding an electron in a certain region of space around the nucleus. This probability can be inferred from the one-electron wave function called the *orbital*. This function depends only on the coordinates of the electron.

It is therefore essential that we understand the subtle differences that exist in the two models:

- Bohr model is valid for only one-electron atoms/ions; an energy value, assigned to each orbit, depends on the principal quantum number n in this model. We know that energy associated with a stationary state of an electron depends on n only, for one-electron atoms/ions. For a multi-electron atom/ion, this is not true.
- The solution of the Schrödinger wave equation, obtained for hydrogen-like atoms/ions, called the wave function, gives information about the probability of finding an electron in various regions around the nucleus. This *orbital* has no resemblance whatsoever with the *orbit* defined for an electron in the Bohr model.

EXAMPLE 12.5

Example 12.5 A 10 kg satellite circles earth once every 2 h in an orbit having a radius of 8000 km. Assuming that Bohr's angular momentum postulate applies to satellites just as it does to an electron in the hydrogen atom, find the quantum number of the orbit of the satellite.

Solution

From Eq. (12.13), we have
 $m v_n r_n = nh/2\pi$

Here $m = 10 \text{ kg}$ and $r_n = 8 \times 10^6 \text{ m}$. We have the time period T of the circling satellite as 2 h. That is $T = 7200 \text{ s}$.

Thus the velocity $v_n = 2\pi r_n/T$.

The quantum number of the orbit of satellite

$$n = (2\pi r_n)^2 \times m / (T \times h)$$

Substituting the values,

$$\begin{aligned} n &= (2\pi \times 8 \times 10^6 \text{ m})^2 \times 10 / (7200 \text{ s} \times 6.64 \times 10^{-34} \text{ J s}) \\ &= 5.3 \times 10^{45} \end{aligned}$$

Note that the quantum number for the satellite motion is extremely large! In fact for such large quantum numbers the results of quantisation conditions tend to those of classical physics.

EXAMPLE 12.5

12.4.1 Energy levels

The energy of an atom is the *least* (largest negative value) when its electron is revolving in an orbit closest to the nucleus i.e., the one for which $n = 1$. For $n = 2, 3, \dots$ the absolute value of the energy E is smaller, hence the energy is progressively larger in the outer orbits. The *lowest* state of the atom, called the *ground state*, is that of the lowest energy, with the electron revolving in the orbit of smallest radius, the Bohr radius, a_0 . The energy of this state ($n = 1$), E_1 is -13.6 eV . Therefore, the minimum energy required to free the electron from the ground state of the hydrogen atom is 13.6 eV . It is called the *ionisation energy* of the hydrogen atom. This prediction of the Bohr's model is in excellent agreement with the experimental value of ionisation energy.

At room temperature, most of the hydrogen atoms are in *ground state*. When a hydrogen atom receives energy by processes such as electron collisions, the atom may acquire sufficient energy to raise the electron to higher energy states. The atom is then said to be in an *excited state*. From Eq. (12.19), for $n = 2$; the energy E_2 is -3.40 eV . It means that the energy required to excite an electron in hydrogen atom to its first excited state, is an energy equal to $E_2 - E_1 = -3.40 \text{ eV} - (-13.6) \text{ eV} = 10.2 \text{ eV}$. Similarly, $E_3 = -1.51 \text{ eV}$ and $E_3 - E_1 = 12.09 \text{ eV}$, or to excite the hydrogen atom from its ground state ($n = 1$) to second excited state ($n = 3$), 12.09 eV energy is required, and so on. From these excited states the electron can then fall back to a state of lower energy, emitting a photon in the process. Thus, as the excitation of hydrogen atom increases (that is as n increases) the value of minimum energy required to free the electron from the excited atom decreases.

The energy level diagram* for the stationary states of a hydrogen atom, computed from Eq. (12.19), is given in

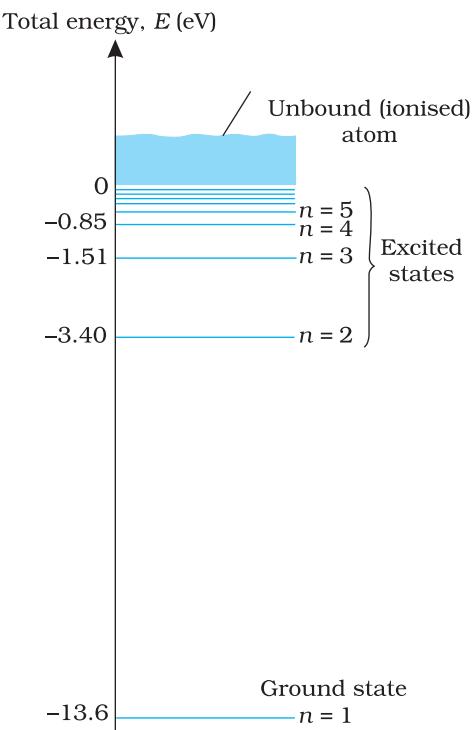


FIGURE 12.8 The energy level diagram for the hydrogen atom. The electron in a hydrogen atom at room temperature spends most of its time in the ground state. To ionise a hydrogen atom an electron from the ground state, 13.6 eV of energy must be supplied. (The horizontal lines specify the presence of allowed energy states.)

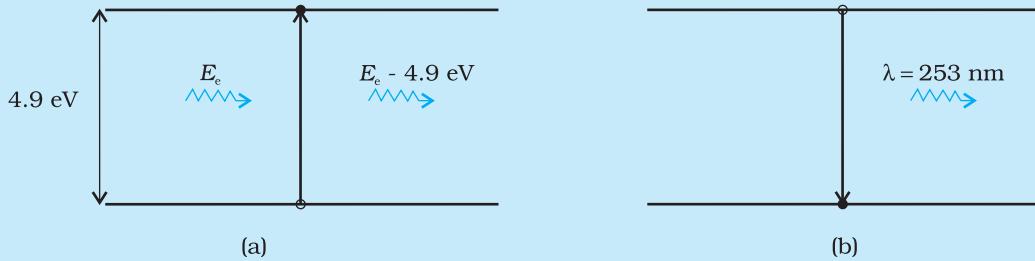
* An electron can have any total energy above $E = 0 \text{ eV}$. In such situations the electron is free. Thus there is a continuum of energy states above $E = 0 \text{ eV}$, as shown in Fig. 12.8.

Physics

Fig. 12.8. The principal quantum number n labels the stationary states in the ascending order of energy. In this diagram, the highest energy state corresponds to $n = \infty$ in Eq. (12.19) and has an energy of 0 eV. This is the energy of the atom when the electron is completely removed ($r = \infty$) from the nucleus and is at rest. Observe how the energies of the excited states come closer and closer together as n increases.

FRANCK – HERTZ EXPERIMENT

The existence of discrete energy levels in an atom was directly verified in 1914 by James Franck and Gustav Hertz. They studied the spectrum of mercury vapour when electrons having different kinetic energies passed through the vapour. The electron energy was varied by subjecting the electrons to electric fields of varying strength. The electrons collide with the mercury atoms and can transfer energy to the mercury atoms. This can only happen when the energy of the electron is higher than the energy difference between an energy level of Hg occupied by an electron and a higher unoccupied level (see Figure). For instance, the difference between an occupied energy level of Hg and a higher unoccupied level is 4.9 eV. If an electron of having an energy of 4.9 eV or more passes through mercury, an electron in mercury atom can absorb energy from the bombarding electron and get excited to the higher level [Fig (a)]. The colliding electron's kinetic energy would reduce by this amount.



The excited electron would subsequently fall back to the ground state by emission of radiation [Fig. (b)]. The wavelength of emitted radiation is:

$$\frac{hc}{E} \frac{6.625 \cdot 10^{-34}}{4.9} \frac{3 \cdot 10^8}{1.6 \cdot 10^{-19}} = 253 \text{ nm}$$

By direct measurement, Franck and Hertz found that the emission spectrum of mercury has a line corresponding to this wavelength. For this experimental verification of Bohr's basic ideas of discrete energy levels in atoms and the process of photon emission, Frank and Hertz were awarded the Nobel prize in 1925.

12.5 THE LINE SPECTRA OF THE HYDROGEN ATOM

According to the third postulate of Bohr's model, when an atom makes a transition from the higher energy state with quantum number n_i to the lower energy state with quantum number n_f ($n_f < n_i$), the difference of energy is carried away by a photon of frequency v_{if} such that

Atoms

$$h\nu_{if} = E_{n_i} - E_{n_f} \quad (12.20)$$

Using Eq. (12.16), for E_{n_f} and E_{n_i} , we get

$$h\nu_{if} = \frac{me^4}{8^2 h^2} \frac{1}{n_f^2} \frac{1}{n_i^2} \quad (12.21)$$

$$\text{or } \nu_{if} = \frac{me^4}{8^2 h^3} \frac{1}{n_f^2} \frac{1}{n_i^2} \quad (12.22)$$

Equation (12.21) is the Rydberg formula, for the spectrum of the hydrogen atom. In this relation, if we take $n_f = 2$ and $n_i = 3, 4, 5, \dots$, it reduces to a form similar to Eq. (12.10) for the Balmer series. The Rydberg constant R is readily identified to be

$$R = \frac{me^4}{8^2 h^3 c} \quad (12.23)$$

If we insert the values of various constants in Eq. (12.23), we get

$$R = 1.03 \times 10^7 \text{ m}^{-1}$$

This is a value very close to the value ($1.097 \times 10^7 \text{ m}^{-1}$) obtained from the empirical Balmer formula. This agreement between the theoretical and experimental values of the Rydberg constant provided a direct and striking confirmation of the Bohr's model.

Since both n_f and n_i are integers, this immediately shows that in transitions between different atomic levels, light is radiated in various discrete frequencies. For hydrogen spectrum, the Balmer formula corresponds to $n_f = 2$ and $n_i = 3, 4, 5, \dots$, etc. The results of the Bohr's model suggested the presence of other series spectra for hydrogen atom—those corresponding to transitions resulting from $n_f = 1$ and $n_i = 2, 3, \dots$; $n_f = 3$ and $n_i = 4, 5, \dots$, etc., and so on. Such series were identified in the course of spectroscopic investigations and are known as the Lyman, Balmer, Paschen, Brackett, and Pfund series. The electronic transitions corresponding to these series are shown in Fig. 12.9.

The various lines in the atomic spectra are produced when electrons jump from higher energy state to a lower energy state and photons are emitted. These spectral lines are called emission lines. But when an atom absorbs a photon that has precisely

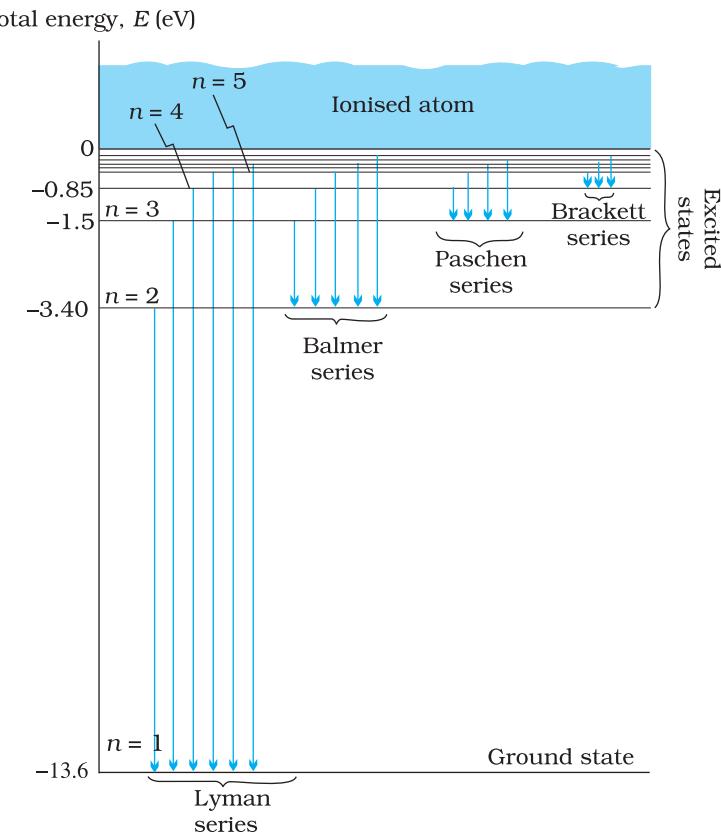


FIGURE 12.9 Line spectra originate in transitions between energy levels.

Physics

the same energy needed by the electron in a lower energy state to make transitions to a higher energy state, the process is called absorption. Thus if photons with a continuous range of frequencies pass through a rarefied gas and then are analysed with a spectrometer, a series of dark spectral absorption lines appear in the continuous spectrum. The dark lines indicate the frequencies that have been absorbed by the atoms of the gas.

The explanation of the hydrogen atom spectrum provided by Bohr's model was a brilliant achievement, which greatly stimulated progress towards the modern quantum theory. In 1922, Bohr was awarded Nobel Prize in Physics.

Example 12.6 Using the Rydberg formula, calculate the wavelengths of the first four spectral lines in the Lyman series of the hydrogen spectrum.

Solution The Rydberg formula is

$$hc/\lambda_{if} = \frac{me^4}{8^2 h^2} \frac{1}{n_f^2} - \frac{1}{n_i^2}$$

The wavelengths of the first four lines in the Lyman series correspond to transitions from $n_i = 2, 3, 4, 5$ to $n_f = 1$. We know that

$$\frac{me^4}{8^2 h^2} = 13.6 \text{ eV} = 21.76 \times 10^{-19} \text{ J}$$

Therefore,

$$\begin{aligned} & \frac{hc}{21.76 \times 10^{-19}} \frac{1}{\frac{1}{n_i^2}} \text{ m} \\ &= \frac{6.625 \times 10^{-34}}{21.76 \times 10^{-19}} \frac{3 \times 10^8}{(n_i^2 - 1)} \text{ m} = \frac{0.9134 n_i^2}{(n_i^2 - 1)} \times 10^7 \text{ m} \\ &= 913.4 \frac{n_i^2}{(n_i^2 - 1)} \text{ Å} \end{aligned}$$

Substituting $n_i = 2, 3, 4, 5$, we get $\lambda_{21} = 1218 \text{ Å}$, $\lambda_{31} = 1028 \text{ Å}$, $\lambda_{41} = 974.3 \text{ Å}$, and $\lambda_{51} = 951.4 \text{ Å}$.

EXAMPLE 12.6

12.6 DE BROGLIE'S EXPLANATION OF BOHR'S SECOND POSTULATE OF QUANTISATION

Of all the postulates, Bohr made in his model of the atom, perhaps the most puzzling is his second postulate. It states that the angular momentum of the electron orbiting around the nucleus is quantised (that is, $L_n = nh/2\pi$; $n = 1, 2, 3 \dots$). Why should the angular momentum have only those values that are integral multiples of $h/2\pi$? The French physicist Louis de Broglie explained this puzzle in 1923, ten years after Bohr proposed his model.

We studied, in Chapter 11, about the de Broglie's hypothesis that material particles, such as electrons, also have a wave nature. C. J. Davisson and L. H. Germer later experimentally verified the wave nature of electrons

in 1927. Louis de Broglie argued that the electron in its circular orbit, as proposed by Bohr, must be seen as a particle wave. In analogy to waves travelling on a string, particle waves too can lead to standing waves under resonant conditions. From Chapter 15 of Class XI Physics textbook, we know that when a string is plucked, a vast number of wavelengths are excited. However only those wavelengths survive which have nodes at the ends and form the standing wave in the string. It means that in a string, standing waves are formed when the total distance travelled by a wave down the string and back is one wavelength, two wavelengths, or any integral number of wavelengths. Waves with other wavelengths interfere with themselves upon reflection and their amplitudes quickly drop to zero. For an electron moving in n^{th} circular orbit of radius r_n , the total distance is the circumference of the orbit, $2\pi r_n$. Thus

$$2\pi r_n = n\lambda, \quad n = 1, 2, 3\dots \quad (12.24)$$

Figure 12.10 illustrates a standing particle wave on a circular orbit for $n = 4$, i.e., $2\pi r_n = 4\lambda$, where λ is the de Broglie wavelength of the electron moving in n^{th} orbit. From Chapter 11, we have $\lambda = h/p$, where p is the magnitude of the electron's momentum. If the speed of the electron is much less than the speed of light, the momentum is mv_n . Thus, $\lambda = h/mv_n$. From Eq. (12.24), we have

$$2\pi r_n = n h/mv_n \quad \text{or} \quad m v_n r_n = nh/2\pi$$

This is the quantum condition proposed by Bohr for the angular momentum of the electron [Eq. (12.13)]. In Section 12.5, we saw that this equation is the basis of explaining the discrete orbits and energy levels in hydrogen atom. Thus de Broglie hypothesis provided an explanation for Bohr's second postulate for the quantisation of angular momentum of the orbiting electron. The quantised electron orbits and energy states are due to the wave nature of the electron and only resonant standing waves can persist.

Bohr's model, involving classical trajectory picture (planet-like electron orbiting the nucleus), correctly predicts the gross features of the hydrogenic atoms*, in particular, the frequencies of the radiation emitted or selectively absorbed. This model however has many limitations. Some are:

- (i) The Bohr model is applicable to hydrogenic atoms. It cannot be extended even to mere two electron atoms such as helium. The analysis of atoms with more than one electron was attempted on the lines of Bohr's model for hydrogenic atoms but did not meet with any success. Difficulty lies in the fact that each electron interacts not only with the positively charged nucleus but also with all other electrons.

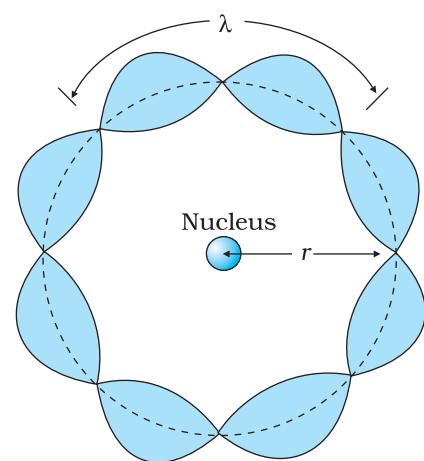


FIGURE 12.10 A standing wave is shown on a circular orbit where four de Broglie wavelengths fit into the circumference of the orbit.

* Hydrogenic atoms are the atoms consisting of a nucleus with positive charge $+Ze$ and a single electron, where Z is the proton number. Examples are hydrogen atom, singly ionised helium, doubly ionised lithium, and so forth. In these atoms more complex electron-electron interactions are nonexistent.

Physics

The formulation of Bohr model involves electrical force between positively charged nucleus and electron. It does not include the electrical forces between electrons which necessarily appear in multi-electron atoms.

- (ii) While the Bohr's model correctly predicts the frequencies of the light emitted by hydrogenic atoms, the model is unable to explain the relative intensities of the frequencies in the spectrum. In emission spectrum of hydrogen, some of the visible frequencies have weak intensity, others strong. Why? Experimental observations depict that some transitions are more favoured than others. Bohr's model is unable to account for the intensity variations.

Bohr's model presents an elegant picture of an atom and cannot be generalised to complex atoms. For complex atoms we have to use a new and radical theory based on Quantum Mechanics, which provides a more complete picture of the atomic structure.

LASER LIGHT

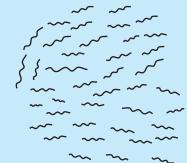
Imagine a crowded market place or a railway platform with people entering a gate and going towards all directions. Their footsteps are random and there is no phase correlation between them. On the other hand, think of a large number of soldiers in a regulated march. Their footsteps are very well correlated. See figure here.

This is similar to the difference between light emitted by an ordinary source like a candle or a bulb and that emitted by a laser. The acronym LASER stands for Light Amplification by Stimulated Emission of Radiation. Since its development in 1960, it has entered into all areas of science and technology. It has found applications in physics, chemistry, biology, medicine, surgery, engineering, etc. There are low power lasers, with a power of 0.5 mW, called pencil lasers, which serve as pointers. There are also lasers of different power, suitable for delicate surgery of eye or glands in the stomach. Finally, there are lasers which can cut or weld steel.

Light is emitted from a source in the form of packets of waves. Light coming out from an ordinary source contains a mixture of many wavelengths. There is also no phase relation between the various waves. Therefore, such light, even if it is passed through an aperture, spreads very fast and the beam size increases rapidly with distance. In the case of laser light, the wavelength of each packet is almost the same. Also the average length of the packet of waves is much larger. This means that there is better phase correlation over a longer duration of time. This results in reducing the divergence of a laser beam substantially.

If there are N atoms in a source, each emitting light with intensity I , then the total intensity produced by an ordinary source is proportional to NI , whereas in a laser source, it is proportional to N^2I . Considering that N is very large, we see that the light from a laser can be much stronger than that from an ordinary source.

When astronauts of the Apollo missions visited the moon, they placed a mirror on its surface, facing the earth. Then scientists on the earth sent a strong laser beam, which was reflected by the mirror on the moon and received back on the earth. The size of the reflected laser beam and the time taken for the round trip were measured. This allowed a very accurate determination of (a) the extremely small divergence of a laser beam and (b) the distance of the moon from the earth.



(a) Light from a bulb



(b) Laser light

SUMMARY

1. Atom, as a whole, is electrically neutral and therefore contains equal amount of positive and negative charges.
2. In *Thomson's model*, an atom is a spherical cloud of positive charges with electrons embedded in it.
3. In *Rutherford's model*, most of the mass of the atom and all its positive charge are concentrated in a tiny nucleus (typically one by ten thousand the size of an atom), and the electrons revolve around it.
4. Rutherford nuclear model has two main difficulties in explaining the structure of atom: (a) It predicts that atoms are unstable because the accelerated electrons revolving around the nucleus must spiral into the nucleus. This contradicts the stability of matter. (b) It cannot explain the characteristic line spectra of atoms of different elements.
5. Atoms of each element are stable and emit characteristic spectrum. The spectrum consists of a set of isolated parallel lines termed as line spectrum. It provides useful information about the atomic structure.
6. The atomic hydrogen emits a line spectrum consisting of various series. The frequency of any line in a series can be expressed as a difference of two terms;

$$\text{Lyman series: } R_c \frac{1}{1^2} - \frac{1}{n^2} ; n = 2, 3, 4, \dots$$

$$\text{Balmer series: } R_c \frac{1}{2^2} - \frac{1}{n^2} ; n = 3, 4, 5, \dots$$

$$\text{Paschen series: } R_c \frac{1}{3^2} - \frac{1}{n^2} ; n = 4, 5, 6, \dots$$

$$\text{Brackett series: } R_c \frac{1}{4^2} - \frac{1}{n^2} ; n = 5, 6, 7, \dots$$

$$\text{Pfund series: } R_c \frac{1}{5^2} - \frac{1}{n^2} ; n = 6, 7, 8, \dots$$

7. To explain the line spectra emitted by atoms, as well as the stability of atoms, Niels Bohr proposed a model for hydrogenic (single electron) atoms. He introduced three postulates and laid the foundations of quantum mechanics:
 - (a) In a hydrogen atom, an electron revolves in certain stable orbits (called stationary orbits) without the emission of radiant energy.
 - (b) The stationary orbits are those for which the angular momentum is some integral multiple of $h/2\pi$. (Bohr's quantisation condition.) That is $L = nh/2\pi$, where n is an integer called a quantum number.
 - (c) The third postulate states that an electron might make a transition from one of its specified non-radiating orbits to another of lower energy. When it does so, a photon is emitted having energy equal to the energy difference between the initial and final states. The frequency (ν) of the emitted photon is then given by

$$h\nu = E_i - E_f$$

An atom absorbs radiation of the same frequency the atom emits, in which case the electron is transferred to an orbit with a higher value of n .

$$E_i + h\nu = E_f$$

8. As a result of the quantisation condition of angular momentum, the electron orbits the nucleus at only specific radii. For a hydrogen atom it is given by

$$r_n = \frac{n^2}{m} \cdot \frac{\hbar^2}{2} \cdot \frac{4\pi^2 e^2}{e^2}$$

The total energy is also quantised:

$$E_n = \frac{me^4}{8n^2 \cdot \pi^2 \hbar^2} \\ = -13.6 \text{ eV}/n^2$$

The $n = 1$ state is called ground state. In hydrogen atom the ground state energy is -13.6 eV. Higher values of n correspond to excited states ($n > 1$). Atoms are excited to these higher states by collisions with other atoms or electrons or by absorption of a photon of right frequency.

9. de Broglie's hypothesis that electrons have a wavelength $\lambda = h/mv$ gave an explanation for Bohr's quantised orbits by bringing in the wave-particle duality. The orbits correspond to circular standing waves in which the circumference of the orbit equals a whole number of wavelengths.
10. Bohr's model is applicable only to hydrogenic (single electron) atoms. It cannot be extended to even two electron atoms such as helium. This model is also unable to explain for the relative intensities of the frequencies emitted even by hydrogenic atoms.

POINTS TO PONDER

- Both the Thomson's as well as the Rutherford's models constitute an unstable system. Thomson's model is unstable electrostatically, while Rutherford's model is unstable because of electromagnetic radiation of orbiting electrons.
- What made Bohr quantise angular momentum (second postulate) and not some other quantity? Note, \hbar has dimensions of angular momentum, and for circular orbits, angular momentum is a very relevant quantity. The second postulate is then so natural!
- The orbital picture in Bohr's model of the hydrogen atom was inconsistent with the uncertainty principle. It was replaced by modern quantum mechanics in which Bohr's orbits are regions where the electron may be found with large probability.
- Unlike the situation in the solar system, where planet-planet gravitational forces are very small as compared to the gravitational force of the sun on each planet (because the mass of the sun is so much greater than the mass of any of the planets), the electron-electron electric force interaction is comparable in magnitude to the electron-nucleus electrical force, because the charges and distances are of the same order of magnitude. This is the reason why the Bohr's model with its planet-like electron is not applicable to many electron atoms.
- Bohr laid the foundation of the quantum theory by postulating specific orbits in which electrons do not radiate. Bohr's model include only

one quantum number n . The new theory called quantum mechanics supports Bohr's postulate. However in quantum mechanics (more generally accepted), a given energy level may not correspond to just one quantum state. For example, a state is characterised by four quantum numbers (n , l , m , and s), but for a pure Coulomb potential (as in hydrogen atom) the energy depends only on n .

6. In Bohr model, contrary to ordinary classical expectation, the frequency of revolution of an electron in its orbit is not connected to the frequency of spectral line. The later is the difference between two orbital energies divided by h . For transitions between large quantum numbers (n to $n - 1$, n very large), however, the two coincide as expected.
7. Bohr's semiclassical model based on some aspects of classical physics and some aspects of modern physics also does not provide a true picture of the simplest hydrogenic atoms. The true picture is quantum mechanical affair which differs from Bohr model in a number of fundamental ways. But then if the Bohr model is not strictly correct, why do we bother about it? The reasons which make Bohr's model still useful are:
 - (i) The model is based on just three postulates but accounts for almost all the general features of the hydrogen spectrum.
 - (ii) The model incorporates many of the concepts we have learnt in classical physics.
 - (iii) The model demonstrates how a theoretical physicist occasionally must quite literally ignore certain problems of approach in hopes of being able to make some predictions. If the predictions of the theory or model agree with experiment, a theoretician then must somehow hope to explain away or rationalise the problems that were ignored along the way.

EXERCISES

- 12.1** Choose the correct alternative from the clues given at the end of the each statement:
- (a) The size of the atom in Thomson's model is the atomic size in Rutherford's model. (much greater than/no different from/much less than.)
 - (b) In the ground state of electrons are in stable equilibrium, while in electrons always experience a net force. (Thomson's model/ Rutherford's model.)
 - (c) A classical atom based on is doomed to collapse. (Thomson's model/ Rutherford's model.)
 - (d) An atom has a nearly continuous mass distribution in a but has a highly non-uniform mass distribution in (Thomson's model/ Rutherford's model.)
 - (e) The positively charged part of the atom possesses most of the mass in (Rutherford's model/both the models.)
- 12.2** Suppose you are given a chance to repeat the alpha-particle scattering experiment using a thin sheet of solid hydrogen in place of the gold foil. (Hydrogen is a solid at temperatures below 14 K.) What results do you expect?

Physics

- 12.3** What is the shortest wavelength present in the Paschen series of spectral lines?
- 12.4** A difference of 2.3 eV separates two energy levels in an atom. What is the frequency of radiation emitted when the atom make a transition from the upper level to the lower level?
- 12.5** The ground state energy of hydrogen atom is -13.6 eV. What are the kinetic and potential energies of the electron in this state?
- 12.6** A hydrogen atom initially in the ground level absorbs a photon, which excites it to the $n = 4$ level. Determine the wavelength and frequency of photon.
- 12.7** (a) Using the Bohr's model calculate the speed of the electron in a hydrogen atom in the $n = 1, 2$, and 3 levels. (b) Calculate the orbital period in each of these levels.
- 12.8** The radius of the innermost electron orbit of a hydrogen atom is 5.3×10^{-11} m. What are the radii of the $n = 2$ and $n = 3$ orbits?
- 12.9** A 12.5 eV electron beam is used to bombard gaseous hydrogen at room temperature. What series of wavelengths will be emitted?
- 12.10** In accordance with the Bohr's model, find the quantum number that characterises the earth's revolution around the sun in an orbit of radius 1.5×10^{11} m with orbital speed 3×10^4 m/s. (Mass of earth = 6.0×10^{24} kg.)

ADDITIONAL EXERCISES

- 12.11** Answer the following questions, which help you understand the difference between Thomson's model and Rutherford's model better.
- Is the average angle of deflection of α -particles by a thin gold foil predicted by Thomson's model much less, about the same, or much greater than that predicted by Rutherford's model?
 - Is the probability of backward scattering (i.e., scattering of α -particles at angles greater than 90°) predicted by Thomson's model much less, about the same, or much greater than that predicted by Rutherford's model?
 - Keeping other factors fixed, it is found experimentally that for small thickness t , the number of α -particles scattered at moderate angles is proportional to t . What clue does this linear dependence on t provide?
 - In which model is it completely wrong to ignore multiple scattering for the calculation of average angle of scattering of α -particles by a thin foil?
- 12.12** The gravitational attraction between electron and proton in a hydrogen atom is weaker than the coulomb attraction by a factor of about 10^{-40} . An alternative way of looking at this fact is to estimate the radius of the first Bohr orbit of a hydrogen atom if the electron and proton were bound by gravitational attraction. You will find the answer interesting.
- 12.13** Obtain an expression for the frequency of radiation emitted when a hydrogen atom de-excites from level n to level $(n-1)$. For large n , show that this frequency equals the classical frequency of revolution of the electron in the orbit.

Atoms

12.14 Classically, an electron can be in any orbit around the nucleus of an atom. Then what determines the typical atomic size? Why is an atom not, say, thousand times bigger than its typical size? The question had greatly puzzled Bohr before he arrived at his famous model of the atom that you have learnt in the text. To simulate what he might well have done before his discovery, let us play as follows with the basic constants of nature and see if we can get a quantity with the dimensions of length that is roughly equal to the known size of an atom ($\sim 10^{-10}\text{m}$).

- Construct a quantity with the dimensions of length from the fundamental constants e , m_e , and c . Determine its numerical value.
- You will find that the length obtained in (a) is many orders of magnitude smaller than the atomic dimensions. Further, it involves c . But energies of atoms are mostly in non-relativistic domain where c is not expected to play any role. This is what may have suggested Bohr to discard c and look for ‘something else’ to get the right atomic size. Now, the Planck’s constant h had already made its appearance elsewhere. Bohr’s great insight lay in recognising that h , m_e , and e will yield the right atomic size. Construct a quantity with the dimension of length from h , m_e , and e and confirm that its numerical value has indeed the correct order of magnitude.

12.15 The total energy of an electron in the first excited state of the hydrogen atom is about -3.4 eV .

- What is the kinetic energy of the electron in this state?
- What is the potential energy of the electron in this state?
- Which of the answers above would change if the choice of the zero of potential energy is changed?

12.16 If Bohr’s quantisation postulate (angular momentum = $nh/2\pi$) is a basic law of nature, it should be equally valid for the case of planetary motion also. Why then do we never speak of quantisation of orbits of planets around the sun?

12.17 Obtain the first Bohr’s radius and the ground state energy of a *muonic hydrogen atom* [i.e., an atom in which a negatively charged muon (μ^-) of mass about $207m_e$ orbits around a proton].

Chapter Thirteen

NUCLEI

13.1 INTRODUCTION

In the previous chapter, we have learnt that in every atom, the positive charge and mass are densely concentrated at the centre of the atom forming its nucleus. The overall dimensions of a nucleus are much smaller than those of an atom. Experiments on scattering of α -particles demonstrated that the radius of a nucleus was smaller than the radius of an atom by a factor of about 10^4 . This means the volume of a nucleus is about 10^{-12} times the volume of the atom. In other words, an atom is almost empty. If an atom is enlarged to the size of a classroom, the nucleus would be of the size of pinhead. Nevertheless, the nucleus contains most (more than 99.9%) of the mass of an atom.

Does the nucleus have a structure, just as the atom does? If so, what are the constituents of the nucleus? How are these held together? In this chapter, we shall look for answers to such questions. We shall discuss various properties of nuclei such as their size, mass and stability, and also associated nuclear phenomena such as radioactivity, fission and fusion.

13.2 ATOMIC MASSES AND COMPOSITION OF NUCLEUS

The mass of an atom is very small, compared to a kilogram; for example, the mass of a carbon atom, ^{12}C , is 1.992647×10^{-26} kg. Kilogram is not a very convenient unit to measure such small quantities. Therefore, a

different mass unit is used for expressing atomic masses. This unit is the atomic mass unit (u), defined as $1/12^{\text{th}}$ of the mass of the carbon (^{12}C) atom. According to this definition

$$\begin{aligned} 1\text{u} &= \frac{\text{mass of one } ^{12}\text{C atom}}{12} \\ &= \frac{1.992647 \times 10^{-26} \text{ kg}}{12} \\ &= 1.660539 \times 10^{-27} \text{ kg} \end{aligned} \quad (13.1)$$

The atomic masses of various elements expressed in atomic mass unit (u) are close to being integral multiples of the mass of a hydrogen atom. There are, however, many striking exceptions to this rule. For example, the atomic mass of chlorine atom is 35.46 u.

Accurate measurement of atomic masses is carried out with a mass spectrometer. The measurement of atomic masses reveals the existence of different types of atoms of the same element, which exhibit the same chemical properties, but differ in mass. Such atomic species of the same element differing in mass are called *isotopes*. (In Greek, isotope means the same place, i.e. they occur in the same place in the periodic table of elements.) It was found that practically every element consists of a mixture of several isotopes. The relative abundance of different isotopes differs from element to element. Chlorine, for example, has two isotopes having masses 34.98 u and 36.98 u, which are nearly integral multiples of the mass of a hydrogen atom. The relative abundances of these isotopes are 75.4 and 24.6 per cent, respectively. Thus, the average mass of a chlorine atom is obtained by the weighted average of the masses of the two isotopes, which works out to be

$$\begin{aligned} &= \frac{75.4 \times 34.98 + 24.6 \times 36.98}{100} \\ &= 35.47 \text{ u} \end{aligned}$$

which agrees with the atomic mass of chlorine.

Even the lightest element, hydrogen has three isotopes having masses 1.0078 u, 2.0141 u, and 3.0160 u. The nucleus of the lightest atom of hydrogen, which has a relative abundance of 99.985%, is called the proton. The mass of a proton is

$$m_p = 1.00727 \text{ u} = 1.67262 \times 10^{-27} \text{ kg} \quad (13.2)$$

This is equal to the mass of the hydrogen atom (= 1.00783u), minus the mass of a single electron ($m_e = 0.00055 \text{ u}$). The other two isotopes of hydrogen are called deuterium and tritium. Tritium nuclei, being unstable, do not occur naturally and are produced artificially in laboratories.

The positive charge in the nucleus is that of the protons. A proton carries one unit of fundamental charge and is stable. It was earlier thought that the nucleus may contain electrons, but this was ruled out later using arguments based on quantum theory. All the electrons of an atom are outside the nucleus. We know that the number of these electrons outside the nucleus of the atom is Z , the atomic number. The total charge of the

Physics

atomic electrons is thus ($-Ze$), and since the atom is neutral, the charge of the nucleus is ($+Ze$). The number of protons in the nucleus of the atom is, therefore, exactly Z , the atomic number.

Discovery of Neutron

Since the nuclei of deuterium and tritium are isotopes of hydrogen, they must contain only one proton each. But the masses of the nuclei of hydrogen, deuterium and tritium are in the ratio of 1:2:3. Therefore, the nuclei of deuterium and tritium must contain, in addition to a proton, some neutral matter. The amount of neutral matter present in the nuclei of these isotopes, expressed in units of mass of a proton, is approximately equal to one and two, respectively. This fact indicates that the nuclei of atoms contain, in addition to protons, neutral matter in multiples of a basic unit. This hypothesis was verified in 1932 by James Chadwick who observed emission of neutral radiation when beryllium nuclei were bombarded with alpha-particles. (α -particles are helium nuclei, to be discussed in a later section). It was found that this neutral radiation could knock out protons from light nuclei such as those of helium, carbon and nitrogen. The only neutral radiation known at that time was photons (electromagnetic radiation). Application of the principles of conservation of energy and momentum showed that if the neutral radiation consisted of photons, the energy of photons would have to be much higher than is available from the bombardment of beryllium nuclei with α -particles. The clue to this puzzle, which Chadwick satisfactorily solved, was to assume that the neutral radiation consists of a new type of neutral particles called *neutrons*. From conservation of energy and momentum, he was able to determine the mass of new particle 'as very nearly the same as mass of proton'.

The mass of a neutron is now known to a high degree of accuracy. It is

$$m_n = 1.00866 \text{ u} = 1.6749 \times 10^{-27} \text{ kg} \quad (13.3)$$

Chadwick was awarded the 1935 Nobel Prize in Physics for his discovery of the neutron.

A free neutron, unlike a free proton, is unstable. It decays into a proton, an electron and a antineutrino (another elementary particle), and has a mean life of about 1000s. It is, however, stable inside the nucleus.

