I. V. Savelyev

PHYSICS

A General Course

ELECTRICITY
& MAGNETISM

WAVES

OPTICS

Mir Publishers
Moscow

II

I. V. SAVELYEV

PHYSICS

A GENERAL COURSE

(In three volumes)

VOLUME II

ELECTRICITY
AND MAGNETISM

WAVES

OPTICS

MIR PUBLISHERS
MOSCOW

Translated from Russian by G. Leib

First published 1980
Revised from the 1978 Russian edition
Second printing 1985
Third printing 1989

Printed in the Union of Soviet Socialist Republics

ISBN 5-03-000902-7, 1978
ISBN 5-03-000900-0, 1980

PREFACE

The main content of the present volume is the science of electromagnetism and the science of waves (elastic, electromagnetic, and light).

The International System of Units (SI) has been used throughout the book, although the reader is simultaneously acquainted with the Gaussian system. In addition to a list of symbols, the appendices at the end of the book give the units of electrical and magnetic quantities in the SI and in the Gaussian system of units, and also compare the form of the basic formulas of electromagnetism in both systems.

The course is the result of twenty five year's work in the Department of General Physics of the Moscow Institute of Engineering Physics. I am grateful to my colleagues and friends for their helpful discussions, criticism and advice in the course of the preparation of the book.

The present course is intended above all for higher technical schools with an extended syllabus in physics. The material has been arranged, however, so that the book can be used as a teaching aid for higher technical schools with an ordinary syllabus simply by omitting some sections.

Igor Savelyev

Moscow, November, 1979

Contents

PART I

ELECTRICITY AND MAGNETISM

Chapter 1
ELECTRIC FIELD IN A VACUUM

1.1. Electric Charge

All bodies in nature are capable of becoming electrified, *i.e.*, acquiring an electric charge. The presence of such a charge manifests itself in that a charged body interacts with other charged bodies. Two kinds of electric charges exist. They are conventionally called positive and negative. Like charges repel each other, and unlike charges attract each other.

An electric charge is an integral part of certain elementary particles[1]. The charge of all elementary particles (if it is not absent) is identical in magnitude. It can be called an **elementary charge**. We shall use the symbol e to denote a positive elementary charge.

The elementary particles include, in particular, the electron (carrying the negative charge $-e$), the proton (carrying the positive charge $+e$), and the neutron (carrying no charge). These particles are the bricks which the atoms and molecules of any substance are built of, therefore all bodies contain electric charges. The particles carrying charges of different signs are usually present in a body in equal numbers and are distributed over it with the same density. The algebraic sum of the charges in any elementary volume of the body equals zero in this case, and each such volume (as well as the body as a whole) will be neutral. If in some way or other we create a surplus of particles of one sign in a body (and, correspondingly, a shortage of particles of the opposite sign), the body will be charged. It is also possible, without changing the total number of positive and negative particles, to cause them to be redistributed in a body so that one part of it has a surplus of charges of one sign and the other part a surplus of charges of the opposite sign.

[1]Elementary particles are defined as such microparticles whose internal structure at the present level of development of physics cannot be conceived as a combination of other particles.

This can be done by bringing a charged body close to an uncharged metal one.

Since a charge q is formed by a plurality of elementary charges, it is an integral multiple of e:

$$q = \pm Ne. \tag{1.1}$$

An elementary charge is so small, however, that macroscopic charges may be considered to have continuously changing magnitudes.

If a physical quantity can take on only definite discrete values, it is said to be quantized. The fact expressed by Eq. (1.1) signifies that an electric charge is quantized.

The magnitude of a charge measured in different inertial reference frames will be found to be the same. Hence, an electric charge is relativistically invariant. It thus follows that the magnitude of a charge does not depend on whether the charge is moving or at rest.

Electric charges can vanish and appear again. Two elementary charges of opposite signs always appear or vanish simultaneously, however. For example, an electron and a positron (a positive electron) meeting each other annihilate, *i.e.*, transform into neutral gamma-photons. This is attended by vanishing of the charges $-e$ and $+e$. In the course of the process called the birth of a pair, a gamma-photon getting into the field of an atomic nucleus transforms into a pair of particles—an electron and a positron. This process causes the charges $-e$ and $+e$ to appear.

Thus, the total charge of an electrically isolated system[2] cannot change. This statement forms the **law of electric charge conservation**.

We must note that the law of electric charge conservation is associated very closely with the relativistic invariance of a charge. Indeed, if the magnitude of a charge depended on its velocity, then by bringing charges of one sign into motion we would change the total charge of the relevant isolated system.

1.2. Coulomb's Law

The law obeyed by the force of interaction of point charges was established experimentally in 1785 by the French physicist Charles A. de Coulomb (1736-1806). A **point charge** is defined as a charged body whose dimensions may be disregarded in comparison with the distances from this body to other bodies carrying an electric charge.

Using a torsion balance (Fig. 1.1) similar to that employed by H. Cavendish to determine the gravitational constant (see Vol. I, Sec. 6.1), Coulomb measured the force of interaction of two charged spheres depending on the magnitude of the

[2] A system is referred to as electrically isolated if no charged particles can penetrate through the surface confining it.

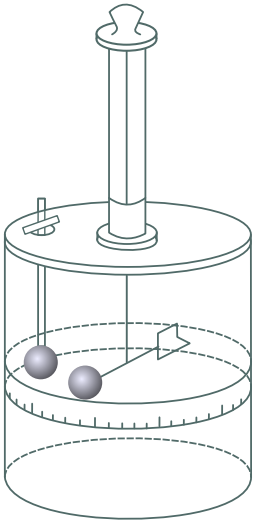


Fig. 1.1

charges on them and on the distance between them. He proceeded from the fact that when a charged metal sphere was touched by an identical uncharged sphere, the charge would be distributed equally between the two spheres.

As a result of his experiments, Coulomb arrived at the conclusion that *the force of interaction between two stationary point charges is proportional to the magnitude of each of them and inversely proportional to the square of the distance between them*. The direction of the force coincides with the straight line connecting the charges.

It must be noted that the direction of the force of interaction along the straight line connecting the point charges follows from considerations of symmetry. An empty space is assumed to be homogeneous and isotropic. Consequently, the only direction distinguished in the space by stationary point charges introduced into it is that from one charge to the other. Assume that the force $\boldsymbol{F}$ acting on the charge q_i (Fig. 1.2) makes the angle α with the direction from q_1 to q_2, and that α differs from 0 or π. But owing to axial symmetry, there are no grounds to set the force $\boldsymbol{F}$ aside from the multitude of forces of other directions making the same angle α with the axis q_1-q_2 (the directions of these forces form a cone with a cone angle of 2α). The difficulty appearing as a result of this vanishes when α equals 0 or π.

Coulomb's law can be expressed by the formula

$$\boldsymbol{F}_{12} = -k\frac{q_1 q_2}{r^2}\,\hat{\boldsymbol{e}}_{12}. \tag{1.2}$$

Here, k is a proportionality constant assumed to be positive, q_1 and q_2 are magnitudes of the interacting charges, r is the distance between the charges, $\hat{\boldsymbol{e}}_{12}$ is the unit

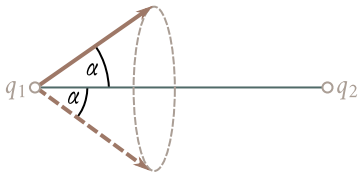


Fig. 1.2

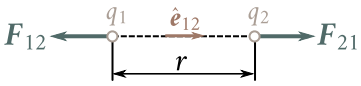


Fig. 1.3

vector directed from the charge q_1 to q_2 and $\boldsymbol{F}_{12}$ is the force acting on the charge q_1 (Fig. 1.3; the figure corresponds to the case of like charges).

The force $\boldsymbol{F}_{21}$ differs from $\boldsymbol{F}_{12}$ in its sign:

$$\boldsymbol{F}_{21} = k\frac{q_1 q_2}{r^2}\,\hat{\boldsymbol{e}}_{12}. \tag{1.3}$$

The magnitude of the interaction force, which is the same for both charges, can be written in the form

$$F = k\frac{|q_1 q_2|}{r^2}. \tag{1.4}$$

Experiments show that the force of interaction between two given charges does not change if other charges are placed near them. Assume that we have the charge q_a and, in addition, N other charges $q_1, q_2, \ldots, q_N$. It can be seen from the above that the resultant force $\boldsymbol{F}$ with which all the N charges q_i act on q_a is

$$\boldsymbol{F} = \sum_{i=1}^{N} \boldsymbol{F}_{\mathrm{a},\,i} \tag{1.5}$$

where $\boldsymbol{F}_{\mathrm{a},\,i}$ is the force with which the charge q_i acts on q_a in the absence of the other $N-1$ charges.

The fact expressed by Eq. (1.5) permits us to calculate the force of interaction between charges concentrated on bodies of finite dimensions, knowing the law of interaction between point charges. For this purpose, we must divide each charge into so small charges $\mathrm{d}q$ that they can be considered as point ones, use Eq. (1.2) to calculate the force of interaction between the charges $\mathrm{d}q$ taken in pairs, and then perform vector summation of these forces. Mathematically, this procedure coincides completely with the calculation of the force of gravitational attraction between bodies of finite dimensions (see Vol. I, Sec. 6.1).

All experimental facts available lead to the conclusion that Coulomb's law holds for distances from 10^{-15} m to at least several kilometres. There are grounds to presume that for distances smaller than 10^{-16} m the law stops being correct. For very great distances, there are no experimental confirmations of Coulomb's law. But there are also no reasons to expect that this law stops being obeyed with very great distances between charges.

1.3. Systems of Units

We can make the proportionality constant in Eq. (1.2) equal unity by properly choosing the unit of charge (the units for F and r were established in mechanics). The relevant unit of charge (when F and r are measured in cgs units) is called the **absolute electrostatic unit** of charge (cgse_q). It is the magnitude of a charge that interacts with a force of 1 dyn in a vacuum with an equal charge at a distance of 1 cm from it.

Careful measurements (they are described in Sec. 10.3) showed that an elementary charge is

$$e = 4.80 \times 10^{-10}\ \text{cgse}_q. \tag{1.6}$$

Adopting the units of length, mass, time, and charge as the basic ones, we can construct a system of units of electrical and magnetic quantities. The system based on the centimetre, gramme, second, and the cgse_q unit is called the **absolute electrostatic system of units** (the cgse system). It is founded on Coulomb's law, *i.e.*, the law of interaction between charges at rest. On a later page, we shall become acquainted with the **absolute electromagnetic system of units** (the cgsm system) based on the law of interaction between conductors carrying an electric current. The Gaussian system in which the units of electrical quantities coincide with those of the cgse system, and of magnetic quantities with those of the cgsm system, is also an absolute system.

Equation (1.4) in the cgse system becomes

$$F = \frac{|q_1 q_2|}{r^2}. \tag{1.7}$$

This equation is correct if the charges are in a vacuum. It has to be determined more accurately for charges in a medium (see Sec. 2.8).

USSR State Standard GOST 9867-61, which came into force on January 1, 1963, prescribes the preferable use of the International System of Units (SI). The basic units of this system are the metre, kilogramme, second, ampere, kelvin, candela, and mole. The SI unit of force is the newton (N) equal to 10^5 dynes.

In establishing the units of electrical and magnetic quantities, the SI system, like the cgsm one, proceeds from the law of interaction of current-carrying conductors instead of charges. Consequently, the proportionality constant in the equation of Coulomb's law is a quantity with a dimension and differing from unity.

The SI unit of charge is the coulomb (C). It has been found experimentally that

$$1\ \text{C} = 2.998 \times 10^9 \approx 3 \times 10^9\ \text{cgse}_q. \tag{1.8}$$

To form an idea of the magnitude of a charge of 1 C, let us calculate the force with which two point charges of 1 C each would interact with each other if they

were 1 m apart. By Eq. (1.7)

$$F = \frac{3 \times 10^9 \times 3 \times 10^9}{100^2} \text{ cgse}_F = 9 \times 10^{14} \text{ dyn} = 9 \times 10^9 \text{ N} \approx 10^9 \text{ kgf}. \quad (1.9)$$

An elementary charge expressed in coulombs is

$$e = 1.60 \times 10^{-19} \text{ C}. \quad (1.10)$$

1.4. Rationalized Form of Writing Formulas

Many formulas of electrodynamics when written in the cgs systems (in particular, in the Gaussian one) include as factors 4π and the so-called electromagnetic constant c equal to the speed of light in a vacuum. To eliminate these factors in the formulas that are most important in practice, the proportionality constant in Coulomb's law is taken equal to $1/4\pi\varepsilon_0$. The equation of the law for charges in a vacuum will thus become

$$F = \frac{1}{4\pi\varepsilon_0} \frac{|q_1 q_2|}{r^2}. \quad (1.11)$$

The other formulas change accordingly. This modified way of writing formulas is called **rationalized**. Systems of units constructed with the use of rationalized formulas are also called **rationalized**. They include the SI system.