The composition of a nucleus can now be described using the following terms and symbols:

$$Z - \text{atomic number} = \text{number of protons} \quad [13.4(a)]$$

$$N - \text{neutron number} = \text{number of neutrons} \quad [13.4(b)]$$

$$A - \text{mass number} = Z + N$$

$$= \text{total number of protons and neutrons} \quad [13.4(c)]$$

One also uses the term nucleon for a proton or a neutron. Thus the number of nucleons in an atom is its mass number A .

Nuclear species or nuclides are shown by the notation ${}^A_Z X$ where X is the chemical symbol of the species. For example, the nucleus of gold is denoted by ${}^{197}_{79} \text{Au}$. It contains 197 nucleons, of which 79 are protons and the rest 118 are neutrons.

The composition of isotopes of an element can now be readily explained. The nuclei of isotopes of a given element contain the same number of protons, but differ from each other in their number of neutrons. Deuterium, ${}^2\text{H}$, which is an isotope of hydrogen, contains one proton and one neutron. Its other isotope tritium, ${}^3\text{H}$, contains one proton and two neutrons. The element gold has 32 isotopes, ranging from $A=173$ to $A=204$. We have already mentioned that chemical properties of elements depend on their electronic structure. As the atoms of isotopes have identical electronic structure they have identical chemical behaviour and are placed in the same location in the periodic table.

All nuclides with same mass number A are called *isobars*. For example, the nuclides ${}^3\text{H}$ and ${}^3\text{He}$ are isobars. Nuclides with same neutron number N but different atomic number Z , for example ${}^{198}_{80}\text{Hg}$ and ${}^{197}_{79}\text{Au}$, are called *isotones*.

13.3 SIZE OF THE NUCLEUS

As we have seen in Chapter 12, Rutherford was the pioneer who postulated and established the existence of the atomic nucleus. At Rutherford's suggestion, Geiger and Marsden performed their classic experiment: on the scattering of α -particles from thin gold foils. Their experiments revealed that the distance of closest approach to a gold nucleus of an α -particle of kinetic energy 5.5 MeV is about 4.0×10^{-14} m. The scattering of α -particle by the gold sheet could be understood by Rutherford by assuming that the coulomb repulsive force was solely responsible for scattering. Since the positive charge is confined to the nucleus, the actual size of the nucleus has to be less than 4.0×10^{-14} m.

If we use α -particles of higher energies than 5.5 MeV, the distance of closest approach to the gold nucleus will be smaller and at some point the scattering will begin to be affected by the short range nuclear forces, and differ from Rutherford's calculations. Rutherford's calculations are based on pure coulomb repulsion between the positive charges of the α -particle and the gold nucleus. From the distance at which deviations set in, nuclear sizes can be inferred.

By performing scattering experiments in which fast electrons, instead of α -particles, are projectiles that bombard targets made up of various elements, the sizes of nuclei of various elements have been accurately measured.

It has been found that a nucleus of mass number A has a radius

$$R = R_0 A^{1/3} \quad (13.5)$$

where $R_0 = 1.2 \times 10^{-15}$ m. This means the volume of the nucleus, which is proportional to R^3 is proportional to A . Thus the density of nucleus is a constant, independent of A , for all nuclei. Different nuclei are like drops of liquid of constant density. The density of nuclear matter is approximately 2.3×10^{17} kg m $^{-3}$. This density is very large compared to ordinary matter, say water, which is 10^3 kg m $^{-3}$. This is understandable, as we have already seen that most of the atom is empty. Ordinary matter consisting of atoms has a large amount of empty space.

Physics

EXAMPLE 13.1

Example 13.1 Given the mass of iron nucleus as 55.85u and A=56, find the nuclear density?

Solution

$$m_{\text{Fe}} = 55.85, \quad u = 9.27 \times 10^{-26} \text{ kg}$$

$$\text{Nuclear density} = \frac{\text{mass}}{\text{volume}} = \frac{9.27 \times 10^{-26}}{(4\pi/3)(1.2 \times 10^{-15})^3} \times \frac{1}{56} \\ = 2.29 \times 10^{17} \text{ kg m}^{-3}$$

The density of matter in neutron stars (an astrophysical object) is comparable to this density. This shows that matter in these objects has been compressed to such an extent that they resemble a *big nucleus*.

13.4 MASS-ENERGY AND NUCLEAR BINDING ENERGY

13.4.1 Mass – Energy

Einstein showed from his theory of special relativity that it is necessary to treat mass as another form of energy. Before the advent of this theory of special relativity it was presumed that mass and energy were conserved separately in a reaction. However, Einstein showed that mass is another form of energy and one can convert mass-energy into other forms of energy, say kinetic energy and vice-versa.

Einstein gave the famous mass-energy equivalence relation

$$E = mc^2 \quad (13.6)$$

Here the energy equivalent of mass m is related by the above equation and c is the velocity of light in vacuum and is approximately equal to $3 \times 10^8 \text{ m s}^{-1}$.

EXAMPLE 13.2

Example 13.2 Calculate the energy equivalent of 1 g of substance.

Solution

$$\text{Energy, } E = 10^{-3} \times (3 \times 10^8)^2 \text{ J}$$

$$E = 10^{-3} \times 9 \times 10^{16} = 9 \times 10^{13} \text{ J}$$

Thus, if one gram of matter is converted to energy, there is a release of enormous amount of energy.

Experimental verification of the Einstein's mass-energy relation has been achieved in the study of nuclear reactions amongst nucleons, nuclei, electrons and other more recently discovered particles. In a reaction the conservation law of energy states that the initial energy and the final energy are equal provided the energy associated with mass is also included. This concept is important in understanding nuclear masses and the interaction of nuclei with one another. They form the subject matter of the next few sections.

13.4.2 Nuclear binding energy

In Section 13.2 we have seen that the nucleus is made up of neutrons and protons. Therefore it may be expected that the mass of the nucleus is equal to the total mass of its individual protons and neutrons. However,

Nuclei

the nuclear mass M is found to be always less than this. For example, let us consider $^{16}_8\text{O}$; a nucleus which has 8 neutrons and 8 protons. We have

$$\text{Mass of 8 neutrons} = 8 \times 1.00866 \text{ u}$$

$$\text{Mass of 8 protons} = 8 \times 1.00727 \text{ u}$$

$$\text{Mass of 8 electrons} = 8 \times 0.00055 \text{ u}$$

$$\begin{aligned}\text{Therefore the expected mass of } &^{16}_8\text{O nucleus} \\ &= 8 \times 2.01593 \text{ u} = 16.12744 \text{ u.}\end{aligned}$$

The atomic mass of $^{16}_8\text{O}$ found from mass spectroscopy experiments is seen to be 15.99493 u. Subtracting the mass of 8 electrons (8×0.00055 u) from this, we get the experimental mass of $^{16}_8\text{O}$ nucleus to be 15.99053 u.

Thus, we find that the mass of the $^{16}_8\text{O}$ nucleus is less than the total mass of its constituents by 0.13691 u. The difference in mass of a nucleus and its constituents, ΔM , is called the *mass defect*, and is given by

$$\Delta M = [Zm_p + (A - Z)m_n] - M \quad (13.7)$$

What is the meaning of the mass defect? It is here that Einstein's equivalence of mass and energy plays a role. Since the mass of the oxygen nucleus is less than the sum of the masses of its constituents (8 protons and 8 neutrons, in the unbound state), the equivalent energy of the oxygen nucleus is less than that of the sum of the equivalent energies of its constituents. If one wants to break the oxygen nucleus into 8 protons and 8 neutrons, this extra energy $\Delta M c^2$, has to be supplied. This energy required E_b is related to the mass defect by

$$E_b = \Delta M c^2 \quad (13.8)$$

Example 13.3 Find the energy equivalent of one atomic mass unit, first in Joules and then in MeV. Using this, express the mass defect of $^{16}_8\text{O}$ in MeV/c^2 .

Solution

$$1\text{u} = 1.6605 \times 10^{-27} \text{ kg}$$

To convert it into energy units, we multiply it by c^2 and find that
energy equivalent = $1.6605 \times 10^{-27} \times (2.9979 \times 10^8)^2 \text{ kg m}^2/\text{s}^2$

$$= 1.4924 \times 10^{-10} \text{ J}$$

$$= \frac{1.4924 \times 10^{-10}}{1.602 \times 10^{-19}} \text{ eV}$$

$$= 0.9315 \times 10^9 \text{ eV}$$

$$= 931.5 \text{ MeV}$$

or, $1\text{u} = 931.5 \text{ MeV}/c^2$

$$\begin{aligned}\text{For } &^{16}_8\text{O}, \quad \Delta M = 0.13691 \text{ u} = 0.13691 \times 931.5 \text{ MeV}/c^2 \\ &= 127.5 \text{ MeV}/c^2\end{aligned}$$

The energy needed to separate $^{16}_8\text{O}$ into its constituents is thus 127.5 MeV/c^2 .

EXAMPLE 13.3

If a certain number of neutrons and protons are brought together to form a nucleus of a certain charge and mass, an energy E_b will be released

Physics

in the process. The energy E_b is called the *binding energy* of the nucleus. If we separate a nucleus into its nucleons, we would have to supply a total energy equal to E_b , to those particles. Although we cannot tear apart a nucleus in this way, the nuclear binding energy is still a convenient measure of how well a nucleus is held together. A more useful measure of the binding between the constituents of the nucleus is the *binding energy per nucleon*, E_{bn} , which is the ratio of the binding energy E_b of a nucleus to the number of the nucleons, A , in that nucleus:

$$E_{bn} = E_b / A \quad (13.9)$$

We can think of binding energy per nucleon as the average energy per nucleon needed to separate a nucleus into its individual nucleons.

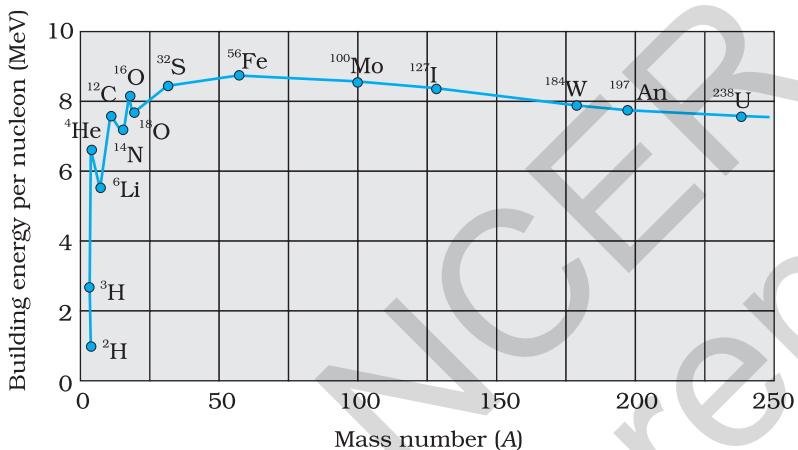


FIGURE 13.1 The binding energy per nucleon as a function of mass number.

Figure 13.1 is a plot of the binding energy per nucleon E_{bn} versus the mass number A for a large number of nuclei. We notice the following main features of the plot:

- (i) the binding energy per nucleon, E_{bn} , is practically constant, i.e. practically independent of the atomic number for nuclei of middle mass number ($30 < A < 170$). The curve has a maximum of about 8.75 MeV for $A = 56$ and has a value of 7.6 MeV for $A = 238$.
- (ii) E_{bn} is lower for both light nuclei ($A < 30$) and heavy nuclei ($A > 170$).

We can draw some conclusions from these two observations:

- (i) The force is attractive and sufficiently strong to produce a binding energy of a few MeV per nucleon.
- (ii) The constancy of the binding energy in the range $30 < A < 170$ is a consequence of the fact that the nuclear force is short-ranged. Consider a particular nucleon inside a sufficiently large nucleus. It will be under the influence of only some of its neighbours, which come within the range of the nuclear force. If any other nucleon is at a distance more than the range of the nuclear force from the particular nucleon it will have no influence on the binding energy of the nucleon under consideration. If a nucleon can have a maximum of p neighbours within the range of nuclear force, its binding energy would be proportional to p . Let the binding energy of the nucleus be pk , where k is a constant having the dimensions of energy. If we increase A by adding nucleons they will not change the binding energy of a nucleon inside. Since most of the nucleons in a large nucleus reside inside it and not on the surface, the change in binding energy per nucleon would be small. The binding energy per nucleon is a constant and is approximately equal to pk . The property that a given nucleon

influences only nucleons close to it is also referred to as saturation property of the nuclear force.

- (iii) A very heavy nucleus, say $A = 240$, has lower binding energy per nucleon compared to that of a nucleus with $A = 120$. Thus if a nucleus $A = 240$ breaks into two $A = 120$ nuclei, nucleons get more tightly bound. This implies energy would be released in the process. It has very important implications for energy production through *fission*, to be discussed later in Section 13.7.1.
- (iv) Consider two very light nuclei ($A \leq 10$) joining to form a heavier nucleus. The binding energy per nucleon of the fused heavier nuclei is more than the binding energy per nucleon of the lighter nuclei. This means that the final system is more tightly bound than the initial system. Again energy would be released in such a process of *fusion*. This is the energy source of sun, to be discussed later in Section 13.7.3.

13.5 NUCLEAR FORCE

The force that determines the motion of atomic electrons is the familiar Coulomb force. In Section 13.4, we have seen that for average mass nuclei the binding energy per nucleon is approximately 8 MeV, which is much larger than the binding energy in atoms. Therefore, to bind a nucleus together there must be a strong attractive force of a totally different kind. It must be strong enough to overcome the repulsion between the (positively charged) protons and to bind both protons and neutrons into the tiny nuclear volume. We have already seen that the constancy of binding energy per nucleon can be understood in terms of its short-range. Many features of the nuclear binding force are summarised below. These are obtained from a variety of experiments carried out during 1930 to 1950.

- (i) The nuclear force is much stronger than the Coulomb force acting between charges or the gravitational forces between masses. The nuclear binding force has to dominate over the Coulomb repulsive force between protons inside the nucleus. This happens only because the nuclear force is much stronger than the coulomb force. The gravitational force is much weaker than even Coulomb force.
- (ii) The nuclear force between two nucleons falls rapidly to zero as their distance is more than a few femtometres. This leads to *saturation of forces* in a medium or a large-sized nucleus, which is the reason for the constancy of the binding energy per nucleon.

A rough plot of the potential energy between two nucleons as a function of distance is shown in the Fig. 13.2. The potential energy is a minimum at a distance r_0 of about 0.8 fm. This means that the force is attractive for distances larger than 0.8 fm and repulsive if they are separated by distances less than 0.8 fm.

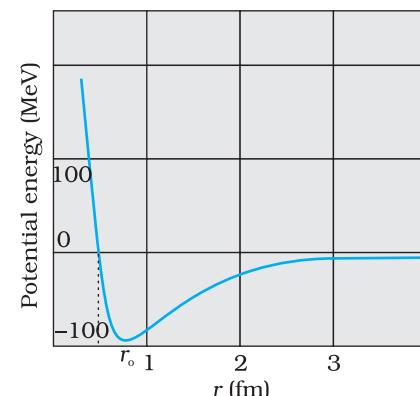


FIGURE 13.2 Potential energy of a pair of nucleons as a function of their separation.

For a separation greater than r_0 , the force is attractive and for separations less than r_0 , the force is strongly repulsive.

Physics

(iii) The nuclear force between neutron-neutron, proton-neutron and proton-proton is approximately the same. The nuclear force does not depend on the electric charge.

Unlike Coulomb's law or the Newton's law of gravitation there is no simple mathematical form of the nuclear force.

13.6 RADIOACTIVITY

A. H. Becquerel discovered radioactivity in 1896 purely by accident. While studying the fluorescence and phosphorescence of compounds irradiated with visible light, Becquerel observed an interesting phenomenon. After illuminating some pieces of uranium-potassium sulphate with visible light, he wrapped them in black paper and separated the package from a photographic plate by a piece of silver. When, after several hours of exposure, the photographic plate was developed, it showed blackening due to something that must have been emitted by the compound and was able to penetrate both black paper and the silver.

Experiments performed subsequently showed that radioactivity was a nuclear phenomenon in which an unstable nucleus undergoes a decay. This is referred to as *radioactive decay*. Three types of radioactive decay occur in nature :

- (i) α -decay in which a helium nucleus ${}^4_2\text{He}$ is emitted;
- (ii) β -decay in which electrons or positrons (particles with the same mass as electrons, but with a charge exactly opposite to that of electron) are emitted;
- (iii) γ -decay in which high energy (hundreds of keV or more) photons are emitted.

Each of these decay will be considered in subsequent sub-sections.

13.6.1 Law of radioactive decay

In any radioactive sample, which undergoes α , β or γ -decay, it is found that the number of nuclei undergoing the decay per unit time is proportional to the total number of nuclei in the sample. If N is the number of nuclei in the sample and ΔN undergo decay in time Δt then

$$\frac{\Delta N}{\Delta t} \propto N$$

or, $\Delta N/\Delta t = \lambda N$, (13.10)

where λ is called the radioactive *decay constant* or *disintegration constant*.

The change in the number of nuclei in the sample* is $dN = -\Delta N$ in time Δt . Thus the rate of change of N is (in the limit $\Delta t \rightarrow 0$)

$$\frac{dN}{dt} = -\lambda N$$

* ΔN is the number of nuclei that decay, and hence is always positive. dN is the change in N , which may have either sign. Here it is negative, because out of original N nuclei, ΔN have decayed, leaving $(N-\Delta N)$ nuclei.

$$\text{or, } \frac{dN}{N} = -\lambda dt$$

Now, integrating both sides of the above equation, we get,

$$\int_{N_0}^N \frac{dN}{N} = -\lambda \int_{t_0}^t dt \quad (13.11)$$

$$\text{or, } \ln N - \ln N_0 = -\lambda (t - t_0) \quad (13.12)$$

Here N_0 is the number of radioactive nuclei in the sample at some arbitrary time t_0 and N is the number of radioactive nuclei at any subsequent time t . Setting $t_0 = 0$ and rearranging Eq. (13.12) gives us

$$\ln \frac{N}{N_0} = -\lambda t \quad (13.13)$$

which gives

$$N(t) = N_0 e^{-\lambda t} \quad (13.14)$$

Note, for example, the light bulbs follow no such exponential decay law. If we test 1000 bulbs for their life (time span before they burn out or fuse), we expect that they will 'decay' (that is, burn out) at more or less the same time. The decay of radionuclides follows quite a different law, the *law of radioactive decay* represented by Eq. (13.14).

The total decay rate R of a sample is the number of nuclei disintegrating per unit time. Suppose in a time interval dt , the decay count measured is ΔN . Then $dN = -\Delta N$.

The positive quantity R is then defined as

$$R = -\frac{dN}{dt}$$

Differentiating Eq. (13.14), we get

$$R = \lambda N_0 e^{-\lambda t}$$

$$\text{or, } R = R_0 e^{-\lambda t} \quad (13.15)$$

This is equivalent to the law of radioactivity decay, since you can integrate Eq. (13.15) to get back Eq. (13.14). Clearly, $R_0 = \lambda N_0$ is the decay rate at $t=0$. The decay rate R at a certain time t and the number of undecayed nuclei N at the same time are related by

$$R = \lambda N \quad (13.16)$$

The decay rate of a sample, rather than the number of radioactive nuclei, is a more direct experimentally measurable quantity and is given a specific name: *activity*. The SI unit for activity is becquerel, named after the discoverer of radioactivity, Henry Becquerel.

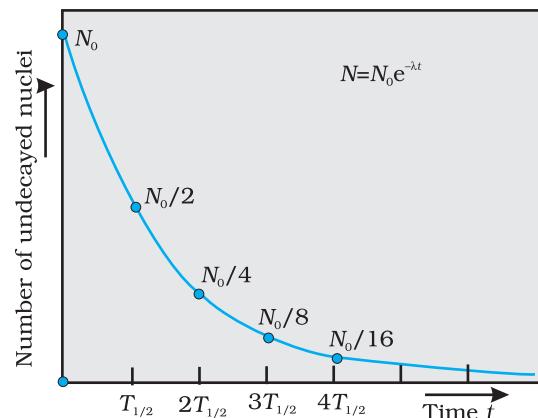


FIGURE 13.3 Exponential decay of a radioactive species. After a lapse of $T_{1/2}$, population of the given species drops by a factor of 2.

Physics

1 becquerel is simply equal to 1 disintegration or decay per second. There is also another unit named “curie” that is widely used and is related to the SI unit as:

$$1 \text{ curie} = 1 \text{ Ci} = 3.7 \times 10^{10} \text{ decays per second}$$
$$= 3.7 \times 10^{10} \text{ Bq}$$

Different radionuclides differ greatly in their rate of decay. A common way to characterize this feature is through the notion of *half-life*. Half-life of a radionuclide (denoted by $T_{1/2}$) is the time it takes for a sample that has initially, say N_0 radionuclei to reduce to $N_0/2$. Putting $N = N_0/2$ and $t = T_{1/2}$ in Eq. (13.14), we get

$$T_{1/2} = \frac{\ln 2}{\lambda} = \frac{0.693}{\lambda} \quad (13.17)$$

Clearly if N_0 reduces to half its value in time $T_{1/2}$, R_0 will also reduce to half its value in the same time according to Eq. (13.16).

Another related measure is the *average* or *mean life* τ . This again can be obtained from Eq. (13.14). The number of nuclei which decay in the time interval t to $t + \Delta t$ is $R(t)\Delta t$ ($= \lambda N_0 e^{-\lambda t} \Delta t$). Each of them has lived for time t . Thus the total life of all these nuclei would be $t \lambda N_0 e^{-\lambda t} \Delta t$. It is clear that some nuclei may live for a short time while others may live longer. Therefore to obtain the mean life, we have to sum (or integrate) this expression over all times from 0 to ∞ , and divide by the total number N_0 of nuclei at $t = 0$. Thus,

$$\tau = \frac{\int_0^\infty t \lambda N_0 e^{-\lambda t} dt}{N_0} = \lambda \int_0^\infty t e^{-\lambda t} dt$$

One can show by performing this integral that

$$\tau = 1/\lambda$$

We summarise these results with the following:

$$T_{1/2} = \frac{\ln 2}{\lambda} = \tau \ln 2 \quad (13.18)$$

Radioactive elements (e.g., tritium, plutonium) which are short-lived i.e., have half-lives much less than the age of the universe (≈ 15 billion years) have obviously decayed long ago and are not found in nature. They can, however, be produced artificially in nuclear reactions.

Example 13.4 The half-life of $^{238}_{92}\text{U}$ undergoing α -decay is 4.5×10^9 years. What is the activity of 1g sample of $^{238}_{92}\text{U}$?

Solution

$$T_{1/2} = 4.5 \times 10^9 \text{ y}$$
$$= 4.5 \times 10^9 \text{ y} \times 3.16 \times 10^7 \text{ s/y}$$
$$= 1.42 \times 10^{17} \text{ s}$$

One k mol of any isotope contains Avogadro's number of atoms, and so 1g of $^{238}_{92}\text{U}$ contains

$$\frac{1}{238 \times 10^{-3}} \text{ kmol} \times 6.025 \times 10^{26} \text{ atoms/kmol}$$

$$= 25.3 \times 10^{20} \text{ atoms.}$$

The decay rate R is

$$R = \lambda N$$

$$\begin{aligned} R &= \frac{0.693}{T_{1/2}} N = \frac{0.693 \times 25.3 \times 10^{20}}{1.42 \times 10^{17}} \text{ s}^{-1} \\ &= 1.23 \times 10^4 \text{ s}^{-1} \\ &= 1.23 \times 10^4 \text{ Bq} \end{aligned}$$

EXAMPLE 13.4

Example 13.5 Tritium has a half-life of 12.5 y undergoing beta decay. What fraction of a sample of pure tritium will remain undecayed after 25 y.

Solution

By definition of half-life, half of the initial sample will remain undecayed after 12.5 y. In the next 12.5 y, one-half of these nuclei would have decayed. Hence, one fourth of the sample of the initial pure tritium will remain undecayed.

EXAMPLE 13.5

13.6.2 Alpha decay

A well-known example of alpha decay is the decay of uranium $^{238}_{92}\text{U}$ to thorium $^{234}_{90}\text{Th}$ with the emission of a helium nucleus ^4_2He



In α -decay, the mass number of the product nucleus (daughter nucleus) is four less than that of the decaying nucleus (parent nucleus), while the atomic number decreases by two. In general, α -decay of a parent nucleus ^A_ZX results in a daughter nucleus $^{A-4}_{Z-2}\text{Y}$



From Einstein's mass-energy equivalence relation [Eq. (13.6)] and energy conservation, it is clear that this spontaneous decay is possible only when the total mass of the decay products is less than the mass of the initial nucleus. This difference in mass appears as kinetic energy of the products. By referring to a table of nuclear masses, one can check that the total mass of $^{234}_{90}\text{Th}$ and ^4_2He is indeed less than that of $^{238}_{92}\text{U}$.

The disintegration energy or the Q -value of a nuclear reaction is the difference between the initial mass energy and the total mass energy of the decay products. For α -decay

$$Q = (m_X - m_Y - m_{\text{He}}) c^2 \quad (13.21)$$

Q is also the net kinetic energy gained in the process or, if the initial nucleus X is at rest, the kinetic energy of the products. Clearly, $Q > 0$ for exothermic processes such as α -decay.

EXAMPLE 13.6

Example 13.6 We are given the following atomic masses:

$${}_{92}^{238}\text{U} = 238.05079 \text{ u} \quad {}_2^4\text{He} = 4.00260 \text{ u}$$

$${}_{90}^{234}\text{Th} = 234.04363 \text{ u} \quad {}_1^1\text{H} = 1.00783 \text{ u}$$

$${}_{91}^{237}\text{Pa} = 237.05121 \text{ u}$$

Here the symbol Pa is for the element protactinium ($Z = 91$).

(a) Calculate the energy released during the alpha decay of ${}_{92}^{238}\text{U}$.

(b) Show that ${}_{92}^{238}\text{U}$ can not spontaneously emit a proton.

Solution

(a) The alpha decay of ${}_{92}^{238}\text{U}$ is given by Eq. (13.20). The energy released in this process is given by

$$Q = (M_{\text{U}} - M_{\text{Th}} - M_{\text{He}}) c^2$$

Substituting the atomic masses as given in the data, we find

$$Q = (238.05079 - 234.04363 - 4.00260) \text{ u} \times c^2$$

$$= (0.00456 \text{ u}) c^2$$

$$= (0.00456 \text{ u}) (931.5 \text{ MeV/u})$$

$$= 4.25 \text{ MeV.}$$

(b) If ${}_{92}^{238}\text{U}$ spontaneously emits a proton, the decay process would be



The Q for this process to happen is

$$= (M_{\text{U}} - M_{\text{Pa}} - M_{\text{H}}) c^2$$

$$= (238.05079 - 237.05121 - 1.00783) \text{ u} \times c^2$$

$$= (-0.00825 \text{ u}) c^2$$

$$= - (0.00825 \text{ u})(931.5 \text{ MeV/u})$$

$$= - 7.68 \text{ MeV}$$

Thus, the Q of the process is negative and therefore it cannot proceed spontaneously. We will have to supply an energy of 7.68 MeV to a ${}_{92}^{238}\text{U}$ nucleus to make it emit a proton.

13.6.3 Beta decay

In beta decay, a nucleus spontaneously emits an electron (β^- decay) or a positron (β^+ decay). A common example of β^- decay is



and that of β^+ decay is



The decays are governed by the Eqs. (13.14) and (13.15), so that one can never predict which nucleus will undergo decay, but one can characterize the decay by a half-life $T_{1/2}$. For example, $T_{1/2}$ for the decays above is respectively 14.3 d and 2.6 y. The emission of electron in β^- decay is accompanied by the emission of an antineutrino ($\bar{\nu}$); in β^+ decay, instead, a neutrino (ν) is generated. Neutrinos are neutral particles with very small (possibly, even zero) mass compared to electrons. They have only weak interaction with other particles. They are, therefore, very difficult to detect, since they can penetrate large quantity of matter (even earth) without any interaction.

In both β^- and β^+ decay, the mass number A remains unchanged. In β^- decay, the atomic number Z of the nucleus goes up by 1, while in β^+ decay Z goes down by 1. The basic nuclear process underlying β^- decay is the conversion of neutron to proton



while for β^+ decay, it is the conversion of proton into neutron



Note that while a free neutron decays to proton, the decay of proton to neutron [Eq. (13.25)] is possible only inside the nucleus, since proton has smaller mass than neutron.

13.6.4 Gamma decay

Like an atom, a nucleus also has discrete energy levels - the ground state and excited states. The scale of energy is, however, very different. Atomic energy level spacings are of the order of eV, while the difference in nuclear energy levels is of the order of MeV. When a nucleus in an excited state spontaneously decays to its ground state (or to a lower energy state), a photon is emitted with energy equal to the difference in the two energy levels of the nucleus. This is the so-called *gamma decay*. The energy (MeV) corresponds to radiation of extremely short wavelength, shorter than the hard X-ray region.

Typically, a gamma ray is emitted when a α or β decay results in a daughter nucleus in an excited state. This then returns to the ground state by a single photon transition or successive transitions involving more than one photon. A familiar example is the successive emission of gamma rays of energies 1.17 MeV and 1.33 MeV from the deexcitation of $^{60}_{28}\text{Ni}$ nuclei formed from β^- decay of $^{60}_{27}\text{Co}$.

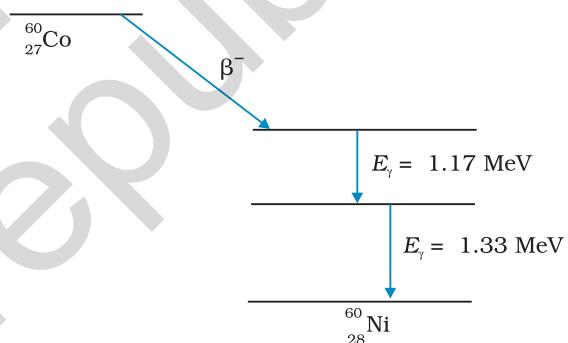


FIGURE 13.4 β -decay of $^{60}_{27}\text{Co}$ followed by emission of two γ rays from deexcitation of the daughter nucleus $^{60}_{28}\text{Ni}$.

13.7 NUCLEAR ENERGY

The curve of binding energy per nucleon E_{bn} , given in Fig. 13.1, has a long flat middle region between $A = 30$ and $A = 170$. In this region the binding energy per nucleon is nearly constant (8.0 MeV). For the lighter nuclei region, $A < 30$, and for the heavier nuclei region, $A > 170$, the binding energy per nucleon is less than 8.0 MeV, as we have noted earlier. Now, the greater the binding energy, the less is the total mass of a bound system, such as a nucleus. Consequently, if nuclei with less total binding energy transform to nuclei with greater binding energy, there will be a net energy release. This is what happens when a heavy nucleus decays into two or more intermediate mass fragments (*fission*) or when light nuclei fuse into a heavier nucleus (*fusion*.)

Exothermic chemical reactions underlie conventional energy sources such as coal or petroleum. Here the energies involved are in the range of

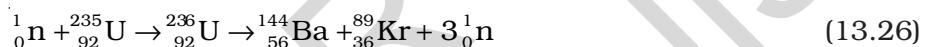
Physics

electron volts. On the other hand, in a nuclear reaction, the energy release is of the order of MeV. Thus for the same quantity of matter, nuclear sources produce a million times more energy than a chemical source. Fission of 1 kg of uranium, for example, generates 10^{14} J of energy; compare it with burning of 1 kg of coal that gives 10^7 J.

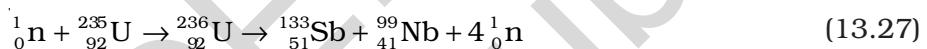
13.7.1 Fission

New possibilities emerge when we go beyond natural radioactive decays and study nuclear reactions by bombarding nuclei with other nuclear particles such as proton, neutron, α -particle, etc.

A most important neutron-induced nuclear reaction is fission. An example of fission is when a uranium isotope $^{235}_{92}\text{U}$ bombarded with a neutron breaks into two intermediate mass nuclear fragments



The same reaction can produce other pairs of intermediate mass fragments



Or, as another example,



The fragment products are radioactive nuclei; they emit β particles in succession to achieve stable end products.

The energy released (the Q value) in the fission reaction of nuclei like uranium is of the order of 200 MeV per fissioning nucleus. This is estimated as follows:

Let us take a nucleus with $A = 240$ breaking into two fragments each of $A = 120$. Then

E_{bn} for $A = 240$ nucleus is about 7.6 MeV,

E_{bn} for the two $A = 120$ fragment nuclei is about 8.5 MeV.

\therefore Gain in binding energy for nucleon is about 0.9 MeV.

Hence the total gain in binding energy is 240×0.9 or 216 MeV.

The disintegration energy in fission events first appears as the kinetic energy of the fragments and neutrons. Eventually it is transferred to the surrounding matter appearing as heat. The source of energy in nuclear reactors, which produce electricity, is nuclear fission. The enormous energy released in an atom bomb comes from uncontrolled nuclear fission. We discuss some details in the next section how a nuclear reactor functions.

13.7.2 Nuclear reactor

Notice one fact of great importance in the fission reactions given in Eqs. (13.26) to (13.28). There is a release of *extra* neutron (s) in the fission process. Averagely, $2\frac{1}{2}$ neutrons are released per fission of uranium nucleus. It is a fraction since in some fission events 2 neutrons are

INDIA'S ATOMIC ENERGY PROGRAMME

The atomic energy programme in India was launched around the time of independence under the leadership of Homi J. Bhabha (1909-1966). An early historic achievement was the design and construction of the first nuclear reactor in India (named Apsara) which went critical on August 4, 1956. It used enriched uranium as fuel and water as moderator. Following this was another notable landmark: the construction of CIRUS (Canada India Research U.S.) reactor in 1960. This 40 MW reactor used natural uranium as fuel and heavy water as moderator. Apsara and CIRUS spurred research in a wide range of areas of basic and applied nuclear science. An important milestone in the first two decades of the programme was the indigenous design and construction of the plutonium plant at Trombay, which ushered in the technology of fuel reprocessing (separating useful fissile and fertile nuclear materials from the spent fuel of a reactor) in India. Research reactors that have been subsequently commissioned include ZERLINA, PURNIMA (I, II and III), DHRUVA and KAMINI. KAMINI is the country's first large research reactor that uses U-233 as fuel. As the name suggests, the primary objective of a research reactor is not generation of power but to provide a facility for research on different aspects of nuclear science and technology. Research reactors are also an excellent source for production of a variety of radioactive isotopes that find application in diverse fields: industry, medicine and agriculture.

The main objectives of the Indian Atomic Energy programme are to provide safe and reliable electric power for the country's social and economic progress and to be self-reliant in all aspects of nuclear technology. Exploration of atomic minerals in India undertaken since the early fifties has indicated that India has limited reserves of uranium, but fairly abundant reserves of thorium. Accordingly, our country has adopted a three-stage strategy of nuclear power generation. The first stage involves the use of natural uranium as a fuel, with heavy water as moderator. The Plutonium-239 obtained from reprocessing of the discharged fuel from the reactors then serves as a fuel for the second stage — the fast breeder reactors. They are so called because they use fast neutrons for sustaining the chain reaction (hence no moderator is needed) and, besides generating power, also breed more fissile species (plutonium) than they consume. The third stage, most significant in the long term, involves using fast breeder reactors to produce fissile Uranium-233 from Thorium-232 and to build power reactors based on them.

India is currently well into the second stage of the programme and considerable work has also been done on the third — the thorium utilisation — stage. The country has mastered the complex technologies of mineral exploration and mining, fuel fabrication, heavy water production, reactor design, construction and operation, fuel reprocessing, etc. Pressurised Heavy Water Reactors (PHWRs) built at different sites in the country mark the accomplishment of the first stage of the programme. India is now more than self-sufficient in heavy water production. Elaborate safety measures both in the design and operation of reactors, as also adhering to stringent standards of radiological protection are the hallmark of the Indian Atomic Energy Programme.

produced, in some 3, etc. The extra neutrons in turn can initiate fission processes, producing still more neutrons, and so on. This leads to the possibility of a chain reaction, as was first suggested by Enrico Fermi. If the chain reaction is controlled suitably, we can get a steady energy

Physics

PHYSICS

output. This is what happens in a nuclear reactor. If the chain reaction is uncontrolled, it leads to explosive energy output, as in a nuclear bomb.

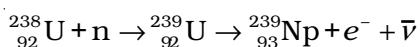
There is, however, a hurdle in sustaining a chain reaction, as described here. It is known experimentally that slow neutrons (thermal neutrons) are much more likely to cause fission in $^{235}_{92}\text{U}$ than fast neutrons. Also fast neutrons liberated in fission would escape instead of causing another fission reaction.

The average energy of a neutron produced in fission of $^{235}_{92}\text{U}$ is 2 MeV. These neutrons unless slowed down will escape from the reactor without interacting with the uranium nuclei, unless a very large amount of fissionable material is used for sustaining the chain reaction. What one needs to do is to slow down the fast neutrons by elastic scattering with light nuclei. In fact, Chadwick's experiments showed that in an elastic collision with hydrogen the neutron almost comes to rest and proton carries away the energy. This is the same situation as when a marble hits head-on an identical marble at rest. Therefore, in reactors, light nuclei called *moderators* are provided along with the fissionable nuclei for slowing down fast neutrons. The moderators commonly used are water, heavy water (D_2O) and graphite. The Apsara reactor at the Bhabha Atomic Research Centre (BARC), Mumbai, uses water as moderator. The other Indian reactors, which are used for power production, use heavy water as moderator.

Because of the use of moderator, it is possible that the ratio, K , of number of fission produced by a given generation of neutrons to the number of fission of the preceding generation may be greater than one. This ratio is called the *multiplication factor*; it is the measure of the growth rate of the neutrons in the reactor. For $K=1$, the operation of the reactor is said to be *critical*, which is what we wish it to be for steady power operation. If K becomes greater than one, the reaction rate and the reactor power increases exponentially. Unless the factor K is brought down very close to unity, the reactor will become supercritical and can even explode. The explosion of the Chernobyl reactor in Ukraine in 1986 is a sad reminder that accidents in a nuclear reactor can be catastrophic.

The reaction rate is controlled through control-rods made out of neutron-absorbing material such as cadmium. In addition to control rods, reactors are provided with *safety rods* which, when required, can be inserted into the reactor and K can be reduced rapidly to less than unity.

The more abundant isotope $^{238}_{92}\text{U}$ in naturally occurring uranium is non-fissionable. When it captures a neutron, it produces the highly radioactive plutonium through these reactions



Plutonium undergoes fission with slow neutrons.

Figure 13.5 shows the schematic diagram of a nuclear reactor based on thermal neutron fission. The *core* of the reactor is the site of nuclear

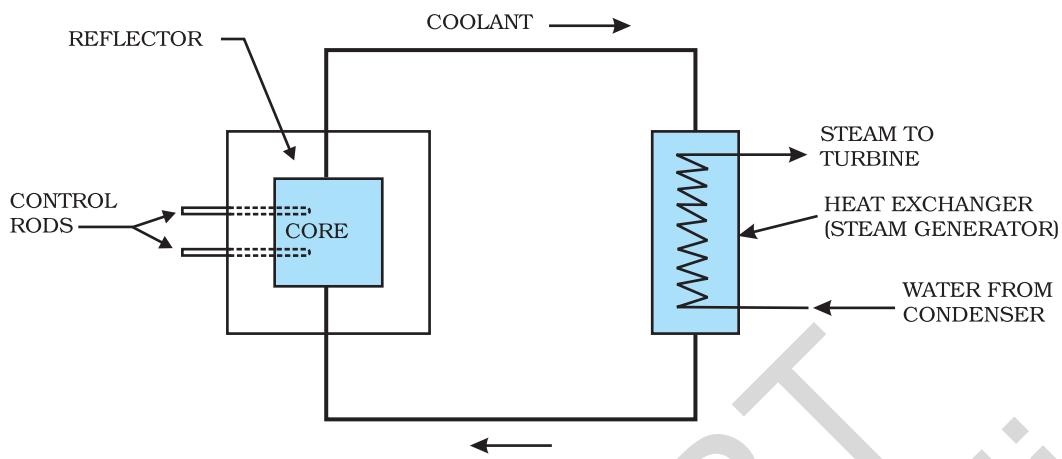


FIGURE 13.5 Schematic diagram of a nuclear reactor based on thermal neutron fission.

fission. It contains the fuel elements in suitably fabricated form. The fuel may be say enriched uranium (i.e., one that has greater abundance of $^{235}_{92}\text{U}$ than naturally occurring uranium). The core contains a moderator to slow down the neutrons. The core is surrounded by a reflector to reduce leakage. The energy (heat) released in fission is continuously removed by a suitable coolant. A containment vessel prevents the escape of radioactive fission products. The whole assembly is shielded to check harmful radiation from coming out. The reactor can be shut down by means of rods (made of, for example, cadmium) that have high absorption of neutrons. The coolant transfers heat to a working fluid which in turn may produce steam. The steam drives turbines and generates electricity.

Like any power reactor, nuclear reactors generate considerable waste products. But nuclear wastes need special care for treatment since they are radioactive and hazardous. Elaborate safety measures, both for reactor operation as well as handling and reprocessing the spent fuel, are required. These safety measures are a distinguishing feature of the Indian Atomic Energy programme. An appropriate plan is being evolved to study the possibility of converting radioactive waste into less active and short-lived material.

13.7.3 Nuclear fusion – energy generation in stars

When two light nuclei fuse to form a larger nucleus, energy is released, since the larger nucleus is more tightly bound, as seen from the binding energy curve in Fig. 13.1. Some examples of such energy liberating nuclear fusion reactions are :



Physics

In the first reaction, two protons combine to form a deuteron and a positron with a release of 0.42 MeV energy. In reaction [13.29(b)], two deuterons combine to form the light isotope of helium. In reaction (13.29c), two deuterons combine to form a triton and a proton. For fusion to take place, the two nuclei must come close enough so that attractive short-range nuclear force is able to affect them. However, since they are both positively charged particles, they experience coulomb repulsion. They, therefore, must have enough energy to overcome this coulomb barrier. The height of the barrier depends on the charges and radii of the two interacting nuclei. It can be shown, for example, that the barrier height for two protons is ~ 400 keV, and is higher for nuclei with higher charges. We can estimate the temperature at which two protons in a proton gas would (averagely) have enough energy to overcome the coulomb barrier:

$$(3/2)k T = K \square 400 \text{ keV}, \text{ which gives } T \sim 3 \times 10^9 \text{ K.}$$

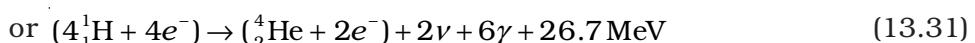
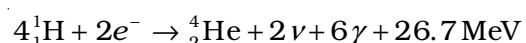
When fusion is achieved by raising the temperature of the system so that particles have enough kinetic energy to overcome the coulomb repulsive behaviour, it is called *thermonuclear fusion*.

Thermonuclear fusion is the source of energy output in the interior of stars. The interior of the sun has a temperature of 1.5×10^7 K, which is considerably less than the estimated temperature required for fusion of particles of average energy. Clearly, fusion in the sun involves protons whose energies are much above the average energy.

The fusion reaction in the sun is a multi-step process in which the hydrogen is burned into helium. Thus, the fuel in the sun is the hydrogen in its core. The *proton-proton (p, p) cycle* by which this occurs is represented by the following sets of reactions:



For the fourth reaction to occur, the first three reactions must occur twice, in which case two light helium nuclei unite to form ordinary helium nucleus. If we consider the combination 2(i) + 2(ii) + 2(iii) +(iv), the net effect is



Thus, four hydrogen atoms combine to form an ${}_2^4\text{He}$ atom with a release of 26.7 MeV of energy.

Helium is not the only element that can be synthesized in the interior of a star. As the hydrogen in the core gets depleted and becomes helium, the core starts to cool. The star begins to collapse under its own gravity

which increases the temperature of the core. If this temperature increases to about 10^8 K, fusion takes place again, this time of helium nuclei into carbon. This kind of process can generate through fusion higher and higher mass number elements. But elements more massive than those near the peak of the binding energy curve in Fig. 13.1 cannot be so produced.

The age of the sun is about 5×10^9 y and it is estimated that there is enough hydrogen in the sun to keep it going for another 5 billion years. After that, the hydrogen burning will stop and the sun will begin to cool and will start to collapse under gravity, which will raise the core temperature. The outer envelope of the sun will expand, turning it into the so called *red giant*.

NUCLEAR HOLOCAUST

In a single uranium fission about 0.9×235 MeV (≈ 200 MeV) of energy is liberated. If each nucleus of about 50 kg of ^{235}U undergoes fission the amount of energy involved is about 4×10^{15} J. This energy is equivalent to about 20,000 tons of TNT, enough for a superexplosion. Uncontrolled release of large nuclear energy is called an atomic explosion. On August 6, 1945 an atomic device was used in warfare for the first time. The US dropped an atom bomb on Hiroshima, Japan. The explosion was equivalent to 20,000 tons of TNT. Instantly the radioactive products devastated 10 sq km of the city which had 3,43,000 inhabitants. Of this number 66,000 were killed and 69,000 were injured; more than 67% of the city's structures were destroyed.

High temperature conditions for fusion reactions can be created by exploding a fission bomb. Super-explosions equivalent to 10 megatons of explosive power of TNT were tested in 1954. Such bombs which involve fusion of isotopes of hydrogen, deuterium and tritium are called hydrogen bombs. It is estimated that a nuclear arsenal sufficient to destroy every form of life on this planet several times over is in position to be triggered by the press of a button. Such a nuclear holocaust will not only destroy the life that exists now but its radioactive fallout will make this planet unfit for life for all times. Scenarios based on theoretical calculations predict a long *nuclear winter*, as the radioactive waste will hang like a cloud in the earth's atmosphere and will absorb the sun's radiation.

13.7.4 Controlled thermonuclear fusion

The natural thermonuclear fusion process in a star is replicated in a thermonuclear fusion device. In controlled fusion reactors, the aim is to generate steady power by heating the nuclear fuel to a temperature in the range of 10^8 K. At these temperatures, the fuel is a mixture of positive ions and electrons (plasma). The challenge is to confine this plasma, since no container can stand such a high temperature. Several countries around the world including India are developing techniques in this connection. If successful, fusion reactors will hopefully supply almost unlimited power to humanity.

Example 13.7 Answer the following questions:

- Are the equations of nuclear reactions (such as those given in Section 13.7) ‘balanced’ in the sense a chemical equation (e.g., $2\text{H}_2 + \text{O}_2 \rightarrow 2\text{H}_2\text{O}$) is? If not, in what sense are they balanced on both sides?
- If both the number of protons and the number of neutrons are conserved in each nuclear reaction, in what way is mass converted into energy (or vice-versa) in a nuclear reaction?
- A general impression exists that mass-energy interconversion takes place only in nuclear reaction and never in chemical reaction. This is strictly speaking, incorrect. Explain.

Solution

- A chemical equation is balanced in the sense that the number of atoms of each element is the same on both sides of the equation. A chemical reaction merely alters the original combinations of atoms. In a nuclear reaction, elements may be transmuted. Thus, the number of atoms of each element is not necessarily conserved in a nuclear reaction. However, the number of protons and the number of neutrons are both separately conserved in a nuclear reaction. [Actually, even this is not strictly true in the realm of very high energies – what is strictly conserved is the total charge and total ‘baryon number’. We need not pursue this matter here.] In nuclear reactions (e.g., Eq. 13.26), the number of protons and the number of neutrons are the same on the two sides of the equation.
- We know that the binding energy of a nucleus gives a negative contribution to the mass of the nucleus (mass defect). Now, since proton number and neutron number are conserved in a nuclear reaction, the total rest mass of neutrons and protons is the same on either side of a reaction. But the total binding energy of nuclei on the left side need not be the same as that on the right hand side. The difference in these binding energies appears as energy released or absorbed in a nuclear reaction. Since binding energy contributes to mass, we say that the difference in the total mass of nuclei on the two sides get converted into energy or vice-versa. It is in these sense that a nuclear reaction is an example of mass-energy interconversion.
- From the point of view of mass-energy interconversion, a chemical reaction is similar to a nuclear reaction *in principle*. The energy released or absorbed in a chemical reaction can be traced to the difference in chemical (not nuclear) binding energies of atoms and molecules on the two sides of a reaction. Since, strictly speaking, chemical binding energy also gives a negative contribution (mass defect) to the total mass of an atom or molecule, we can equally well say that the difference in the total mass of atoms or molecules, on the two sides of the chemical reaction gets converted into energy or vice-versa. However, the mass defects involved in a chemical reaction are almost a million times smaller than those in a nuclear reaction. This is the reason for the general impression, (which is *incorrect*) that mass-energy interconversion does not take place in a chemical reaction.

SUMMARY

- An atom has a nucleus. The nucleus is positively charged. The radius of the nucleus is smaller than the radius of an atom by a factor of 10^4 . More than 99.9% mass of the atom is concentrated in the nucleus.
- On the atomic scale, mass is measured in atomic mass units (u). By definition, 1 atomic mass unit (1u) is $1/12^{\text{th}}$ mass of one atom of ^{12}C ; $1\text{u} = 1.660563 \times 10^{-27} \text{ kg}$.
- A nucleus contains a neutral particle called neutron. Its mass is almost the same as that of proton
- The atomic number Z is the number of protons in the atomic nucleus of an element. The mass number A is the total number of protons and neutrons in the atomic nucleus; $A = Z+N$; Here N denotes the number of neutrons in the nucleus.

A nuclear species or a nuclide is represented as ${}^A_Z\text{X}$, where X is the chemical symbol of the species.

Nuclides with the same atomic number Z , but different neutron number N are called *isotopes*. Nuclides with the same A are *isobars* and those with the same N are *isotones*.

Most elements are mixtures of two or more isotopes. The atomic mass of an element is a weighted average of the masses of its isotopes. The masses are the relative abundances of the isotopes.

- A nucleus can be considered to be spherical in shape and assigned a radius. Electron scattering experiments allow determination of the nuclear radius; it is found that radii of nuclei fit the formula

$$R = R_0 A^{1/3},$$

where R_0 = a constant = 1.2 fm. This implies that the nuclear density is independent of A . It is of the order of 10^{17} kg/m^3 .

- Neutrons and protons are bound in a nucleus by the short-range strong nuclear force. The nuclear force does not distinguish between neutron and proton.
- The nuclear mass M is always less than the total mass, Σm , of its constituents. The difference in mass of a nucleus and its constituents is called the *mass defect*,

$$\Delta M = (Z m_p + (A - Z)m_n) - M$$

Using Einstein's mass energy relation, we express this mass difference in terms of energy as

$$\Delta E_b = \Delta M c^2$$

The energy ΔE_b represents the *binding energy* of the nucleus. In the mass number range $A = 30$ to 170 , the binding energy per nucleon is nearly constant, about 8 MeV/nucleon.

- Energies associated with nuclear processes are about a million times larger than chemical process.
- The Q -value of a nuclear process is

$$Q = \text{final kinetic energy} - \text{initial kinetic energy}.$$

Due to conservation of mass-energy, this is also,

$$Q = (\text{sum of initial masses} - \text{sum of final masses})c^2$$

- Radioactivity is the phenomenon in which nuclei of a given species transform by giving out α or β or γ rays; α -rays are helium nuclei;

Physics

β -rays are electrons. γ -rays are electromagnetic radiation of wavelengths shorter than X-rays;

11. Law of radioactive decay : $N(t) = N(0) e^{-\lambda t}$
where λ is the decay constant or disintegration constant.
The half-life $T_{1/2}$ of a radionuclide is the time in which N has been reduced to one-half of its initial value. The mean life τ is the time at which N has been reduced to e^{-1} of its initial value

$$T_{1/2} = \frac{\ln 2}{\lambda} = \tau \ln 2$$
12. Energy is released when less tightly bound nuclei are transmuted into more tightly bound nuclei. In fission, a heavy nucleus like $^{235}_{92}\text{U}$ breaks into two smaller fragments, e.g., $^{235}_{92}\text{U} + {}^1_0\text{n} \rightarrow {}^{133}_{51}\text{Sb} + {}^{99}_{41}\text{Nb} + 4 {}^1_0\text{n}$
13. The fact that more neutrons are produced in fission than are consumed gives the possibility of a chain reaction with each neutron that is produced triggering another fission. The chain reaction is uncontrolled and rapid in a nuclear bomb explosion. It is controlled and steady in a nuclear reactor. In a reactor, the value of the neutron multiplication factor k is maintained at 1.
14. In fusion, lighter nuclei combine to form a larger nucleus. Fusion of hydrogen nuclei into helium nuclei is the source of energy of all stars including our sun.

Physical Quantity	Symbol	Dimensions	Units	Remarks
Atomic mass unit		[M]	u	Unit of mass for expressing atomic or nuclear masses. One atomic mass unit equals $1/12^{\text{th}}$ of the mass of ^{12}C atom.
Disintegration or decay constant	λ	$[\text{T}^{-1}]$	s^{-1}	
Half-life	$T_{1/2}$	$[\text{T}]$	s	Time taken for the decay of one-half of the initial number of nuclei present in a radioactive sample.
Mean life	τ	$[\text{T}]$	s	Time at which number of nuclei has been reduced to e^{-1} of its initial value
Activity of a radioactive sample	R	$[\text{T}^{-1}]$	Bq	Measure of the activity of a radioactive source.

POINTS TO PONDER

1. The density of nuclear matter is independent of the size of the nucleus. The mass density of the atom does not follow this rule.
2. The radius of a nucleus determined by electron scattering is found to be slightly different from that determined by alpha-particle scattering.

Nuclei

This is because electron scattering senses the charge distribution of the nucleus, whereas alpha and similar particles sense the nuclear matter.

3. After Einstein showed the equivalence of mass and energy, $E = mc^2$, we cannot any longer speak of separate laws of conservation of mass and conservation of energy, but we have to speak of a unified law of conservation of mass and energy. The most convincing evidence that this principle operates in nature comes from nuclear physics. It is central to our understanding of nuclear energy and harnessing it as a source of power. Using the principle, Q of a nuclear process (decay or reaction) can be expressed also in terms of initial and final masses.
4. The nature of the binding energy (per nucleon) curve shows that exothermic nuclear reactions are possible, when two light nuclei fuse or when a heavy nucleus undergoes fission into nuclei with intermediate mass.
5. For fusion, the light nuclei must have sufficient initial energy to overcome the coulomb potential barrier. That is why fusion requires very high temperatures.
6. Although the binding energy (per nucleon) curve is smooth and slowly varying, it shows peaks at nuclides like ${}^4\text{He}$, ${}^{16}\text{O}$ etc. This is considered as evidence of atom-like shell structure in nuclei.
7. Electrons and positron are a particle-antiparticle pair. They are identical in mass; their charges are equal in magnitude and opposite. (It is found that when an electron and a positron come together, they annihilate each other giving energy in the form of gamma-ray photons.)
8. In α -decay (electron emission), the particle emitted along with electron is anti-neutrino ($\bar{\nu}$). On the other hand, the particle emitted in β^+ -decay (positron emission) is neutrino (ν). Neutrino and anti-neutrino are a particle-antiparticle pair. There are anti particles associated with every particle. What should be antiproton which is the anti particle of the proton?
9. A free neutron is unstable ($n \rightarrow p + e^- + \bar{\nu}$). But a similar free proton decay is not possible, since a proton is (slightly) lighter than a neutron.
10. Gamma emission usually follows alpha or beta emission. A nucleus in an excited (higher) state goes to a lower state by emitting a gamma photon. A nucleus may be left in an excited state after alpha or beta emission. Successive emission of gamma rays from the same nucleus (as in case of ${}^{60}\text{Ni}$, Fig. 13.4) is a clear proof that nuclei also have discrete energy levels as do the atoms.
11. Radioactivity is an indication of the instability of nuclei. Stability requires the ratio of neutron to proton to be around 1:1 for light nuclei. This ratio increases to about 3:2 for heavy nuclei. (More neutrons are required to overcome the effect of repulsion among the protons.) Nuclei which are away from the stability ratio, i.e., nuclei which have an excess of neutrons or protons are unstable. In fact, only about 10% of known isotopes (of all elements), are stable. Others have been either artificially produced in the laboratory by bombarding α , p , d , n or other particles on targets of stable nuclear species or identified in astronomical observations of matter in the universe.

EXERCISES

You may find the following data useful in solving the exercises:

$$e = 1.6 \times 10^{-19} \text{ C} \quad N = 6.023 \times 10^{23} \text{ per mole}$$

$$1/(4\pi\epsilon_0) = 9 \times 10^9 \text{ N m}^2/\text{C}^2 \quad k = 1.381 \times 10^{-23} \text{ J}^0 \text{ K}^{-1}$$

$$1 \text{ MeV} = 1.6 \times 10^{-13} \text{ J} \quad 1 \text{ u} = 931.5 \text{ MeV}/c^2$$

$$1 \text{ year} = 3.154 \times 10^7 \text{ s}$$

$$m_{\text{H}} = 1.007825 \text{ u} \quad m_{\text{n}} = 1.008665 \text{ u}$$

$$m({}_2^4\text{He}) = 4.002603 \text{ u} \quad m_{\text{e}} = 0.000548 \text{ u}$$

- 13.1** (a) Two stable isotopes of lithium ${}^6_3\text{Li}$ and ${}^7_3\text{Li}$ have respective abundances of 7.5% and 92.5%. These isotopes have masses 6.01512 u and 7.01600 u, respectively. Find the atomic mass of lithium.
- (b) Boron has two stable isotopes, ${}^{10}_5\text{B}$ and ${}^{11}_5\text{B}$. Their respective masses are 10.01294 u and 11.00931 u, and the atomic mass of boron is 10.811 u. Find the abundances of ${}^{10}_5\text{B}$ and ${}^{11}_5\text{B}$.
- 13.2** The three stable isotopes of neon: ${}^{20}_{10}\text{Ne}$, ${}^{21}_{10}\text{Ne}$ and ${}^{22}_{10}\text{Ne}$ have respective abundances of 90.51%, 0.27% and 9.22%. The atomic masses of the three isotopes are 19.99 u, 20.99 u and 21.99 u, respectively. Obtain the average atomic mass of neon.
- 13.3** Obtain the binding energy (in MeV) of a nitrogen nucleus (${}^{14}_7\text{N}$), given $m({}^{14}_7\text{N}) = 14.00307 \text{ u}$
- 13.4** Obtain the binding energy of the nuclei ${}^{56}_{26}\text{Fe}$ and ${}^{209}_{83}\text{Bi}$ in units of MeV from the following data:
 $m({}^{56}_{26}\text{Fe}) = 55.934939 \text{ u}$ $m({}^{209}_{83}\text{Bi}) = 208.980388 \text{ u}$
- 13.5** A given coin has a mass of 3.0 g. Calculate the nuclear energy that would be required to separate all the neutrons and protons from each other. For simplicity assume that the coin is entirely made of ${}^{63}_{29}\text{Cu}$ atoms (of mass 62.92960 u).
- 13.6** Write nuclear reaction equations for
 (i) α -decay of ${}^{226}_{88}\text{Ra}$ (ii) α -decay of ${}^{242}_{94}\text{Pu}$
 (iii) β^- -decay of ${}^{32}_{15}\text{P}$ (iv) β^- -decay of ${}^{210}_{83}\text{Bi}$
 (v) β^+ -decay of ${}^{11}_{6}\text{C}$ (vi) β^+ -decay of ${}^{97}_{43}\text{Tc}$
 (vii) Electron capture of ${}^{120}_{54}\text{Xe}$
- 13.7** A radioactive isotope has a half-life of T years. How long will it take the activity to reduce to a) 3.125%, b) 1% of its original value?
- 13.8** The normal activity of living carbon-containing matter is found to be about 15 decays per minute for every gram of carbon. This activity arises from the small proportion of radioactive ${}^{14}_{6}\text{C}$ present with the stable carbon isotope ${}^{12}_{6}\text{C}$. When the organism is dead, its interaction with the atmosphere (which maintains the above equilibrium activity) ceases and its activity begins to drop. From the known half-life (5730 years) of ${}^{14}_{6}\text{C}$, and the measured activity, the age of the specimen can be approximately estimated. This is the principle of ${}^{14}_{6}\text{C}$ dating

used in archaeology. Suppose a specimen from Mohenjodaro gives an activity of 9 decays per minute per gram of carbon. Estimate the approximate age of the Indus-Valley civilisation.

- 13.9** Obtain the amount of $^{60}_{27}\text{Co}$ necessary to provide a radioactive source of 8.0 mCi strength. The half-life of $^{60}_{27}\text{Co}$ is 5.3 years.

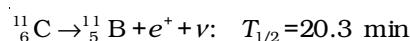
- 13.10** The half-life of $^{90}_{38}\text{Sr}$ is 28 years. What is the disintegration rate of 15 mg of this isotope?

- 13.11** Obtain approximately the ratio of the nuclear radii of the gold isotope $^{197}_{79}\text{Au}$ and the silver isotope $^{107}_{47}\text{Ag}$.

- 13.12** Find the Q-value and the kinetic energy of the emitted α -particle in the α -decay of (a) $^{226}_{88}\text{Ra}$ and (b) $^{220}_{86}\text{Rn}$.

$$\text{Given } m(^{226}_{88}\text{Ra}) = 226.02540 \text{ u}, \quad m(^{222}_{86}\text{Rn}) = 222.01750 \text{ u}, \\ m(^{222}_{86}\text{Rn}) = 220.01137 \text{ u}, \quad m(^{216}_{84}\text{Po}) = 216.00189 \text{ u}.$$

- 13.13** The radionuclide $^{11}_6\text{C}$ decays according to



The maximum energy of the emitted positron is 0.960 MeV.

Given the mass values:

$$m(^{11}_6\text{C}) = 11.011434 \text{ u} \text{ and } m(^{11}_5\text{B}) = 11.009305 \text{ u},$$

calculate Q and compare it with the maximum energy of the positron emitted.

- 13.14** The nucleus $^{23}_{10}\text{Ne}$ decays by β^- emission. Write down the β -decay equation and determine the maximum kinetic energy of the electrons emitted. Given that:

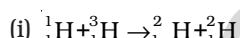
$$m(^{23}_{10}\text{Ne}) = 22.994466 \text{ u}$$

$$m(^{23}_{11}\text{Na}) = 22.089770 \text{ u}.$$

- 13.15** The Q value of a nuclear reaction $A + b \rightarrow C + d$ is defined by

$$Q = [m_A + m_b - m_C - m_d]c^2$$

where the masses refer to the respective nuclei. Determine from the given data the Q-value of the following reactions and state whether the reactions are exothermic or endothermic.



Atomic masses are given to be

$$m(^2_1\text{H}) = 2.014102 \text{ u}$$

$$m(^3_1\text{H}) = 3.016049 \text{ u}$$

$$m(^{12}_6\text{C}) = 12.000000 \text{ u}$$

$$m(^{20}_{10}\text{Ne}) = 19.992439 \text{ u}$$

- 13.16** Suppose, we think of fission of a $^{56}_{26}\text{Fe}$ nucleus into two equal fragments, $^{28}_{13}\text{Al}$. Is the fission energetically possible? Argue by working out Q of the process. Given $m(^{56}_{26}\text{Fe}) = 55.93494 \text{ u}$ and $m(^{28}_{13}\text{Al}) = 27.98191 \text{ u}$.

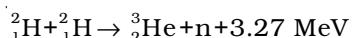
- 13.17** The fission properties of $^{239}_{94}\text{Pu}$ are very similar to those of $^{235}_{92}\text{U}$. The average energy released per fission is 180 MeV. How much energy,

Physics

in MeV, is released if all the atoms in 1 kg of pure $^{239}_{94}\text{Pu}$ undergo fission?

- 13.18** A 1000 MW fission reactor consumes half of its fuel in 5.00 y. How much $^{235}_{92}\text{U}$ did it contain initially? Assume that the reactor operates 80% of the time, that all the energy generated arises from the fission of $^{235}_{92}\text{U}$ and that this nuclide is consumed only by the fission process.

- 13.19** How long can an electric lamp of 100W be kept glowing by fusion of 2.0 kg of deuterium? Take the fusion reaction as



- 13.20** Calculate the height of the potential barrier for a head on collision of two deuterons. (Hint: The height of the potential barrier is given by the Coulomb repulsion between the two deuterons when they just touch each other. Assume that they can be taken as hard spheres of radius 2.0 fm.)

- 13.21** From the relation $R = R_0 A^{1/3}$, where R_0 is a constant and A is the mass number of a nucleus, show that the nuclear matter density is nearly constant (i.e. independent of A).

- 13.22** For the β^+ (positron) emission from a nucleus, there is another competing process known as electron capture (electron from an inner orbit, say, the K-shell, is captured by the nucleus and a neutrino is emitted).



Show that if β^+ emission is energetically allowed, electron capture is necessarily allowed but not vice-versa.

ADDITIONAL EXERCISES

- 13.23** In a periodic table the average atomic mass of magnesium is given as 24.312 u. The average value is based on their relative natural abundance on earth. The three isotopes and their masses are $^{24}_{12}\text{Mg}$ (23.98504u), $^{25}_{12}\text{Mg}$ (24.98584u) and $^{26}_{12}\text{Mg}$ (25.98259u). The natural abundance of $^{24}_{12}\text{Mg}$ is 78.99% by mass. Calculate the abundances of other two isotopes.

- 13.24** The neutron separation energy is defined as the energy required to remove a neutron from the nucleus. Obtain the neutron separation energies of the nuclei ${}^{41}_{20}\text{Ca}$ and ${}^{27}_{13}\text{Al}$ from the following data:

$$m({}^{40}_{20}\text{Ca}) = 39.962591 \text{ u}$$

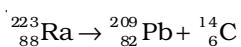
$$m({}^{41}_{20}\text{Ca}) = 40.962278 \text{ u}$$

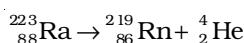
$$m({}^{26}_{13}\text{Al}) = 25.986895 \text{ u}$$

$$m({}^{27}_{13}\text{Al}) = 26.981541 \text{ u}$$

- 13.25** A source contains two phosphorous radio nuclides ${}^{32}_{15}\text{P}$ ($T_{1/2} = 14.3\text{d}$) and ${}^{33}_{15}\text{P}$ ($T_{1/2} = 25.3\text{d}$). Initially, 10% of the decays come from ${}^{33}_{15}\text{P}$. How long one must wait until 90% do so?

- 13.26** Under certain circumstances, a nucleus can decay by emitting a particle more massive than an α -particle. Consider the following decay processes:





Calculate the Q -values for these decays and determine that both are energetically allowed.

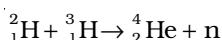
- 13.27** Consider the fission of $^{238}_{92}\text{U}$ by fast neutrons. In one fission event, no neutrons are emitted and the final end products, after the beta decay of the primary fragments, are $^{140}_{58}\text{Ce}$ and $^{99}_{44}\text{Ru}$. Calculate Q for this fission process. The relevant atomic and particle masses are

$$m(^{238}_{92}\text{U}) = 238.05079 \text{ u}$$

$$m(^{140}_{58}\text{Ce}) = 139.90543 \text{ u}$$

$$m(^{99}_{44}\text{Ru}) = 98.90594 \text{ u}$$

- 13.28** Consider the D-T reaction (deuterium-tritium fusion)



- (a) Calculate the energy released in MeV in this reaction from the data:

$$m(^2_1\text{H}) = 2.014102 \text{ u}$$

$$m(^3_1\text{H}) = 3.016049 \text{ u}$$

- (b) Consider the radius of both deuterium and tritium to be approximately 2.0 fm. What is the kinetic energy needed to overcome the coulomb repulsion between the two nuclei? To what temperature must the gas be heated to initiate the reaction?

(Hint: Kinetic energy required for one fusion event = average thermal kinetic energy available with the interacting particles = $2(3kT/2)$; k = Boltzman's constant, T = absolute temperature.)

- 13.29** Obtain the maximum kinetic energy of β -particles, and the radiation frequencies of γ decays in the decay scheme shown in Fig. 13.6. You are given that

$$m(^{198}\text{Au}) = 197.968233 \text{ u}$$

$$m(^{198}\text{Hg}) = 197.966760 \text{ u}$$

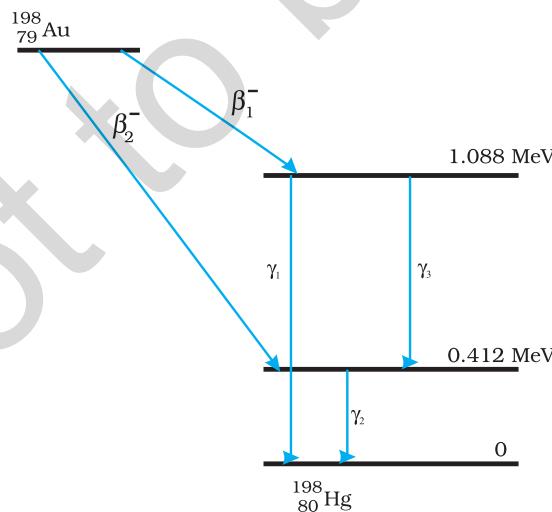


FIGURE 13.6

■ Physics

- 13.30** Calculate and compare the energy released by a) fusion of 1.0 kg of hydrogen deep within Sun and b) the fission of 1.0 kg of ^{235}U in a fission reactor.
- 13.31** Suppose India had a target of producing by 2020 AD, 200,000 MW of electric power, ten percent of which was to be obtained from nuclear power plants. Suppose we are given that, on an average, the efficiency of utilization (i.e. conversion to electric energy) of thermal energy produced in a reactor was 25%. How much amount of fissionable uranium would our country need per year by 2020? Take the heat energy per fission of ^{235}U to be about 200MeV.

Chapter Fourteen

SEMICONDUCTOR

ELECTRONICS:

MATERIALS, DEVICES

AND SIMPLE CIRCUITS

14.1 INTRODUCTION

Devices in which a controlled flow of electrons can be obtained are the basic *building blocks* of all the electronic circuits. Before the discovery of transistor in 1948, such devices were mostly vacuum tubes (also called valves) like the vacuum diode which has two electrodes, viz., anode (often called plate) and cathode; triode which has three electrodes – cathode, plate and grid; tetrode and pentode (respectively with 4 and 5 electrodes). In a vacuum tube, the electrons are supplied by a heated cathode and the controlled flow of these electrons *in vacuum* is obtained by varying the voltage between its different electrodes. Vacuum is required in the inter-electrode space; otherwise the moving electrons may lose their energy on collision with the air molecules in their path. In these devices the electrons can flow only from the cathode to the anode (i.e., only in one direction). Therefore, such devices are generally referred to as *valves*. These vacuum tube devices are bulky, consume high power, operate generally at high voltages (~ 100 V) and have limited life and low reliability. The seed of the development of modern *solid-state semiconductor electronics* goes back to 1930's when it was realised that some solid-state semiconductors and their junctions offer the possibility of controlling the number and the direction of flow of charge carriers through them. Simple excitations like light, heat or small applied voltage can change the number of mobile charges in a semiconductor. Note that the supply

Physics

and flow of charge carriers in the semiconductor devices are *within the solid itself*, while in the earlier vacuum tubes/valves, the mobile electrons were obtained from a heated cathode and they were made to flow in an *evacuated* space or vacuum. No external heating or large evacuated space is required by the semiconductor devices. They are small in size, consume low power, operate at low voltages and have long life and high reliability. Even the Cathode Ray Tubes (CRT) used in television and computer monitors which work on the principle of vacuum tubes are being replaced by Liquid Crystal Display (LCD) monitors with supporting solid state electronics. Much before the full implications of the semiconductor devices was formally understood, a naturally occurring crystal of *galena* (Lead sulphide, PbS) with a metal point contact attached to it was used as *detector* of radio waves.

In the following sections, we will introduce the basic concepts of semiconductor physics and discuss some semiconductor devices like junction diodes (a 2-electrode device) and bipolar junction transistor (a 3-electrode device). A few circuits illustrating their applications will also be described.

14.2 CLASSIFICATION OF METALS, CONDUCTORS AND SEMICONDUCTORS

On the basis of conductivity

On the basis of the relative values of electrical conductivity (σ) or resistivity ($\rho = 1/\sigma$), the solids are broadly classified as:

(i) **Metals:** They possess very low resistivity (or high conductivity).

$$\rho \sim 10^{-2} - 10^{-8} \Omega \text{ m}$$

$$\sigma \sim 10^2 - 10^8 \text{ S m}^{-1}$$

(ii) **Semiconductors:** They have resistivity or conductivity intermediate to metals and insulators.

$$\rho \sim 10^{-5} - 10^6 \Omega \text{ m}$$

$$\sigma \sim 10^5 - 10^{-6} \text{ S m}^{-1}$$

(iii) **Insulators:** They have high resistivity (or low conductivity).

$$\rho \sim 10^{11} - 10^{19} \Omega \text{ m}$$

$$\sigma \sim 10^{-11} - 10^{-19} \text{ S m}^{-1}$$

The values of ρ and σ given above are indicative of magnitude and could well go outside the ranges as well. Relative values of the resistivity are not the only criteria for distinguishing metals, insulators and semiconductors from each other. There are some other differences, which will become clear as we go along in this chapter.

Our interest in this chapter is in the study of semiconductors which could be:

(i) **Elemental semiconductors:** Si and Ge

(ii) **Compound semiconductors:** Examples are:

- Inorganic: CdS, GaAs, CdSe, InP, etc.

- Organic: anthracene, doped phthalocyanines, etc.

- Organic polymers: polypyrrole, polyaniline, polythiophene, etc.

Most of the currently available semiconductor devices are based on elemental semiconductors Si or Ge and compound *inorganic*

semiconductors. However, after 1990, a few semiconductor devices using organic semiconductors and semiconducting polymers have been developed signalling the birth of a futuristic technology of polymer-electronics and molecular-electronics. In this chapter, we will restrict ourselves to the study of inorganic semiconductors, particularly elemental semiconductors Si and Ge. The general concepts introduced here for discussing the elemental semiconductors, by-and-large, apply to most of the compound semiconductors as well.

On the basis of energy bands

According to the Bohr atomic model, in an *isolated atom* the energy of any of its electrons is decided by the orbit in which it revolves. But when the atoms come together to form a solid they are close to each other. So the outer orbits of electrons from neighbouring atoms would come very close or could even overlap. This would make the nature of electron motion in a solid very different from that in an isolated atom.

Inside the crystal each electron has a unique position and no two electrons see exactly the same pattern of surrounding charges. Because of this, each electron will have a different *energy level*. These different energy levels with continuous energy variation form what are called *energy bands*. The energy band which includes the energy levels of the valence electrons is called the *valence band*. The energy band above the valence band is called the *conduction band*. With no external energy, all the valence electrons will reside in the valence band. If the lowest level in the conduction band happens to be lower than the highest level of the valence band, the electrons from the valence band can easily move into the conduction band. Normally the conduction band is empty. But when it overlaps on the valence band electrons can move freely into it. This is the case with metallic conductors.

If there is some gap between the conduction band and the valence band, electrons in the valence band all remain bound and no free electrons are available in the conduction band. This makes the material an insulator. But some of the electrons from the valence band may gain external energy to cross the gap between the conduction band and the valence band. Then these electrons will move into the conduction band. At the same time they will create vacant energy levels in the valence band where other valence electrons can move. Thus the process creates the possibility of conduction due to electrons in conduction band as well as due to vacancies in the valence band.

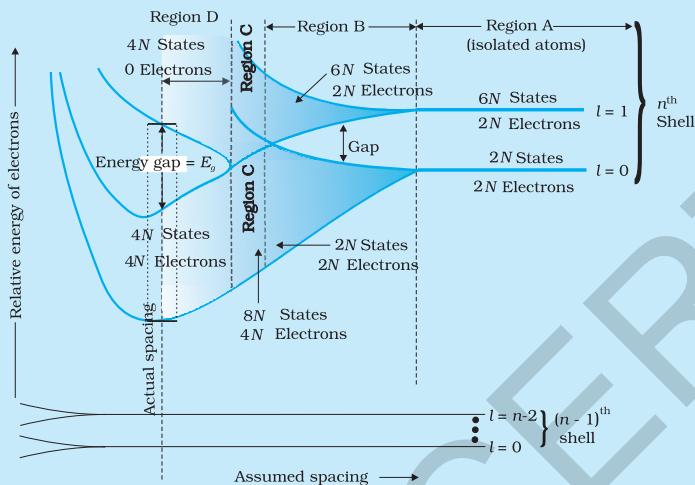
Let us consider what happens in the case of Si or Ge crystal containing N atoms. For Si, the outermost orbit is the third orbit ($n = 3$), while for Ge it is the fourth orbit ($n = 4$). The number of electrons in the outermost orbit is 4 (2s and 2p electrons). Hence, the total number of outer electrons in the crystal is $4N$. The maximum possible number of electrons in the outer orbit is 8 (2s + 6p electrons). So, for the $4N$ valence electrons there are $8N$ available energy states. These $8N$ discrete energy levels can either form a continuous band or they may be grouped in different bands depending upon the distance between the atoms in the crystal (see box on Band Theory of Solids).