The quantity ε_0 is called the **electric constant**. It has the dimension of capacitance divided by length. It is accordingly expressed in units called the farad per metre. To find the numerical value of ε_0, we shall introduce the values of the quantities corresponding to the case of two charges of 1 C each and 1 m apart into Eq. (1.11). By Eq. (1.9), the force of interaction in this case is 9×10^9 N. Using this value of the force, and also $q_1 = q_2 = 1$ C and $r = 1$ m in Eq. (1.11), we get

$$9 \times 10^9 = \frac{1}{4\pi\varepsilon_0} \frac{|1 \times 1|}{1^2}$$

whence

$$\varepsilon_0 = \frac{1}{4\pi \times 9 \times 10^9} = 0.885 \times 10^{-11} \text{ F m}^{-1}. \quad (1.12)$$

The Gaussian system of units was widely used and is continuing to be used in physical publications. We therefore consider it essential to acquaint our reader with both the SI and the Gaussian system. We shall set out the material in the SI units showing at the same time how the formulas look in the Gaussian system. The fundamental formulas of electrodynamics written in the SI and the Gaussian system are compared in Appendix. A.3.

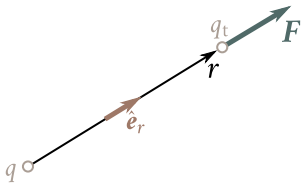


Fig. 1.4

1.5. Electric Field. Field Strength

Charges at rest interact through an electric field[3]. A charge alters the properties of the space surrounding it—it sets up an electric field in it. This field manifests itself in that an electric charge placed at a point of it experiences the action of a force. Hence, to see whether there is an electric field at a given place, we must place a charged body (in the following we shall say simply a charge for brevity) at it and determine whether or not it experiences the action of an electric force. We can evidently assess the "strength" of the field according to the magnitude of the force exerted on the given charge.

Thus, to detect and study an electric field, we must use a "test" charge. For the force acting on our test charge to characterize the field "at the given point", the test charge must be a point one. Otherwise, the force acting on the charge will characterize the properties of the field averaged over the volume occupied by the body that carries the test charge.

Let us study the field set up by the stationary point charge q with the aid of the point test charge q_t. We place the test charge at a point whose position relative to the charge q is determined by the position vector $\boldsymbol{r}$ (Fig. 1.4). We see that the test charge experiences the force

$$\boldsymbol{F} = q_t \left(\frac{1}{4\pi\varepsilon_0} \frac{q}{r^2} \hat{\boldsymbol{e}}_r \right) \tag{1.13}$$

[see Eqs. (1.3) and (1.11)]. Here $\hat{\boldsymbol{e}}_r$ is the unit vector of the position vector $\boldsymbol{r}$.

A glance at Eq. (1.13) shows that the force acting on our test charge depends not only on the quantities determining the field (on q and $\boldsymbol{r}$), but also on the magnitude of the test charge q_t. If we take different test charges q_t', q_t'', etc., then the forces $\boldsymbol{F}'$, $\boldsymbol{F}''$, etc. which they experience at the given point of the field will be different. We can see from Eq. (1.13), however, that the ratio F/q_t for all the test charges will be the same and depend only on the values of q and $\boldsymbol{r}$ determining the field at the given point. It is therefore natural to adopt this ratio as the quantity characterizing an

[3] We shall see in Sec. 6.2 that when considering moving charges, their interaction in addition to an electric field is due to a magnetic field.

electric field:

$$\boldsymbol{E} = \frac{\boldsymbol{F}}{q_\mathrm{t}}. \tag{1.14}$$

This vector quantity is called the **electric field strength** (or **intensity**) at a given point (*i.e.*, at the point where the test charge q_t experiences the action of the force $\boldsymbol{F}$).

According to Eq. (1.14), the electric field strength numerically equals the force acting on a unit point charge at the given point of the field. The direction of the vector $\boldsymbol{E}$ coincides with that of the force acting on a positive charge.

It must be noted that Eq. (1.14) also holds when the test charge is negative ($q_\mathrm{t} < 0$). In this case, the vectors $\boldsymbol{E}$ and $\boldsymbol{F}$ have opposite directions.

We have arrived at the concept of electric field strength when studying the field of a stationary point charge. Definition (1.14), however, also covers the case of a field set up by any collection of stationary charges, but here the following clarification is needed. The arrangement of the charges setting up the field being studied may change under the action of the test charge. This will happen, for example, when the charges producing the field are on a conductor and can freely move within its limits. Therefore, to avoid appreciable alterations in the field being studied, a sufficiently small test charge must be taken.

It follows from Eqs. (1.13) and (1.14) that the field strength of a point charge varies directly with the magnitude of the charge q and inversely with the square of the distance r from the charge to the given point of the field:

$$\boldsymbol{E} = \frac{1}{4\pi\varepsilon_0}\frac{q}{r^2}\,\hat{\boldsymbol{e}}_r. \tag{1.15}$$

The vector $\boldsymbol{E}$ is directed along the radial straight line passing through the charge and the given point of the field, from the charge if the latter is positive and toward the charge if it is negative.

In the Gaussian system, the equation for the field strength of a point charge in a vacuum has the form

$$\boldsymbol{E} = \frac{q}{r^2}\,\hat{\boldsymbol{e}}_r. \tag{1.16}$$

The unit of electric field strength is the strength at a point where unit force (1 N in the SI and 1 dyn in the Gaussian system) acts on unit charge (1 C in the SI and 1 cgse_q in the Gaussian system). This unit has no special name in the Gaussian system. The SI unit of electric field strength is called the volt per metre ($\mathrm{V\,m^{-1}}$) [see Eq. (1.44)].

According to Eq. (1.15), a charge of 1 C produces the following field strength in a

vacuum at a distance of 1 m from this charge:

$$E = \frac{1}{4\pi\left(1/4\pi \times 9 \times 10^9\right)} \frac{1}{1^2} = 9 \times 10^9 \text{ V m}^{-1}.$$

This strength in the Gaussian system is

$$E = \frac{q}{r^2} = \frac{3 \times 10^9}{100^2} = 3 \times 10^5 \text{ cgse}_E.$$

Comparing these two results, we find that

$$1 \text{ cgse}_E = 3 \times 10^4 \text{ V m}^{-1}. \tag{1.17}$$

According to Eq. (1.14), the force exerted on a test charge is

$$\boldsymbol{F} = q_t \boldsymbol{E}.$$

It is obvious that any point charge q[4] at a point of a field with the strength $\boldsymbol{E}$ will experience the force

$$\boldsymbol{F} = q\boldsymbol{E}. \tag{1.18}$$

If the charge q is positive, the direction of the force coincides with that of the vector $\boldsymbol{E}$. If q is negative, the vectors $\boldsymbol{F}$ and $\boldsymbol{E}$ are directed oppositely.

We mentioned in Sec. 1.2 that the force with which a system of charges acts on a charge not belonging to the system equals the vector sum of the forces which each of the charges of the system exerts separately on the given charge [see Eq. (1.15)]. Hence it follows that *the field strength of a system of charges equals the vector sum of the field strengths that would be produced by each of the charges of the system separately*:

$$\boldsymbol{E} = \sum_i \boldsymbol{E}_i. \tag{1.19}$$

This statement is called the **principle of electric field superposition**.

The superposition principle allows us to calculate the field strength of any system of charges. By dividing extended charges into sufficiently small fractions dq, we can reduce any system of charges to a collection of point charges. We calculate the contribution of each of such charges to the resultant field by Eq. (1.15).

An electric field can be described by indicating the magnitude and direction of the vector $\boldsymbol{E}$ for each of its points. The combination of these vectors forms the field of the electric field strength vector (compare with the field of the velocity vector, Vol. I, Sec. 9.1). The velocity vector field can be represented very illustratively with the aid of flow lines. Similarly, an electric field can be described with the aid of strength lines, which we shall call for short $\boldsymbol{E}$ lines or field lines. These lines are drawn so that a tangent to them at every point coincides with the direction of the

[4] In Eq. (1.15), q stands for the charge setting up the field. In Eq. (1.18), q stands for the charge experiencing the force $\boldsymbol{F}$ at a point of strength $\boldsymbol{E}$.

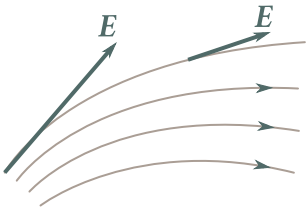


Fig. 1.5

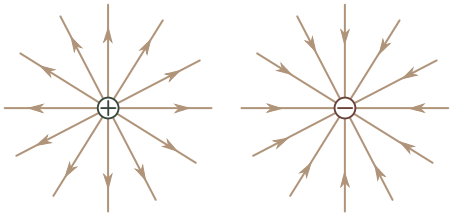


Fig. 1.6

vector $\boldsymbol{E}$. The density of the lines is selected so that their number passing through a unit area at right angles to the lines equals the numerical value of the vector $\boldsymbol{E}$. Hence, the pattern of field lines permits us to assess the direction and magnitude of the vector $\boldsymbol{E}$ at various points of space (Fig. 1.5).

The $\boldsymbol{E}$ lines of a point charge field are a collection of radial straight lines directed away from the charge if it is positive and toward it if it is negative (Fig. 1.6). One end of each line is at the charge, and the other extends to infinity. Indeed, the total number of lines intersecting a spherical surface of arbitrary radius r will equal the product of the density of the lines and the surface area of the sphere $4\pi r^2$. We have assumed that the density of the lines numerically equals $E = (1/4\pi\varepsilon_0)(q/r^2)$. Hence, the number of lines is $(1/4\pi\varepsilon_0)(q/r^2)4\pi r^2 = q/\varepsilon_0$. This result signifies that the number of lines at any distance from a charge will be the same. It thus follows hat the lines do not begin and do not terminate anywhere except tor the charge. Beginning at the charge, they extend to infinity (the charge is positive), or arriving from infinity, they terminate at the charge (the latter is negative). This property of the $\boldsymbol{E}$ lines is common for all electrostatic fields, *i.e.*, fields set up by any system of stationary charges: the field lines can begin or terminate only at charges or extend to infinity.

1.6. Potential

Let us consider the field produced by a stationary point charge q. At any point of this field, the point charge q' experiences the force

$$\boldsymbol{F} = \frac{1}{4\pi\varepsilon_0}\frac{qq'}{r^2}\hat{\boldsymbol{e}}_r = F(r)\hat{\boldsymbol{e}}_r. \tag{1.20}$$

Here $F(r)$ is the magnitude of the force $\boldsymbol{F}$, and $\hat{\boldsymbol{e}}_r$ is the unit vector of the position vector $\boldsymbol{r}$ determining the position of the charge q' relative to the charge q.

The force (1.20) is a central one (see Vol. I, Sec. 3.4). A central field of forces is conservative. Consequently, the work done by the forces of the field on the charge q' when it is moved from one point to another does not depend on the path. This

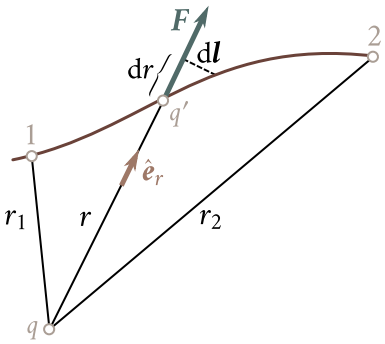


Fig. 1.7

work is

$$A_{12} = \int_1^2 F(r)\hat{\boldsymbol{e}}_r \, \mathrm{d}\boldsymbol{l} \tag{1.21}$$

where $\mathrm{d}\boldsymbol{l}$ is the elementary displacement of the charge q'. Inspection of Fig. 1.7 shows that the scalar product $\hat{\boldsymbol{e}}_r \, \mathrm{d}\boldsymbol{l}$ equals the increment of the magnitude of the position vector $\boldsymbol{r}$, *i.e.*, $\mathrm{d}r$. Equation (1.21) can therefore be written in the form

$$A_{12} = \int_1^2 F(r) \, \mathrm{d}r$$

[compare with Eq. (3.24) of Vol. I]. Introduction of the expression for $F(r)$ yields

$$A_{12} = \frac{qq'}{4\pi\varepsilon_0} \int_{r_1}^{r_2} \frac{\mathrm{d}r}{r^2} = \frac{1}{4\pi\varepsilon_0} \left(\frac{qq'}{r_1} - \frac{qq'}{r_2} \right). \tag{1.22}$$

The work of the forces of a conservative field can be represented as a decrement of the potential energy:

$$A_{12} = W_{\mathrm{p},1} - W_{\mathrm{p},2}. \tag{1.23}$$

A comparison of Eqs. (1.22) and (1.23) leads to the following expression for the potential energy of the charge q' in the field of the charge q:

$$W_{\mathrm{p}} = \frac{1}{4\pi\varepsilon_0} \frac{qq'}{r_2} + \text{constant}.$$

The value of the constant in the expression for the potential energy is usually chosen so that when the charge moves away to infinity (*i.e.*, when $r = \infty$), the potential energy vanishes. When this condition is observed, we get

$$W_{\mathrm{p}} = \frac{1}{4\pi\varepsilon_0} \frac{qq'}{r_2}. \tag{1.24}$$

Let us use the charge q' as a test charge for studying the field. By Eq. (1.24), the potential energy which the test charge has depends not only on its magnitude q', but also on the quantities q and r determining the field. Thus, we can use this energy

to describe the field just like we used the force acting on the test charge for this purpose.