At the distance between the atoms in the crystal lattices of Si and Ge, the energy band of these $8N$ states is split apart into two which are separated by an *energy gap* E_g (Fig. 14.1). The lower band which is

Physics

completely occupied by the $4N$ valence electrons at temperature of absolute zero is the *valence band*. The other band consisting of $4N$ energy states, called the *conduction band*, is completely empty at absolute zero.

BAND THEORY OF SOLIDS



Consider that the Si or Ge crystal contains N atoms. Electrons of each atom will have discrete energies in different orbits. The electron energy will be same if all the atoms are *isolated*, i.e., separated from each other by a large distance. However, in a crystal, the atoms are close to each other (2 to 3 Å) and therefore the electrons interact with each other and also with the neighbouring atomic cores. The overlap (or interaction) will be more felt by the electrons in the outermost orbit while the inner orbit or core electron energies may remain unaffected. Therefore, for understanding electron energies in Si or Ge crystal, we need to consider the changes in the energies of the electrons in the outermost orbit only. For Si, the outermost orbit is the third orbit ($n = 3$), while for Ge it is the fourth orbit ($n = 4$). The number of electrons in the outermost orbit is 4 (2s and 2p electrons). Hence, the total number of outer electrons in the crystal is $4N$. The maximum possible number of outer electrons in the orbit is 8 (2 s + 6 p electrons). So, out of the $4N$ electrons, $2N$ electrons are in the *2Ns-states* (orbital quantum number $l = 0$) and $2N$ electrons are in the available *6Np-states*. Obviously, some p-electron states are empty as shown in the extreme right of Figure. This is the case of well separated or isolated atoms [region A of Figure].

Suppose these atoms start coming nearer to each other to form a solid. The energies of these electrons in the outermost orbit may change (both increase and decrease) due to the interaction between the electrons of different atoms. The $6N$ states for $l = 1$, which originally had identical energies in the isolated atoms, spread out and form an *energy band* [region B in Figure]. Similarly, the $2N$ states for $l = 0$, having identical energies in the isolated atoms, split into a second band (carefully see the region B of Figure) separated from the first one by an *energy gap*.

At still smaller spacing, however, there comes a region in which the bands merge with each other. The lowest energy state that is a split from the upper atomic level appears to drop below the upper state that has come from the lower atomic level. In this region (region C in Figure), *no energy gap exists where the upper and lower energy states get mixed*.

Finally, if the distance between the atoms further decreases, the energy bands again split apart and are separated by an *energy gap* E_g (region D in Figure). The total number of available energy states $8N$ has been *re-apportioned* between the two bands ($4N$ states each in the lower and upper energy bands). Here the significant point is that there are exactly as many states in the lower band ($4N$) as there are available valence electrons from the atoms ($4N$).

Therefore, this band (called the *valence band*) is completely filled while the upper band is completely empty. The upper band is called the *conduction band*.

The lowest energy level in the conduction band is shown as E_c and highest energy level in the valence band is shown as E_v . Above E_c and below E_v there are a large number of closely spaced energy levels, as shown in Fig. 14.1.

The gap between the top of the valence band and bottom of the conduction band is called the *energy band gap* (Energy gap E_g). It may be large, small, or zero, depending upon the material. These different situations, are depicted in Fig. 14.2 and discussed below:

Case I: This refers to a situation, as shown in Fig. 14.2(a). One can have a metal either when the conduction band is partially filled and the balanced band is partially empty or when the conduction and valance bands overlap. When there is overlap electrons from valence band can easily move into the conduction band. This situation makes a large number of electrons available for electrical conduction. When the valence band is partially empty, electrons from its lower level can move to higher level making conduction possible. Therefore, the resistance of such materials is low or the conductivity is high.

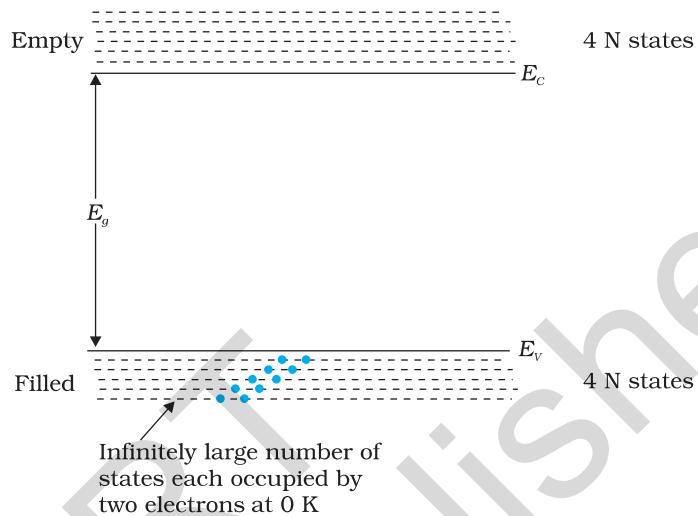


FIGURE 14.1 The energy band positions in a semiconductor at 0 K. The upper band, called the conduction band, consists of infinitely large number of closely spaced energy states. The lower band, called the valence band, consists of closely spaced completely filled energy states.

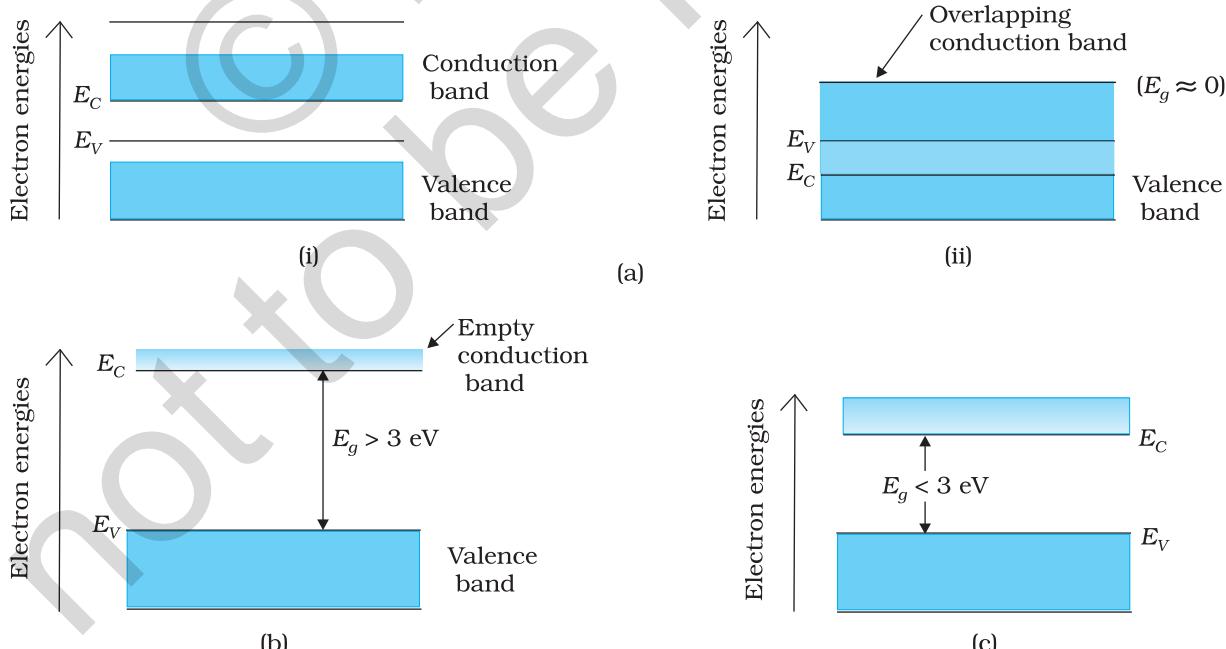


FIGURE 14.2 Difference between energy bands of (a) metals, (b) insulators and (c) semiconductors.

Case II: In this case, as shown in Fig. 14.2(b), a large band gap E_g exists ($E_g > 3$ eV). There are no electrons in the conduction band, and therefore no electrical conduction is possible. Note that the energy gap is so large that electrons cannot be excited from the valence band to the conduction band by thermal excitation. This is the case of *insulators*.

Case III: This situation is shown in Fig. 14.2(c). Here a finite but small band gap ($E_g < 3$ eV) exists. Because of the small band gap, at room temperature some electrons from valence band can acquire enough energy to cross the energy gap and enter the *conduction band*. These electrons (though small in numbers) can move in the conduction band. Hence, the resistance of *semiconductors* is not as high as that of the insulators.

In this section we have made a broad classification of metals, conductors and semiconductors. In the section which follows you will learn the conduction process in semiconductors.

14.3 INTRINSIC SEMICONDUCTOR

We shall take the most common case of Ge and Si whose lattice structure is shown in Fig. 14.3. These structures are called the diamond-like structures. Each atom is surrounded by four nearest neighbours. We know that Si and Ge have four valence electrons. In its crystalline structure, every Si or Ge atom tends to *share* one of its four valence electrons with each of its four nearest neighbour atoms, and also to *take share* of one electron from each such neighbour. These shared electron pairs are referred to as forming a *covalent bond* or simply a *valence bond*. The two shared electrons can be assumed to shuttle back-and-forth between the associated atoms holding them together strongly. Figure 14.4 schematically shows the 2-dimensional representation of Si or Ge structure shown in Fig. 14.3 which overemphasises the covalent bond. It shows an idealised picture in which no bonds are broken (all bonds are intact).

Such a situation arises at low temperatures. As the temperature increases, more thermal energy becomes available to these electrons and some of these electrons may break-away (becoming *free electrons* contributing to conduction). The thermal energy effectively ionises only a few atoms in the crystalline lattice and creates a *vacancy* in the bond as shown in Fig. 14.5(a). The neighbourhood, from which the free electron (with charge $-q$) has come out leaves a vacancy with an effective charge ($+q$). This *vacancy* with the effective positive electronic charge is called a *hole*. The hole behaves as an *apparent free particle* with effective positive charge.

In intrinsic semiconductors, the number of free electrons, n_e is equal to the number of holes, n_h . That is

$$n_e = n_h = n_i \quad (14.1)$$

where n_i is called *intrinsic carrier concentration*.

Semiconductors possess the unique property in which, apart from electrons, the holes also move.

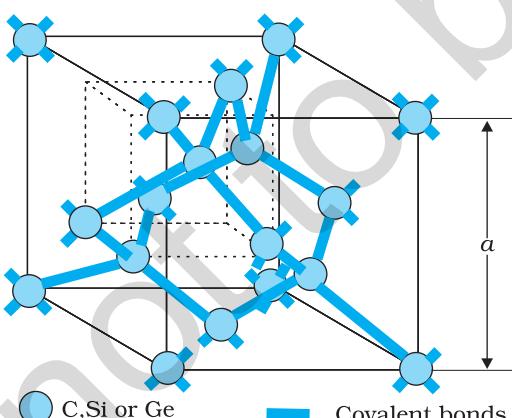


FIGURE 14.3 Three-dimensional diamond-like crystal structure for Carbon, Silicon or Germanium with respective lattice spacing a equal to 3.56, 5.43 and 5.66 Å.

Suppose there is a hole at site 1 as shown in Fig. 14.5(a). The movement of holes can be visualised as shown in Fig. 14.5(b). An electron from the covalent bond at site 2 may jump to the vacant site 1 (hole). Thus, after such a jump, the hole is at site 2 and the site 1 has now an electron. Therefore, apparently, the hole has moved from site 1 to site 2. Note that the electron originally set free [Fig. 14.5(a)] is not involved in this process of hole motion. The free electron moves completely independently as conduction electron and gives rise to an electron current, I_e under an applied electric field. Remember that the motion of hole is only a convenient way of describing the actual motion of *bound* electrons, whenever there is an empty bond anywhere in the crystal. Under the action of an electric field, these holes move towards negative potential giving the hole current, I_h . The total current, I is thus the sum of the electron current I_e and the hole current I_h :

$$I = I_e + I_h \quad (14.2)$$

It may be noted that apart from the process of generation of conduction electrons and holes, a simultaneous process of recombination occurs in which the electrons *recombine* with the holes. At equilibrium, the rate of generation is equal to the rate of recombination of charge carriers. The recombination occurs due to an electron colliding with a hole.

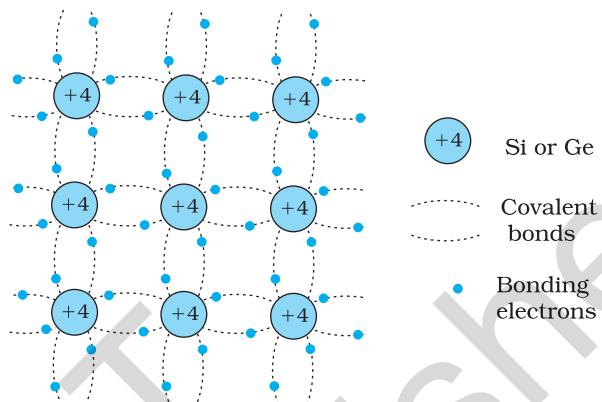
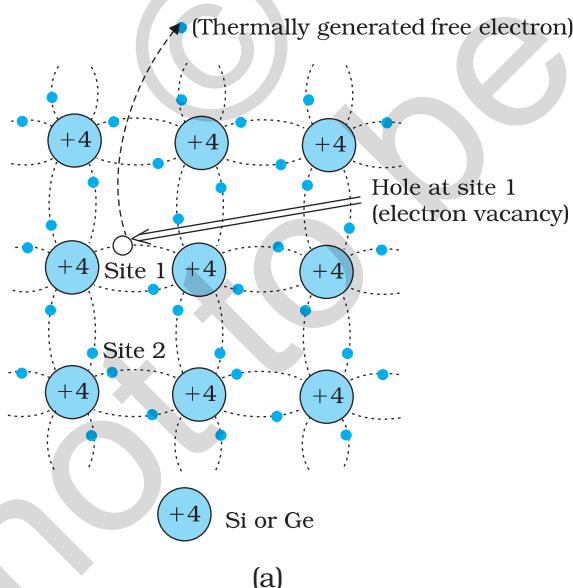
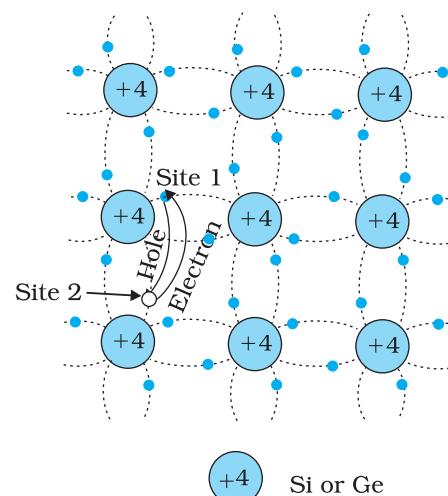


FIGURE 14.4 Schematic two-dimensional representation of Si or Ge structure showing covalent bonds at low temperature (all bonds intact). +4 symbol indicates inner cores of Si or Ge.



(a)



(b)

FIGURE 14.5 (a) Schematic model of generation of hole at site 1 and conduction electron due to thermal energy at moderate temperatures. (b) Simplified representation of possible thermal motion of a hole. The electron from the lower left hand covalent bond (site 2) goes to the earlier hole site 1, leaving a hole at its site indicating an apparent movement of the hole from site 1 to site 2.

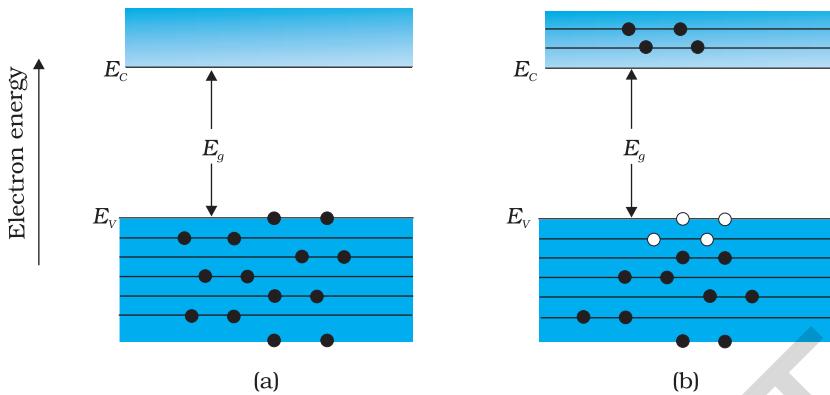


FIGURE 14.6 (a) An intrinsic semiconductor at $T = 0\text{ K}$ behaves like insulator. (b) At $T > 0\text{ K}$, four thermally generated electron-hole pairs. The filled circles (●) represent electrons and empty fields (○) represent holes.

An intrinsic semiconductor will behave like an insulator at $T = 0\text{ K}$ as shown in Fig. 14.6(a). It is the thermal energy at higher temperatures ($T > 0\text{ K}$), which excites some electrons from the valence band to the conduction band. These thermally excited electrons at $T > 0\text{ K}$, partially occupy the conduction band. Therefore, the energy-band diagram of an intrinsic semiconductor will be as shown in Fig. 14.6(b). Here, some electrons are shown in the conduction band. These have come from the valence band leaving equal number of holes there.

EXAMPLE 14.1

Example 14.1 C, Si and Ge have same lattice structure. Why is C insulator while Si and Ge intrinsic semiconductors?

Solution The 4 bonding electrons of C, Si or Ge lie, respectively, in the second, third and fourth orbit. Hence, energy required to take out an electron from these atoms (i.e., ionisation energy E_g) will be least for Ge, followed by Si and highest for C. Hence, number of free electrons for conduction in Ge and Si are significant but negligibly small for C.

14.4 EXTRINSIC SEMICONDUCTOR

The conductivity of an intrinsic semiconductor depends on its temperature, but at room temperature its conductivity is very low. As such, no important electronic devices can be developed using these semiconductors. Hence there is a necessity of improving their conductivity. This can be done by making use of impurities.

When a small amount, say, a few parts per million (ppm), of a suitable impurity is added to the pure semiconductor, the conductivity of the semiconductor is increased manifold. Such materials are known as *extrinsic semiconductors* or *impurity semiconductors*. The deliberate addition of a desirable impurity is called *doping* and the impurity atoms are called *dopants*. Such a material is also called a *doped semiconductor*. The dopant has to be such that it does not distort the original pure semiconductor lattice. It occupies only a very few of the original semiconductor atom sites in the crystal. A necessary condition to attain this is that the sizes of the dopant and the semiconductor atoms should be nearly the same.

There are two types of dopants used in doping the tetravalent Si or Ge:

- Pentavalent (valency 5); like Arsenic (As), Antimony (Sb), Phosphorous (P), etc.

- (ii) Trivalent (valency 3); like Indium (In), Boron (B), Aluminium (Al), etc.

We shall now discuss how the doping changes the number of charge carriers (and hence the conductivity) of semiconductors. Si or Ge belongs to the fourth group in the Periodic table and, therefore, we choose the dopant element from nearby fifth or third group, expecting and taking care that the size of the dopant atom is nearly the same as that of Si or Ge. Interestingly, the pentavalent and trivalent dopants in Si or Ge give two entirely different types of semiconductors as discussed below.

(i) n-type semiconductor

Suppose we dope Si or Ge with a pentavalent element as shown in Fig. 14.7. When an atom of +5 valency element occupies the position of an atom in the crystal lattice of Si, four of its electrons bond with the four silicon neighbours while the fifth remains very weakly bound to its parent atom. This is because the four electrons participating in bonding are seen as part of the effective core of the atom by the fifth electron. As a result the ionisation energy required to set this electron free is very small and even at room temperature it will be free to move in the lattice of the semiconductor. For example, the energy required is ~ 0.01 eV for germanium, and 0.05 eV for silicon, to separate this electron from its atom. This is in contrast to the energy required to jump the forbidden band (about 0.72 eV for germanium and about 1.1 eV for silicon) at room temperature in the intrinsic semiconductor. Thus, the pentavalent dopant is donating one extra electron for conduction and hence is known as *donor impurity*. The number of electrons made available for conduction by dopant atoms depends strongly upon the doping level and is independent of any increase in ambient temperature. On the other hand, the number of free electrons (with an equal number of holes) generated by Si atoms, increases weakly with temperature.

In a doped semiconductor the total number of conduction electrons n_e is due to the electrons contributed by donors and those generated intrinsically, while the total number of holes n_h is only due to the holes from the intrinsic source. But the rate of recombination of holes would increase due to the increase in the number of electrons. As a result, the number of holes would get reduced further.

Thus, with proper level of doping the number of conduction electrons can be made much larger than the number of holes. Hence in an extrinsic

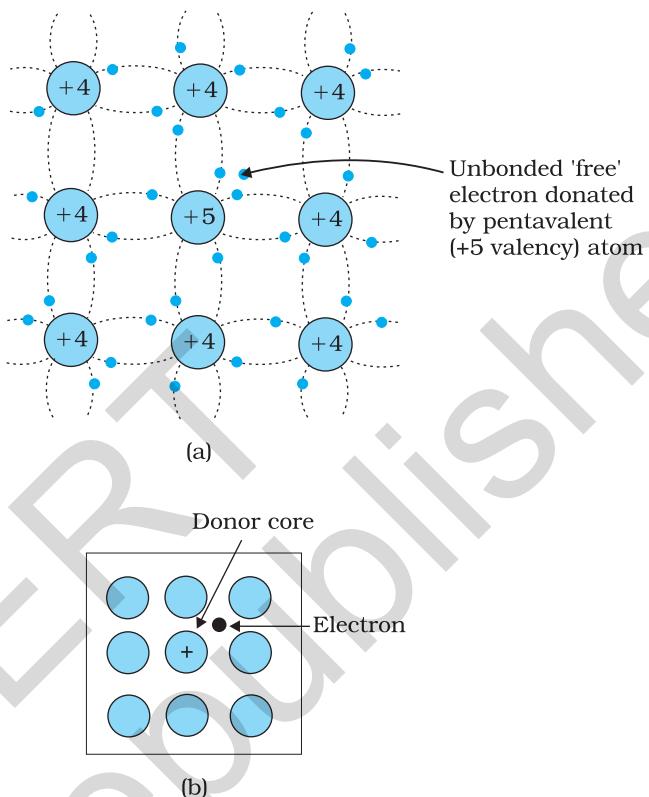
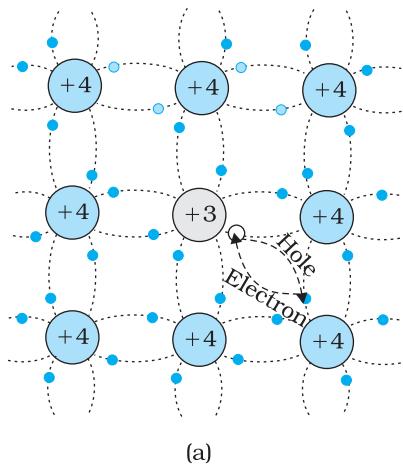
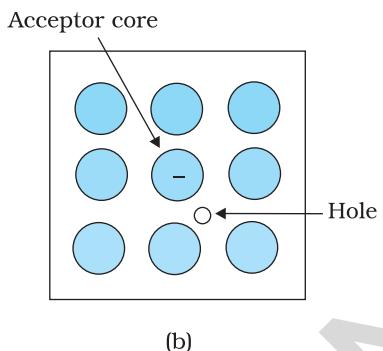


FIGURE 14.7 (a) Pentavalent donor atom (As, Sb, P, etc.) doped for tetravalent Si or Ge giving n-type semiconductor, and (b) Commonly used schematic representation of n-type material which shows only the fixed cores of the substituent donors with one additional effective positive charge and its associated extra electron.



(a)



(b)

FIGURE 14.8 (a) Trivalent acceptor atom (In, Al, B etc.) doped in tetravalent Si or Ge lattice giving p-type semiconductor. (b) Commonly used schematic representation of p-type material which shows only the fixed core of the substituent acceptor with one effective additional negative charge and its associated hole.

semiconductor doped with pentavalent impurity, electrons become the *majority carriers* and holes the *minority carriers*. These semiconductors are, therefore, known as *n-type semiconductors*. For n-type semiconductors, we have,

$$n_e \gg n_h \quad (14.3)$$

(ii) p-type semiconductor

This is obtained when Si or Ge is doped with a trivalent impurity like Al, B, In, etc. The dopant has one valence electron less than Si or Ge and, therefore, this atom can form covalent bonds with neighbouring three Si atoms but does not have any electron to offer to the fourth Si atom. So the bond between the fourth neighbour and the trivalent atom has a vacancy or hole as shown in Fig. 14.8. Since the neighbouring Si atom in the lattice wants an electron in place of a hole, an electron in the outer orbit of an atom in the neighbourhood may jump to fill this vacancy, leaving a vacancy or hole at its own site. Thus the *hole* is available for conduction. Note that the trivalent foreign atom becomes effectively negatively charged when it shares fourth electron with neighbouring Si atom. Therefore, the dopant atom of p-type material can be treated as *core of one negative charge* along with its associated hole as shown in Fig. 14.8(b). It is obvious that one acceptor atom gives one hole. These holes are in addition to the intrinsically generated holes while the source of conduction electrons is only intrinsic generation. Thus, for such a material, the holes are the majority carriers and electrons are minority carriers. Therefore, extrinsic semiconductors doped with trivalent impurity are called *p-type semiconductors*. For p-type semiconductors, the recombination process will reduce the number (n) of intrinsically generated electrons to n_e . We have, for p-type semiconductors

$$n_h \gg n_e \quad (14.4)$$

Note that *the crystal maintains an overall charge neutrality as the charge of additional charge carriers is just equal and opposite to that of the ionised cores in the lattice*.

In extrinsic semiconductors, because of the abundance of majority current carriers, the minority carriers produced thermally have more chance of meeting majority carriers and thus getting destroyed. Hence, the dopant, by adding a large number of current carriers of one type, which become the majority carriers, indirectly helps to reduce the intrinsic concentration of minority carriers.

The semiconductor's energy band structure is affected by doping. In the case of extrinsic semiconductors, additional energy states due to donor impurities (E_D) and acceptor impurities (E_A) also exist. In the energy band diagram of n-type Si semiconductor, the donor energy level E_D is slightly below the bottom E_C of the conduction band and electrons from this level move into the conduction band with very small supply of energy. At room temperature, most of the donor atoms get ionised but very few ($\sim 10^{-12}$) atoms of Si get ionised. So the conduction band will have most electrons coming from the donor impurities, as shown in Fig. 14.9(a). Similarly,

for p-type semiconductor, the acceptor energy level E_A is slightly above the top E_V of the valence band as shown in Fig. 14.9(b). With very small supply of energy an electron from the valence band can jump to the level E_A and ionise the acceptor negatively. (Alternately, we can also say that with very small supply of energy the hole from level E_A sinks down into the valence band. Electrons rise up and holes fall down when they gain external energy.) At room temperature, most of the acceptor atoms get ionised leaving holes in the valence band. Thus at room temperature the density of holes in the valence band is predominantly due to impurity in the extrinsic semiconductor. The electron and hole concentration in a semiconductor *in thermal equilibrium* is given by

$$n_e n_h = n_i^2 \quad (14.5)$$

Though the above description is grossly approximate and hypothetical, it helps in understanding the difference between metals, insulators and semiconductors (extrinsic and intrinsic) in a simple manner. The difference in the resistivity of C, Si and Ge depends upon the energy gap between their conduction and valence bands. For C (diamond), Si and Ge, the energy gaps are 5.4 eV, 1.1 eV and 0.7 eV, respectively. Sn also is a group IV element but it is a metal because the energy gap in its case is 0 eV.

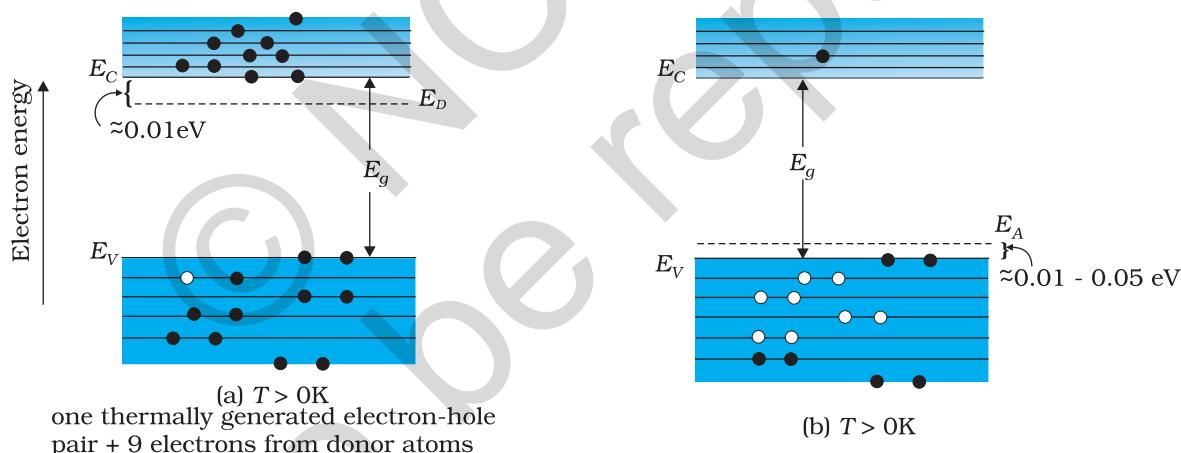


FIGURE 14.9 Energy bands of (a) n-type semiconductor at $T > 0\text{K}$, (b) p-type semiconductor at $T > 0\text{K}$.

Example 14.2 Suppose a pure Si crystal has $5 \times 10^{28} \text{ atoms m}^{-3}$. It is doped by 1 ppm concentration of pentavalent As. Calculate the number of electrons and holes. Given that $n_i = 1.5 \times 10^{16} \text{ m}^{-3}$.

Solution Note that thermally generated electrons ($n_i \sim 10^{16} \text{ m}^{-3}$) are negligibly small as compared to those produced by doping.

Therefore, $n_e \approx N_D$.

Since $n_e n_h = n_i^2$, The number of holes

$$n_h = (2.25 \times 10^{32}) / (5 \times 10^{22})$$

$$\sim 4.5 \times 10^9 \text{ m}^{-3}$$

Physics

14.5 p-n JUNCTION

A p-n junction is the basic building block of many semiconductor devices like diodes, transistor, etc. A clear understanding of the junction behaviour is important to analyse the working of other semiconductor devices. We will now try to understand how a junction is formed and how the junction behaves under the influence of external applied voltage (also called *bias*).

14.5.1 p-n junction formation

Consider a thin p-type silicon (p-Si) semiconductor wafer. By adding precisely a small quantity of pentavalent impurity, part of the p-Si wafer can be converted into n-Si. There are several processes by which a semiconductor can be formed. The wafer now contains p-region and n-region and a metallurgical junction between p-, and n- region.

Two important processes occur during the formation of a p-n junction: *diffusion* and *drift*. We know that in an n-type semiconductor, the concentration of electrons (number of electrons per unit volume) is more compared to the concentration of holes. Similarly, in a p-type semiconductor, the concentration of holes is more than the concentration of electrons. During the formation of p-n junction, and due to the concentration gradient across p-, and n- sides, holes diffuse from p-side to n-side ($p \rightarrow n$) and electrons diffuse from n-side to p-side ($n \rightarrow p$). This motion of charge carries gives rise to diffusion current across the junction.

When an electron diffuses from $n \rightarrow p$, it leaves behind an ionised donor on n-side. This ionised donor (positive charge) is immobile as it is bonded to the surrounding atoms. As the electrons continue to diffuse from $n \rightarrow p$, a layer of positive charge (or positive space-charge region) on n-side of the junction is developed.

Similarly, when a hole diffuses from $p \rightarrow n$ due to the concentration gradient, it leaves behind an ionised acceptor (negative charge) which is immobile. As the holes continue to diffuse, a layer of negative charge (or negative space-charge region) on the p-side of the junction is developed. This space-charge region on either side of the junction together is known as *depletion region* as the electrons and holes taking part in the initial

movement across the junction *depleted* the region of its free charges (Fig. 14.10). The thickness of depletion region is of the order of one-tenth of a micrometre. Due to the positive space-charge region on n-side of the junction and negative space charge region on p-side of the junction, an electric field directed from positive charge towards negative charge develops. Due to this field, an electron on p-side of the junction moves to n-side and a hole on n-side of the junction moves to p-side. The motion of charge carriers due to the electric field is called drift. Thus a drift current, which is opposite in direction to the diffusion current (Fig. 14.10) starts.

PHYSICS

Formation and working of p-n junction diode
<http://hyperphysics.phy-astr.gsu.edu/hbase/solids/pnjun.html>

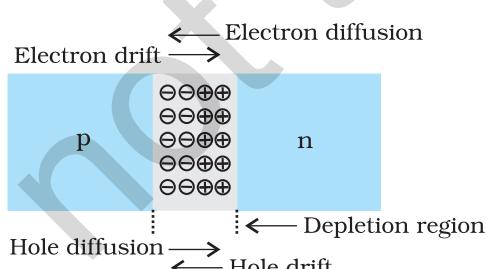


FIGURE 14.10 p-n junction formation process.

Initially, diffusion current is large and drift current is small. As the diffusion process continues, the space-charge regions on either side of the junction extend, thus increasing the electric field strength and hence drift current. This process continues until the diffusion current equals the drift current. Thus a p-n junction is formed. In a p-n junction under equilibrium there is *no net current*.

The loss of electrons from the n-region and the gain of electron by the p-region causes a difference of potential across the junction of the two regions. The polarity of this potential is such as to oppose further flow of carriers so that a condition of equilibrium exists. Figure 14.11 shows the p-n junction at equilibrium and the potential across the junction. The n-material has lost electrons, and p material has acquired electrons. The n material is thus positive relative to the p material. Since this potential tends to prevent the movement of electron from the n region into the p region, it is often called a *barrier potential*.

Example 14.3 Can we take one slab of p-type semiconductor and physically join it to another n-type semiconductor to get p-n junction?

Solution No! Any slab, howsoever flat, will have roughness much larger than the inter-atomic crystal spacing (~ 2 to 3 \AA) and hence *continuous contact* at the atomic level will not be possible. The junction will behave as a *discontinuity* for the flowing charge carriers.

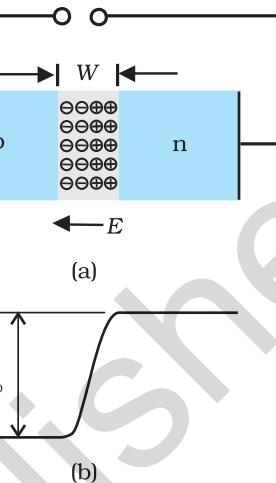


FIGURE 14.11 (a) Diode under equilibrium ($V = 0$), (b) Barrier potential under no bias.

EXAMPLE 14.3

14.6 SEMICONDUCTOR DIODE

A semiconductor diode [Fig. 14.12(a)] is basically a p-n junction with metallic contacts provided at the ends for the application of an external voltage. It is a two terminal device. A p-n junction diode is symbolically represented as shown in Fig. 14.12(b).

The direction of arrow indicates the conventional direction of current (when the diode is under forward bias). The equilibrium barrier potential can be altered by applying an external voltage V across the diode. The situation of p-n junction diode under equilibrium (without bias) is shown in Fig. 14.11(a) and (b).

14.6.1 p-n junction diode under forward bias

When an external voltage V is applied across a semiconductor diode such that p-side is connected to the positive terminal of the battery and n-side to the negative terminal [Fig. 14.13(a)], it is said to be *forward biased*.

The applied voltage mostly drops across the depletion region and the voltage drop across the p-side and n-side of the junction is negligible. (This is because the resistance of the depletion region – a region where there are no charges – is very high compared to the resistance of n-side and p-side.) The direction of the applied voltage (V) is opposite to the

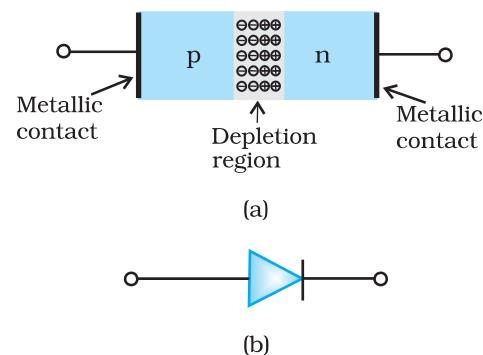


FIGURE 14.12 (a) Semiconductor diode,
(b) Symbol for p-n junction diode.

Physics

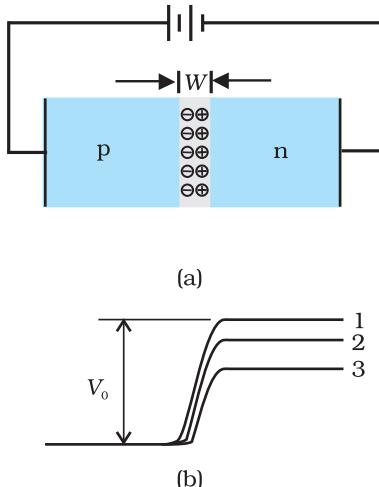


FIGURE 14.13 (a) p-n junction diode under forward bias, (b) Barrier potential (1) without battery, (2) Low battery voltage, and (3) High voltage battery.

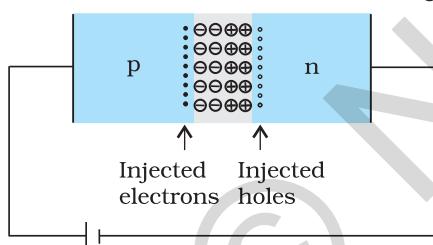


FIGURE 14.14 Forward bias minority carrier injection.

built-in potential V_0 . As a result, the depletion layer width decreases and the barrier height is reduced [Fig. 14.13(b)]. The effective barrier height under forward bias is $(V_0 - V)$.