Different test charges q_t', q_t'', etc. will have different energies W_p', W_p'', etc. at the same point of a field. But the ratio W_p/q_t will be the same for all the charges [see Eq. (1.24)]. The quantity

$$\varphi = \frac{W_p}{q_t} \tag{1.25}$$

is called the **field potential** at a given point and is used together with the field strength $\boldsymbol{E}$ to describe electric fields.

It can be seen from Eq. (1.25) that the potential numerically equals the potential energy which a unit positive charge would have at the given point of the field. Substituting for the potential energy in Eq. (1.25) its value from (1.24), we get the following expression for the potential of a point charge:

$$\varphi = \frac{1}{4\pi\varepsilon_0}\frac{q}{r}. \tag{1.26}$$

In the Gaussian system, the potential of the field of a point charge in a vacuum is determined by the formula

$$\varphi = \frac{q}{r}. \tag{1.27}$$

Let us consider the field produced by a system of N point charges $q_1, q_2, \ldots, q_N$. Let $r_1, r_2, \ldots, r_N$ be the distances from each of the charges to the given point of the field. The work done by the forces of this field on the charge q' will equal the algebraic sum of the work done by the forces set up by each of the charges separately:

$$A_{12} = \sum_{i=1}^{N} A_i.$$

By Eq. (1.22), each work A_i equals

$$A_i = \frac{1}{4\pi\varepsilon_0}\left(\frac{q_i q'}{r_{i,1}} - \frac{q_i q'}{r_{i,2}}\right)$$

where $r_{i,1}$ is the distance from the charge q_i to the initial position of the charge q', and $r_{i,2}$ is the distance from q_i to the final position of the charge q'. Hence,

$$A_{i2} = \frac{1}{4\pi\varepsilon_0}\sum_{i=1}^{N}\frac{q_i q'}{r_{i,1}} - \frac{1}{4\pi\varepsilon_0}\sum_{i=1}^{N}\frac{q_i q'}{r_{i,2}}.$$

Comparing this equation with Eq. (1.23), we get the following expression for the

potential energy of the charge q' in the field of a system of charges:

$$W_{\mathrm{p}} = \frac{1}{4\pi\varepsilon_0} \sum_{i=1}^{N} \frac{q_i q'}{r_i}$$

from which it can be seen that

$$\varphi = \frac{1}{4\pi\varepsilon_0} \sum_{i=1}^{N} \frac{q_i}{r_i}. \tag{1.28}$$

Comparing this formula with Eq. (1.26), we arrive at the conclusion that *the potential of the field produced by a system of charge equals the algebraic sum of the potentials produced by each of the charges separately*. Whereas the field strengths are added vectorially in the superposition of fields, the potentials are added algebraically. This is why it is usually much simpler to calculate the potentials than the electric field strengths.

Examination of Eq. (1.25) shows that the charge q at a point of a field with the potential φ has the potential energy

$$W_{\mathrm{p}} = q\varphi. \tag{1.29}$$

Hence, the work of the field forces on the charge q can be expressed through the potential difference:

$$A_{12} = W_{\mathrm{p},1} - W_{\mathrm{p},2} = q\,(\varphi_1 - \varphi_2)\,. \tag{1.30}$$

Thus, the work done on a charge by the forces of a field equals the product of the magnitude of the charge and the difference between the potentials at the initial and final points (*i.e.*, the potential decrement).

If the charge q is removed from a point having the potential φ to infinity (where by convention the potential vanishes), then the work of the field forces will be

$$A_{\infty} = q\varphi. \tag{1.31}$$

Here, it follows that the potential numerically equals the work done by the forces of a field on a unit positive charge when the latter is removed from the given point to infinity. Work of the same magnitude must be done against the electric field forces to move a unit positive charge from infinity to the given point of a field.

Equation (1.31) can be used to establish the units of potential. The unit of potential is taken equal to the potential at a point of a field when work equal to unity is required to move unit positive charge from infinity to this point. The SI unit of potential called the volt (V) is taken equal to the potential at a point when work of 1 joule has to be done to move a charge of 1 coulomb from infinity to this point:

$$1\,\mathrm{J} = 1\,\mathrm{C} \times 1\,\mathrm{V}, \quad \text{thus,} \quad 1\,\mathrm{V} = \frac{1\,\mathrm{J}}{1\,\mathrm{C}}. \tag{1.32}$$

The absolute electrostatic unit of potential (cgse_{φ}) is taken equal to the potential

at a point when work of 1 erg has to be done to move a charge of 1 cgse_q from infinity to this point. Expressing 1 J and 1 C in Eq. (1.32) through cgse units, we shall find the relation between the volt and the cgse potential unit:

$$1\text{ V} = \frac{1\text{ J}}{1\text{ C}} = \frac{10^7\text{ erg}}{3\times 10^9\text{ cgse}_q} = \frac{1}{300}\text{ cgse}_\varphi. \tag{1.33}$$

Thus, 1 cgse_φ equals 300 V.

A unit of energy and work called the **electron-volt** (eV) is frequently used in physics. An electron-volt is defined as the work done by the forces of a field on a charge equal to that of an electron (*i.e.*, on the elementary charge e) when it passes through a potential difference of 1 V:

$$1\text{ eV} = 1.60\times 10^{-19}\text{ C}\times 1\text{ V} = 1.60\times 10^{-19}\text{ J} = 1.60\times 10^{-12}\text{ erg}. \tag{1.34}$$

Multiple units of the electron-volt are also used:

$$1\text{ keV (kiloelectron-volt)} = 10^3\text{ eV},$$
$$1\text{ MeV (megaelectron-volt)} = 10^6\text{ eV},$$
$$1\text{ GeV (gigaelectron-volt)} = 10^9\text{ eV}.$$

1.7. Interaction Energy of a System of Charges

Equation (1.24) can be considered as the mutual potential energy of the charges q and q'. Using the symbols q_1 and q_2 for these charges, we get the following formula for their interaction energy:

$$W_\text{p} = \frac{1}{4\pi\varepsilon_0}\frac{q_1 q_2}{r_{12}}. \tag{1.35}$$

The symbol r_{12} stands for the distance between the charges.

Let us consider a system consisting of N point charges $q_1, q_2, \ldots, q_N$. We showed in Sec. 3.6 of Vol. I that the energy of interaction of such a system equals the sum of the energies of interaction of the charges taken in pairs:

$$W_\text{p} = \frac{1}{2}\sum_{(i\neq k)} W_{\text{p},ik}(r_{ik}) \tag{1.36}$$

[see Eq. (3.60) of Vol. I].

According to Eq. (1.35)

$$W_{\text{p},ik} = \frac{1}{4\pi\varepsilon_0}\frac{q_i q_k}{r_{ik}}.$$

Using this equation in (1.36), we find that

$$W_\text{p} = \frac{1}{2}\sum_{(i\neq k)}\frac{1}{4\pi\varepsilon_0}\frac{q_i q_k}{r_{ik}}. \tag{1.37}$$

In the Gaussian system, the factor $1/(4\pi\varepsilon_0)$ is absent in this equation.

In Eq. (1.37), summation is performed over the subscripts i and k. Both subscripts pass independently through all the values from 1 to N. Addends for which the value of the subscript i coincides with that of k are not taken into consideration. Let us write Eq. (1.37) as follows:

$$W_{\mathrm{p}} = \frac{1}{2}\sum_{i=1}^{N} q_i \sum_{\substack{i=1\\(i\neq k)}}^{N} \frac{1}{4\pi\varepsilon_0}\frac{q_k}{r_{ik}}. \tag{1.38}$$

The expression

$$\varphi_i = \frac{1}{4\pi\varepsilon_0}\sum_{\substack{i=1\\(i\neq k)}}^{N} \frac{q_k}{r_{ik}}$$

is the potential produced by all the charges except q_i at the point where the charge q_i is. With this in view, we get the following formula for the interaction energy:

$$W_{\mathrm{p}} = \frac{1}{2}\sum_{i=1}^{N} q_i\varphi_i. \tag{1.39}$$

1.8. Relation Between Electric Field Strength and Potential

An electric field can be described either with the aid of the vector quantity $\boldsymbol{E}$, or with the aid of the scalar quantity φ. There must evidently be a definite relation between these quantities. If we bear in mind that $\boldsymbol{E}$ is proportional to the force acting on a charge and φ to the potential energy of the charge, it is easy to see that this relation must be similar to that between the potential energy and the force.

The force $\boldsymbol{F}$ is related to the potential energy by the expression

$$\boldsymbol{F} = -\nabla W_{\mathrm{p}} \tag{1.40}$$

[see Eq. (3.32) of Vol. I]. For a charged particle in an electrostatic field, we have $\boldsymbol{F} = q\boldsymbol{E}$ and $W_{\mathrm{p}} = q\varphi$. Introducing these values into Eq. (1.40), we find that

$$q\boldsymbol{E} = -\nabla(q\varphi).$$

The constant q can be put outside the gradient sign. Doing this and then cancelling q, we arrive at the formula

$$\boldsymbol{E} = -\nabla\varphi \tag{1.41}$$

establishing the relation between the field strength and potential.

Taking into account the definition of the gradient [see Eq. (3.31) of Vol. 1], we

can write that

$$\boldsymbol{E} = -\frac{\partial\varphi}{\partial x}\hat{\boldsymbol{e}}_x - \frac{\partial\varphi}{\partial y}\hat{\boldsymbol{e}}_y - \frac{\partial\varphi}{\partial z}\hat{\boldsymbol{e}}_z. \tag{1.42}$$

Hence, Eq. (1.41) has the following form in projections onto the coordinate axes:

$$E_x = -\frac{\partial\varphi}{\partial x}, \quad E_y = -\frac{\partial\varphi}{\partial y}, \quad E_z = -\frac{\partial\varphi}{\partial z}. \tag{1.43}$$

Similarly, the projection of the vector $\boldsymbol{E}$ onto an arbitrary direction l equals the derivative of φ with respect to l taken with the opposite sign, *i.e.*, the rate of diminishing of the potential when moving along the direction l:

$$E_l = -\frac{\partial\varphi}{\partial l}. \tag{1.44}$$

It is easy to see that Eq. (1.44) is correct by choosing l as one of the coordinate axes and taking Eq. (1.43) into account.

Let us explain Eq. (1.41) using as an example the field of a point charge. The potential of this field is expressed by Eq. (1.26). Passing over to Cartesian coordinates, we get the expression

$$\varphi = \frac{1}{4\pi\varepsilon_0}\frac{q}{r} = \frac{1}{4\pi\varepsilon_0}\frac{q}{(x^2+y^2+z^2)^{1/2}}.$$

The partial derivative of this function with respect to x is

$$\frac{\partial\varphi}{\partial x} = -\frac{1}{4\pi\varepsilon_0}\frac{qx}{(x^2+y^2+z^2)^{3/2}} = -\frac{1}{4\pi\varepsilon_0}\frac{qx}{r^3}.$$

Similarly,

$$\frac{\partial\varphi}{\partial y} = -\frac{1}{4\pi\varepsilon_0}\frac{qy}{r^3}, \quad \frac{\partial\varphi}{\partial y} = -\frac{1}{4\pi\varepsilon_0}\frac{qz}{r^3}.$$

Using the found values of the derivatives in Eq. (1.42), we arrive at the expression

$$\boldsymbol{E} = \frac{1}{4\pi\varepsilon_0}\frac{q\left(x\hat{\boldsymbol{e}}_x + y\hat{\boldsymbol{e}}_y + z\hat{\boldsymbol{e}}_z\right)}{r^3} = \frac{1}{4\pi\varepsilon_0}\frac{q\boldsymbol{r}}{r^3} = \frac{1}{4\pi\varepsilon_0}\frac{q}{r^2}\hat{\boldsymbol{e}}_r$$

that coincides with Eq. (1.15).