If the applied voltage is small, the barrier potential will be reduced only slightly below the equilibrium value, and only a small number of carriers in the material—those that happen to be in the uppermost energy levels—will possess enough energy to cross the junction. So the current will be small. If we increase the applied voltage significantly, the barrier height will be reduced and more number of carriers will have the required energy. Thus the current increases.

Due to the applied voltage, electrons from n-side cross the depletion region and reach p-side (where they are minority carriers). Similarly, holes from p-side cross the junction and reach the n-side (where they are minority carriers). This process under forward bias is known as minority carrier injection. At the junction boundary, on each side, the minority carrier concentration increases significantly compared to the locations far from the junction.

Due to this concentration gradient, the injected electrons on p-side diffuse from the junction edge of p-side to the other end of p-side. Likewise, the injected holes on n-side diffuse from the junction edge of n-side to the other end of n-side (Fig. 14.14). This motion of charged carriers on either side gives rise to current. The total diode forward current is sum of hole diffusion current and conventional current due to electron diffusion. The magnitude of this current is usually in mA.

14.6.2 p-n junction diode under reverse bias

When an external voltage (V) is applied across the diode such that n-side is positive and p-side is negative, it is said to be *reverse biased* [Fig. 14.15(a)]. The applied voltage mostly drops across the depletion region. The direction of applied voltage is same as the direction of barrier potential. As a result, the barrier height increases and the depletion region widens due to the change in the electric field. The effective barrier height under reverse bias is $(V_0 + V)$, [Fig. 14.15(b)]. This suppresses the flow of electrons from $n \rightarrow p$ and holes from $p \rightarrow n$. Thus, diffusion current, decreases enormously compared to the diode under forward bias.

The electric field direction of the junction is such that if electrons on p-side or holes on n-side in their random motion come close to the junction, they will be swept to its majority zone. This drift of carriers gives rise to current. The drift current is of the order of a few μA . This is quite low because it is due to the motion of carriers from their minority side to their majority side across the junction. The drift current is also there under forward bias but it is negligible (μA) when compared with current due to injected carriers which is usually in mA.

The diode reverse current is not very much dependent on the applied voltage. Even a small voltage is sufficient to sweep the minority carriers from one side of the junction to the other side of the junction. The current

is not limited by the magnitude of the applied voltage but is limited due to the concentration of the minority carrier on either side of the junction.

The current under reverse bias is essentially voltage independent upto a critical reverse bias voltage, known as breakdown voltage (V_{br}). When $V = V_{br}$, the diode reverse current increases sharply. Even a slight increase in the bias voltage causes large change in the current. If the reverse current is not limited by an external circuit below the rated value (specified by the manufacturer) the p-n junction will get destroyed. Once it exceeds the rated value, the diode gets destroyed due to overheating. This can happen even for the diode under forward bias, if the forward current exceeds the rated value.

The circuit arrangement for studying the V - I characteristics of a diode, (i.e., the variation of current as a function of applied voltage) are shown in Fig. 14.16(a) and (b). The battery is connected to the diode through a potentiometer (or rheostat) so that the applied voltage to the diode can be changed. For different values of voltages, the value of the current is noted. A graph between V and I is obtained as in Fig. 14.16(c). Note that in forward bias measurement, we use a milliammeter since the expected current is large (as explained in the earlier section) while a micrometer is used in reverse bias to measure the current. You can see in Fig. 14.16(c) that in forward

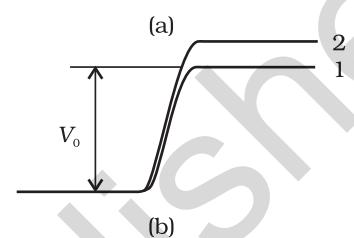
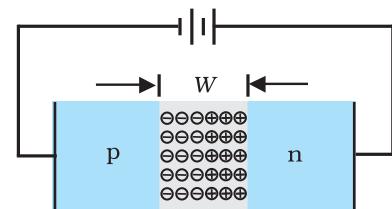


FIGURE 14.15 (a) Diode under reverse bias,
(b) Barrier potential under reverse bias.

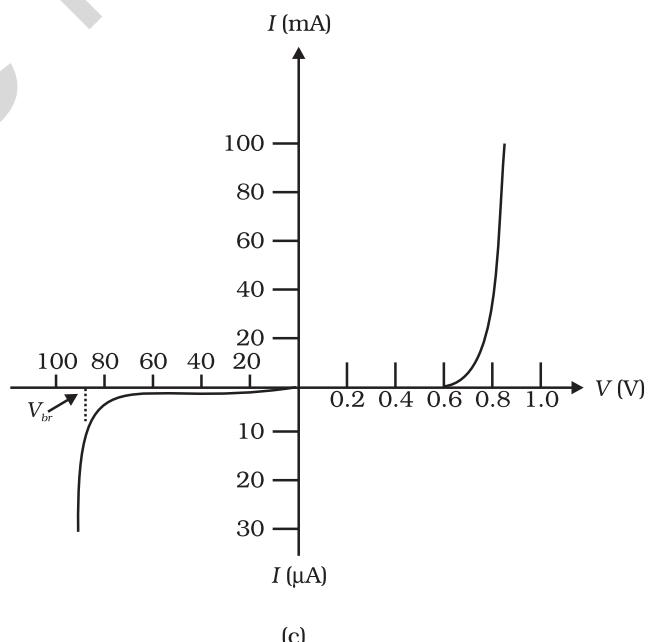
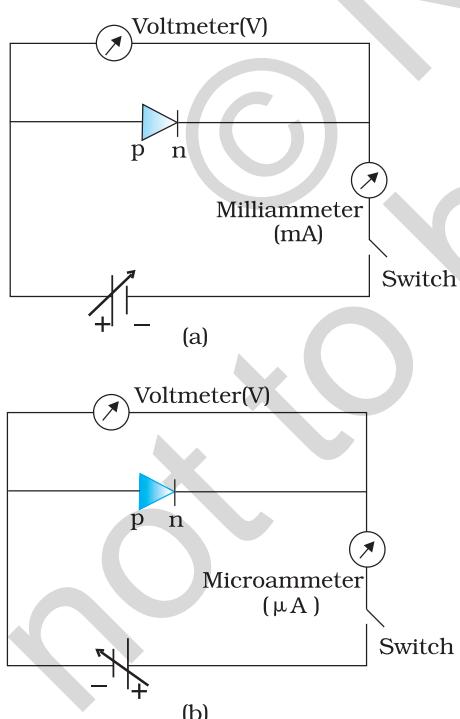


FIGURE 14.16 Experimental circuit arrangement for studying V - I characteristics of a p-n junction diode (a) in forward bias , (b) in reverse bias. (c) Typical V - I characteristics of a silicon diode.

Physics

bias, the current first increases very slowly, almost negligibly, till the voltage across the diode crosses a certain value. After the characteristic voltage, the diode current increases significantly (exponentially), even for a very small increase in the diode bias voltage. This voltage is called the *threshold voltage* or cut-in voltage ($\sim 0.2\text{V}$ for germanium diode and $\sim 0.7\text{ V}$ for silicon diode).

For the diode in reverse bias, the current is very small ($\sim \mu\text{A}$) and almost remains constant with change in bias. It is called *reverse saturation current*. However, for special cases, at very high reverse bias (break down voltage), the current suddenly increases. This special action of the diode is discussed later in Section 14.8. The general purpose diode are not used beyond the reverse saturation current region.

The above discussion shows that the p-n junction diode primerly allows the flow of current only in one direction (forward bias). The forward bias resistance is low as compared to the reverse bias resistance. This property is used for rectification of ac voltages as discussed in the next section. For diodes, we define a quantity called *dynamic resistance* as the ratio of small change in voltage ΔV to a small change in current ΔI :

$$r_d = \frac{\Delta V}{\Delta I} \quad (14.6)$$

Example 14.4 The V - I characteristic of a silicon diode is shown in the Fig. 14.17. Calculate the resistance of the diode at (a) $I_D = 15\text{ mA}$ and (b) $V_D = -10\text{ V}$.

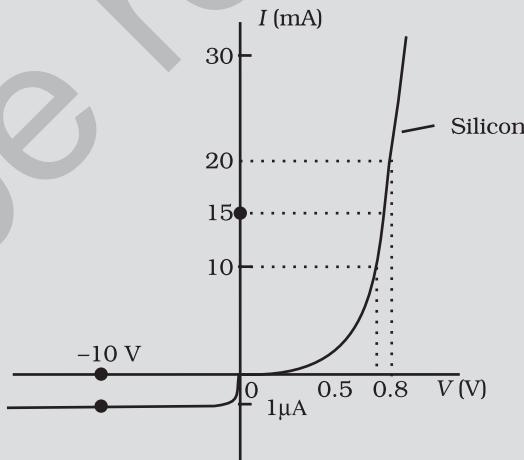


FIGURE 14.17

Solution Considering the diode characteristics as a straight line between $I = 10\text{ mA}$ to $I = 20\text{ mA}$ passing through the origin, we can calculate the resistance using Ohm's law.

- From the curve, at $I = 20\text{ mA}$, $V = 0.8\text{ V}$, $I = 10\text{ mA}$, $V = 0.7\text{ V}$
 $r_{fb} = \Delta V / \Delta I = 0.1\text{V}/10\text{ mA} = 10\Omega$
- From the curve at $V = -10\text{ V}$, $I = -1\mu\text{A}$,
Therefore,
 $r_{rb} = 10\text{ V}/1\mu\text{A} = 1.0 \times 10^7\Omega$

14.7 APPLICATION OF JUNCTION DIODE AS A RECTIFIER

From the $V-I$ -characteristic of a junction diode we see that it allows current to pass only when it is forward biased. So if an alternating voltage is applied across a diode the current flows only in that part of the cycle when the diode is forward biased. This property is used to *rectify* alternating voltages and the circuit used for this purpose is called a *rectifier*.

If an alternating voltage is applied across a diode in series with a load, a pulsating voltage will appear across the load only during the half cycles of the ac input during which the diode is forward biased. Such rectifier circuit, as shown in Fig. 14.18, is called a *half-wave rectifier*. The secondary of a transformer supplies the desired ac voltage across terminals A and B. When the voltage at A is positive, the diode is forward biased and it conducts. When A is negative, the diode is reverse-biased and it does not conduct. The reverse saturation current of a diode is negligible and can be considered equal to zero for practical purposes. (The reverse breakdown voltage of the diode must be sufficiently higher than the peak ac voltage at the secondary of the transformer to protect the diode from reverse breakdown.)

Therefore, in the positive *half-cycle* of ac there is a current through the load resistor R_L and we get an output voltage, as shown in Fig. 14.18(b), whereas there is no current in the negative half-cycle. In the next positive half-cycle, again we get the output voltage. Thus, the output voltage, though still varying, is restricted to *only one direction* and is said to be *rectified*. Since the rectified output of this circuit is only for half of the input ac wave it is called as *half-wave rectifier*.

The circuit using two diodes, shown in Fig. 14.19(a), gives output rectified voltage corresponding to both the positive as well as negative half of the ac cycle. Hence, it is known as *full-wave rectifier*. Here the p-side of the two diodes are connected to the ends of the secondary of the transformer. The n-side of the diodes are connected together and the output is taken between this common point of diodes and the midpoint of the secondary of the transformer. So for a full-wave rectifier the secondary of the transformer is provided with a centre tapping and so it is called *centre-tap transformer*. As can be seen from Fig. 14.19(c) the voltage rectified by each diode is only half the total secondary voltage. Each diode rectifies only for half the cycle, but the two do so for alternate cycles. Thus, the output between their common terminals and the centre-tap of the transformer becomes a full-wave rectifier output. (Note that there is another circuit of full wave rectifier which does not need a centre-tap transformer but needs four diodes.) Suppose the input voltage to A

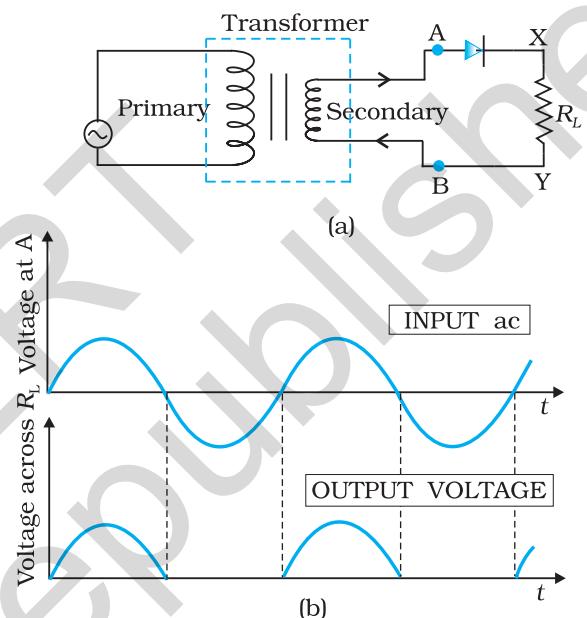


FIGURE 14.18 (a) Half-wave rectifier circuit, (b) Input ac voltage and output voltage waveforms from the rectifier circuit.

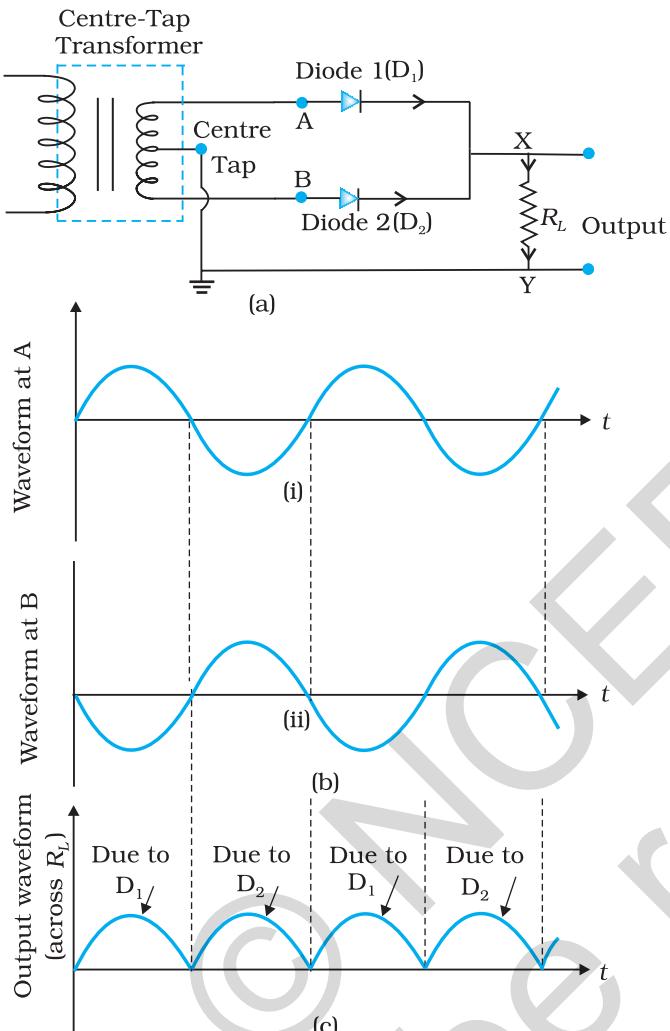


FIGURE 14.19 (a) A Full-wave rectifier circuit; (b) Input wave forms given to the diode D₁ at A and to the diode D₂ at B; (c) Output waveform across the load R_L connected in the full-wave rectifier circuit.

with respect to the centre tap at any instant is positive. It is clear that, at that instant, voltage at B being out of phase will be negative as shown in Fig. 14.19(b). So, diode D₁ gets forward biased and conducts (while D₂ being reverse biased is not conducting). Hence, during this positive half cycle we get an output current (and a output voltage across the load resistor R_L) as shown in Fig. 14.19(c). In the course of the ac cycle when the voltage at A becomes negative with respect to centre tap, the voltage at B would be positive. In this part of the cycle diode D₁ would not conduct but diode D₂ would, giving an output current and output voltage (across R_L) during the negative half cycle of the input ac. Thus, we get output voltage during both the positive as well as the negative half of the cycle. Obviously, this is a more efficient circuit for getting rectified voltage or current than the half-wave rectifier.

The rectified voltage is in the form of pulses of the shape of half sinusoids. Though it is unidirectional it does not have a steady value. To get steady dc output from the pulsating voltage normally a capacitor is connected across the output terminals (parallel to the load R_L). One can also use an inductor in series with R_L for the same purpose. Since these additional circuits appear to filter out the ac ripple and give a pure dc voltage, so they are called filters.

Now we shall discuss the role of capacitor in filtering. When the voltage across the capacitor is rising, it gets charged. If there is no external load, it remains charged to the peak voltage of the rectified output. When there is a load, it gets discharged through the load and the voltage across it begins to fall. In the next half-cycle of rectified output it again gets charged to the peak value (Fig. 14.20). The rate of fall of the voltage across the capacitor depends upon the inverse product of capacitor C and the effective resistance R_L used in the circuit and is called the *time constant*. To make the time constant large value of C should be large. So capacitor input filters use large capacitors. The *output voltage* obtained by using capacitor input filter is nearer to the *peak voltage* of the rectified voltage. This type of filter is most widely used in power supplies.

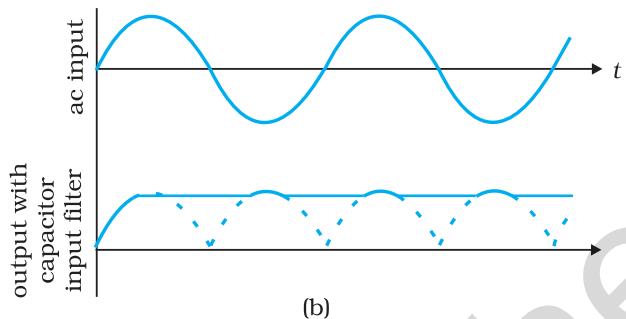
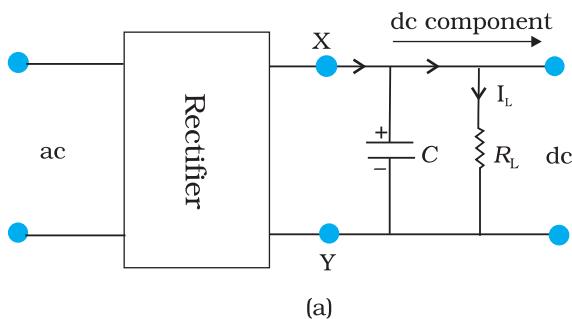


FIGURE 14.20 (a) A full-wave rectifier with capacitor filter, (b) Input and output voltage of rectifier in (a).

14.8 SPECIAL PURPOSE p-n JUNCTION DIODES

In the section, we shall discuss some devices which are basically junction diodes but are developed for different applications.

14.8.1 Zener diode

It is a special purpose semiconductor diode, named after its inventor C. Zener. It is designed to operate under reverse bias in the breakdown region and used as a voltage regulator. The symbol for Zener diode is shown in Fig. 14.21(a).

Zener diode is fabricated by heavily doping both p-, and n- sides of the junction. Due to this, depletion region formed is very thin ($<10^{-6}$ m) and the electric field of the junction is extremely high ($\sim 5 \times 10^6$ V/m) even for a small reverse bias voltage of about 5V. The I-V characteristics of a Zener diode is shown in Fig. 14.21(b). It is seen that when the applied reverse bias voltage(V) reaches the breakdown voltage (V_z) of the Zener diode, there is a large change in the current. Note that after the breakdown voltage V_z , a large change in the current can be produced by almost insignificant change in the reverse bias voltage. In other words, Zener voltage remains constant, even though current through the Zener diode varies over a wide range. This property of the Zener diode is used for regulating supply voltages so that they are constant.

Let us understand how reverse current suddenly increases at the breakdown voltage. We know that reverse current is due to the flow of electrons (minority carriers) from p \rightarrow n and holes from n \rightarrow p. As the reverse bias voltage is increased, the electric field at the junction becomes significant. When the reverse bias voltage $V = V_z$, then the electric field strength is high enough to pull valence electrons from the host atoms on the p-side which are accelerated to n-side. These electrons account for high current observed at the breakdown. The emission of electrons from the host atoms due to the high electric field is known as internal field emission or field ionisation. The electric field required for field ionisation is of the order of 10^6 V/m.

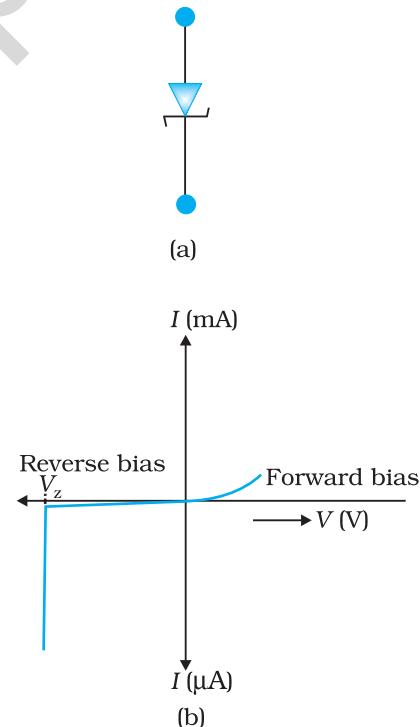


FIGURE 14.21 Zener diode,
(a) symbol, (b) I-V
characteristics.

Zener diode as a voltage regulator

We know that when the ac input voltage of a rectifier fluctuates, its rectified output also fluctuates. To get a constant dc voltage from the dc unregulated output of a rectifier, we use a Zener diode. The circuit diagram of a voltage regulator using a Zener diode is shown in Fig. 14.22.

The unregulated dc voltage (filtered output of a rectifier) is connected to the Zener diode through a series resistance R_s such that the Zener diode is reverse biased. If the input voltage increases, the current through R_s and Zener diode also increases. This increases the voltage drop across R_s without any change in the voltage across the Zener diode. This is because in the breakdown region, Zener voltage remains constant even though the current through the Zener diode changes. Similarly, if the input voltage decreases, the current through R_s and Zener diode also decreases. The voltage drop across R_s decreases without any change in the voltage across the Zener diode. Thus any increase/decrease in the input voltage results in, increase/decrease of the voltage drop across R_s without any change in voltage across the Zener diode. Thus the Zener diode acts as a voltage regulator. We have to select the Zener diode according to the required output voltage and accordingly the series resistance R_s .

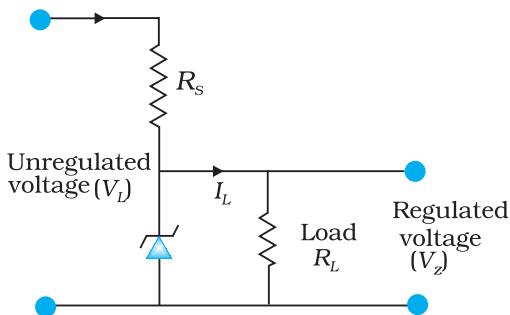


FIGURE 14.22 Zener diode as DC voltage regulator

change in voltage across the Zener diode. Thus the Zener diode acts as a voltage regulator. We have to select the Zener diode according to the required output voltage and accordingly the series resistance R_s .

Example 14.5 In a Zener regulated power supply a Zener diode with $V_Z = 6.0$ V is used for regulation. The load current is to be 4.0 mA and the unregulated input is 10.0 V. What should be the value of series resistor R_s ?

Solution

The value of R_s should be such that the current through the Zener diode is much larger than the load current. This is to have good load regulation. Choose Zener current as five times the load current, i.e., $I_Z = 20$ mA. The total current through R_s is, therefore, 24 mA. The voltage drop across R_s is $10.0 - 6.0 = 4.0$ V. This gives $R_s = 4.0\text{V}/(24 \times 10^{-3})\text{ A} = 167\Omega$. The nearest value of carbon resistor is 150Ω . So, a series resistor of 150Ω is appropriate. Note that slight variation in the value of the resistor does not matter, what is important is that the current I_z should be sufficiently larger than I_L .

EXAMPLE 14.5

14.8.2 Optoelectronic junction devices

We have seen so far, how a semiconductor diode behaves under applied electrical inputs. In this section, we learn about semiconductor diodes in which carriers are generated by photons (photo-excitation). All these devices are called *optoelectronic devices*. We shall study the functioning of the following optoelectronic devices:

- Photodiodes used for detecting optical signal (photodetectors).
- Light emitting diodes (LED) which convert electrical energy into light.
- Photovoltaic devices which convert optical radiation into electricity (solar cells).

(i) Photodiode

A Photodiode is again a special purpose p-n junction diode fabricated with a transparent window to allow light to fall on the diode. It is operated under reverse bias. When the photodiode is illuminated with light (photons) with energy ($h\nu$) greater than the energy gap (E_g) of the semiconductor, then electron-hole pairs are generated due to the absorption of photons. The diode is fabricated such that the generation of e-hpairs takes place in or near the depletion region of the diode. Due to electric field of the junction, electrons and holes are separated before they recombine. The direction of the electric field is such that electrons reach n-side and holes reach p-side. Electrons are collected on n-side and holes are collected on p-side giving rise to an emf. When an external load is connected, current flows. The magnitude of the photocurrent depends on the intensity of incident light (photocurrent is proportional to incident light intensity).

It is easier to observe the change in the current with change in the light intensity, if a reverse bias is applied. Thus photodiode can be used as a photodetector to detect optical signals. The circuit diagram used for the measurement of I-V characteristics of a photodiode is shown in Fig. 14.23(a) and a typical I-V characteristics in Fig. 14.23(b).

Example 14.6 The current in the forward bias is known to be more (~mA) than the current in the reverse bias (~ μA). What is the reason then to operate the photodiodes in reverse bias?

Solution Consider the case of an n-type semiconductor. Obviously, the majority carrier density (n) is considerably larger than the minority hole density p (i.e., $n \gg p$). On illumination, let the excess electrons and holes generated be Δn and Δp , respectively:

$$n' = n + \Delta n$$

$$p' = p + \Delta p$$

Here n' and p' are the electron and hole concentrations* at any particular illumination and n and p are carriers concentration when there is no illumination. Remember $\Delta n = \Delta p$ and $n \gg p$. Hence, the

EXAMPLE 14.6

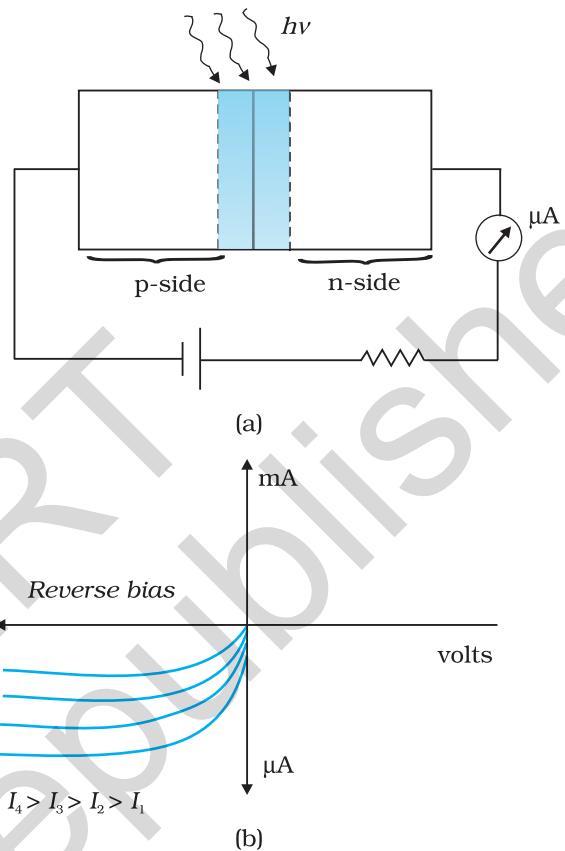


FIGURE 14.23 (a) An illuminated photodiode under reverse bias , (b) I-V characteristics of a photodiode for different illumination intensities $I_4 > I_3 > I_2 > I_1$.

* Note that, to create an e-h pair, we spend some energy (photoexcitation, thermal excitation, etc.). Therefore when an electron and hole recombine the energy is released in the form of light (radiative recombination) or heat (non-radiative recombination). It depends on semiconductor and the method of fabrication of the p-n junction. For the fabrication of LEDs, semiconductors like GaAs, GaAs-GaP are used in which radiative recombination dominates.

EXAMPLE 14.6

fractional change in the majority carriers (i.e., $\Delta n/n$) would be much less than that in the minority carriers (i.e., $\Delta p/p$). In general, we can state that the fractional change due to the photo-effects on the *minority carrier dominated reverse bias current* is more easily measurable than the fractional change in the forward bias current. Hence, photodiodes are preferably used in the reverse bias condition for measuring light intensity.

(ii) Light emitting diode

It is a heavily doped p-n junction which under forward bias emits spontaneous radiation. The diode is encapsulated with a transparent cover so that emitted light can come out.

When the diode is forward biased, electrons are sent from $n \rightarrow p$ (where they are minority carriers) and holes are sent from $p \rightarrow n$ (where they are minority carriers). At the junction boundary the concentration of minority carriers increases compared to the equilibrium concentration (i.e., when there is no bias). Thus at the junction boundary on either side of the junction, excess minority carriers are there which recombine with majority carriers near the junction. On recombination, the energy is released in the form of photons. Photons with energy equal to or slightly less than the band gap are emitted. When the forward current of the diode is small, the intensity of light emitted is small. As the forward current increases, intensity of light increases and reaches a maximum. Further increase in the forward current results in decrease of light intensity. LEDs are biased such that the light emitting efficiency is maximum.

The V-I characteristics of a LED is similar to that of a Si junction diode. But the threshold voltages are much higher and slightly different for each colour. The reverse breakdown voltages of LEDs are very low, typically around 5V. So care should be taken that high reverse voltages do not appear across them.

LEDs that can emit red, yellow, orange, green and blue light are commercially available. The semiconductor used for fabrication of visible LEDs must at least have a band gap of 1.8 eV (spectral range of visible light is from about $0.4\text{ }\mu\text{m}$ to $0.7\text{ }\mu\text{m}$, i.e., from about 3 eV to 1.8 eV). The compound semiconductor Gallium Arsenide – Phosphide ($\text{GaAs}_{1-x}\text{P}_x$) is used for making LEDs of different colours. $\text{GaAs}_{0.6}\text{P}_{0.4}$ ($E_g \sim 1.9\text{ eV}$) is used for red LED. GaAs ($E_g \sim 1.4\text{ eV}$) is used for making infrared LED. These LEDs find extensive use in remote controls, burglar alarm systems, optical communication, etc. Extensive research is being done for developing white LEDs which can replace incandescent lamps.

LEDs have the following advantages over conventional incandescent low power lamps:

- (i) Low operational voltage and less power.
- (ii) Fast action and no warm-up time required.
- (iii) The bandwidth of emitted light is 100 \AA to 500 \AA or in other words it is nearly (but not exactly) monochromatic.
- (iv) Long life and ruggedness.
- (v) Fast on-off switching capability.

(iii) Solar cell

A solar cell is basically a p-n junction which generates emf when solar radiation falls on the p-n junction. It works on the same principle (photovoltaic effect) as the photodiode, except that no external bias is applied and the junction area is kept much larger for solar radiation to be incident because we are interested in more power.

A simple p-n junction solar cell is shown in Fig. 14.24.

A p-Si wafer of about $300\text{ }\mu\text{m}$ is taken over which a thin layer ($\sim 0.3\text{ }\mu\text{m}$) of n-Si is grown on one-side by diffusion process. The other side of p-Si is coated with a metal (back contact). On the top of n-Si layer, metal finger electrode (or metallic grid) is deposited. This acts as a front contact. The metallic grid occupies only a very small fraction of the cell area ($<15\%$) so that light can be incident on the cell from the top.

The generation of emf by a solar cell, when light falls on, it is due to the following three basic processes: generation, separation and collection—
(i) generation of e-h pairs due to light (with $h\nu > E_g$) close to the junction; (ii) separation of electrons and holes due to electric field of the depletion region. Electrons are swept to n-side and holes to p-side; (iii) the electrons reaching the n-side are collected by the front contact and holes reaching p-side are collected by the back contact. Thus p-side becomes positive and n-side becomes negative giving rise to *photovoltage*.

When an external load is connected as shown in the Fig. 14.25(a) a photocurrent I_L flows through the load. A typical I-V characteristics of a solar cell is shown in the Fig. 14.25(b).

Note that the $I - V$ characteristics of solar cell is drawn in the fourth quadrant of the coordinate axes. This is because a solar cell does not draw current but supplies the same to the load.

Semiconductors with band gap close to 1.5 eV are ideal materials for solar cell fabrication. Solar cells are made with semiconductors like Si ($E_g = 1.1\text{ eV}$), GaAs ($E_g = 1.43\text{ eV}$), CdTe ($E_g = 1.45\text{ eV}$), CuInSe₂ ($E_g = 1.04\text{ eV}$), etc. The important criteria for the selection of a material for solar cell fabrication are (i) band gap (~ 1.0 to 1.8 eV), (ii) high optical absorption ($\sim 10^4\text{ cm}^{-1}$), (iii) electrical conductivity, (iv) availability of the raw material, and (v) cost. Note that sunlight is not always required for a solar cell. Any light with photon energies greater than the bandgap will do. Solar cells are used to power electronic devices in satellites and space vehicles and also as power supply to some calculators. Production of low-cost photovoltaic cells for large-scale solar energy is a topic for research.

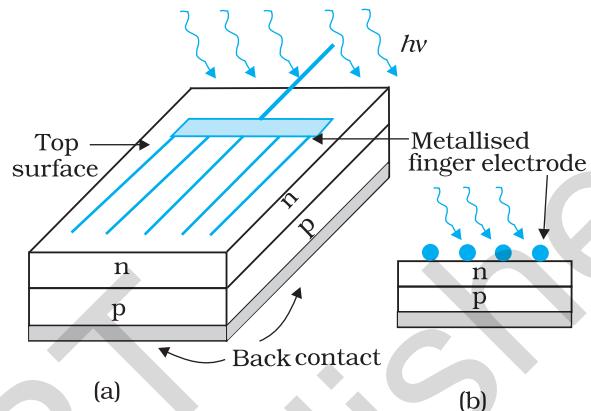


FIGURE 14.24 (a) Typical p-n junction solar cell; (b) Cross-sectional view.

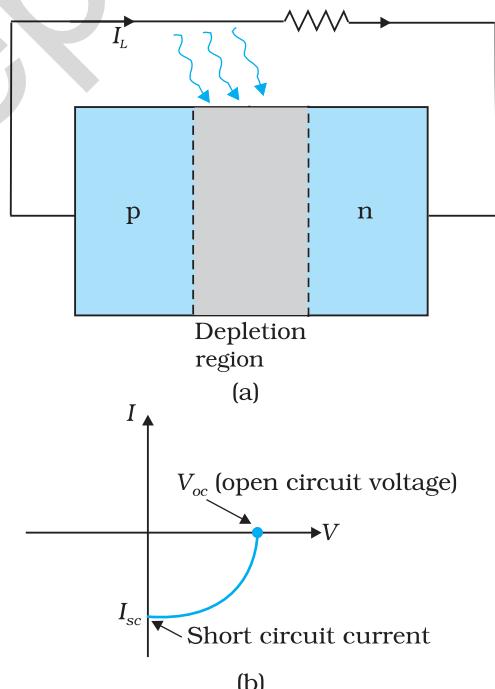


FIGURE 14.25 (a) A typical illuminated p-n junction solar cell; (b) I-V characteristics of a solar cell.

EXAMPLE 14.7

Example 14.7 Why are Si and GaAs are preferred materials for solar cells?

Solution The solar radiation spectrum received by us is shown in Fig. 14.26.

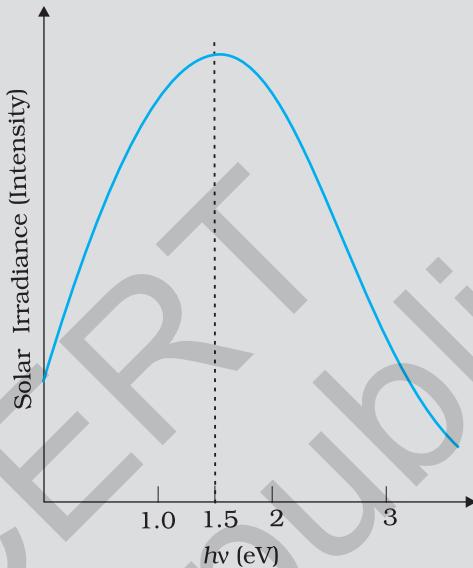


FIGURE 14.26

The maxima is near 1.5 eV. For photo-excitation, $h\nu > E_g$. Hence, semiconductor with band gap ~ 1.5 eV or lower is likely to give better solar conversion efficiency. Silicon has $E_g \sim 1.1$ eV while for GaAs it is ~ 1.53 eV. In fact, GaAs is better (in spite of its higher band gap) than Si because of its relatively higher absorption coefficient. If we choose materials like CdS or CdSe ($E_g \sim 2.4$ eV), we can use only the high energy component of the solar energy for photo-conversion and a significant part of energy will be of no use.

The question arises: why we do not use material like PbS ($E_g \sim 0.4$ eV) which satisfy the condition $h\nu > E_g$ for v maxima corresponding to the solar radiation spectra? If we do so, most of the solar radiation will be absorbed on the *top-layer* of solar cell and will not reach in or near the depletion region. For effective electron-hole separation, due to the junction field, we want the photo-generation to occur in the junction region only.

14.9 JUNCTION TRANSISTOR

The credit of inventing the transistor in the year 1947 goes to J. Bardeen and W.H. Brattain of Bell Telephone Laboratories, U.S.A. That transistor was a point-contact transistor. The first junction transistor consisting of two back-to-back p-n junctions was invented by William Shockley in 1951.

As long as only the junction transistor was known, it was known simply as transistor. But over the years new types of transistors were invented and to differentiate it from the new ones it is now called the Bipolar Junction Transistor (BJT). Even now, often the word transistor

is used to mean BJT when there is no confusion. Since our study is limited to only BJT, we shall use the word transistor for BJT without any ambiguity.

14.9.1 Transistor: structure and action

A transistor has three doped regions forming two p-n junctions between them. Obviously, there are two types of transistors, as shown in Fig. 14.27.

(i) **n-p-n transistor:** Here two segments of n-type semiconductor (emitter and collector) are separated by a segment of p-type semiconductor (base).

(ii) **p-n-p transistor:** Here two segments of p-type semiconductor (termed as emitter and collector) are separated by a segment of n-type semiconductor (termed as base).