Equation (1.41) allows us to find the field strength at every point from the known values of φ. We can also solve the reverse problem, *i.e.*, find the potential difference between two arbitrary points of a field according to the given values of $\boldsymbol{E}$. For this purpose, we shall take advantage of the circumstance that the work done by the forces of a field on the charge q when it is moved from point 1 to point 2 can be calculated as

$$A_{12} = \int_1^2 q\boldsymbol{E}\,\mathrm{d}\boldsymbol{l}.$$

At the same time in accordance with Eq. (1.30), this work can be written as

$$A_{12} = q\,(\varphi_1 - \varphi_2)\,.$$

Equating these two expressions and cancelling q, we obtain

$$\varphi_1 - \varphi_2 = \int_1^2 \boldsymbol{E}\,\mathrm{d}\boldsymbol{l}. \tag{1.45}$$

The integral can be taken along any line joining points 1 and 2 because the work of the field forces is independent of the path. For circumvention along a closed contour, $\varphi_1 = \varphi_2$, and Eq. (1.45) becomes

$$\oint \boldsymbol{E}\,\mathrm{d}\boldsymbol{l} = 0 \tag{1.46}$$

(the circle on the integral sign indicates that integration is performed over a closed contour). It must be noted that this relation holds only for an electrostatic field. We shall see on a later page that the field of moving charges (*i.e.*, a field changing with time) is not a potential one. Therefore, condition (1.46) is not observed for it.

An imaginary surface all of whose points have the same potential is called an equipotential surface. Its equation has the form

$$\varphi(x, y, z) = \text{constant}.$$

The potential does not change in movement along an equipotential surface over the distance $\mathrm{d}l$ ($\mathrm{d}\varphi = 0$). Hence, according to Eq. (1.44), the tangential component of the vector $\boldsymbol{E}$ to the surface equals zero. We thus conclude that the vector $\boldsymbol{E}$ at every point is directed along a normal to the equipotential surface passing through the given point. Bearing in mind that the vector $\boldsymbol{E}$ is directed along a tangent to an $\boldsymbol{E}$ line, we can easily see that the field lines at every point are orthogonal to the equipotential surfaces.

An equipotential surface can be drawn through any point of a field. Consequently, we can construct an infinitely great number of such surfaces. They are conventionally drawn so that the potential difference for two adjacent surfaces is the same everywhere. Thus, the density of the equipotential surfaces allows us to assess the magnitude of the field strength. Indeed, the denser are the equipotential surfaces, the more rapidly does the potential change when moving along a normal to the surface. Hence, $\nabla\varphi$ is greater at the given place, and, therefore, $\boldsymbol{E}$ is greater too.

Figure 1.8 shows equipotential surfaces (more exactly, their intersections with the plane of the drawing) for the field of a point charge. In accordance with the nature of the dependence of E on r, equipotential surfaces become the denser, the nearer we approach a charge.

Equipotential surfaces for a homogeneous field are a collection of equispaced

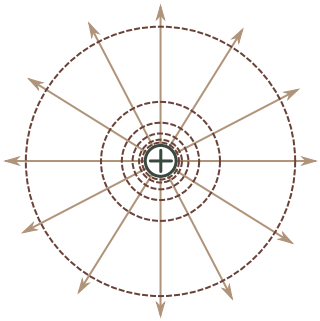


Fig. 1.8

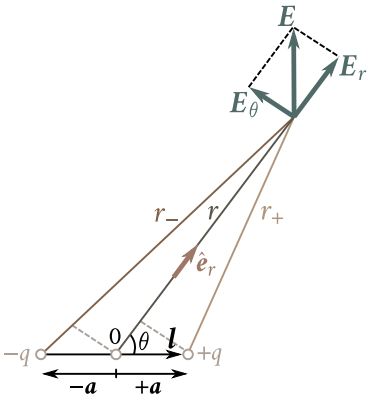


Fig. 1.9

planes at right angles to the direction of the field.

1.9. Dipole

An **electric dipole** is defined as a system of two point charges $+q$ and $-q$ identical in value and opposite in sign, the distance between which is much smaller than that to the points at which the field of the system is being determined. The straight line passing through both charges is called the **dipole axis**.

Let us first calculate the potential and then the field strength of a dipole. This field has axial symmetry. Therefore, the pattern of the field in any plane passing through the dipole axis will be the same, the vector $\boldsymbol{E}$ being in this plane. The position of a point relative to the dipole will be characterized with the aid of the position vector $\boldsymbol{r}$ or with the aid of the polar coordinates r and θ (Fig. 1.9). We shall introduce the vector $\boldsymbol{l}$ passing from the negative charge to the positive one. The position of the charge $+q$ relative to the centre of the dipole is determined by the vector $\boldsymbol{a}$, and of the charge $-q$ by the vector $-\boldsymbol{a}$. It is obvious that $\boldsymbol{l} = 2\boldsymbol{a}$. We shall designate the distances to a given point from the charges $+q$ and $-q$ by r_+ and r_-, respectively.

Owing to the smallness of a in comparison with r, we can assume approximately that

$$\begin{aligned} r_+ &= r - a\cos\theta = r - \boldsymbol{a}\cdot\hat{\boldsymbol{e}}_r, \\ r_- &= r + a\cos\theta = r + \boldsymbol{a}\cdot\hat{\boldsymbol{e}}_r. \end{aligned} \tag{1.47}$$

The potential at a point determined by the position vector $\boldsymbol{r}$ is

$$\varphi(\boldsymbol{r}) = \frac{1}{4\pi\varepsilon_0}\left(\frac{q}{r_+} - \frac{q}{r_-}\right) = \frac{1}{4\pi\varepsilon_0}\frac{q(r_- - r_+)}{r_+ r_-}.$$

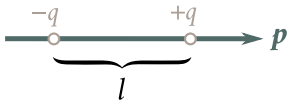


Fig. 1.10

The product r_+r_- can be replaced with r^2. The difference $r_- - r_+$ according to Eqs. (1.47), is $2(\boldsymbol{a}\cdot\hat{\boldsymbol{e}}_r) = \boldsymbol{l}\cdot\hat{\boldsymbol{e}}_r$. Hence,

$$\varphi(\boldsymbol{r}) = \frac{1}{4\pi\varepsilon_0}\frac{q(\boldsymbol{l}\cdot\hat{\boldsymbol{e}}_r)}{r^2} = \frac{1}{4\pi\varepsilon_0}\frac{(\boldsymbol{p}\cdot\hat{\boldsymbol{e}}_r)}{r^2} \tag{1.48}$$

where

$$\boldsymbol{p} = q\boldsymbol{l} \tag{1.49}$$

is a characteristic of a dipole called its **electric moment**. The vector $\boldsymbol{p}$ is directed along the dipole axis from the negative charge to the positive one (Fig. 1.10).

A glance at Eq. (1.48) shows that the field of a dipole is determined by its electric moment $\boldsymbol{p}$. We shall see below that the behaviour of a dipole in an external electric field is also determined by its electric moment $\boldsymbol{p}$. A comparison with Eq. (1.26) shows that the potential of a dipole field diminishes with the distance more rapidly (as $1/r^2$) than the potential of a point charge field (which changes in proportion to $1/r$).

It can be seen from Fig. 1.9 that $\boldsymbol{p}\cdot\hat{\boldsymbol{e}}_r = p\cos\theta$. Therefore, Eq. (1.48) can be written as follows:

$$\varphi(r,\theta) = \frac{1}{4\pi\varepsilon_0}\frac{p\cos\theta}{r^2}. \tag{1.50}$$

To find the field strength of a dipole, let us calculate the projections of the vector $\boldsymbol{E}$ onto two mutually perpendicular directions by Eq. (1.44). One of them is determined by the motion of a point due to the change in the distance r (with θ fixed), the other by the motion of the point due to the change in the angle θ (with r fixed, see Fig. 1.9). The first projection is obtained by differentiation of Eq. (1.50) with respect to r:

$$E_r = -\frac{\partial\varphi}{\partial r} = \frac{1}{4\pi\varepsilon_0}\frac{2p\cos\theta}{r^3}. \tag{1.51}$$

We shall find the second projection (let us designate it by E_θ) by taking the ratio of the increment of the potential φ obtained when the angle θ grows by $\mathrm{d}\theta$ to the distance $r\,\mathrm{d}\theta$ over which the end of the segment r moves (in this case the quantity $\mathrm{d}l$ in Eq. (1.44) equals $r\,\mathrm{d}\theta$]. Thus,

$$E_\theta = -\frac{1}{r}\frac{\partial\varphi}{\partial\theta}.$$

Introducing the value of the derivative of function (1.50) with respect to θ we get

$$E_\theta = \frac{1}{4\pi\varepsilon_0}\frac{p\sin\theta}{r^3}. \tag{1.52}$$

The sum of the squares of Eqs. (1.51) and (1.52) gives the square of the vector $\boldsymbol{E}$ (see Fig. 1.9):

$$E^2 = E_r^2 + E_\theta^2 = \left(\frac{1}{4\pi\varepsilon_0}\right)^2 \left(\frac{p}{r^3}\right)^2 \left(4\cos^2\theta + \sin^2\theta\right)$$
$$= \left(\frac{1}{4\pi\varepsilon_0}\right)^2 \left(\frac{p}{r^3}\right)^2 \left(1 + 3\cos^2\theta\right).$$

Hence

$$E = \frac{1}{4\pi\varepsilon_0}\frac{p}{r^3}\left(1 + 3\cos^2\theta\right)^{1/2}. \tag{1.53}$$

Assuming in Eq. (1.53) that $\theta = 0$, we get the strength on the dipole axis:

$$E_\parallel = \frac{1}{4\pi\varepsilon_0}\frac{2p}{r^3}. \tag{1.54}$$

The vector $\boldsymbol{E}_\parallel$ is directed along the dipole axis. This is in agreement with the axial symmetry of the problem. Examination of Eq. (1.51) shows that $E_r > 0$ when $\theta = 0$, and $E_r < 0$ when $\theta = \pi$. This signifies that in any case the vector $\boldsymbol{E}_\parallel$ has a direction coinciding with that from $-q$ to $+q$ (*i.e.*, with the direction of $\boldsymbol{p}$). Equation (1.54) can therefore be written in the vector form:

$$\boldsymbol{E}_\parallel = \frac{1}{4\pi\varepsilon_0}\frac{2\boldsymbol{p}}{r^3}. \tag{1.55}$$

Assuming in Eq. (1.53) that $\theta = \pi/2$, we get the field strength on the straight line passing through the centre of the dipole and perpendicular to its axis:

$$E_\perp = \frac{1}{4\pi\varepsilon_0}\frac{p}{r^3}. \tag{1.56}$$

By Eq. (1.51), when $\theta = \pi/2$, the projection E_r equals zero. Hence, the vector $\boldsymbol{E}_\perp$ is parallel to the dipole axis. It follows from Eq. (1.52) that when $\theta = \pi/2$, the projection E_θ is positive. This signifies that the vector $\boldsymbol{E}_\perp$ is directed toward the growth of the angle θ, *i.e.*, antiparallel to the vector $\boldsymbol{p}$.

The field strength of a dipole is characterized by the circumstance that it diminishes with the distance from the dipole in proportion to $1/r^3$, *i.e.*, more rapidly than the field strength of a point charge (which diminishes in proportion to $1/r^2$).

Figure 1.11 shows $\boldsymbol{E}$ lines (the solid lines) and equipotential surfaces (the dash lines) of the field of a dipole. According to Eq. (1.50), when $\theta = \pi/2$, the potential vanishes for all the r's. Thus, all the points of a plane at right angles to the dipole axis and passing through its middle have a zero potential. This could have been predicted because the distances from the charges $+q$ and $-q$ to any point of this plane are identical.

Now let us turn to the behaviour of a dipole in an external electric field. If a dipole is placed in a homogeneous electric field, the charges $+q$ and $-q$ forming

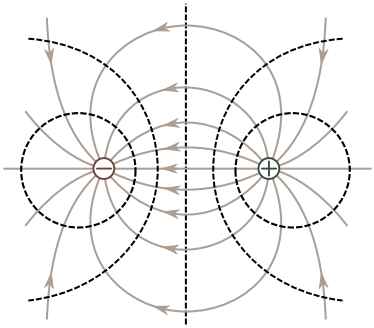


Fig. 1.11

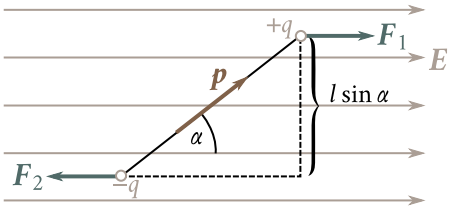


Fig. 1.12

the dipole will be under the action of the forces $\boldsymbol{F}_1$ and $\boldsymbol{F}_2$ equal in magnitude, but opposite in direction (Fig. 1.12). These forces form a couple whose arm is $l \sin \alpha$, *i.e.*, depends on the orientation of the dipole relative to the field. The magnitude of each of the forces is qE. Multiplying it by the arm, we get the magnitude of the torque acting on a dipole:

$$T = qEl \sin \alpha = pE \sin \alpha \tag{1.57}$$

(p is the electric moment of the dipole). It is easy to see that Eq. (1.57) can be written in the vector form

$$\boldsymbol{T} = \boldsymbol{p} \times \boldsymbol{E}. \tag{1.58}$$

The torque (1.58) tends to turn a dipole so that its electric moment $\boldsymbol{p}$ is in the direction of the field.