The schematic representations of an n-p-n and a p-n-p configuration are shown in Fig. 14.27(a). All the three segments of a transistor have different thickness and their doping levels are also different. In the schematic symbols used for representing p-n-p and n-p-n transistors [Fig. 14.27(b)] the arrowhead shows the direction of conventional current in the transistor. A brief description of the three segments of a transistor is given below:

- **Emitter:** This is the segment on one side of the transistor shown in Fig. 14.27(a). It is of *moderate size* and *heavily doped*. It supplies a large number of majority carriers for the current flow through the transistor.
- **Base:** This is the central segment. It is *very thin* and *lightly doped*.
- **Collector:** This segment collects a *major portion* of the majority carriers supplied by the emitter. The collector side is *moderately doped* and *larger* in size as compared to the *emitter*.

We have seen earlier in the case of a p-n junction, that there is a formation of depletion region across the junction. In case of a transistor depletion regions are formed at the emitter-base junction and the base-collector junction. For understanding the action of a transistor, we have to consider the nature of depletion regions formed at these junctions. The charge carriers move across different regions of the transistor when proper voltages are applied across its terminals.

The biasing of the transistor is done differently for different uses. The transistor can be used in two distinct ways. Basically, it was invented to function as an amplifier, a device which produces an enlarged copy of a signal. But later its use as a switch acquired equal importance. We shall study both these functions and the ways the transistor is biased to achieve these mutually exclusive functions.

First we shall see what gives the transistor its amplifying capabilities. The transistor works as an amplifier, with its emitter-base junction forward biased and the base-collector junction reverse biased. This situation is shown in Fig. 14.28, where V_{CC} and V_{EE} are used for creating the respective biasing. When the transistor is biased in this way it is said to be in *active state*. We represent the voltage between emitter and base as V_{EB} and that between the collector and the base as V_{CB} . In

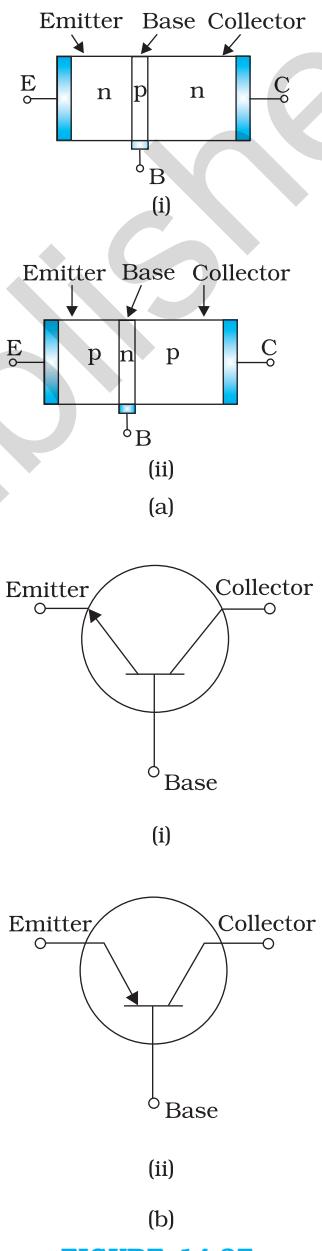


FIGURE 14.27

(a) Schematic representations of a n-p-n transistor and p-n-p transistor, and (b) Symbols for n-p-n and p-n-p transistors.

Physics

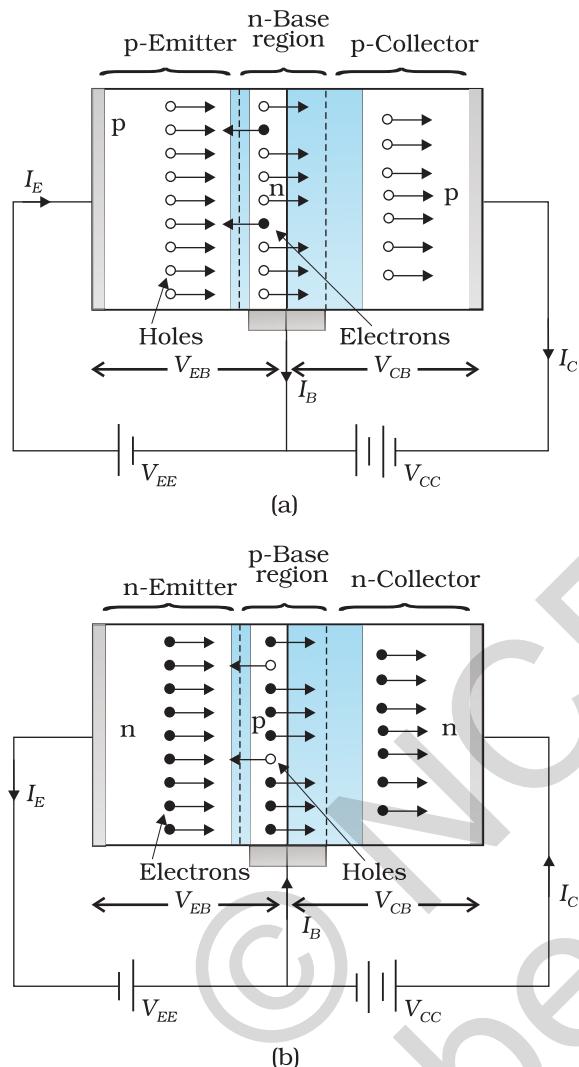


FIGURE 14.28 Bias Voltage applied on: (a) p-n-p transistor and (b) n-p-n transistor.

Fig. 14.28, base is a common terminal for the two power supplies whose other terminals are connected to emitter and collector, respectively. So the two power supplies are represented as V_{EE} , and V_{CC} , respectively. In circuits, where emitter is the common terminal, the power supply between the base and the emitter is represented as V_{BB} and that between collector and emitter as V_{CE} .

Let us see now the paths of current carriers in the transistor with emitter-base junction forward biased and base-collector junction reverse biased. The heavily doped emitter has a high concentration of majority carriers, which will be holes in a p-n-p transistor and electrons in an n-p-n transistor. These majority carriers enter the base region in large numbers. The base is thin and lightly doped. So the majority carriers there would be few. In a p-n-p transistor the majority carriers in the base are electrons since base is of n-type semiconductor. The large number of holes entering the base from the emitter swamps the small number of electrons there. As the base collector-junction is reverse-biased, these holes, which appear as minority carriers at the junction, can easily cross the junction and enter the collector. The holes in the base could move either towards the base terminal to combine with the electrons entering from outside or cross the junction to enter into the collector and reach the collector terminal. The base is made thin so that most of the holes find themselves near the reverse-biased base-collector junction and so cross the junction instead of moving to the base terminal.

It is interesting to note that due to forward bias a large current enters the emitter-base junction, but most of it is diverted to adjacent reverse-biased base-collector junction and the current coming out of the base becomes a very small fraction of the current that entered the junction. If we represent the hole current and the electron current crossing the forward biased junction by I_h and I_e respectively then the total current in a forward biased diode is the sum $I_h + I_e$. We see that the emitter current $I_E = I_h + I_e$ but the base current $I_B \ll I_h + I_e$, because a major part of I_E goes to collector instead of coming out of the base terminal. The base current is thus a small fraction of the emitter current.

The current entering into the emitter from outside is equal to the emitter current I_E . Similarly the current emerging from the base terminal is I_B and that from collector terminal is I_C . It is obvious from the above description and also from a straight forward application of Kirchhoff's law to Fig. 14.28(a) that the emitter current is the sum of collector current and base current:

$$I_E = I_C + I_B \quad (14.7)$$

We also see that $I_C \approx I_E$.

Our description of the direction of motion of the holes is identical with the direction of the conventional current. But the direction of motion of electrons is just opposite to that of the current. Thus in a p-n-p transistor the current enters from emitter into base whereas in a n-p-n transistor it enters from the base into the emitter. The arrowhead in the emitter shows the direction of the conventional current.

The description about the paths followed by the majority and minority carriers in a n-p-n is exactly the same as that for the p-n-p transistor. But the current paths are exactly opposite, as shown in Fig. 14.28. In Fig. 14.28(b) the electrons are the majority carriers supplied by the n-type emitter region. They cross the thin p-base region and are able to reach the collector to give the collector current, I_C . From the above description we can conclude that in the active state of the transistor the emitter-base junction acts as a low resistance while the base collector acts as a high resistance.

14.9.2 Basic transistor circuit configurations and transistor characteristics

In a transistor, only three terminals are available, viz., *Emitter (E)*, *Base (B)* and *Collector (C)*. Therefore, in a circuit the input/output connections have to be such that one of these (E, B or C) is common to both the input and the output. Accordingly, the transistor can be *connected* in either of the following three configurations:

Common Emitter (CE), Common Base (CB), Common Collector (CC)

The transistor is most widely used in the CE configuration and we shall restrict our discussion to only this configuration. Since more commonly used transistors are n-p-n Si transistors, we shall confine our discussion to such transistors only. With p-n-p transistors the polarities of the external power supplies are to be inverted.

Common emitter transistor characteristics

When a transistor is used in CE configuration, the input is between the base and the emitter and the output is between the collector and the emitter. The variation of the base current I_B with the base-emitter voltage V_{BE} is called the *input characteristic*. Similarly, the variation of the collector current I_C with the collector-emitter voltage V_{CE} is called the *output characteristic*. You will see that the output characteristics are controlled by the input characteristics. This implies that the collector current changes with the base current.

The input and the output characteristics of an n-p-n transistors can be studied by using the circuit shown in Fig. 14.29.

To study the input characteristics of the transistor in C_E configuration, a curve is plotted between the base current I_B against the base-emitter voltage V_{BE} . The

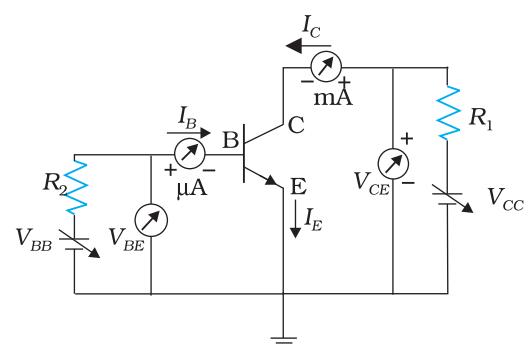


FIGURE 14.29 Circuit arrangement for studying the input and output characteristics of n-p-n transistor in CE configuration.

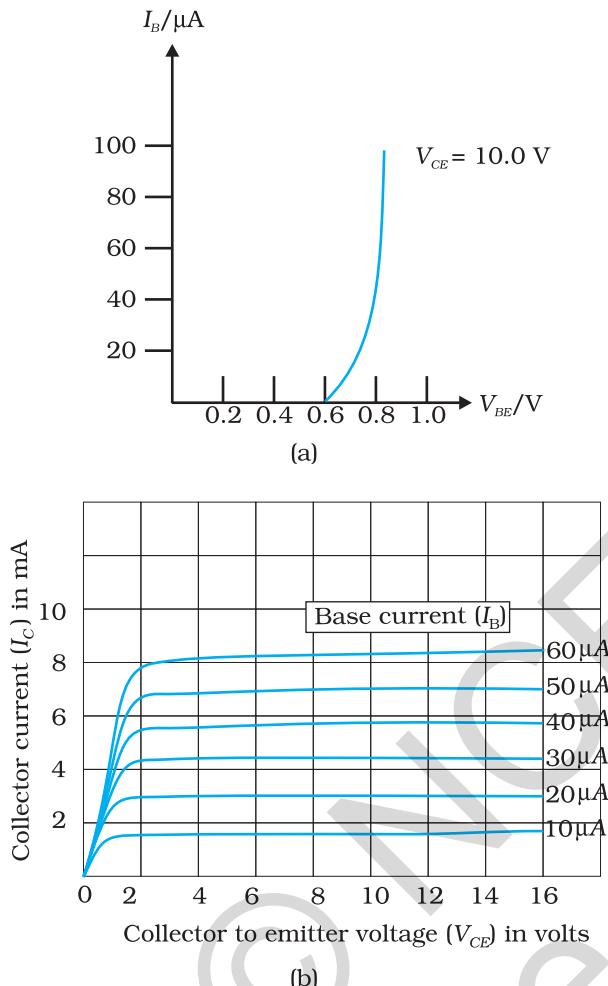


FIGURE 14.30 (a) Typical input characteristics, and (b) Typical output characteristics.

output characteristic. So there will be different output characteristics corresponding to different values of I_B as shown in Fig. 14.30(b).

The linear segments of both the input and output characteristics can be used to calculate some important ac parameters of transistors as shown below.

(i) **Input resistance (r_i):** This is defined as the ratio of change in base-emitter voltage (ΔV_{BE}) to the resulting change in base current (ΔI_B) at constant collector-emitter voltage (V_{CE}). This is dynamic (ac resistance) and as can be seen from the input characteristic, its value varies with the operating current in the transistor:

$$r_i = \frac{\Delta V_{BE}}{\Delta I_B} \quad v_{ce} \quad (14.8)$$

The value of r_i can be anything from a few hundreds to a few thousand ohms.

collector-emitter voltage V_{CE} is kept fixed while studying the dependence of I_B on V_{BE} . We are interested to obtain the input characteristic when the transistor is in active state. So the collector-emitter voltage V_{CE} is kept large enough to make the base collector junction reverse biased. Since $V_{CE} = V_{CB} + V_{BE}$ and for Si transistor V_{BE} is 0.6 to 0.7 V, V_{CE} must be sufficiently larger than 0.7 V. Since the transistor is operated as an amplifier over large range of V_{CE} the reverse bias across the base-collector junction is high most of the time. Therefore, the input characteristics may be obtained for V_{CE} somewhere in the range of 3 V to 20 V. Since the increase in V_{CE} appears as increase in V_{CB} , its effect on I_B is negligible. As a consequence, input characteristics for various values of V_{CE} will give almost identical curves. Hence, it is enough to determine only one input characteristic. The input characteristic of a transistor is as shown in Fig. 14.30(a).

The output characteristic is obtained by observing the variation of I_C as V_{CE} is varied keeping I_B constant. It is obvious that if V_{BE} is increased by a small amount, both hole current from the emitter region and the electron current from the base region will increase. As a consequence both I_B and I_C will increase proportionately. This shows that when I_B increases I_C also increases. The plot of I_C versus V_{CE} for different fixed values of I_B gives one

(ii) **Output resistance (r_o):** This is defined as the ratio of change in collector-emitter voltage (ΔV_{CE}) to the change in collector current (ΔI_C) at a constant base current I_B .

$$r_o = \frac{\Delta V_{CE}}{\Delta I_C} \quad (14.9)$$

The output characteristics show that initially for very small values of V_{CE} , I_C increases almost linearly. This happens because the base-collector junction is not reverse biased and the transistor is not in active state. In fact, the transistor is in the saturation state and the current is controlled by the supply voltage V_{CC} ($=V_{CE}$) in this part of the characteristic. When V_{CE} is more than that required to reverse bias the base-collector junction, I_C increases very little with V_{CE} . The reciprocal of the slope of the linear part of the output characteristic gives the values of r_o . The output resistance of the transistor is mainly controlled by the bias of the base-collector junction. The high magnitude of the output resistance (of the order of 100 k Ω) is due to the reverse-biased state of this diode. This also explains why the resistance at the initial part of the characteristic, when the transistor is in saturation state, is very low.

(iii) **Current amplification factor (β):** This is defined as the ratio of the change in collector current to the change in base current at a constant collector-emitter voltage (V_{CE}) when the transistor is in active state.

$$\beta_{ac} = \frac{\Delta I_C}{\Delta I_B} \quad (14.10)$$

This is also known as *small signal current gain* and its value is very large.

If we simply find the ratio of I_C and I_B we get what is called dc β of the transistor. Hence,

$$\beta_{dc} = \frac{I_C}{I_B} \quad (14.11)$$

Since I_C increases with I_B almost linearly and $I_C = 0$ when $I_B = 0$, the values of both β_{dc} and β_{ac} are nearly equal. So, for most calculations β_{dc} can be used. Both β_{ac} and β_{dc} vary with V_{CE} and I_B (or I_C) slightly.

Example 14.8 From the output characteristics shown in Fig. 14.30(b), calculate the values of β_{ac} and β_{dc} of the transistor when V_{CE} is 10 V and $I_C = 4.0$ mA.

Solution

$$\beta_{ac} = \frac{\Delta I_C}{\Delta I_B} \quad , \quad \beta_{dc} = \frac{I_C}{I_B}$$

For determining β_{ac} and β_{dc} at the stated values of V_{CE} and I_C one can proceed as follows. Consider any two characteristics for two values of I_B which lie above and below the given value of I_C . Here $I_C = 4.0$ mA. (Choose characteristics for $I_B = 30$ and $20 \mu\text{A}$.) At $V_{CE} = 10$ V we read the two values of I_C from the graph. Then

EXAMPLE 14.8

EXAMPLE 14.8

$$\Delta I_B = (30 - 20) \mu\text{A} = 10 \mu\text{A}, \Delta I_C = (4.5 - 3.0) \text{ mA} = 1.5 \text{ mA}$$

$$\text{Therefore, } \beta_{ac} = 1.5 \text{ mA} / 10 \mu\text{A} = 150$$

For determining β_{dc} , either estimate the value of I_B corresponding to $I_C = 4.0 \text{ mA}$ at $V_{CE} = 10 \text{ V}$ or calculate the two values of β_{dc} for the two characteristics chosen and find their mean.

Therefore, for $I_C = 4.5 \text{ mA}$ and $I_B = 30 \mu\text{A}$,

$$\beta_{dc} = 4.5 \text{ mA} / 30 \mu\text{A} = 150$$

and for $I_C = 3.0 \text{ mA}$ and $I_B = 20 \mu\text{A}$

$$\beta_{dc} = 3.0 \text{ mA} / 20 \mu\text{A} = 150$$

$$\text{Hence, } \beta_{dc} = (150 + 150) / 2 = 150$$

14.9.3 Transistor as a device

The transistor can be used as a device application depending on the configuration used (namely CB, CC and CE), the biasing of the E-B and B-C junction and the operation region namely cutoff, active region and saturation. As mentioned earlier we have confined only to the CE configuration and will be concentrating on the biasing and the operation region to understand the working of a device.

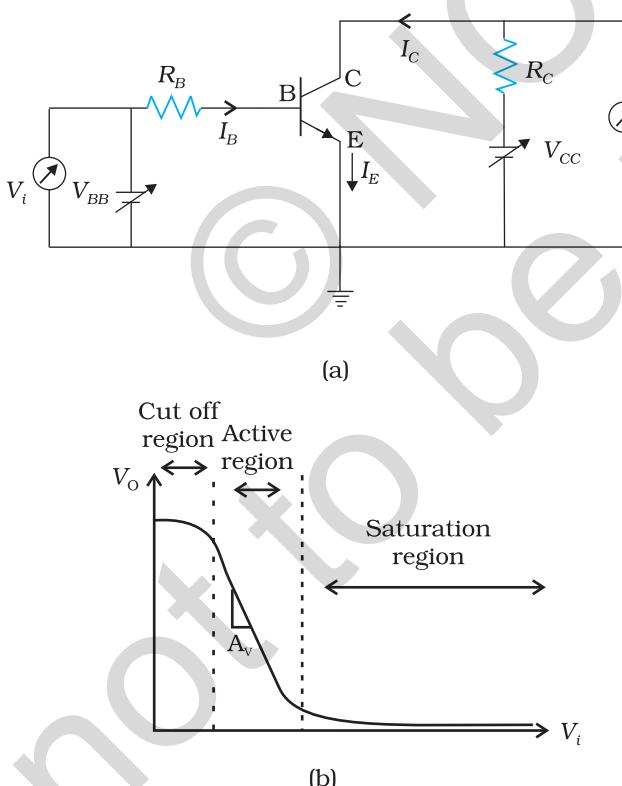


FIGURE 14.31 (a) Base-biased transistor in CE configuration, (b) Transfer characteristic.

When the transistor is used in the cutoff or saturation state it acts as a *switch*. On the other hand for using the transistor as an *amplifier*, it has to operate in the active region.

(i) Transistor as a switch

We shall try to understand the operation of the transistor as a switch by analysing the behaviour of the base-biased transistor in CE configuration as shown in Fig. 14.31(a).

Applying Kirchhoff's voltage rule to the input and output sides of this circuit, we get

$$V_{BB} = I_B R_B + V_{BE} \quad (14.12)$$

and

$$V_{CE} = V_{CC} - I_C R_C \quad (14.13)$$

We shall treat V_{BB} as the dc input voltage V_i and V_{CE} as the dc output voltage V_o . So, we have

$$V_i = I_B R_B + V_{BE}$$

$$V_o = V_{CC} - I_C R_C$$

Let us see how V_o changes as V_i increases from zero onwards. In the case of Si transistor, as long as input V_i is less than 0.6 V, the transistor will be in cut off state and current I_C will be zero.

Hence $V_o = V_{CC}$

When V_i becomes greater than 0.6 V the transistor is in active state with some current I_C in the output path and the output V_o decrease as the

term $I_C R_C$ increases. With increase of V_i , I_C increases almost linearly and so V_o decreases linearly till its value becomes less than about 1.0 V.

Beyond this, the change becomes non linear and transistor goes into saturation state. With further increase in V_i the output voltage is found to decrease further towards zero though it may never become zero. If we plot the V_o vs V_i curve, [also called the transfer characteristics of the base-biased transistor (Fig. 14.31(b))], we see that between cut off state and active state and also between active state and saturation state there are regions of non-linearity showing that the transition from cutoff state to active state and from active state to saturation state are not sharply defined.

Let us see now how the transistor is operated as a switch. As long as V_i is *low* and unable to forward-bias the transistor, V_o is *high* (at V_{CC}). If V_i is *high* enough to drive the transistor into saturation, then V_o is *low*, very near to zero. When the transistor is not conducting it is said to be *switched off* and when it is driven into saturation it is said to be *switched on*. This shows that if we define low and high states as below and above certain voltage levels corresponding to cutoff and saturation of the transistor, then we can say that a *low* input switches the transistor off and a *high* input switches it on. Alternatively, we can say that a *low* input to the transistor gives a *high* output and a *high* input gives a *low* output. The switching circuits are designed in such a way that the transistor does not remain in active state.

(ii) Transistor as an amplifier

For using the transistor as an amplifier we will use the active region of the V_o versus V_i curve. The slope of the linear part of the curve represents the rate of change of the output with the input. It is negative because the output is $V_{CC} - I_C R_C$ and not $I_C R_C$. That is why as input voltage of the CE amplifier increases its output voltage decreases and the output is said to be out of phase with the input. If we consider ΔV_o and ΔV_i as small changes in the output and input voltages then $\Delta V_o / \Delta V_i$ is called the small signal voltage gain A_v of the amplifier.

If the V_{BB} voltage has a fixed value corresponding to the mid point of the active region, the circuit will behave as a CE amplifier with voltage gain $\Delta V_o / \Delta V_i$. We can express the voltage gain A_v in terms of the resistors in the circuit and the current gain of the transistor as follows.

We have, $V_o = V_{CC} - I_C R_C$

Therefore, $\Delta V_o = 0 - R_C \Delta I_C$

Similarly, from $V_i = I_B R_B + V_{BE}$

$$\Delta V_i = R_B \Delta I_B + \Delta V_{BE}$$

But ΔV_{BE} is negligibly small in comparison to $\Delta I_B R_B$ in this circuit.

So, the *voltage gain* of this *CE amplifier* (Fig. 14.32) is given by

$$A_v = -R_C \Delta I_C / R_B \Delta I_B = -\beta_{ac} (R_C / R_B) \quad (14.14)$$

where β_{ac} is equal to $\Delta I_C / \Delta I_B$ from Eq. (14.10). Thus the linear portion of the active region of the transistor can be exploited for the use in amplifiers. Transistor as an amplifier (CE configuration) is discussed in detail in the next section.

14.9.4 Transistor as an Amplifier (CE-Configuration)

To operate the transistor as an amplifier it is necessary to fix its operating point somewhere in the middle of its active region. If we fix the value of V_{BB} corresponding to a point in the middle of the linear part of the transfer curve then the dc base current I_B would be constant and corresponding collector current I_C will also be constant. The dc voltage $V_{CE} = V_{CC} - I_C R_C$ would also remain constant. The operating values of V_{CE} and I_B determine the operating point, of the amplifier.

If a small sinusoidal voltage with amplitude v_s is superposed on the dc base bias by connecting the source of that signal in series with the V_{BB} supply, then the base current will have sinusoidal variations superimposed on the value of I_B . As a consequence the collector current also will have sinusoidal variations superimposed on the value of I_C , producing in turn corresponding change in the value of V_O . We can measure the ac variations across the input and output terminals by blocking the dc voltages by large capacitors.

In the description of the amplifier given above we have not considered any ac signal. In general, amplifiers are used to amplify alternating signals. Now let us superimpose an ac input signal v_i (to be amplified) on the bias V_{BB} (dc) as shown in Fig. 14.32. The output is taken between the collector and the ground.

The working of an amplifier can be easily understood, if we first assume that $v_i = 0$. Then applying Kirchhoff's law to the output loop, we get

$$V_{cc} = V_{CE} + I_C R_L \quad (14.15)$$

Likewise, the input loop gives

$$V_{BB} = V_{BE} + I_B R_B \quad (14.16)$$

When v_i is not zero, we get

$$V_{BE} + v_i = V_{BE} + I_B R_B + \Delta I_B (R_B + r_i)$$

The change in V_{BE} can be related to the input resistance r_i [see Eq. (14.8)] and the change in I_B . Hence

$$\begin{aligned} v_i &= \Delta I_B (R_B + r_i) \\ &= r_i \Delta I_B \end{aligned}$$

The change in I_B causes a change in I_C . We define a parameter β_{ac} , which is similar to the β_{dc} defined in Eq. (14.11), as

$$\beta_{ac} = \frac{\Delta I_C}{\Delta I_B} = \frac{i_c}{i_b} \quad (14.17)$$

which is also known as the *ac current gain* A_i . Usually β_{ac} is close to β_{dc} in the linear region of the output characteristics.

The change in I_C due to a change in I_B causes a change in V_{CE} and the voltage drop across the resistor R_L because V_{CC} is fixed.

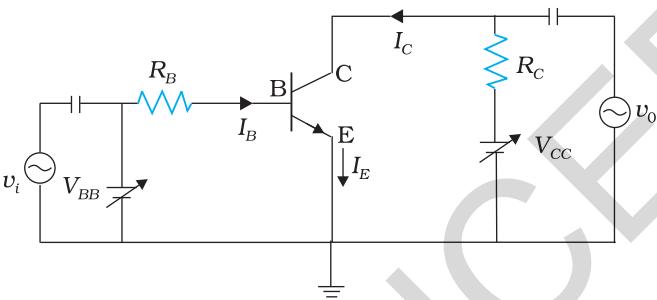


FIGURE 14.32 A simple circuit of a CE-transistor amplifier.

These changes can be given by Eq. (14.15) as

$$\Delta V_{CC} = \Delta V_{CE} + R_L \Delta I_C = 0$$

$$\text{or } \Delta V_{CE} = -R_L \Delta I_C$$

The change in V_{CE} is the output voltage v_o . From Eq. (14.10), we get

$$v_o = \Delta V_{CE} = -\beta_{ac} R_L \Delta I_B$$

The voltage gain of the amplifier is

$$\begin{aligned} A_v &= \frac{v_o}{v_i} = \frac{\Delta V_{CE}}{r \Delta I_B} \\ &= -\frac{\beta_{ac} R_L}{r} \end{aligned} \quad (14.18)$$

The negative sign represents that output voltage is opposite with phase with the input voltage.

From the discussion of the transistor characteristics you have seen that there is a current gain β_{ac} in the CE configuration. Here we have also seen the voltage gain A_v . Therefore the power gain A_p can be expressed as the product of the current gain and voltage gain. Mathematically

$$A_p = \beta_{ac} \times A_v \quad (14.19)$$

Since β_{ac} and A_v are greater than 1, we get ac power gain. However it should be realised that transistor is not a power generating device. The energy for the higher ac power at the output is supplied by the battery.

Example 14.9 In Fig. 14.31(a), the V_{BB} supply can be varied from 0V to 5.0 V. The Si transistor has $\beta_{dc} = 250$ and $R_B = 100 \text{ k}\Omega$, $R_C = 1 \text{ k}\Omega$, $V_{CC} = 5.0 \text{ V}$. Assume that when the transistor is saturated, $V_{CE} = 0 \text{ V}$ and $V_{BE} = 0.8 \text{ V}$. Calculate (a) the minimum base current, for which the transistor will reach saturation. Hence, (b) determine V_1 when the transistor is 'switched on'. (c) find the ranges of V_1 for which the transistor is 'switched off' and 'switched on'.

Solution

Given at saturation $V_{CE} = 0 \text{ V}$, $V_{BE} = 0.8 \text{ V}$

$$V_{CE} = V_{CC} - I_C R_C$$

$$I_C = V_{CC}/R_C = 5.0 \text{ V}/1.0 \text{ k}\Omega = 5.0 \text{ mA}$$

$$\text{Therefore } I_B = I_C/\beta = 5.0 \text{ mA}/250 = 20 \mu\text{A}$$

The input voltage at which the transistor will go into saturation is given by

$$\begin{aligned} V_{IH} &= V_{BB} = I_B R_B + V_{BE} \\ &= 20 \mu\text{A} \times 100 \text{ k}\Omega + 0.8 \text{ V} = 2.8 \text{ V} \end{aligned}$$

The value of input voltage below which the transistor remains cutoff is given by

$$V_{IL} = 0.6 \text{ V}, V_{IH} = 2.8 \text{ V}$$

Between 0.0V and 0.6V, the transistor will be in the 'switched off' state. Between 2.8V and 5.0V, it will be in 'switched on' state.

Note that the transistor is in active state when I_B varies from 0.0mA to 20mA. In this range, $I_C = \beta I_B$ is valid. In the saturation range, $I_C \leq \beta I_B$.

EXAMPLE 14.10

Example 14.10 For a CE transistor amplifier, the audio signal voltage across the collector resistance of $2.0\text{ k}\Omega$ is 2.0 V . Suppose the current amplification factor of the transistor is 100. What should be the value of R_B in series with V_{BB} supply of 2.0 V if the dc base current has to be 10 times the signal current. Also calculate the dc drop across the collector resistance. (Refer to Fig. 14.33).

Solution The output ac voltage is 2.0 V . So, the ac collector current $i_C = 2.0/2000 = 1.0\text{ mA}$. The signal current through the base is, therefore given by $i_B = i_C/\beta = 1.0\text{ mA}/100 = 0.010\text{ mA}$. The dc base current has to be $10 \times 0.010 = 0.10\text{ mA}$.

From Eq. 14.16, $R_B = (V_{BB} - V_{BE}) / I_B$. Assuming $V_{BE} = 0.6\text{ V}$, $R_B = (2.0 - 0.6)/0.10 = 14\text{ k}\Omega$.

The dc collector current $I_C = 100 \times 0.10 = 10\text{ mA}$.

14.9.5 Feedback amplifier and transistor oscillator

In an amplifier, we have seen that a sinusoidal input is given which appears as an amplified signal in the output. This means that an *external input is necessary to sustain ac signal in the output for an amplifier*. In an oscillator, we get ac output without any external input signal. In other words, the output in an oscillator is *self-sustained*. To attain this, an amplifier is taken. A portion of the output power is returned back (feedback) to the input *in phase* with the starting power (this process is termed *positive feedback*) as shown in Fig. 14.33(a). The feedback can be achieved by inductive coupling (through mutual inductance) or *LC* or *RC* networks. Different types of oscillators essentially use different methods of coupling the output to the input (feedback network), apart from the resonant circuit for obtaining oscillation at a particular frequency. For understanding the oscillator action, we consider the circuit shown in Fig. 14.33(b) in which the feedback is accomplished by *inductive coupling* from one coil winding (T_1) to another coil winding (T_2). Note that the coils T_2 and T_1 are wound on the same core and hence are inductively coupled through their mutual inductance. As in an amplifier, the base-emitter junction is forward biased while the base-collector junction is reverse biased. Detailed biasing circuits actually used have been omitted for simplicity.

Let us try to understand how oscillations are built. Suppose switch S_1 is put on to

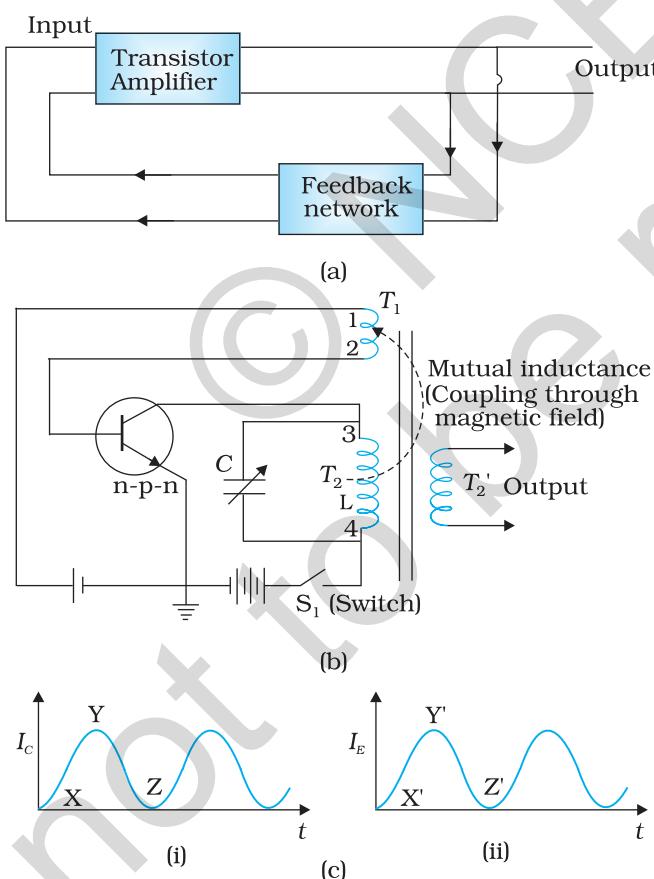


FIGURE 14.33 (a) Principle of a transistor amplifier with positive feedback working as an oscillator and (b) Tuned collector oscillator, (c) Rise and fall (or built up) of current I_c and I_e due to the inductive coupling.

apply proper bias for the first time. Obviously, a *surge* of collector current flows in the transistor. This current flows through the coil T_2 where terminals are numbered 3 and 4 [Fig. 14.33(b)]. This current does not reach full amplitude instantaneously but increases from X to Y, as shown in Fig. [14.33(c)(i)]. The inductive coupling between coil T_2 and coil T_1 now causes a current to flow in the emitter circuit (note that this actually is the 'feedback' from input to output). As a result of this positive feedback, this current (in T_1 ; emitter current) also increases from X to Y [Fig. 14.33(c)(ii)]. The current in T_2 (collector current) connected in the collector circuit acquires the value Y when the transistor becomes *saturated*. This means that maximum collector current is flowing and can increase no further. Since there is no further change in collector current, the magnetic field around T_2 ceases to grow. As soon as the field becomes static, there will be no further feedback from T_2 to T_1 . Without continued feedback, the emitter current begins to fall. Consequently, collector current decreases from Y towards Z [Fig. 14.33(c)(i)]. However, a decrease of collector current causes the magnetic field to decay around the coil T_2 . Thus, T_1 is now seeing a decaying field in T_2 (opposite from what it saw when the field was growing at the initial *start* operation). This causes a further decrease in the emitter current till it reaches Z' when the transistor is *cut-off*. This means that both I_E and I_C cease to flow. Therefore, the transistor has reverted back to its original state (when the power was first switched on). The whole process now repeats itself. That is, the transistor is driven to saturation, then to cut-off, and then back to saturation. The time for change from saturation to cut-off and back is determined by the constants of the tank circuit or tuned circuit (inductance L of coil T_2 and C connected in parallel to it). The resonance frequency (v) of this tuned circuit determines the frequency at which the oscillator will oscillate.

$$v = \frac{1}{2\pi\sqrt{LC}} \quad (14.20)$$

In the circuit of Fig. 14.33(b), the tank or tuned circuit is connected in the collector side. Hence, it is known as *tuned collector oscillator*. If the tuned circuit is on the base side, it will be known as *tuned base oscillator*. There are many other types of tank circuits (say *RC*) or feedback circuits giving different types of oscillators like Colpitt's oscillator, Hartley oscillator, *RC*-oscillator.

14.10 DIGITAL ELECTRONICS AND LOGIC GATES

In electronics circuits like amplifiers, oscillators, introduced to you in earlier sections, the signal (current or voltage) has been in the form of continuous, time-varying voltage or current. Such signals are called continuous or *analogue signals*. A typical analogue signal is shown in Figure. 14.34(a). Fig. 14.34(b) shows a *pulse waveform* in which only discrete values of voltages are possible. It is convenient to use binary numbers to represent such signals. A binary number has only two digits '0' (say, 0V) and '1' (say, 5V). In digital electronics we use only these two levels of voltage as shown in Fig. 14.34(b). Such signals are called *Digital Signals*. In digital circuits only two values (represented by 0 or 1) of the input and output voltage are permissible.

This section is intended to provide the first step in our understanding of digital electronics. We shall restrict our study to some basic building blocks of digital electronics (called *Logic Gates*) which process the digital signals in a specific manner. Logic gates are used in calculators, digital watches, computers, robots, industrial control systems, and in telecommunications.

A light switch in your house can be used as an example of a digital circuit. The light is either ON or OFF depending on the switch position. When the light is ON, the output value is '1'. When the light is OFF the output value is '0'. The inputs are the position of the light switch. The switch is placed either in the ON or OFF position to activate the light.