Let us find the potential energy belonging to a dipole in an external electric field. By Eq. (1.29), this energy is

$$W_\mathrm{p} = q\varphi_+ - q\varphi_- = q(\varphi_+ - \varphi_-). \tag{1.59}$$

Here φ_+ and φ_- are the values of the potential of the external field at the points where the charges $+q$ and $-q$ are placed.

The potential of a homogeneous field diminishes linearly in the direction of the vector $\boldsymbol{E}$. Assuming that the x-axis is this direction (Fig. 1.13), we can write that $E = E_x = -\mathrm{d}\varphi/\mathrm{d}x$. A glance at Fig. 1.13 shows that the difference $\varphi_+ - \varphi_-$ equals the increment of the potential on the segment $\Delta x = l \cos \alpha$:

$$\varphi_+ - \varphi_- = \frac{\mathrm{d}\varphi}{\mathrm{d}x} l \cos \alpha = -El \cos \alpha.$$

Introducing this value into Eq. (1.59), we find that

$$W_\mathrm{p} = -qEl \cos \alpha = -pE \cos \alpha. \tag{1.60}$$

Here α is the angle between the vectors $\boldsymbol{p}$ and $\boldsymbol{E}$. We can therefore write Eq. (1.60)

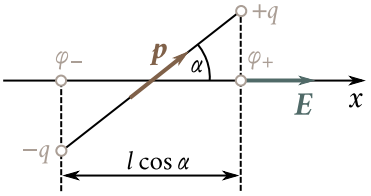


Fig. 1.13

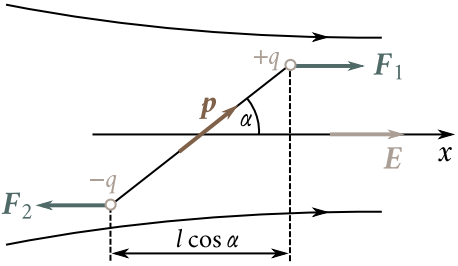


Fig. 1.14

in the form

$$W_{\mathrm{p}} = -\boldsymbol{p} \cdot \boldsymbol{E}. \tag{1.61}$$

We must note that this expression takes no account of the energy of interaction of the charges $+q$ and $-q$ forming a dipole.

We have obtained Eq. (1.61) assuming for simplicity's sake that the field is homogeneous. This equation also holds, however, for an inhomogeneous field.

Let us consider a dipole in an inhomogeneous field that is symmetrical relative to the x-axis[5]. Let the centre of the dipole be on this axis, the dipole electric moment making with the axis an angle α, differing from $\pi/2$ (Fig. 1.14). In this case, the forces acting on the dipole charges are not identical in magnitude. Therefore, apart from the rotational moment (torque), the dipole will experience a force tending to move it in the direction of the x-axis. To find the value of this force, we shall use Eq. (1.40), according to which

$$F_x = -\frac{\partial W_{\mathrm{p}}}{\partial x}, \quad F_y = -\frac{\partial W_{\mathrm{p}}}{\partial y}, \quad F_z = -\frac{\partial W_{\mathrm{p}}}{\partial z}.$$

In view of Eq. (1.60), we can written

$$W_{\mathrm{p}}(x, y, z) = -pE(x, y, z) \cos \alpha$$

(we consider the orientation of the dipole relative to the vector $\boldsymbol{E}$ to be constant, $\alpha = \text{constant}$).

For points on the x-axis, the derivatives of E with respect to y and z are zero. Accordingly, $\partial W_{\mathrm{p}}/\partial y = \partial W_{\mathrm{p}}/\partial z = 0$. Thus, only the force component F_x differs from zero. It is

$$F_x = -\frac{\partial W_{\mathrm{p}}}{\partial x} = p\frac{\partial E}{\partial x} \cos \alpha. \tag{1.62}$$

This result can be obtained if we take account of the fact that the field strength at the points where the charges $+q$ and $-q$ are (see Fig. 1.14) differs by the amount

[5]A particular case of such a field is that of a point charge if we take a straight line passing through the charge as the x-axis.

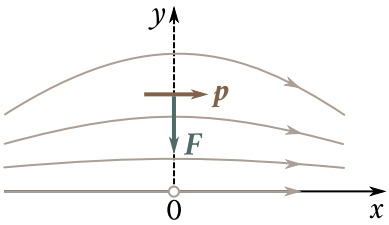


Fig. 1.15

$(\partial E/\partial x)l\cos\alpha$. Accordingly, the difference between the forces acting on the charges is $q(\partial E/\partial x)l\cos\alpha$, which coincides with Eq. (1.62).

When α is less than $\pi/2$, the value of F_x determined by Eq. (1.62) is positive. This signifies that under the action of the force the dipole is pulled into the region of a stronger field (see Fig. 1.14). When α is greater than $\pi/2$, the dipole is pushed out of the field.

In the case shown in Fig. 1.15, only the derivative $\partial E/\partial y$ differs from zero for points on the y-axis. Therefore, the force acting on the dipole is determined by the component

$$F_y = -\frac{\partial W_p}{\partial y} = p\frac{\partial E}{\partial y}, \quad (\cos\alpha = 1).$$

The derivative $\partial E/\partial y$ is negative. Consequently, the force is directed as shown in the figure. Thus, in this case too, the dipole is pulled into the field.

We shall note that like $-\partial W_p/\partial x$ gives the projection of the force acting on the system onto the x-axis, so does the derivative of Eq. (1.60) with respect to α taken with the opposite sign give the projection of the torque onto the α-"axis": $T_\alpha = -pE\sin\alpha$. The minus sign was obtained because the α-"axis" and the torque T are directed oppositely (see Fig. 1.12).

1.10. Field of a System of Charges at Great Distances

Let us take a system of N charges $q_1, q_2, \ldots, q_N$ in a volume having linear dimensions of the order of l, and study the field set up by this system at distances r that are great in comparison with l ($r > l$). We take the origin of coordinates 0 inside the volume occupied by the system and shall determine the positions of the charges with the aid of the position vectors $\boldsymbol{r}_i$, (Fig. 1.16; to simplify the figure, we have shown only the position vector of the i-th charge).

The potential at the point determined by the position vector $\boldsymbol{r}$ is

$$\varphi(\boldsymbol{r}) = \frac{1}{4\pi\varepsilon_0}\sum_{i=1}^{N}\frac{q_i}{|\boldsymbol{r}-\boldsymbol{r}_i|}. \tag{1.63}$$

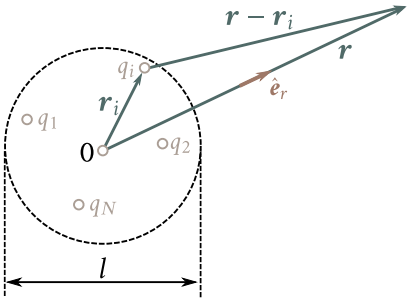


Fig. 1.16

Owing to the smallness of r_i in comparison with r, we can assume that

$$|\boldsymbol{r} - \boldsymbol{r}_i| = r - \boldsymbol{r}_i \cdot \hat{\boldsymbol{e}}_r = r\left(1 - \frac{\boldsymbol{r}_i \cdot \hat{\boldsymbol{e}}_r}{r}\right)$$

[compare with Eqs. (1.47)]. Introduction of this expression into Eq. (1.63) yields

$$\varphi(\boldsymbol{r}) = \frac{1}{4\pi\varepsilon_0} \sum_{i=1}^{N} \frac{q_i}{r}\left[\frac{1}{1 - (\boldsymbol{r}_i \cdot \hat{\boldsymbol{e}}_r / r)}\right]. \tag{1.64}$$

Using the formula

$$\frac{1}{1-x} \approx 1 + x$$

which holds when $x \ll 1$, we can transform Eq. (1.64) as follows:

$$\begin{aligned}\varphi(\boldsymbol{r}) &= \frac{1}{4\pi\varepsilon_0} \sum_{i=1}^{N} \frac{q_i}{r}\left(1 + \frac{\boldsymbol{r}_i \cdot \hat{\boldsymbol{e}}_r}{r}\right) \\ &= \frac{1}{4\pi\varepsilon_0}\frac{1}{r}\sum_{i=1}^{N} q_i + \frac{1}{4\pi\varepsilon_0}\frac{1}{r^2}\left(\sum_{i=1}^{N} q_i \boldsymbol{r}_i\right) \cdot \hat{\boldsymbol{e}}_r. \end{aligned} \tag{1.65}$$

The first term of the expression obtained is the potential of the field of a point charge having the value $q = \sum_i q_i$ [compare with Eq. (1.26)]. The second term has the same form as the expression determining the potential of a dipole field, the part of the electric moment of the dipole being played by the quantity

$$\boldsymbol{p} = \sum_{i=1}^{N} q_i \boldsymbol{r}_i. \tag{1.66}$$

This quantity is called the **dipole electric moment** of a system of charges. It is easy to verify that for a dipole Eq. (1.66) transforms into the expression $\boldsymbol{p} = q\boldsymbol{l}$ which we are already familiar with.

If the total charge of a system is zero ($\sum_i q_i = 0$), the value of the dipole moment does not depend on our choice of the origin of coordinates. To convince ourselves

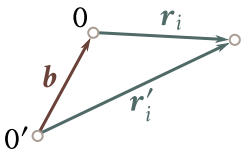
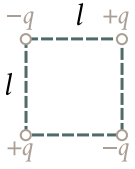


Fig. 1.17



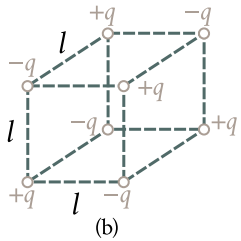


Fig. 1.18

that this is true, let us take two arbitrary origins of coordinates 0 and 0′ (Fig. 1.17). The position vectors of the i-th charge conducted from these points are related as follows:

$$\boldsymbol{r}'_i = \boldsymbol{b} + \boldsymbol{r}_i \tag{1.67}$$

(what the vector $\boldsymbol{b}$ is clear from the figure). With account taken of Eq. (1.67), the dipole moment in the system with the origin 0′ is

$$\boldsymbol{p}' = \sum_i q_i \boldsymbol{r}_i = \sum_i q_i(\boldsymbol{b} + \boldsymbol{r}_i) = \boldsymbol{b} \sum_i q_i + \sum_i q_i \boldsymbol{r}_i.$$

The first addend equals zero (because $\sum_i q_i = 0$). The second one is $\boldsymbol{p}$—the dipole moment in a coordinate system with its origin at 0. We have thus obtained that $\boldsymbol{p}' = \boldsymbol{p}$.

Equation (1.65) is in essence the first two terms of the series expansion of function (1.63) by powers of r_i/r. When $\sum_i q_i \neq 0$, the first term of Eq. (1.65) makes the main contribution to the potential (the second term diminishes in proportion to $1/r^2$ and is therefore much smaller than the first one). For an electrically neutral system ($\sum_i q_i = 0$), the first term equals zero, and the potential is determined mainly by the second term of Eq. (1.65). This is how matters stand, in particular, for the field of a dipole.

For the system of charges depicted in Fig. 1.18a and called a **quadrupole**, both $\sum_i q_i$ and $\boldsymbol{p}$ equal zero so that Eq. (1.65) gives a zero value of the potential. Actually, however, the field of a quadrupole, although it is much weaker than that of a dipole (with the same values of q and l), differs from zero. The potential of the field set up by a quadrupole is determined mainly by the third term of the expansion that is proportional to $1/r^3$. To obtain this term, we must take into consideration quantities of the order of $(r_i/r)^2$ which we disregarded in deriving Eq. (1.65). For the system of charges shown in Fig. 1.18b and called an **octupole**, the third term of the expansion also equals zero. The potential of the field of such a system is determined by the fourth term of the expansion, which is proportional to $1/r^4$.

It must be noted that the quantity equal to $\sum_i q_i$ in the numerator of the first

term of Eq. (1.65) is called a **monopole** or a **zero-order multipole**, a dipole is also called a **first-order multipole**, a quadrupole is called a **second-order multipole**, and so on.

Thus, in the general case, the field of a system of charges at great distances can be represented as the superposition of fields set up by multipoles of different orders—a monopole, dipole, quadrupole, octupole, etc.