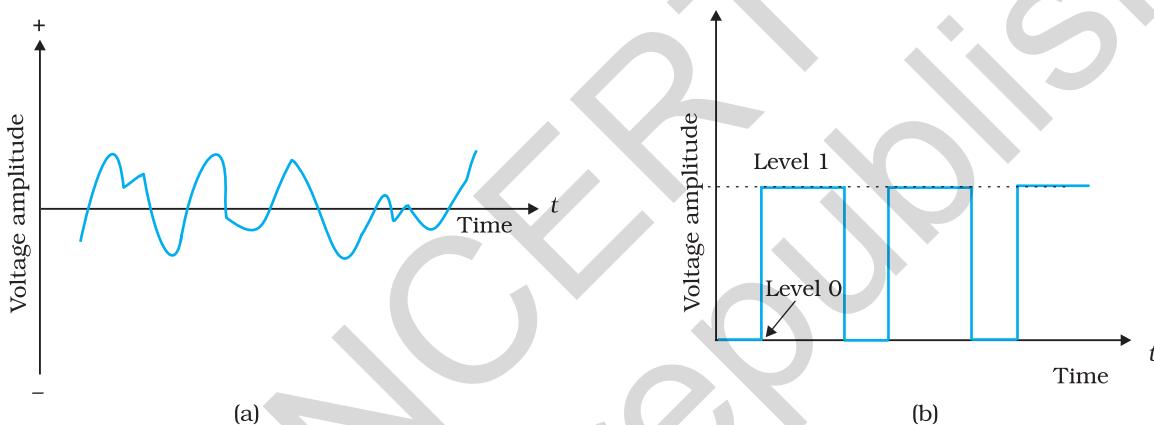
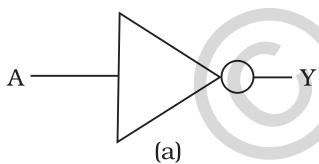


FIGURE 14.34 (a) Analogue signal, (b) Digital signal.



Input	Output
A	Y
0	1
1	0

FIGURE 14.35
 (a) Logic symbol,
 (b) Truth table of
 NOT gate.

14.10.1 Logic gates

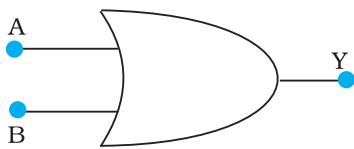
A gate is a digital circuit that follows certain *logical* relationship between the input and output voltages. Therefore, they are generally known as *logic gates* — gates because they control the flow of information. The five common logic gates used are NOT, AND, OR, NAND, NOR. Each logic gate is indicated by a symbol and its function is defined by a *truth table* that shows all the possible input logic level combinations with their respective output logic levels. Truth tables help understand the behaviour of logic gates. These logic gates can be realised using semiconductor devices.

(i) NOT gate

This is the most basic gate, with one input and one output. It produces a '1' output if the input is '0' and vice-versa. That is, it produces an inverted version of the input at its output. This is why it is also known as an *inverter*. The commonly used symbol together with the truth table for this gate is given in Fig. 14.35.

(ii) OR Gate

An *OR* gate has two or more inputs with one output. The logic symbol and truth table are shown in Fig. 14.36. The output Y is 1 when either input A or input B or both are 1s, that is, if any of the input is high, the output is high.



(a)

Input		Output
A	B	Y
0	0	0
0	1	1
1	0	1
1	1	1

(b)

FIGURE 14.36 (a) Logic symbol (b) Truth table of OR gate.

Apart from carrying out the above mathematical logic operation, this gate can be used for modifying the pulse waveform as explained in the following example.

Example 14.11 Justify the output waveform (Y) of the OR gate for the following inputs A and B given in Fig. 14.37.

Solution Note the following:

- At $t < t_1$; $A = 0, B = 0$; Hence $Y = 0$
- For t_1 to t_2 ; $A = 1, B = 0$; Hence $Y = 1$
- For t_2 to t_3 ; $A = 1, B = 1$; Hence $Y = 1$
- For t_3 to t_4 ; $A = 0, B = 1$; Hence $Y = 1$
- For t_4 to t_5 ; $A = 0, B = 0$; Hence $Y = 0$
- For t_5 to t_6 ; $A = 1, B = 0$; Hence $Y = 1$
- For $t > t_6$; $A = 0, B = 1$; Hence $Y = 1$

Therefore the waveform Y will be as shown in the Fig. 14.37.

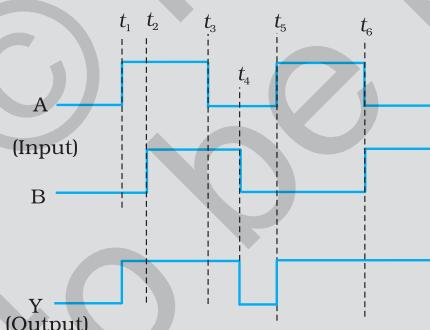
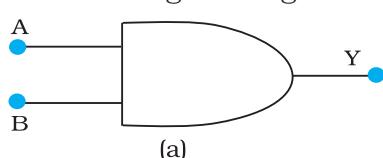


FIGURE 14.37

EXAMPLE 14.11

(iii) AND Gate

An AND gate has two or more inputs and one output. The output Y of AND gate is 1 only when input A and input B are both 1. The logic symbol and truth table for this gate are given in Fig. 14.38



(a)

Input		Output
A	B	Y
0	0	0
0	1	0
1	0	0
1	1	1

(b)

FIGURE 14.38 (a) Logic symbol, (b) Truth table of AND gate.

EXAMPLE 14.12

Example 14.12 Take A and B input waveforms similar to that in Example 14.11. Sketch the output waveform obtained from AND gate.

Solution

- For $t \leq t_1$; $A = 0, B = 0$; Hence $Y = 0$
- For t_1 to t_2 ; $A = 1, B = 0$; Hence $Y = 0$
- For t_2 to t_3 ; $A = 1, B = 1$; Hence $Y = 1$
- For t_3 to t_4 ; $A = 0, B = 1$; Hence $Y = 0$
- For t_4 to t_5 ; $A = 0, B = 0$; Hence $Y = 0$
- For t_5 to t_6 ; $A = 1, B = 0$; Hence $Y = 0$
- For $t > t_6$; $A = 0, B = 1$; Hence $Y = 0$

Based on the above, the output waveform for AND gate can be drawn as given below.

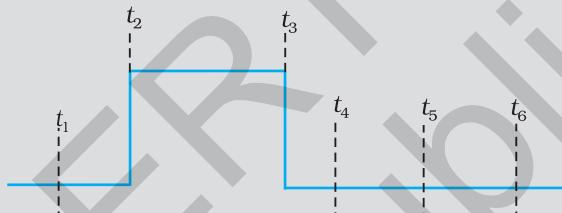
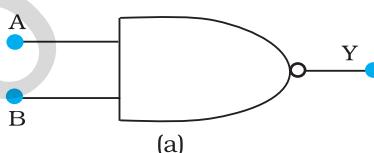


FIGURE 14.39

(iv) NAND Gate

This is an AND gate followed by a NOT gate. If inputs A and B are both '1', the output Y is *not* '1'. The gate gets its name from this NOT AND behaviour. Figure 14.40 shows the symbol and truth table of NAND gate.

NAND gates are also called *Universal Gates* since by using these gates you can realise other basic gates like OR, AND and NOT (Exercises 14.16 and 14.17).



(a)

Input		Output
A	B	Y
0	0	1
0	1	1
1	0	1
1	1	0

(b)

FIGURE 14.40 (a) Logic symbol, (b) Truth table of NAND gate.

Example 14.13 Sketch the output Y from a NAND gate having inputs A and B given below:

Solution

- For $t < t_1$; $A = 1, B = 1$; Hence $Y = 0$
- For t_1 to t_2 ; $A = 0, B = 0$; Hence $Y = 1$
- For t_2 to t_3 ; $A = 0, B = 1$; Hence $Y = 1$
- For t_3 to t_4 ; $A = 1, B = 0$; Hence $Y = 1$

EXAMPLE 14.13

- For t_4 to t_5 ; A = 1, B = 1; Hence Y = 0
- For t_5 to t_6 ; A = 0, B = 0; Hence Y = 1
- For $t > t_6$; A = 0, B = 1; Hence Y = 1

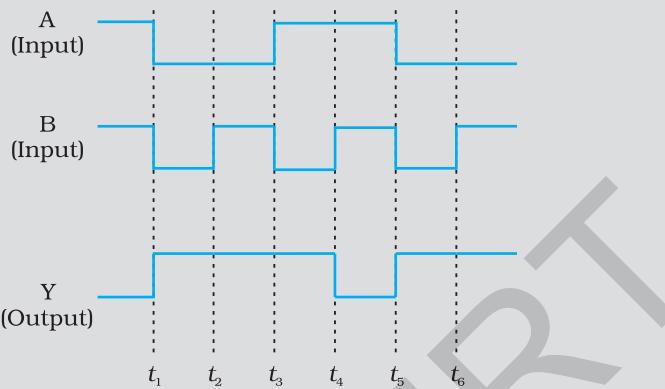
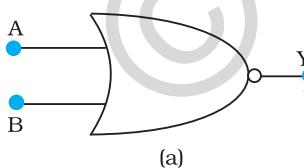


FIGURE 14.41

EXAMPLE 14.13

(v) NOR Gate

It has two or more inputs and one output. A NOT- operation applied after OR gate gives a NOT-OR gate (or simply NOR gate). Its output Y is '1' only when both inputs A and B are '0', i.e., neither one input nor the other is '1'. The symbol and truth table for NOR gate is given in Fig. 14.42.



Input		Output
A	B	Y
0	0	1
0	1	0
1	0	0
1	1	0

(b)

FIGURE 14.42 (a) Logic symbol, (b) Truth table of NOR gate.

NOR gates are considered as *universal* gates because you can obtain all the gates like AND, OR, NOT by using only NOR gates (Exercises 14.18 and 14.19).

14.11 INTEGRATED CIRCUITS

The conventional method of making circuits is to choose components like diodes, transistor, R , L , C etc., and connect them by soldering wires in the desired manner. Inspite of the miniaturisation introduced by the discovery of transistors, such circuits were still bulky. Apart from this, such circuits were less reliable and less shock proof. The concept of fabricating *an entire circuit* (consisting of many passive components like R and C and active devices like diode and transistor) on a small single block (or chip) of a semiconductor has revolutionised the electronics technology. Such a circuit is known as *Integrated Circuit* (IC). The most widely used technology is the *Monolithic Integrated Circuit*. The word

Physics

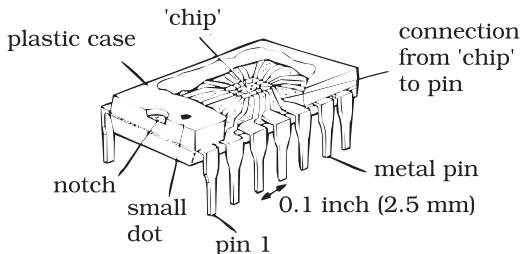


FIGURE 14.43 The casing and connection of a 'chip'.

monolithic is a combination of two greek words, *monos* means single and *lithos* means stone. This, in effect, means that the entire circuit is formed on a single silicon crystal (or *chip*). The *chip* dimensions are as small as $1\text{mm} \times 1\text{mm}$ or it could even be smaller. Figure 14.43 shows a chip in its protective plastic case, partly removed to reveal the connections coming out from the 'chip' to the pins that enable it to make external connections.

Depending on nature of input signals, IC's can be grouped in two categories: (a) *linear* or *analogue* IC's and (b) digital IC's. The linear IC's process analogue signals which change smoothly and continuously over a range of values between a maximum and a minimum. The output is more or less directly proportional to the input, i.e., it varies *linearly* with the input. One of the most useful linear IC's is the operational amplifier.

The digital IC's process signals that have only two values. They contain circuits such as logic gates. Depending upon the level of integration (i.e., the number of circuit components or logic gates), the ICs are termed as Small Scale Integration, SSI (logic gates ≤ 10); Medium Scale Integration, MSI (logic gates ≤ 100); Large Scale Integration, LSI (logic gates ≤ 1000); and Very Large Scale Integration, VLSI (logic gates > 1000). The technology of fabrication is very involved but large scale industrial production has made them very inexpensive.

FASTER AND SMALLER: THE FUTURE OF COMPUTER TECHNOLOGY

The *Integrated Chip* (IC) is at the heart of all computer systems. In fact ICs are found in almost all electrical devices like cars, televisions, CD players, cell phones etc. The miniaturisation that made the modern personal computer possible could never have happened without the IC. ICs are electronic devices that contain many transistors, resistors, capacitors, connecting wires – all in one package. You must have heard of the *microprocessor*. The microprocessor is an IC that processes all information in a computer, like keeping track of what keys are pressed, running programmes, games etc. The IC was first invented by Jack Kilby at Texas Instruments in 1958 and he was awarded Nobel Prize for this in 2000. ICs are produced on a piece of semiconductor crystal (or chip) by a process called *photolithography*. Thus, the entire Information Technology (IT) industry hinges on semiconductors. Over the years, the complexity of ICs has increased while the size of its features continued to shrink. In the past five decades, a dramatic miniaturisation in computer technology has made modern day computers *faster and smaller*. In the 1970s, Gordon Moore, co-founder of INTEL, pointed out that the memory capacity of a chip (IC) approximately doubled every one and a half years. This is popularly known as *Moore's law*. The number of transistors per chip has risen exponentially and each year computers are becoming more powerful, yet cheaper than the year before. It is intimated from current trends that the computers available in 2020 will operate at 40 GHz (40,000 MHz) and would be much smaller, more efficient and less expensive than present day computers. The explosive growth in the semiconductor industry and computer technology is best expressed by a famous quote from Gordon Moore: "If the auto industry advanced as rapidly as the semiconductor industry, a Rolls Royce would get half a million miles per gallon, and it would be cheaper to throw it away than to park it".

SUMMARY

1. Semiconductors are the basic materials used in the present solid state electronic devices like diode, transistor, ICs, etc.
2. Lattice structure and the atomic structure of constituent elements decide whether a particular material will be insulator, metal or semiconductor.
3. Metals have low resistivity (10^{-2} to $10^{-8} \Omega\text{m}$), insulators have very high resistivity ($>10^8 \Omega \text{ m}^{-1}$), while semiconductors have intermediate values of resistivity.
4. Semiconductors are elemental (Si, Ge) as well as compound (GaAs, CdS, etc.).
5. Pure semiconductors are called 'intrinsic semiconductors'. The presence of charge carriers (electrons and holes) is an 'intrinsic' property of the material and these are obtained as a result of thermal excitation. The number of electrons (n_e) is equal to the number of holes (n_h) in intrinsic conductors. Holes are essentially electron vacancies with an effective positive charge.
6. The number of charge carriers can be changed by 'doping' of a suitable impurity in pure semiconductors. Such semiconductors are known as extrinsic semiconductors. These are of two types (n-type and p-type).
7. In n-type semiconductors, $n_e \gg n_h$ while in p-type semiconductors $n_h \gg n_e$.
8. n-type semiconducting Si or Ge is obtained by doping with pentavalent atoms (donors) like As, Sb, P, etc., while p-type Si or Ge can be obtained by doping with trivalent atom (acceptors) like B, Al, In etc.
9. $n_e n_h = n_i^2$ in all cases. Further, the material possesses an *overall charge neutrality*.
10. There are two distinct band of energies (called valence band and conduction band) in which the electrons in a material lie. Valence band energies are low as compared to conduction band energies. All energy levels in the valence band are filled while energy levels in the conduction band may be fully empty or partially filled. The electrons in the conduction band are free to move in a solid and are responsible for the conductivity. The extent of conductivity depends upon the energy gap (E_g) between the top of valence band (E_V) and the bottom of the conduction band E_C . The electrons from valence band can be excited by heat, light or electrical energy to the conduction band and thus, produce a change in the current flowing in a semiconductor.
11. For insulators $E_g > 3 \text{ eV}$, for semiconductors E_g is 0.2 eV to 3 eV, while for metals $E_g \approx 0$.
12. p-n junction is the 'key' to all semiconductor devices. When such a junction is made, a 'depletion layer' is formed consisting of immobile ion-cores devoid of their electrons or holes. This is responsible for a junction potential barrier.
13. By changing the external applied voltage, junction barriers can be changed. In forward bias (n-side is connected to negative terminal of the battery and p-side is connected to the positive), the barrier is decreased while the barrier increases in reverse bias. Hence, forward bias current is more (mA) while it is very small (μA) in a p-n junction diode.
14. Diodes can be used for rectifying an ac voltage (restricting the ac voltage to one direction). With the help of a capacitor or a suitable filter, a dc voltage can be obtained.
15. There are some special purpose diodes.

Physics

16. Zener diode is one such special purpose diode. In reverse bias, after a certain voltage, the current suddenly increases (breakdown voltage) in a Zener diode. This property has been used to obtain *voltage regulation*.
17. p-n junctions have also been used to obtain many photonic or optoelectronic devices where one of the participating entity is 'photon': (a) Photodiodes in which photon excitation results in a change of reverse saturation current which helps us to measure light intensity; (b) Solar cells which convert photon energy into electricity; (c) Light Emitting Diode and Diode Laser in which electron excitation by a bias voltage results in the generation of light.
18. Transistor is an n-p-n or p-n-p junction device. The central block (thin and lightly doped) is called 'Base' while the other electrodes are 'Emitter' and 'Collectors'. The emitter-base junction is forward biased while collector-base junction is reverse biased.
19. The transistors can be connected in such a manner that either C or E or B is common to both the input and output. This gives the three configurations in which a transistor is used: Common Emitter (CE), Common Collector (CC) and Common Base (CB). The plot between I_C and V_{CE} for fixed I_B is called output characteristics while the plot between I_B and V_{BE} with fixed V_{CE} is called input characteristics. The important transistor parameters for CE-configuration are:

$$\text{input resistance, } r_i = \frac{\Delta V_{BE}}{\Delta I_B} \Big|_{V_{CE}}$$

$$\text{output resistance, } r_o = \frac{\Delta V_{CE}}{\Delta I_C} \Big|_{I_B}$$

$$\text{current amplification factor, } \beta = \frac{\Delta I_C}{\Delta I_B} \Big|_{V_{CE}}$$

20. Transistor can be used as an amplifier and oscillator. In fact, an oscillator can also be considered as a self-sustained amplifier in which a part of output is fed-back to the input in the same phase (positive feed back). The voltage gain of a transistor amplifier in common emitter configuration is: $A_v = \frac{v_o}{v_i} = \beta \frac{R_C}{R_B}$, where R_C and R_B are respectively the resistances in collector and base sides of the circuit.
21. When the transistor is used in the cutoff or saturation state, it acts as a switch.
22. There are some special circuits which handle the digital data consisting of 0 and 1 levels. This forms the subject of Digital Electronics.
23. The important digital circuits performing special logic operations are called logic gates. These are: OR, AND, NOT, NAND, and NOR gates.
24. In modern day circuit, many logical gates or circuits are integrated in one single 'Chip'. These are known as Integrated circuits (IC).

POINTS TO PONDER

1. The energy bands (E_c or E_v) in the semiconductors are space delocalised which means that these are not located in any specific place inside the solid. The energies are the overall averages. When you see a picture in which E_c or E_v are drawn as straight lines, then they should be respectively taken simply as the *bottom* of conduction band energy levels and *top* of valence band energy levels.

2. In elemental semiconductors (Si or Ge), the n-type or p-type semiconductors are obtained by introducing 'dopants' as defects. In compound semiconductors, the change in relative stoichiometric ratio can also change the type of semiconductor. For example, in ideal GaAs the ratio of Ga:As is 1:1 but in Ga-rich or As-rich GaAs it could respectively be $\text{Ga}_{1.1} \text{As}_{0.9}$ or $\text{Ga}_{0.9} \text{As}_{1.1}$. In general, the presence of defects control the properties of semiconductors in many ways.
3. In transistors, the base region is both narrow and lightly doped, otherwise the electrons or holes coming from the input side (say, emitter in CE-configuration) will not be able to reach the collector.
4. We have described an oscillator as a positive feedback amplifier. For stable oscillations, the voltage feedback (V_{fb}) from the output voltage (V_o) should be such that after amplification (A) it should again become V_o . If a fraction β' is feedback, then $V_{fb} = V_o \cdot \beta'$ and after amplification its value $A(v_o \cdot \beta')$ should be equal to V_o . This means that the criteria for stable oscillations to be sustained is $A \beta' = 1$. This is known as Barkhausen's Criteria.
5. In an oscillator, the feedback is in the same phase (positive feedback). If the feedback voltage is in opposite phase (negative feedback), the gain is less than 1 and it can never work as oscillator. It will be an amplifier with reduced gain. However, the negative feedback also reduces noise and distortion in an amplifier which is an advantageous feature.

EXERCISES

- 14.1** In an n-type silicon, which of the following statement is true:
- (a) Electrons are majority carriers and trivalent atoms are the dopants.
 - (b) Electrons are minority carriers and pentavalent atoms are the dopants.
 - (c) Holes are minority carriers and pentavalent atoms are the dopants.
 - (d) Holes are majority carriers and trivalent atoms are the dopants.
- 14.2** Which of the statements given in Exercise 14.1 is true for p-type semiconductors.
- 14.3** Carbon, silicon and germanium have four valence electrons each. These are characterised by valence and conduction bands separated by energy band gap respectively equal to $(E_g)_C$, $(E_g)_{\text{Si}}$ and $(E_g)_{\text{Ge}}$. Which of the following statements is true?
- (a) $(E_g)_{\text{Si}} < (E_g)_{\text{Ge}} < (E_g)_C$
 - (b) $(E_g)_C < (E_g)_{\text{Ge}} > (E_g)_{\text{Si}}$
 - (c) $(E_g)_C > (E_g)_{\text{Si}} > (E_g)_{\text{Ge}}$
 - (d) $(E_g)_C = (E_g)_{\text{Si}} = (E_g)_{\text{Ge}}$
- 14.4** In an unbiased p-n junction, holes diffuse from the p-region to n-region because
- (a) free electrons in the n-region attract them.
 - (b) they move across the junction by the potential difference.
 - (c) hole concentration in p-region is more as compared to n-region.
 - (d) All the above.

Physics

- 14.5** When a forward bias is applied to a p-n junction, it
- raises the potential barrier.
 - reduces the majority carrier current to zero.
 - lowers the potential barrier.
 - None of the above.
- 14.6** For transistor action, which of the following statements are correct:
- Base, emitter and collector regions should have similar size and doping concentrations.
 - The base region must be very thin and lightly doped.
 - The emitter junction is forward biased and collector junction is reverse biased.
 - Both the emitter junction as well as the collector junction are forward biased.
- 14.7** For a transistor amplifier, the voltage gain
- remains constant for all frequencies.
 - is high at high and low frequencies and constant in the middle frequency range.
 - is low at high and low frequencies and constant at mid frequencies.
 - None of the above.
- 14.8** In half-wave rectification, what is the output frequency if the input frequency is 50 Hz. What is the output frequency of a full-wave rectifier for the same input frequency.
- 14.9** For a CE-transistor amplifier, the audio signal voltage across the collected resistance of $2\text{ k}\Omega$ is 2 V. Suppose the current amplification factor of the transistor is 100, find the input signal voltage and base current, if the base resistance is $1\text{ k}\Omega$.
- 14.10** Two amplifiers are connected one after the other in series (cascaded). The first amplifier has a voltage gain of 10 and the second has a voltage gain of 20. If the input signal is 0.01 volt, calculate the output ac signal.
- 14.11** A p-n photodiode is fabricated from a semiconductor with band gap of 2.8 eV. Can it detect a wavelength of 6000 nm?

ADDITIONAL EXERCISES

- 14.12** The number of silicon atoms per m^3 is 5×10^{28} . This is doped simultaneously with 5×10^{22} atoms per m^3 of Arsenic and 5×10^{20} per m^3 atoms of Indium. Calculate the number of electrons and holes. Given that $n_i = 1.5 \times 10^{16} \text{ m}^{-3}$. Is the material n-type or p-type?
- 14.13** In an intrinsic semiconductor the energy gap E_g is 1.2eV. Its hole mobility is much smaller than electron mobility and independent of temperature. What is the ratio between conductivity at 600K and that at 300K? Assume that the temperature dependence of intrinsic carrier concentration n_i is given by

$$n_i = n_0 \exp -\frac{E_g}{2k_B T}$$

where n_0 is a constant.

14.14 In a p-n junction diode, the current I can be expressed as

$$I = I_0 \exp \frac{eV}{2k_B T} - 1$$

where I_0 is called the reverse saturation current, V is the voltage across the diode and is positive for forward bias and negative for reverse bias, and I is the current through the diode, k_B is the Boltzmann constant (8.6×10^{-5} eV/K) and T is the absolute temperature. If for a given diode $I_0 = 5 \times 10^{-12}$ A and $T = 300$ K, then

- (a) What will be the forward current at a forward voltage of 0.6 V?
- (b) What will be the increase in the current if the voltage across the diode is increased to 0.7 V?
- (c) What is the dynamic resistance?
- (d) What will be the current if reverse bias voltage changes from 1 V to 2 V?

14.15 You are given the two circuits as shown in Fig. 14.44. Show that circuit (a) acts as OR gate while the circuit (b) acts as AND gate.

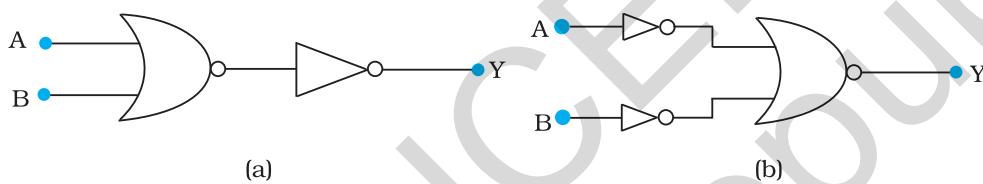


FIGURE 14.44

14.16 Write the truth table for a NAND gate connected as given in Fig. 14.45.



FIGURE 14.45

Hence identify the exact logic operation carried out by this circuit.

14.17 You are given two circuits as shown in Fig. 14.46, which consist of NAND gates. Identify the logic operation carried out by the two circuits.

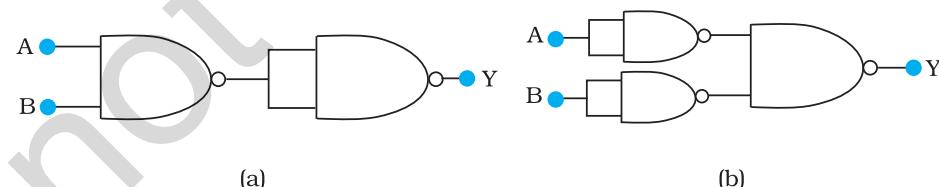


FIGURE 14.46

14.18 Write the truth table for circuit given in Fig. 14.47 below consisting of NOR gates and identify the logic operation (OR, AND, NOT) which this circuit is performing.

Physics

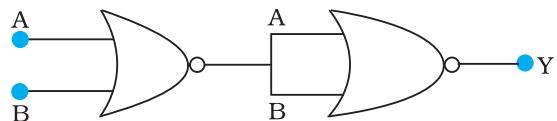


FIGURE 14.47

(Hint: $A = 0, B = 1$ then A and B inputs of second NOR gate will be 0 and hence $Y=1$. Similarly work out the values of Y for other combinations of A and B. Compare with the truth table of OR, AND, NOT gates and find the correct one.)

- 14.19** Write the truth table for the circuits given in Fig. 14.48 consisting of NOR gates only. Identify the logic operations (OR, AND, NOT) performed by the two circuits.

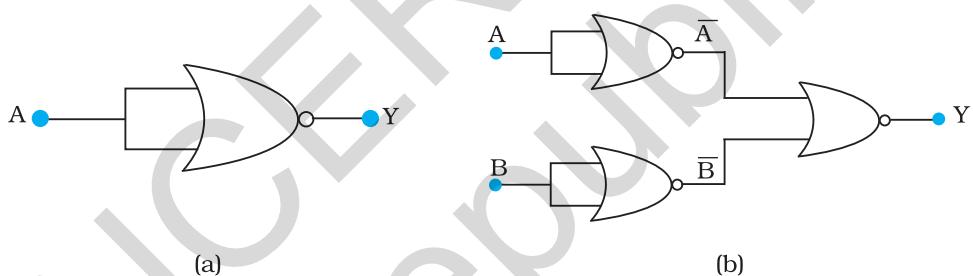


FIGURE 14.48

Chapter Fifteen

COMMUNICATION

SYSTEMS

15.1 INTRODUCTION

Communication is the act of transmission of information. Every living creature in the world experiences the need to impart or receive information almost continuously with others in the surrounding world. For communication to be successful, it is essential that the sender and the receiver understand a common language. Man has constantly made endeavors to improve the quality of communication with other human beings. Languages and methods used in communication have kept evolving from prehistoric to modern times, to meet the growing demands in terms of speed and complexity of information. It would be worthwhile to look at the major milestones in events that promoted developments in communications, as presented in Table 15.1.

Modern communication has its roots in the 19th and 20th century in the work of scientists like J.C. Bose, F.B. Morse, G. Marconi and Alexander Graham Bell. The pace of development seems to have increased dramatically after the first half of the 20th century. We can hope to see many more accomplishments in the coming decades. The aim of this chapter is to introduce the concepts of communication, namely the mode of communication, the need for modulation, production and deduction of amplitude modulation.

15.2 ELEMENTS OF A COMMUNICATION SYSTEM

Communication pervades all stages of life of all living creatures. Irrespective of its nature, every communication system has three essential elements-

Physics

TABLE 15.1 SOME MAJOR MILESTONES IN THE HISTORY OF COMMUNICATION

Year	Event	Remarks
Around 1565 A.D.	The reporting of the delivery of a child by queen using drum beats from a distant place to King Akbar.	It is believed that minister Birbal experimented with the arrangement to decide the number of drummers posted between the place where the queen stayed and the place where the king stayed.
1835	Invention of telegraph by Samuel F.B. Morse and Sir Charles Wheatstone	It resulted in tremendous growth of messages through post offices and reduced physical travel of messengers considerably.
1876	Telephone invented by Alexander Graham Bell and Antonio Meucci	Perhaps the most widely used means of communication in the history of mankind.
1895	Jagadis Chandra Bose and Guglielmo Marconi demonstrated wireless telegraphy.	It meant a giant leap – from an era of communication using wires to communicating without using wires. (wireless)
1936	Television broadcast(John Logi Baird)	First television broadcast by BBC
1955	First radio FAX transmitted across continent.(Alexander Bain)	The idea of FAX transmission was patented by Alexander Bain in 1843.
1968	ARPANET- the first internet came into existence(J.C.R. Licklider)	ARPANET was a project undertaken by the U.S. defence department. It allowed file transfer from one computer to another connected to the network.
1975	Fiber optics developed at Bell Laboratories	Fiber optical systems are superior and more economical compared to traditional communication systems.
1989-91	Tim Berners-Lee invented the World Wide Web .	WWW may be regarded as the mammoth encyclopedia of knowledge accessible to everyone round the clock throughout the year.

Communication System

transmitter, medium/channel and receiver. The block diagram shown in Fig. 15.1 depicts the general form of a communication system.

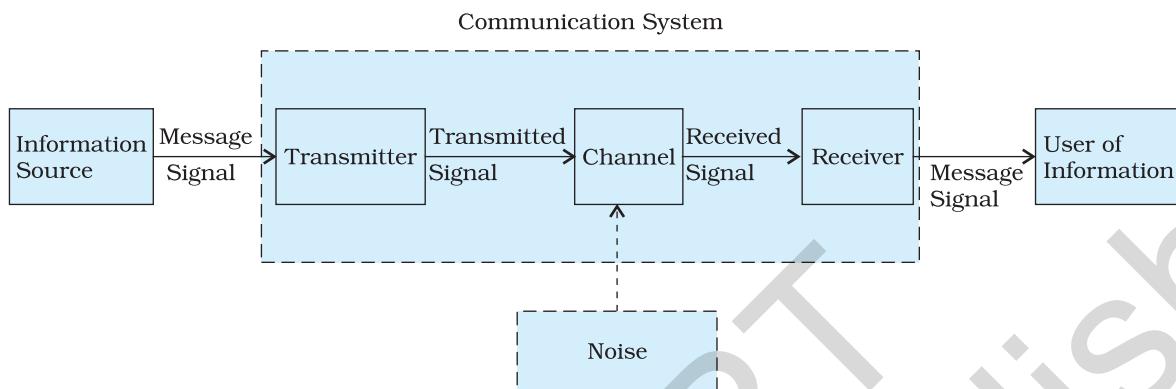


FIGURE 15.1 Block diagram of a generalised communication system.

In a communication system, the transmitter is located at one place, the receiver is located at some other place (far or near) separate from the transmitter and the channel is the physical medium that connects them. Depending upon the type of communication system, a channel may be in the form of wires or cables connecting the transmitter and the receiver or it may be wireless. The purpose of the transmitter is to convert the message signal produced by the source of information into a form suitable for transmission through the channel. If the output of the information source is a non-electrical signal like a voice signal, a transducer converts it to electrical form before giving it as an input to the transmitter. When a transmitted signal propagates along the channel it may get distorted due to channel imperfection. Moreover, noise adds to the transmitted signal and the receiver receives a corrupted version of the transmitted signal. The receiver has the task of operating on the received signal. It reconstructs a recognisable form of the original message signal for delivering it to the user of information.

There are two basic modes of communication: *point-to-point* and *broadcast*.

In point-to-point communication mode, communication takes place over a link between a single transmitter and a receiver. Telephony is an example of such a mode of communication. In contrast, in the broadcast mode, there are a large number of receivers corresponding to a single transmitter. Radio and television are examples of broadcast mode of communication.

15.3 BASIC TERMINOLOGY USED IN ELECTRONIC COMMUNICATION SYSTEMS

By now, we have become familiar with some terms like information source, transmitter, receiver, channel, noise, etc. It would be easy to understand the principles underlying any communication, if we get ourselves acquainted with the following basic terminology.

Physics



Jagadis Chandra Bose (1858 – 1937) He developed an apparatus for generating ultrashort electro-magnetic waves and studied their quasi-optical properties. He was said to be the first to employ a semiconductor like galena as a self-recovering detector of electromagnetic waves. Bose published three papers in the British magazine, 'The Electrician' of 27 Dec. 1895. His invention was published in the 'Proceedings of The Royal Society' on 27 April 1899 over two years before Marconi's first wireless communication on 13 December 1901. Bose also invented highly sensitive instruments for the detection of minute responses by living organisms to external stimuli and established parallelism between animal and plant tissues.

JAGADIS CHANDRA BOSE (1858 – 1937)

- (i) **Transducer:** Any device that converts one form of energy into another can be termed as a transducer. In electronic communication systems, we usually come across devices that have either their inputs or outputs in the electrical form. An electrical transducer may be defined as a device that converts some physical variable (pressure, displacement, force, temperature, etc) into corresponding variations in the electrical signal at its output.
- (ii) **Signal:** Information converted in electrical form and suitable for transmission is called a signal. Signals can be either *analog* or *digital*. Analog signals are continuous variations of voltage or current. *They are essentially single-valued functions of time.* Sine wave is a fundamental analog signal. All other analog signals can be fully understood in terms of their sine wave components. Sound and picture signals in TV are analog in nature. Digital signals are those which can take only discrete stepwise values. Binary system that is extensively used in digital electronics employs just two levels of a signal. '0' corresponds to a low level and '1' corresponds to a high level of voltage/ current. There are several coding schemes useful for digital communication. They employ suitable combinations of number systems such as the binary coded decimal (BCD)*. American Standard Code for Information Interchange (ASCII)** is a universally popular digital code to represent numbers, letters and certain characters.
- (iii) **Noise:** Noise refers to the unwanted signals that tend to disturb the transmission and processing of message signals in a communication system. The source generating the noise may be located inside or outside the system.
- (iv) **Transmitter:** A transmitter processes the incoming message signal so as to make it suitable for transmission through a channel and subsequent reception.
- (v) **Receiver:** A receiver extracts the desired message signals from the received signals at the channel output.
- (vi) **Attenuation:** The loss of strength of a signal while propagating through a medium is known as attenuation.

* In BCD, a digit is usually represented by four binary (0 or 1) bits. For example the numbers 0, 1, 2, 3, 4 in the decimal system are written as 0000, 0001, 0010, 0011 and 0100. 1000 would represent eight.

** It is a character encoding in terms of numbers based on English alphabet since the computer can only understand numbers.

- (vii) **Amplification:** It is the process of *increasing the amplitude* (and consequently the strength) of a signal using an electronic circuit called the amplifier (reference Chapter 14). Amplification is necessary to compensate for the attenuation of the signal in communication systems. The energy needed for additional signal strength is obtained from a DC power source. Amplification is done at a place between the source and the destination wherever signal strength becomes weaker than the required strength.
- (viii) **Range:** It is the largest distance between a source and a destination up to which the signal is received with sufficient strength.
- (ix) **Bandwidth:** Bandwidth refers to the frequency range over which an equipment operates or the portion of the spectrum occupied by the signal.
- (x) **Modulation:** The original low frequency message/information signal cannot be transmitted to long distances because of reasons given in Section 15.7. Therefore, at the transmitter, information contained in the low frequency message signal is superimposed on a high frequency wave, which acts as a carrier of the information. This process is known as modulation. As will be explained later, there are several types of modulation, abbreviated as AM, FM and PM.
- (xi) **Demodulation:** The process of retrieval of information from the carrier wave at the receiver is termed demodulation. This is the reverse process of modulation.
- (xii) **Repeater:** A repeater is a combination of a receiver and a transmitter. A repeater, picks up the signal from the transmitter, amplifies and retransmits it to the receiver sometimes with a change in carrier frequency. Repeaters are used to extend the range of a communication system as shown in Fig. 15.2. A communication satellite is essentially a repeater station in space.

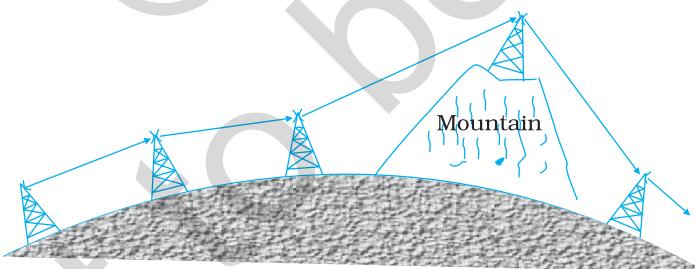


FIGURE 15.2 Use of repeater station to increase the range of communication.

15.4 BANDWIDTH OF SIGNALS

In a communication system, the message signal can be voice, music, picture or computer data. Each of these signals has different ranges of frequencies. The type of communication system needed for a given signal depends on the band of frequencies which is considered essential for the communication process.