1.11. A Description of the Properties of Vector Fields

To continue our study of the electric field, we must acquaint ourselves with the mathematical tools used to describe the properties of vector fields. These tools are called **vector analysis**. In the present section, we shall treat the fundamental concepts and selected formulas of vector analysis, and also prove its two main theorems—the Ostrogradsky-Gauss theorem (sometimes called Gauss's divergence theorem) and Stokes's theorem.

The quantities used in vector analysis can be best illustrated for the field of the velocity vector of a flowing liquid. We shall therefore introduce these quantities while dealing with the flow of an ideal incompressible liquid, and then extend the results obtained to vector fields of any nature.

We are already acquainted with one of the concepts of vector analysis. This is the **gradient**, used to characterize scalar fields. If the value of the scalar quantity $\varphi = \varphi(x, y, z)$ is compared with every point P having the coordinates x, y, z, we say that the scalar field of φ has been set. The gradient of the quantity φ is defined as the vector

$$\operatorname{grad} \varphi = \frac{\partial \varphi}{\partial x} \hat{\boldsymbol{e}}_x + \frac{\partial \varphi}{\partial y} \hat{\boldsymbol{e}}_y + \frac{\partial \varphi}{\partial z} \hat{\boldsymbol{e}}_z. \tag{1.68}$$

The increment of the function φ upon displacement over the length $\mathrm{d}\boldsymbol{l} = \hat{\boldsymbol{e}}_x \, \mathrm{d}x + \hat{\boldsymbol{e}}_y \, \mathrm{d}y + \hat{\boldsymbol{e}}_z \, \mathrm{d}z$ is

$$\mathrm{d}\varphi = \frac{\partial \varphi}{\partial x} \mathrm{d}x + \frac{\partial \varphi}{\partial y} \mathrm{d}y + \frac{\partial \varphi}{\partial z} \mathrm{d}z$$

which can be written in the form

$$\mathrm{d}\varphi = \operatorname{grad} \varphi \cdot \mathrm{d}\boldsymbol{l}. \tag{1.69}$$

Now we shall go over to establishing the characteristics of vector fields.

Vector flux. Assume that the flow of a liquid is characterized by the field of the velocity vector. The volume of liquid flowing in unit time through an imaginary surface S is called the flux of the liquid through this surface. To find the flux, let us divide the surface into elementary sections of the size ΔS. It can be seen from

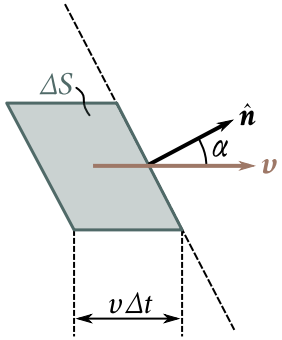


Fig. 1.19

Fig. 1.19 that during the time Δt a volume of liquid equal to

$$\Delta V = (\Delta S \cos\alpha) v \Delta t$$

will pass through section ΔS. Dividing this volume by the time Δt, we shall find the flux through surface ΔS:

$$\Delta\Phi = \frac{\Delta V}{\Delta t} = \Delta S v \cos\alpha.$$

Passing over to differentials, we find that

$$\mathrm{d}\Phi = (v\cos\alpha)\,\mathrm{d}S. \tag{1.70}$$

Equation (1.70) can be written in two other ways. First, if we take into account that $v\cos\alpha$ gives the projection of the velocity vector onto the normal $\hat{\boldsymbol{e}}_n$ to area $\mathrm{d}S$, we can write Eq. (1.70) in the form

$$\mathrm{d}\Phi = v_n\,\mathrm{d}S. \tag{1.71}$$

Second, we can introduce the vector $\mathrm{d}\boldsymbol{S}$ whose magnitude equals that of area $\mathrm{d}S$, while its direction coincides with the direction of a normal $\hat{\boldsymbol{n}}$ to the area:

$$\mathrm{d}\boldsymbol{S} = \mathrm{d}S\,\hat{\boldsymbol{n}}.$$

Since the direction of the vector $\hat{\boldsymbol{n}}$ is chosen arbitrarily (it can be directed to either side of the area), then $\mathrm{d}\boldsymbol{S}$ is not a true vector, but is a pseudo vector. The angle α in Eq. (1.70) is the angle between the vectors $\boldsymbol{v}$ and $\mathrm{d}\boldsymbol{S}$. Hence, this equation can be written in the form

$$\mathrm{d}\Phi = \boldsymbol{v}\cdot\mathrm{d}\boldsymbol{S}. \tag{1.72}$$

By summating the fluxes through all the elementary areas into which we have divided surface S, we get the flux of the liquid through S:

$$\Phi_v = \int_S \boldsymbol{v}\cdot\mathrm{d}\boldsymbol{S} = \int_S v_n\,\mathrm{d}S. \tag{1.73}$$

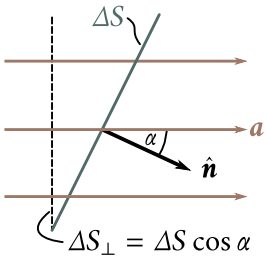


Fig. 1.20

A similar expression written for an arbitrary vector field $\boldsymbol{a}$, *i.e.*, the quantity

$$\Phi_a = \int_S \boldsymbol{a} \cdot d\boldsymbol{S} = \int_S a_n \, dS \tag{1.74}$$

is called the **flux of the vector a through surface** S. In accordance with this definition, the flux of a liquid can be called the flux of the vector $\boldsymbol{v}$ through the relevant surface [see Eq. (1.73)].

The flux of a vector is an algebraic quantity. Its sign depends on the choice of the direction of a normal to the elementary areas into which surface S is divided in calculating the flux. Reversal of the direction of the normal changes the sign of a_n and, therefore, the sign of the quantity (1.74). The customary practice for closed surfaces is calculation of the flux "emerging outward" from the region enclosed by the surface. Accordingly, in the following we shall always implicate that $\hat{\boldsymbol{v}}$ is an outward normal.

We can give an illustrative geometrical interpretation of the vector flux. For this purpose, we shall represent a vector field by a system of lines $\boldsymbol{a}$ constructed so that the density of the lines at every point is numerically equal to the magnitude of the vector $\boldsymbol{a}$ at the same point of the field (compare with the rule for constructing the lines of the vector $\boldsymbol{E}$ set out at the end of Sec. 1.5). Let us find the number ΔN of intersections of the field lines with the imaginary area ΔS. A glance at Fig. 1.20 shows that this number equals the density of the lines (*i.e.*, a) multiplied by $\Delta S_\perp = \Delta S \cos \alpha$:

$$\Delta N \,(=)\, a \Delta S \cos \alpha = a_n \Delta S.$$

We are speaking only about the numerical equality between ΔN and $a_n \Delta S$. This is why the equality sign is confined in parentheses. According to Eq. (1.74), the expression $a_n \Delta S$ is $\Delta\Phi$—the flux of the vector a through area ΔS. Thus,

$$\Delta N \,(=)\, \Delta\Phi_a. \tag{1.75}$$

For the sign of ΔN to coincide with that of $\Delta\Phi_a$, we must consider those intersections to be positive for which the angle α between the positive direction of a field line and a normal to the area is acute. The intersection should be considered negative if the angle α is obtuse. For the area shown in Fig. 1.20, all three intersec-

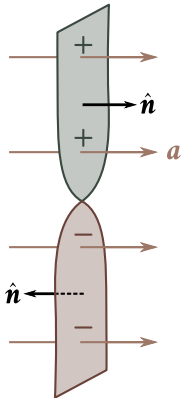


Fig. 1.21

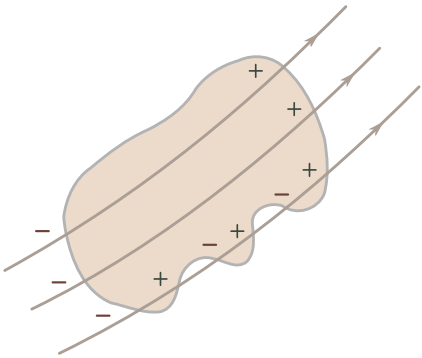


Fig. 1.22

tions are positive: $\Delta N = +3$ ($\Delta\Phi_a$ in this case is also positive because $a_n > 0$). If the direction of the normal in Fig. 1.20 is reversed, the intersections will become negative ($\Delta N = -3$), and the flux $\Delta\Phi_a$ will also be negative.

Summation of Eq. (1.75) over the finite imaginary surface S yields

$$\Delta\Phi_a \,(=)\, \sum \Delta N = N_+ - N_- \tag{1.76}$$

where N_+ and N_- are the total number of positive and negative intersections of the field lines with surface S, respectively.

The reader may be puzzled by the circumstance that since the flux, as a rule, is expressed by a fractional number, the number of intersections of the field lines with a surface compared with the flux will also be fractional. Do not be confused by this, however. Field lines are a purely conditional image deprived of a physical meaning.

Let us take an imaginary surface in the form of a strip of paper whose bottom part is twisted relative to the top one through the angle π (Fig. 1.21). The direction of a normal must be chosen identically for the entire surface. Hence, if in the top part of the strip a positive normal is directed to the right, then in the bottom part a normal will be directed to the left. Accordingly, the intersections of the field lines depicted in Fig. 1.21 with the top half of the surface must be considered positive, and with the bottom half, negative.

An outward normal is considered to be positive for a closed surface (Fig. 1.22). Therefore, the intersections corresponding to outward protrusion of the lines (in this case the angle α is acute) must be taken with the plus sign, and the ones appearing when the lines enter the surface (in this case the angle α is obtuse) must be taken with the minus sign.

Inspection of Fig. 1.22 shows that when the field lines enter a closed surface continuously, each line when intersecting the surface enters it and emerges from it

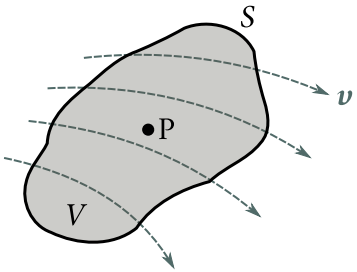


Fig. 1.23

the same number of times. As a result, the flux of the corresponding vector through this surface equals zero. It is easy to see that if field lines end inside a surface, the vector flux through the closed surface will numerically equal the difference between the number of lines beginning inside the surface (N_{beg}) and the number of lines terminating inside the surface (N_{term}):

$$\Phi_a \,(=)\, N_{\text{beg}} - N_{\text{term}}. \tag{1.77}$$

The sign of the flux depends on which of these numbers is greater. When N_{beg} is equal to N_{term}, the flux equals zero.

Divergence. Assume that we are given the field of the velocity vector of an incompressible continuous liquid. Let us take an imaginary closed surface S in the vicinity of point P (Fig. 1.23). If in the volume confined by this surface no liquid appears and no liquid vanishes, then the flux flowing outward through the surface will evidently equal zero. A liquid flux Φ_v other than zero will indicate that there are liquid sources or sinks inside the surface, *i.e.*, points at which the liquid enters the volume (sources) or emerges from it (sinks). The magnitude of the flux determines the total algebraic power of the sources and sinks[6]. When the sources predominate over the sinks, the magnitude of the flux will be positive, and when the sinks predominate, negative.

The quotient obtained when dividing the flux Φ_v by the volume which it flows out from, *i.e.*,

$$\frac{\Phi_v}{V} \tag{1.78}$$

gives the average unit power of the sources confined in the volume V. In the limit when V tends to zero, *i.e.*, when the volume V contracts to point P, expression (1.78) gives the true unit power of the sources at point P, which is called the **divergence**

[6]The power of a source (sink) is defined as the volume of liquid discharged (absorbed) in unit time. A sink can be considered as a source with a negative power.

of the vector $\boldsymbol{v}$ (it is designated by div $\boldsymbol{v}$). Thus, by definition,

$$\operatorname{div} \boldsymbol{v} = \lim_{V \to \mathrm{P}} \frac{\Phi_v}{V}.$$

The divergence of any vector $\boldsymbol{a}$ is determined in a similar way:

$$\operatorname{div} \boldsymbol{a} = \lim_{V \to \mathrm{P}} \frac{\Phi_v}{V} = \lim_{V \to \mathrm{P}} \oint \boldsymbol{a} \cdot \mathrm{d}\boldsymbol{S}. \tag{1.79}$$

The integral is taken over arbitrary closed surface S surrounding point P[7]; V is the volume confined by this surface. Since the transition $V \to \mathrm{P}$ is being performed upon which S tends to zero, we can assume that Eq. (1.79) cannot depend on the shape of the surface. This assumption is confirmed by strict calculations.