For speech signals, frequency range 300 Hz to 3100 Hz is considered adequate. Therefore speech signal requires a bandwidth of 2800 Hz (3100 Hz – 300 Hz) for commercial telephonic communication. To transmit music,

an approximate bandwidth of 20 kHz is required because of the high frequencies produced by the musical instruments. The audible range of frequencies extends from 20 Hz to 20 kHz.

Video signals for transmission of pictures require about 4.2 MHz of bandwidth. A TV signal contains both voice and picture and is usually allocated 6 MHz of bandwidth for transmission.

In the preceding paragraph, we have considered only analog signals. Digital signals are in the form of rectangular waves as shown in Fig. 15.3. One can show that this rectangular wave can be decomposed into a superposition of sinusoidal waves of frequencies v_0 , $2v_0$, $3v_0$, $4v_0 \dots nv_0$ where n is an integer extending to infinity and $v_0 = 1/T_0$. The fundamental (v_0), fundamental (v_0) + second harmonic ($2v_0$), and fundamental (v_0) + second harmonic ($2v_0$) + third harmonic ($3v_0$), are shown in the same figure to illustrate this fact. It is clear that to reproduce the rectangular wave shape exactly we need to superimpose all the harmonics v_0 , $2v_0$, $3v_0$, $4v_0 \dots$, which implies an infinite bandwidth. However, for practical purposes, the contribution from higher harmonics can be neglected, thus limiting the bandwidth. As a result, received waves are a distorted version of the transmitted one. If the bandwidth is large enough to accommodate a few harmonics, the information is not lost and the rectangular signal is more or less recovered. This is so because the higher the harmonic, less is its contribution to the wave form.

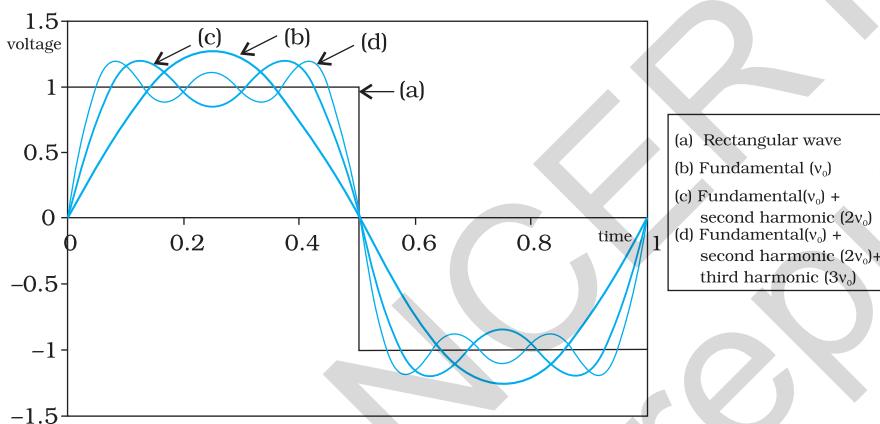


FIGURE 15.3 Approximation of a rectangular wave in terms of a fundamental sine wave and its harmonics.

transmitted one. If the bandwidth is large enough to accommodate a few harmonics, the information is not lost and the rectangular signal is more or less recovered. This is so because the higher the harmonic, less is its contribution to the wave form.

15.5 BANDWIDTH OF TRANSMISSION MEDIUM

Similar to message signals, different types of transmission media offer different bandwidths. The commonly used transmission media are wire, free space and fiber optic cable. Coaxial cable is a widely used wire medium, which offers a bandwidth of approximately 750 MHz. Such cables are normally operated below 18 GHz. Communication through free space using radio waves takes place over a very wide range of frequencies: from a few hundreds of kHz to a few GHz. This range of frequencies is further subdivided and allocated for various services as indicated in Table 15.2. Optical communication using fibers is performed in the frequency range of 1 THz to 1000 THz (microwaves to ultraviolet). An optical fiber can offer a transmission bandwidth in excess of 100 GHz.

Spectrum allocations are arrived at by an international agreement. The International Telecommunication Union (ITU) administers the present system of frequency allocations.

TABLE 15.2 SOME IMPORTANT WIRELESS COMMUNICATION FREQUENCY BANDS

Service	Frequency bands	Comments
Standard AM broadcast	540-1600 kHz	
FM broadcast	88-108 MHz	
Television	54-72 MHz 76-88 MHz 174-216 MHz 420-890 MHz	VHF (very high frequencies) TV UHF (ultra high frequencies) TV
Cellular Mobile Radio	896-901 MHz 840-935 MHz	Mobile to base station Base station to mobile
Satellite Communication	5.925-6.425 GHz 3.7-4.2 GHz	Uplink Downlink

15.6 PROPAGATION OF ELECTROMAGNETIC WAVES

In communication using radio waves, an antenna at the transmitter radiates the Electromagnetic waves (em waves), which travel through the space and reach the receiving antenna at the other end. As the em wave travels away from the transmitter, the strength of the wave keeps on decreasing. Several factors influence the propagation of em waves and the path they follow. At this point, it is also important to understand the composition of the earth's atmosphere as it plays a vital role in the propagation of em waves. A brief discussion on some useful layers of the atmosphere is given in Table 15.3.

15.6.1 Ground wave

To radiate signals with high efficiency, the antennas should have a size comparable to the wavelength λ of the signal (at least $\sim \lambda/4$). At longer wavelengths (i.e., at lower frequencies), the antennas have large physical size and they are located on or very near to the ground. In standard AM broadcast, ground based vertical towers are generally used as transmitting antennas. For such antennas, ground has a strong influence on the propagation of the signal. The mode of propagation is called surface wave propagation and the wave glides over the surface of the earth. A wave induces current in the ground over which it passes and it is attenuated as a result of absorption of energy by the earth. The attenuation of surface waves increases very rapidly with increase in frequency. The maximum range of coverage depends on the transmitted power and frequency (less than a few MHz).

Physics

TABLE 15.3 DIFFERENT LAYERS OF ATMOSPHERE AND THEIR INTERACTION WITH THE PROPAGATING ELECTROMAGNETIC WAVES

Name of the stratum (layer)	Approximate height over earth's surface	Exists during	Frequencies most affected	
Troposphere	10 km	Day and night	VHF (up to several GHz)	
D (part of stratosphere)	P A R T S O F I O N O S P H E R E	65-75 km	Day only	Reflects LF, absorbs MF and HF to some degree
E (part of Stratosphere)		100 km	Day only	Helps surface waves, reflects HF
F ₁ (Part of Mesosphere)		170-190 km	Daytime, merges with F ₂ at night	Partially absorbs HF waves yet allowing them to reach F ₂
F ₂ (Thermosphere)		300 km at night, 250-400 km during daytime	Day and night	Efficiently reflects HF waves, particularly at night

15.6.2 Sky waves

In the frequency range from a few MHz up to 30 to 40 MHz, long distance communication can be achieved by ionospheric reflection of radio waves back towards the earth. This mode of propagation is called *sky wave propagation* and is used by short wave broadcast services. The ionosphere is so called because of the presence of a large number of ions or charged particles. It extends from a height of ~ 65 Km to about 400 Km above the earth's surface. Ionisation occurs due to the absorption of the ultraviolet and other high-energy radiation coming from the sun by air molecules. The ionosphere is further subdivided into several layers, the details of which are given in Table 15.3. The degree of ionisation varies with the height. The density of atmosphere decreases with height. At great heights the solar radiation is intense but there are few molecules to be ionised. Close to the earth, even though the molecular concentration is very high, the radiation intensity is low so that the ionisation is again low. However, at some intermediate heights, there occurs a peak of ionisation density. The ionospheric layer acts as a reflector for a certain range of frequencies (3 to 30 MHz). Electromagnetic waves of frequencies higher than 30 MHz penetrate the ionosphere and escape. These phenomena are shown in the Fig. 15.4. The phenomenon of bending of em waves so that they are diverted towards the earth is similar to total internal reflection in optics*.

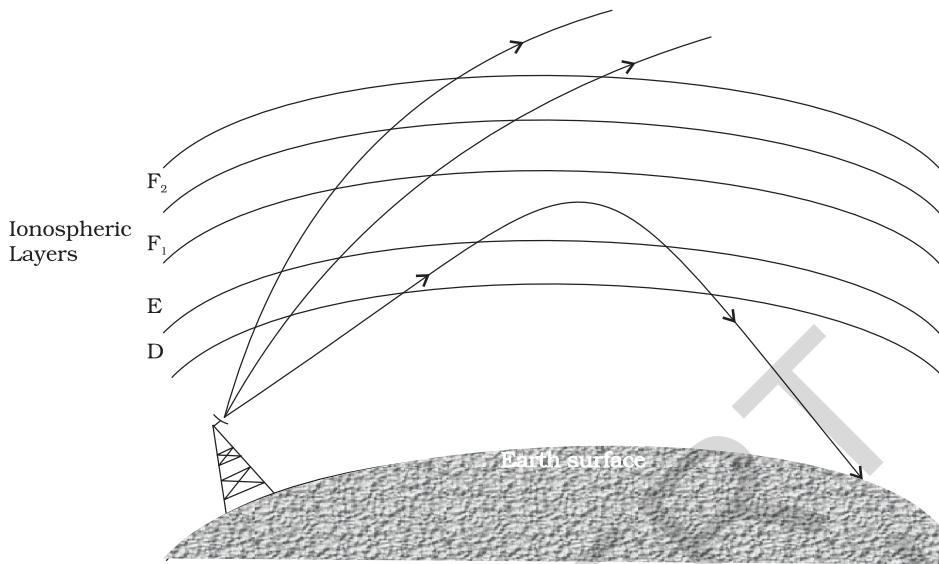


FIGURE 15.4 Sky wave propagation. The layer nomenclature is given in Table 15.3.

15.6.3 Space wave

Another mode of radio wave propagation is by *space waves*. A space wave travels in a straight line from transmitting antenna to the receiving antenna. Space waves are used for line-of-sight (LOS) communication as well as satellite communication. At frequencies above 40 MHz, communication is essentially limited to line-of-sight paths. At these frequencies, the antennas are relatively smaller and can be placed at heights of many wavelengths above the ground. Because of line-of-sight nature of propagation, direct waves get blocked at some point by the curvature of the earth as illustrated in Fig. 15.5. If the signal is to be received beyond the horizon then the receiving antenna must be high enough to intercept the line-of-sight waves.

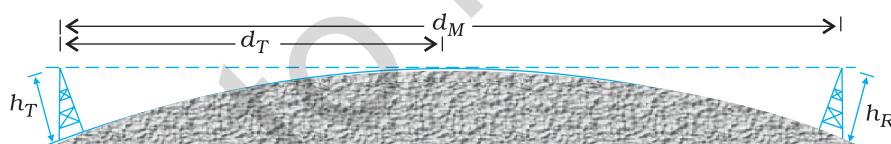


FIGURE 15.5 Line of sight communication by space waves.

If the transmitting antenna is at a height h_T then you can show that the distance to the horizon d_T is given as , where R is the radius of the earth (approximately 6400 km). d_T is also called the radio horizon of the transmitting antenna. With reference to Fig. 15.5 the maximum line-of-sight distance d_M between the two antennas having heights h_T and h_R above the earth is given by

$$d_M = \sqrt{2Rh_T} + \sqrt{2Rh_R} \quad (15.1)$$

where h_R is the height of receiving antenna.

Physics

Television broadcast, microwave links and satellite communication are some examples of communication systems that use space wave mode of propagation. Figure 15.6 summarises the various modes of wave propagation discussed so far.

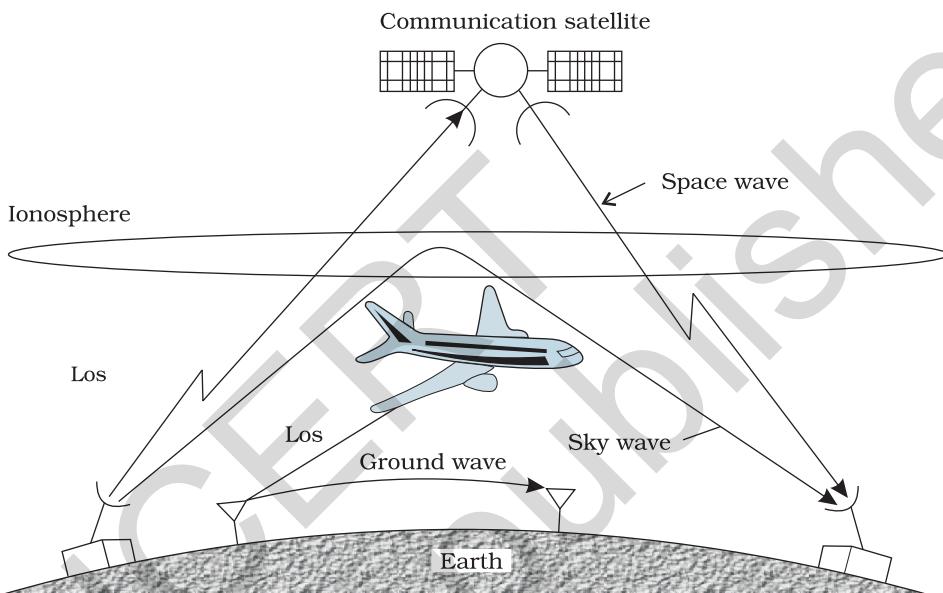


FIGURE 15.6 Various propagation modes for em waves.

EXAMPLE 15.1

Example 15.1 A transmitting antenna at the top of a tower has a height 32 m and the height of the receiving antenna is 50 m. What is the maximum distance between them for satisfactory communication in LOS mode? Given radius of earth 6.4×10^6 m.

Solution

$$\begin{aligned}d_m &= \sqrt{2 \times 64 \times 10^5 \times 32} + \sqrt{2 \times 64 \times 10^5 \times 50} \text{ m} \\&= 64 \times 10^2 \times \sqrt{10} + 8 \times 10^3 \times \sqrt{10} \text{ m} \\&= 144 \times 10^2 \times \sqrt{10} \text{ m} = 45.5 \text{ km}\end{aligned}$$

15.7 MODULATION AND ITS NECESSITY

As already mentioned, the purpose of a communication system is to transmit information or message signals. Message signals are also called *baseband signals*, which essentially designate the band of frequencies representing the original signal, as delivered by the source of information. No signal, in general, is a single frequency sinusoid, but it spreads over a range of frequencies called the *signal bandwidth*. Suppose we wish to transmit an electronic signal in the audio frequency (AF) range (baseband signal frequency less than 20 kHz) over a long distance directly. Let us find what factors prevent us from doing so and how we overcome these factors,

15.7.1 Size of the antenna or aerial

For transmitting a signal, we need an antenna or an aerial. This antenna should have a size comparable to the wavelength of the signal (at least $\lambda/4$ in dimension) so that the antenna properly senses the time variation of the signal. For an electromagnetic wave of frequency 20 kHz, the wavelength λ is 15 km. Obviously, such a long antenna is not possible to construct and operate. Hence direct transmission of such baseband signals is not practical. We can obtain transmission with reasonable antenna lengths if transmission frequency is high (for example, if *viz* 1 MHz, then λ is 300 m). Therefore, there is a need of translating the information contained in our original low frequency baseband signal into high or radio frequencies before transmission.

15.7.2 Effective power radiated by an antenna

A theoretical study of radiation from a linear antenna (length l) shows that the power radiated is proportional to $(l/\lambda)^2$. This implies that for the same antenna length, the power radiated increases with decreasing λ , i.e., increasing frequency. Hence, the effective power radiated by a long wavelength baseband signal would be small. For a good transmission, we need high powers and hence this also points out to the need of using high frequency transmission.

15.7.3 Mixing up of signals from different transmitters

Another important argument against transmitting baseband signals directly is more practical in nature. Suppose many people are talking at the same time or many transmitters are transmitting baseband information signals simultaneously. All these signals will get mixed up and there is no simple way to distinguish between them. This points out towards a possible solution by using communication at high frequencies and allotting a band of frequencies to each message signal for its transmission.

The above arguments suggest that there is a need for translating the original low frequency baseband message or information signal into high frequency wave before transmission such that the translated signal continues to possess the information contained in the original signal. In doing so, we take the help of a high frequency signal, known as the *carrier wave*, and a process known as modulation which attaches information to it. The carrier wave may be continuous (sinusoidal) or in the form of pulses as shown in Fig. 15.7.

A sinusoidal carrier wave can be represented as

$$c(t) = A_c \sin(\omega_c t + \phi) \quad (15.2)$$

where $c(t)$ is the signal strength (voltage or current), A_c is the amplitude, $\omega_c (= 2\pi\nu_c)$ is the angular frequency and ϕ is the initial phase of the carrier wave. During the process of modulation, any of the three parameters, *viz* A_c , ω_c and ϕ , of the carrier wave can be controlled by the message or

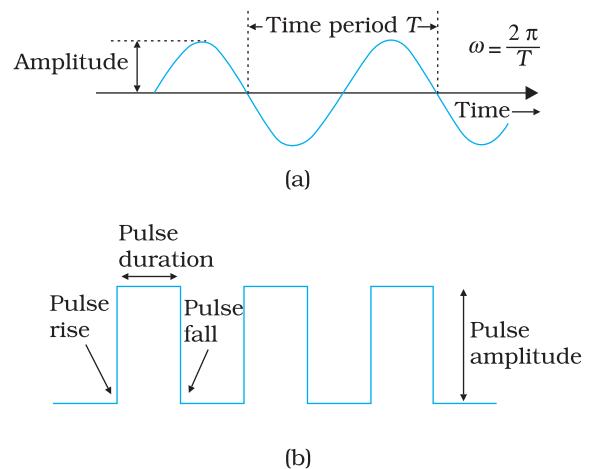


FIGURE 15.7 (a) Sinusoidal, and (b) pulse shaped signals.

Physics

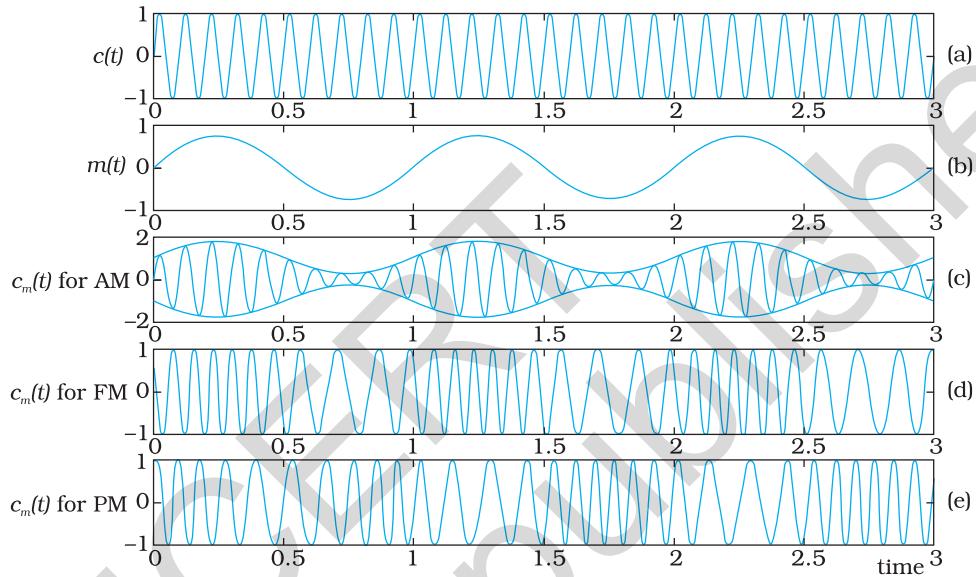


FIGURE 15.8 Modulation of a carrier wave: (a) a sinusoidal carrier wave; (b) a modulating signal; (c) amplitude modulation; (d) frequency modulation; and (e) phase modulation.

Similarly, the significant characteristics of a pulse are: pulse amplitude, pulse duration or pulse width, and pulse position (denoting the time of rise or fall of the pulse amplitude) as shown in Fig. 15.7(b). Hence, different types of pulse modulation are: (a) pulse amplitude modulation (PAM), (b) pulse duration modulation (PDM) or pulse width modulation (PWM), and (c) pulse position modulation (PPM). In this chapter, we shall confine to amplitude modulation only.

15.8 AMPLITUDE MODULATION

In amplitude modulation the amplitude of the carrier is varied in accordance with the information signal. Here we explain amplitude modulation process using a sinusoidal signal as the modulating signal.

Let $c(t) = A_c \sin \omega_c t$ represent carrier wave and $m(t) = A_m \sin \omega_m t$ represent the message or the modulating signal where $\omega_m = 2\pi f_m$ is the angular frequency of the message signal. The modulated signal $c_m(t)$ can be written as

$$\begin{aligned} c_m(t) &= (A_c + A_m \sin \omega_m t) \sin \omega_c t \\ &= A_c \left[1 + \frac{A_m}{A_c} \sin \omega_m t \right] \sin \omega_c t \end{aligned} \quad (15.3)$$

Note that the modulated signal now contains the message signal. This can also be seen from Fig. 15.8(c). From Eq. (15.3), we can write,

$$c_m(t) = A_c \sin \omega_c t + \mu A_c \sin \omega_m t \sin \omega_c t \quad (15.4)$$

Here $\mu = A_m/A_c$ is the *modulation index*; in practice, μ is kept ≤ 1 to avoid distortion.

Using the trigonometric relation $\sin A \sin B = \frac{1}{2} (\cos(A - B) - \cos(A + B))$, we can write $c_m(t)$ of Eq. (15.4) as

$$c_m(t) = A_c \sin \omega_c t + \frac{\mu A_c}{2} \cos(\omega_c - \omega_m) t - \frac{\mu A_c}{2} \cos(\omega_c + \omega_m) t \quad (15.5)$$

Here $\omega_c - \omega_m$ and $\omega_c + \omega_m$ are respectively called the lower side and upper side frequencies. The modulated signal now consists of the carrier wave of frequency ω_c plus two sinusoidal waves each with a frequency slightly different from, known as side bands. The frequency spectrum of the amplitude modulated signal is shown in Fig. 15.9.

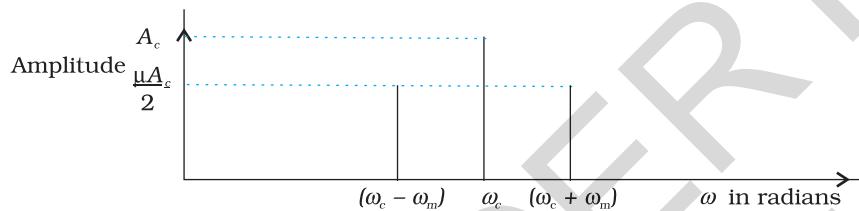


FIGURE 15.9 A plot of amplitude versus ω for an amplitude modulated signal.

As long as the broadcast frequencies (carrier waves) are sufficiently spaced out so that sidebands do not overlap, different stations can operate without interfering with each other.

Example 15.2 A message signal of frequency 10 kHz and peak voltage of 10 volts is used to modulate a carrier of frequency 1 MHz and peak voltage of 20 volts. Determine (a) modulation index, (b) the side bands produced.

Solution

- (a) Modulation index = $10/20 = 0.5$
- (b) The side bands are at $(1000+10 \text{ kHz})=1010 \text{ kHz}$ and $(1000 - 10 \text{ kHz}) = 990 \text{ kHz}$.

EXAMPLE 15.2

15.9 PRODUCTION OF AMPLITUDE MODULATED WAVE

Amplitude modulation can be produced by a variety of methods. A conceptually simple method is shown in the block diagram of Fig. 15.10.

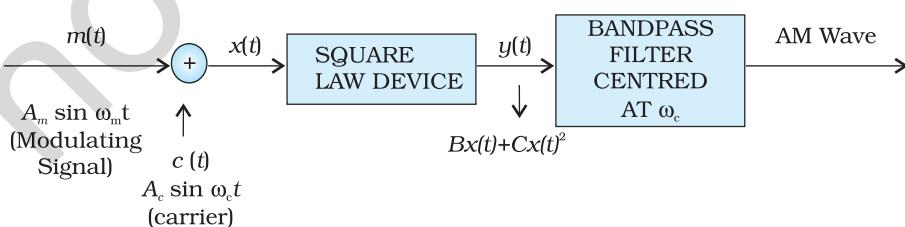


FIGURE 15.10 Block diagram of a simple modulator for obtaining an AM signal.

Physics

Here the modulating signal $A_m \sin \omega_m t$ is added to the carrier signal $A_c \sin \omega_c t$ to produce the signal $x(t) = A_m \sin \omega_m t + A_c \sin \omega_c t$. This signal $x(t) = A_m \sin \omega_m t + A_c \sin \omega_c t$ is passed through a square law device which is a non-linear device which produces an output

$$y(t) = Bx(t) + Cx^2(t) \quad (15.6)$$

where B and C are constants. Thus,

$$\begin{aligned} y(t) &= BA_m \sin \omega_m t + BA_c \sin \omega_c t \\ &+ C A_m^2 \sin^2 \omega_m t + C A_c^2 \sin^2 \omega_c t + 2A_m A_c \sin \omega_m t \sin \omega_c t \\ &= BA_m \sin \omega_m t + BA_c \sin \omega_c t \\ &+ \frac{CA_m^2}{2} + A_c^2 - \frac{CA_m^2}{2} \cos 2\omega_m t - \frac{CA_c^2}{2} \cos 2\omega_c t \end{aligned} \quad (15.7)$$

$$+ CA_m A_c \cos(\omega_c - \omega_m)t - CA_m A_c \cos(\omega_c + \omega_m)t \quad (15.8)$$

where the trigonometric relations $\sin^2 A = (1 - \cos 2A)/2$ and the relation for $\sin A \sin B$ mentioned earlier are used.

In Eq. (15.8), there is a dc term $C/2 (A_m^2 + A_c^2)$ and sinusoids of frequencies ω_m , $2\omega_m$, ω_c , $2\omega_c$, $\omega_c - \omega_m$ and $\omega_c + \omega_m$. As shown in Fig. 15.10 this signal is passed through a band pass filter* which rejects dc and the sinusoids of frequencies ω_m , $2\omega_m$ and $2\omega_c$ and retains the frequencies ω_c , $\omega_c - \omega_m$ and $\omega_c + \omega_m$. The output of the band pass filter therefore is of the same form as Eq. (15.5) and is therefore an AM wave.

It is to be mentioned that the modulated signal cannot be transmitted as such. The modulator is to be followed by a power amplifier which provides the necessary power and then the modulated signal is fed to an antenna of appropriate size for radiation as shown in Fig. 15.11.

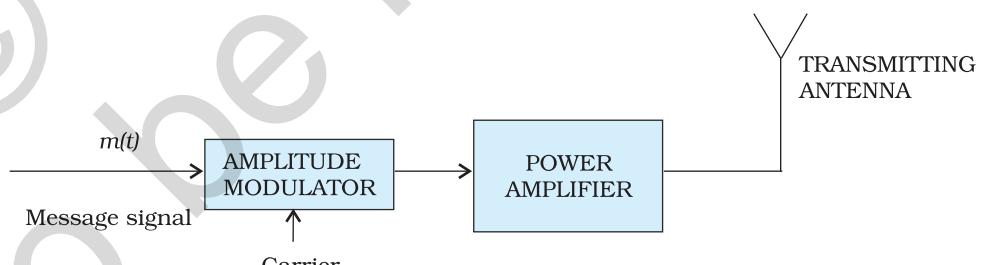


FIGURE 15.11 Block diagram of a transmitter.

15.10 DETECTION OF AMPLITUDE MODULATED WAVE

The transmitted message gets attenuated in propagating through the channel. The receiving antenna is therefore to be followed by an amplifier and a detector. In addition, to facilitate further processing, the carrier frequency is usually changed to a lower frequency by what is called an *intermediate frequency (IF) stage* preceding the detection. The detected signal may not be strong enough to be made use of and hence is required

* A band pass filter rejects low and high frequencies and allows a band of frequencies to pass through.

Communication System

to be amplified. A block diagram of a typical receiver is shown in Fig. 15.12

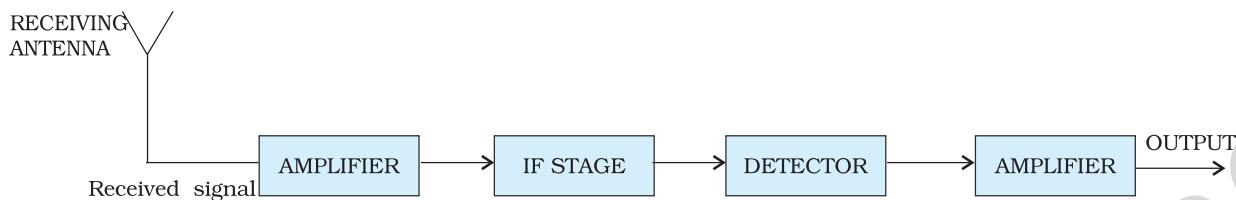


FIGURE 15.12 Block diagram of a receiver.

Detection is the process of recovering the modulating signal from the modulated carrier wave. We just saw that the modulated carrier wave contains the frequencies ω_c and $\omega_c \pm \omega_m$. In order to obtain the original message signal $m(t)$ of angular frequency ω_m , a simple method is shown in the form of a block diagram in Fig. 15.13.

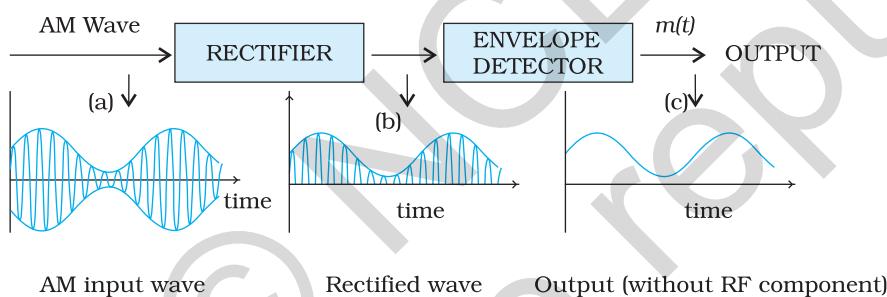


FIGURE 15.13 Block diagram of a detector for AM signal. The quantity on y-axis can be current or voltage.

The modulated signal of the form given in (a) of fig. 15.13 is passed through a rectifier to produce the output shown in (b). This envelope of signal (b) is the message signal. In order to retrieve $m(t)$, the signal is passed through an envelope detector (which may consist of a simple RC circuit).

In the present chapter we have discussed some basic concepts of communication and communication systems. We have also discussed one specific type of analog modulation namely Amplitude Modulation (AM). Other forms of modulation and digital communication systems play an important role in modern communication. These and other exciting developments are taking place everyday.

So far we have restricted our discussion to some basic communication systems. Before we conclude this chapter, it is worth taking a glance at some of the communication systems (see the box) that in recent times have brought major changes in the way we exchange information even in our day-to-day life:

Physics

ADDITIONAL INFORMATION

The Internet

It is a system with billions of users worldwide. It permits communication and sharing of all types of information between any two or more computers connected through a large and complex network. It was started in 1960's and opened for public use in 1990's. With the passage of time it has witnessed tremendous growth and it is still expanding its reach. Its applications include

- (i) *Email* – It permits exchange of text/graphic material using email software. We can write a letter and send it to the recipient through ISP's (Internet Service Providers) who work like the dispatching and receiving post offices.
- (ii) *File transfer* – A FTP (File Transfer Programmes) allows transfer of files/software from one computer to another connected to the Internet.
- (iii) *World Wide Web (WWW)* – Computers that store specific information for sharing with others provide *websites* either directly or through web service providers. Government departments, companies, NGO's (Non-Government Organisations) and individuals can post information about their activities for restricted or free use on their websites. This information becomes accessible to the users. Several search engines like Google, Yahoo! etc. help us in finding information by listing the related websites. *Hypertext* is a powerful feature of the web that automatically links relevant information from one page on the web to another using *HTML (hypertext markup language)*.
- (iv) *E-commerce* – Use of the Internet to promote business using electronic means such as using credit cards is called E-commerce. Customers view images and receive all the information about various products or services of companies through their websites. They can do *on-line shopping* from home/office. Goods are dispatched or services are provided by the company through mail/courier.
- (v) *Chat* – Real time conversation among people with common interests through typed messages is called chat. Everyone belonging to the *chat group* gets the message instantaneously and can respond rapidly.

Facsimile (FAX)

It scans the contents of a document (as an image, not text) to create electronic signals. These signals are then sent to the destination (another FAX machine) in an orderly manner using telephone lines. At the destination, the signals are reconverted into a replica of the original document. Note that FAX provides image of a static document unlike the image provided by television of objects that might be dynamic.

Mobile telephony

The concept of mobile telephony was developed first in 1970's and it was fully implemented in the following decade. The central concept of this system is to divide the service area into a suitable number of *cells* centred on an office called *MTSO (Mobile Telephone Switching Office)*. Each cell contains a low-power transmitter called a *base station* and caters to a large number of mobile receivers (popularly called cell phones). Each cell could have a service area of a few square kilometers or even less depending upon the number of customers. When a mobile receiver crosses the coverage area of one base station, it is necessary for the mobile user to be transferred to another base station. This procedure is called *handover* or *handoff*. This process is carried out very rapidly, to the extent that the consumer does not even notice it. Mobile telephones operate typically in the UHF range of frequencies (about 800-950 MHz).

SUMMARY

1. Electronic communication refers to the faithful transfer of information or message (available in the form of electrical voltage and current) from one point to another point.
2. Transmitter, transmission channel and receiver are three basic units of a communication system.
3. Two important forms of communication system are: Analog and Digital. The information to be transmitted is generally in continuous waveform for the former while for the latter it has only discrete or quantised levels.
4. Every message signal occupies a range of frequencies. The bandwidth of a message signal refers to the band of frequencies, which are necessary for satisfactory transmission of the information contained in the signal. Similarly, any practical communication system permits transmission of a range of frequencies only, which is referred to as the bandwidth of the system.
5. Low frequencies cannot be transmitted to long distances. Therefore, they are superimposed on a high frequency carrier signal by a process known as modulation.
6. In modulation, some characteristic of the carrier signal like amplitude, frequency or phase varies in accordance with the modulating or message signal. Correspondingly, they are called Amplitude Modulated (AM), Frequency Modulated (FM) or Phase Modulated (PM) waves.
7. Pulse modulation could be classified as: Pulse Amplitude Modulation (PAM), Pulse Duration Modulation (PDM) or Pulse Width Modulation (PWM) and Pulse Position Modulation (PPM).
8. For transmission over long distances, signals are radiated into space using devices called antennas. The radiated signals propagate as electromagnetic waves and the mode of propagation is influenced by the presence of the earth and its atmosphere. Near the surface of the earth, electromagnetic waves propagate as surface waves. Surface wave propagation is useful up to a few MHz frequencies.
9. Long distance communication between two points on the earth is achieved through reflection of electromagnetic waves by ionosphere. Such waves are called sky waves. Sky wave propagation takes place up to frequency of about 30 MHz. Above this frequency, electromagnetic waves essentially propagate as space waves. Space waves are used for line-of-sight communication and satellite communication.
10. If an antenna radiates electromagnetic waves from a height h_T , then the range d_T is given by $\sqrt{2Rh_T}$ where R is the radius of the earth.
11. Amplitude modulated signal contains frequencies $(\omega_c - \omega_m)$, ω_c and $(\omega_c + \omega_m)$.
12. Amplitude modulated waves can be produced by application of the message signal and the carrier wave to a non-linear device, followed by a band pass filter.
13. AM detection, which is the process of recovering the modulating signal from an AM waveform, is carried out using a rectifier and an envelope detector.

POINTS TO PONDER

1. In the process of transmission of message/ information signal, noise gets added to the signal anywhere between the information source and the receiving end. Can you think of some sources of noise?
2. In the process of modulation, new frequencies called sidebands are generated on either side (higher and lower than the carrier frequency) of the carrier by an amount equal to the highest modulating frequency. Is it possible to retrieve the message by transmitting (a) only the side bands, (b) only one side band?
3. In amplitude modulation, modulation index $\mu \leq 1$ is used. What will happen if $\mu > 1$?

EXERCISES

- 15.1** Which of the following frequencies will be suitable for beyond-the-horizon communication using sky waves?
(a) 10 kHz
(b) 10 MHz
(c) 1 GHz
(d) 1000 GHz
- 15.2** Frequencies in the UHF range normally propagate by means of:
(a) Ground waves.
(b) Sky waves.
(c) Surface waves.
(d) Space waves.
- 15.3** Digital signals
(i) do not provide a continuous set of values,
(ii) represent values as discrete steps,
(iii) can utilize binary system, and
(iv) can utilize decimal as well as binary systems.
Which of the above statements are true?
(a) (i) and (ii) only
(b) (ii) and (iii) only
(c) (i), (ii) and (iii) but not (iv)
(d) All of (i), (ii), (iii) and (iv).
- 15.4** Is it necessary for a transmitting antenna to be at the same height as that of the receiving antenna for line-of-sight communication? A TV transmitting antenna is 81m tall. How much service area can it cover if the receiving antenna is at the ground level?
- 15.5** A carrier wave of peak voltage 12V is used to transmit a message signal. What should be the peak voltage of the modulating signal in order to have a modulation index of 75%?
- 15.6** A modulating signal is a square wave, as shown in Fig. 15.14.

Communication System

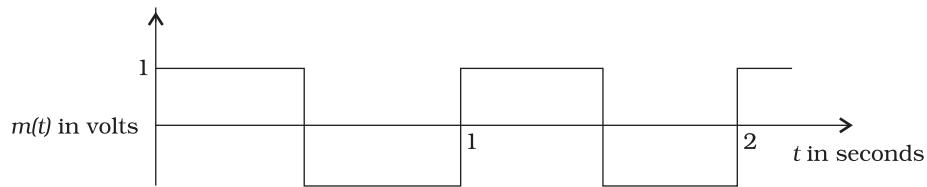


FIGURE 15.14

The carrier wave is given by $c(t) = 2 \sin(8\pi t)$ volts.

- (i) Sketch the amplitude modulated waveform
- (ii) What is the modulation index?

- 15.7** For an amplitude modulated wave, the maximum amplitude is found to be 10V while the minimum amplitude is found to be 2V. Determine the modulation index, μ .
What would be the value of μ if the minimum amplitude is zero volt?
- 15.8** Due to economic reasons, only the upper sideband of an AM wave is transmitted, but at the receiving station, there is a facility for generating the carrier. Show that if a device is available which can multiply two signals, then it is possible to recover the modulating signal at the receiver station.