Let us surround point P with a spherical surface of an extremely small radius r (Fig. 1.24). Owing to the smallness of r, the volume V enclosed by the sphere will also be very small. We can therefore consider with a high degree of accuracy that the value of div $\boldsymbol{a}$ within the limits of the volume V is constant[8]. In this case, we can write in accordance with Eq. (1.79) that

$$\Phi_a \approx (\operatorname{div} \boldsymbol{a}) V$$

where Φ_a is the flux of the vector a through the surface surrounding the volume V. By Eq. (1.77), Φ_a equals N_{beg}, the number of lines of a beginning inside V if div $\boldsymbol{a}$ at point P is positive, or N_{term}, the number of lines of a terminating inside V if div $\boldsymbol{a}$ at point P is negative.

It follows from the above that the lines of the vector $\boldsymbol{a}$ begin in the closest vicinity of a point with a positive divergence. The field lines "diverge" from this point; the latter is the "source" of the field (Fig. 1.24a). On the other hand, in the vicinity of a point with a negative divergence, the lines of the vector $\boldsymbol{a}$ terminate. The field lines "converge" toward this point; the latter is the "sink" of the field (Fig. 1.24b). The greater the absolute value of div $\boldsymbol{a}$, the bigger is the number of lines that begin or terminate in the vicinity of the given point.

It can be seen from definition (1.79) that the divergence is a scalar function of the coordinates determining the positions of points in space (briefly—a point function). Definition (1.79) is the most general one that is independent of the kind of coordinate system used.

Let us find an expression for the divergence in a Cartesian coordinate system. We shall consider a small volume in the form of a parallelepiped with ribs parallel to the coordinate axes in the vicinity of point $P(x, y, z)$ (Fig. 1.25). The vector flux through the surface of the parallelepiped is formed from the fluxes passing through

[7] The circle on the integral sign signifies that integration is performed over a closed surface.

[8] It is assumed that the value of div $\boldsymbol{a}$ changes continuously, without any jumps, when passing from one point of a field to another.

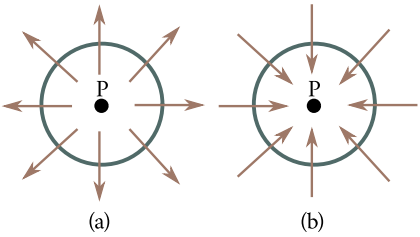


Fig. 1.24

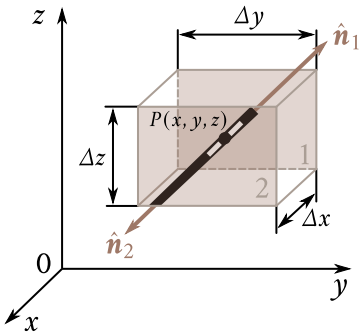


Fig. 1.25

each of the six faces separately.

Let us find the flux through the pair of faces perpendicular to the x-axis (in Fig. 1.25 these faces are designated by shaded areas and by the numbers 1 and 2). The outward normal $\hat{\boldsymbol{n}}_2$ to face 2 coincides with the direction of the x-axis. Hence, for points of this face, $a_{n_2} = a_x$. The outward normal $\hat{\boldsymbol{n}}_1$ to face 1 is directed oppositely to the x-axis. Therefore, for points on this face, $a_{n_2} = -a_x$. The flux through face 2 can be written in the form

$$a_{x,2}\Delta y\Delta z$$

where $a_{x,2}$ is the value of a_x averaged over face 2. The flux through face 1 is

$$-a_{x,1}\Delta y\Delta z$$

where $a_{x,1}$ is the average value of a_x for face 1. The total flux through faces 1 and 2 is determined by the expression

$$(a_{x,2} - a_{x,1})\Delta y\Delta z. \tag{1.80}$$

The difference $a_{x,2} - a_{x,1}$ is the increment of the average (over a face) value of a_x upon a displacement along the x-axis by Δx. Owing to the smallness of the parallelepiped (we remind our reader that we shall let its dimensions shrink to zero), this increment can be written in the form $(\partial a_x/\partial x)\Delta x$, where the value $\partial a_x/\partial x$ is taken at point P[9]. Therefore, Eq. (1.80) becomes

$$\frac{\partial a_x}{\partial x}\Delta x\Delta y\Delta z = \frac{\partial a_x}{\partial x}\Delta V.$$

Similar reasoning allows us to obtain the following expressions for the fluxes through the pairs of faces perpendicular to the y- and z-axes:

$$\frac{\partial a_y}{\partial y}\Delta x\Delta y\Delta z = \frac{\partial a_y}{\partial y}\Delta V, \quad \frac{\partial a_z}{\partial z}\Delta x\Delta y\Delta z = \frac{\partial a_z}{\partial z}\Delta V.$$

[9]The inaccuracy which we tolerate here vanishes when the volume shrinks to point P in the limit transition.

Thus, the total flux through the entire close surface is determined by the expression

$$\Phi_a = \left(\frac{\partial a_x}{\partial x} + \frac{\partial a_y}{\partial y} + \frac{\partial a_z}{\partial z} \right) \Delta V.$$

Dividing this expression by ΔV, we shall find the divergence of the vector a at point $P(x, y, z)$:

$$\operatorname{div} \boldsymbol{a} = \frac{\partial a_x}{\partial x} + \frac{\partial a_y}{\partial y} + \frac{\partial a_z}{\partial z}. \tag{1.81}$$

The Ostrogradsky-Gauss Theorem. If we know the divergence of the vector $\boldsymbol{a}$ at every point of space, we can calculate the flux of this vector through any closed surface of finite dimensions. Let us first do this for the flux of the vector $\boldsymbol{v}$ (a liquid flux). The product of $\operatorname{div} \boldsymbol{v}$ and $\mathrm{d}V$ gives the power of the sources of the liquid confined within the volume $\mathrm{d}V$. The sum of such products, *i.e.*, $\int (\operatorname{div} \boldsymbol{v})\, \mathrm{d}V$, gives the total algebraic power of the sources confined in the volume V over which integration is performed. Owing to incompressibility of the liquid, the total power of the sources must equal the liquid flux emerging through surface S enclosing the volume V. We thus arrive at the equation

$$\oint_S \boldsymbol{v} \cdot \mathrm{d}\boldsymbol{S} = \int_V (\operatorname{div} \boldsymbol{v})\, \mathrm{d}V.$$

A similar equation holds for a vector field of any nature:

$$\oint_S \boldsymbol{a} \cdot \mathrm{d}\boldsymbol{S} = \int_V (\operatorname{div} \boldsymbol{a})\, \mathrm{d}V. \tag{1.82}$$

This relation is called the **Ostrogradsky-Gauss** theorem. The integral in the left-hand side of the equation is calculated over an arbitrary closed surface S, and the integral in the right-hand side over the volume V enclosed by this surface.

Circulation. Let us revert to the flow of an ideal incompressible liquid. Imagine a closed line—the contour Γ. Assume that in some way or other we have instantaneously frozen the liquid in the entire volume except for a very thin closed channel of constant cross section including the contour Γ (Fig. 1.26). Depending on the nature of the velocity vector field, the liquid in the channel formed will either be stationary or move along the contour (circulate) in one of the two possible directions. Let us take the quantity equal to the product of the velocity of the liquid in the channel and the length of the contour l as a measure of this motion. This quantity is called the **circulation** of the vector $\boldsymbol{v}$ around the contour Γ. Thus,

$$\text{circulation of } \boldsymbol{v} \text{ around } \Gamma = vl$$

(since we assumed that the channel has a constant cross section, the magnitude of the velocity, v, is a constant).

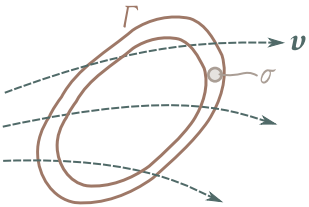


Fig. 1.26

At the moment when the walls freeze, the velocity component perpendicular to a wall will be eliminated in each of the liquid particles, and only the velocity component tangent to the contour will remain, *i.e.*, v_l. The momentum $\mathrm{d}\boldsymbol{p}_l$, is associated with this component. The magnitude of the momentum for a liquid particle contained within a segment of the channel of length $\mathrm{d}l$ is $\rho\sigma v_l\,\mathrm{d}l$ (ρ is the density of the liquid, and σ is the cross-sectional area of the channel). Since the liquid is ideal, the action of the walls can change only the direction of the vector $\mathrm{d}\boldsymbol{p}_l$, but not its magnitude. The interaction between the liquid particles will cause a redistribution of the momentum between them that will level out the velocities of all the particles. The algebraic sum of the tangential components of the momenta cannot change: the momentum acquired by one of the interacting particles equals the momentum lost by the second particle. This signifies that

$$\rho\sigma v l = \oint_\Gamma \rho\sigma v_l\,\mathrm{d}l$$

where v is the circulation velocity, and v_l is the tangential component of the liquid's velocity in the volume $\sigma\,\mathrm{d}l$ at the moment of time preceding the freezing of the channel walls. Cancelling $\rho\sigma$, we get

$$\text{circulation of } \boldsymbol{v} \text{ around } \Gamma = v l = \oint_\Gamma v_l\,\mathrm{d}l.$$

The circulation of any vector $\boldsymbol{a}$ around an arbitrary closed contour Γ is determined in a similar way:

$$\text{circulation of } \boldsymbol{a} \text{ around } \Gamma = \oint_\Gamma \boldsymbol{a}\cdot\mathrm{d}\boldsymbol{l} = \oint_\Gamma a_l\,\mathrm{d}l. \tag{1.83}$$

It may seem that for the circulation to be other than zero the vector lines must be closed or at least bent in some way or other in the direction of circumventing the contour. It is easy to see that this assumption is wrong. Let us consider the laminar flow of water in a river. The velocity of the water directly at the river bottom is zero and grows as we approach the surface of the water (Fig. 1.27). The streamlines (lines of the vector $\boldsymbol{v}$) are straight. Notwithstanding this fact, the circulation of the vector $\boldsymbol{v}$ around the contour depicted by the dash line obviously differs from zero.

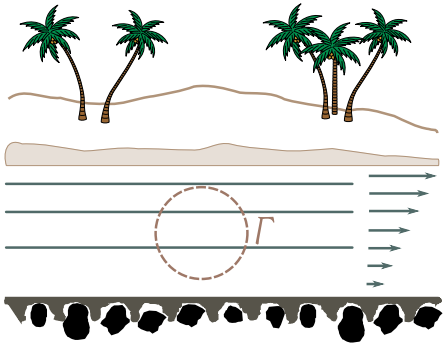


Fig. 1.27

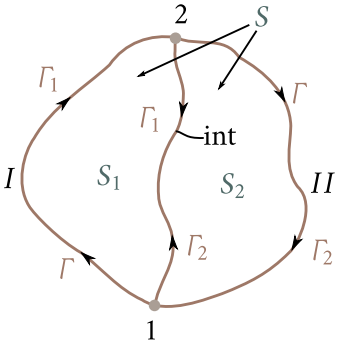


Fig. 1.28

On the other hand, in a field with curved lines, the circulation may equal zero.

Circulation has the property of additivity. This signifies that the sum of the circulations around contours Γ_1 and Γ_2 enclosing neighboring surfaces S_1 and S_2 (Fig. 1.28) equals the circulation around contour Γ enclosing surface S, which is the sum of surfaces S_1 and S_2. Indeed, the circulation C_1 around the contour bounding surface S_1 can be represented as the sum of the integrals

$$C_1 = \oint_{\Gamma_1} \boldsymbol{a} \cdot \mathrm{d}\boldsymbol{l} = \int_{1,(I)}^{2} \boldsymbol{a} \cdot \mathrm{d}\boldsymbol{l} + \int_{2,(\mathrm{int.})}^{1} \boldsymbol{a} \cdot \mathrm{d}\boldsymbol{l}. \tag{1.84}$$

The first integral is taken over section I of the outer contour, the second over the interface between surfaces S_1 and S_2 in direction 2-1.

Similarly, the circulation C_2 around the contour enclosing surface S_2 is

$$C_2 = \oint_{\Gamma_2} \boldsymbol{a} \cdot \mathrm{d}\boldsymbol{l} = \int_{2,(II)}^{1} \boldsymbol{a} \cdot \mathrm{d}\boldsymbol{l} + \int_{1,(\mathrm{int.})}^{2} \boldsymbol{a} \cdot \mathrm{d}\boldsymbol{l}. \tag{1.85}$$

The first integral is taken over section II of the outer contour, the second over the interface between surfaces S_1 and S_2 in direction 1-2.

The circulation around the contour bounding total surface S can be represented in the form

$$C = \oint_{\Gamma} \boldsymbol{a} \cdot \mathrm{d}\boldsymbol{l} = \int_{1,(I)}^{2} \boldsymbol{a} \cdot \mathrm{d}\boldsymbol{l} + \int_{2,(II)}^{1} \boldsymbol{a} \cdot \mathrm{d}\boldsymbol{l}. \tag{1.86}$$

The second addends in Eqs. (1.84) and (1.85) differ only in their sign. Therefore, the sum of these expressions will equal Eq. (1.86). Thus,

$$C = C_1 + C_2. \tag{1.87}$$

Equation (1.87) which we have proved does not depend on the shape of the surfaces and holds for any number of addends. Hence, if we divide an arbitrary open surface S into a great number of elementary surfaces ΔS[10] (Fig. 1.29), then

[10] In the figure, the elementary surfaces are depicted in the form of rectangles. Actually, their shape

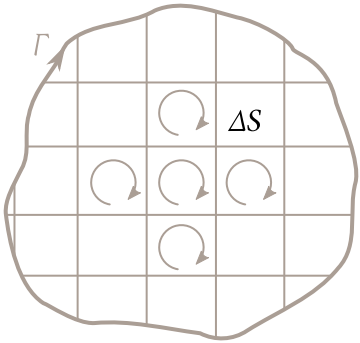


Fig. 1.29

the circulation around the contour enclosing S can be written as the sum of the elementary circulations ΔC around the contours enclosing the ΔS's:

$$C = \sum_i \Delta C_i. \tag{1.88}$$

Curl. The additivity of the circulation permits us to introduce the concept of unit circulation, *i.e.*, consider the ratio of the circulation C to the magnitude of surface S around which the circulation "flows". When surface S is finite, the ratio C/S gives the mean value of the unit circulation. This value characterizes the properties of a field averaged over surface S. To obtain the characteristic of the field at point P, we must reduce the dimensions of the surface, making it shrink to point P. The ratio C/S tends to a limit that characterizes the properties of the field at point P.

Thus, let us take an imaginary contour Γ in a plane passing through point P, and consider the expression

$$\lim_{S \to \mathrm{P}} \frac{C_a}{S} \tag{1.89}$$

where C_a is the circulation of the vector $\boldsymbol{a}$ around the contour Γ and S is the surface area enclosed by the contour.

Limit (1.89) calculated for an arbitrarily oriented plane cannot be an exhaustive characteristic of the field at point P because the magnitude of this limit depends on the orientation of the contour in space in addition to the properties of the field at point P. This orientation can be given by the direction of a positive normal $\hat{\boldsymbol{n}}$ to the plane of the contour (a positive normal is one that is associated with the direction of circumvention of the contour in integration by the right-hand screw rule). In determining limit (1.89) at the same point P for different directions $\hat{\boldsymbol{n}}$, we

may be absolutely arbitrary.

shall obtain different values. For opposite directions, these values will differ only in their sign (reversal of the direction $\hat{\boldsymbol{n}}$ is equivalent to reversing the direction of circumvention of the contour in integration, which only causes a change in the sign of the circulation). For a certain direction of the normal, the magnitude of expression (1.89) at the given point will be maximum.

Thus, quantity (1.89) behaves like the projection of a vector onto the direction of a normal to the plane of the contour around which the circulation is taken. The maximum value of quantity (1.89) determines the magnitude of this vector, and the direction of the positive normal $\hat{\boldsymbol{n}}$ at which the maximum is reached gives the direction of the vector. This vector is called the **curl** of the vector $\boldsymbol{a}$. Its symbol is curl $\boldsymbol{a}$. Using this notation, we can write expression (1.89) in the form

$$(\operatorname{curl} \boldsymbol{a})_n = \lim_{S \to \mathrm{P}} \frac{C_a}{S} = \lim_{S \to \mathrm{P}} \frac{1}{S} \oint_S \boldsymbol{a}\, \mathrm{d}\boldsymbol{l}. \tag{1.90}$$

We can obtain a graphical picture of the curl of the vector $\boldsymbol{v}$ by imagining a small and light fan impeller placed at the given point of a flowing liquid (Fig. 1.30). At the spots where the curl differs from zero, the impeller will rotate, its velocity being the higher, the greater in value is the projection of the curl onto the impeller axis.

Equation (1.90) defines the vector curl $\boldsymbol{a}$. This definition is a most general one that does not depend on the kind of coordinate system used. To find expressions for the projections of the vector curl $\boldsymbol{a}$ onto the axes of a Cartesian coordinate system, we must determine the values of quantity (1.90) for such orientations of area S for which the normal $\hat{\boldsymbol{n}}$ to the area coincides with one of the axes x, y, z. If, for example, we direct $\hat{\boldsymbol{n}}$ along the x-axis, then (1.90) becomes $(\operatorname{curl} \boldsymbol{a})_x$. Contour Γ in this case is arranged in a plane parallel to the coordinate plane yz. Let us take this contour in the form of a rectangle with the sides Δy and Δz (Fig. 1.31, the x-axis is directed toward us in this figure; the direction of circumvention indicated in the figure is associated with the direction of the x-axis by the right-hand screw rule). Section 1 of the contour is opposite in direction to the z-axis. Therefore, a_l on this section coincides with $-a_z$. Similar reasoning shows that a_l on sections 2, 3, and 4 equals a_y, a_z, and $-a_y$, respectively. Hence, the circulation can be written in the form

$$\left(a_{z,3} - a_{z,1}\right) \Delta z - \left(a_{y,4} - a_{y,2}\right) \Delta y \tag{1.91}$$

where $a_{z,3}$ and $a_{z,1}$ are the average values of a_z on sections 3 and 1, respectively, and $a_{y,4}$ and $a_{y,2}$ are the average values of a_y on sections 4 and 2.

The difference $a_{z,3} - a_{z,1}$ is the increment of the average value of a_z on the section Δz when this section is displaced in the direction of the y-axis by Δy. Owing to the smallness of Δy and Δz, this increment can be represented in the form $(\partial a_z / \partial y)\Delta y$,

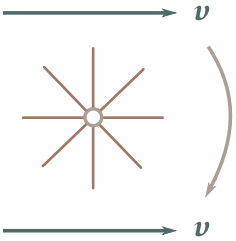


Fig. 1.30

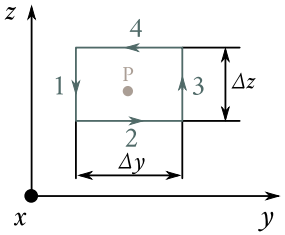


Fig. 1.31

where the value of $\partial a_z/\partial y$ is taken for point P[11]. Similarly, the difference $a_{y,4} - a_{y,2}$ can be represented in the form $(\partial a_y/\partial z)\,\Delta z$. Using these expressions in Eq. (1.91) and putting the common factor outside the parentheses, we get the following expression for the circulation:

$$\left(\frac{\partial a_z}{\partial y} - \frac{\partial a_y}{\partial z}\right)\Delta y\,\Delta z = \left(\frac{\partial a_z}{\partial y} - \frac{\partial a_y}{\partial z}\right)\Delta S$$

where ΔS is the area of the contour. Dividing the circulation by ΔS, we find the expression for the projection of curl $\boldsymbol{a}$ onto the x-axis:

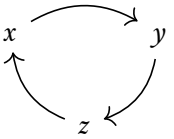
$$(\operatorname{curl}\boldsymbol{a})_x = \frac{\partial a_z}{\partial y} - \frac{\partial a_y}{\partial z}. \tag{1.92}$$

We can find by similar reasoning that

$$(\operatorname{curl}\boldsymbol{a})_y = \frac{\partial a_x}{\partial z} - \frac{\partial a_z}{\partial x}, \tag{1.93}$$

$$(\operatorname{curl}\boldsymbol{a})_z = \frac{\partial a_y}{\partial x} - \frac{\partial a_x}{\partial y}. \tag{1.94}$$

It is easy to see that any of the equations (1.92)-(1.94) can be obtained from the preceding one [Eq. (1.94) should be considered as the preceding one for Eq. (1.94)] by the so-called cyclic transposition of the coordinates, *i.e.*, by replacing the coordinates according to the scheme



Thus, the curl of the vector $\boldsymbol{a}$ is determined in the Cartesian coordinate system by the following expression:

$$\operatorname{curl}\boldsymbol{a} = \hat{\boldsymbol{e}}_x\left(\frac{\partial a_z}{\partial y} - \frac{\partial a_y}{\partial z}\right) + \hat{\boldsymbol{e}}_y\left(\frac{\partial a_x}{\partial z} - \frac{\partial a_z}{\partial x}\right) + \hat{\boldsymbol{e}}_z\left(\frac{\partial a_y}{\partial x} - \frac{\partial a_x}{\partial y}\right). \tag{1.95}$$

[11]The inaccuracy which we tolerate here vanishes when the contour shrinks to point P in the limit transition.

Below we shall indicate a more elegant way of writing this expression.

Stokes' Theorem. Knowing the curl of the vector $\boldsymbol{a}$ at every point of surface S (not necessarily plane), we can calculate the circulation of this vector around contour Γ enclosing S (the contour may also not be plane). For this purpose, we divide the surface into very small elements ΔS. Owing to their smallness, these elements can be considered as plane. Therefore in accordance with Eq. (1.90), the circulation of the vector $\boldsymbol{a}$ around the contour bounding ΔS can be written in the form

$$\Delta C \approx (\operatorname{curl} \boldsymbol{a})_n \Delta S = \operatorname{curl} \boldsymbol{a} \cdot \Delta \boldsymbol{S} \tag{1.96}$$

where $\hat{\boldsymbol{n}}$ is a positive normal to surface element ΔS.

In accordance with Eq. (1.88), summation of expression (1.96) over all the ΔS's yields the circulation of the vector $\boldsymbol{a}$ around contour Γ enclosing S:

$$C = \sum \Delta C \approx \sum \operatorname{curl} \boldsymbol{a} \cdot \Delta \boldsymbol{S}.$$

Performing a limit transition in which all the ΔS's shrink to zero (their number grows unlimitedly), we arrive at the equation

$$\oint_{\Gamma} \boldsymbol{a} \cdot \mathrm{d}\boldsymbol{l} = \int_{S} (\operatorname{curl} \boldsymbol{a}) \cdot \Delta \boldsymbol{S}. \tag{1.97}$$

Equation (1.97) is called **Stokes' theorem**. Its meaning is that the circulation of the vector $\boldsymbol{a}$ around an arbitrary contour Γ equals the flux of the vector curl $\boldsymbol{a}$ through the arbitrary surface S surrounded by the given contour.

The Del Operator. Writing of the formulas of vector analysis is simplified quite considerably if we introduce a vector differential operator designated by the symbol ∇ (nabla or del) and called the **del operator** or the **Hamiltonian operator**. This operator denotes a vector with the components $\partial/\partial x$, $\partial/\partial y$ and $\partial/\partial z$. Consequently,

$$\nabla = \hat{\boldsymbol{e}}_x \frac{\partial}{\partial x} + \hat{\boldsymbol{e}}_y \frac{\partial}{\partial y} + \hat{\boldsymbol{e}}_z \frac{\partial}{\partial z}. \tag{1.98}$$

This vector has no meaning by itself. It acquires a meaning in combination with the scalar or vector function by which it is symbolically multiplied. Thus, if we multiply the vector ∇ by the scalar φ we obtain the vector

$$\nabla\varphi = \hat{\boldsymbol{e}}_x \frac{\partial \varphi}{\partial x} + \hat{\boldsymbol{e}}_y \frac{\partial \varphi}{\partial y} + \hat{\boldsymbol{e}}_z \frac{\partial \varphi}{\partial z} \tag{1.99}$$

which is the gradient of the function φ [see Eq. (1.68)].

The scalar product of the vectors ∇ and $\boldsymbol{a}$ gives the scalar

$$\nabla \cdot \boldsymbol{a} = \nabla_x a_x + \nabla_y a_y + \nabla_z a_z \tag{1.100}$$

which we can see to be the divergence of the vector $\boldsymbol{a}$ [see Eq. (1.81)].

Finally, the vector product of the vectors ∇ and $\boldsymbol{a}$ gives a vector with the components $(\nabla \times \boldsymbol{a})_x = \nabla_y a_z - \nabla_z a_y = \partial a_z/\partial y - \partial a_y/\partial z$, etc., that coincide with

the components of curl $\boldsymbol{a}$ [see Eqs. (1.92)-(1.94)]. Hence, using the writing of a vector product with the aid of a determinant, we have

$$\operatorname{curl} \boldsymbol{a} = \nabla \times \boldsymbol{a} = \begin{vmatrix} \hat{\boldsymbol{e}}_x & \hat{\boldsymbol{e}}_y & \hat{\boldsymbol{e}}_z \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ a_x & a_y & a_z \end{vmatrix}. \tag{1.101}$$

Thus, there are two ways of denoting the gradient, divergence, and curl:

$$\nabla\varphi \equiv \operatorname{grad} \varphi, \quad \nabla \cdot \boldsymbol{a} \equiv \operatorname{div} \boldsymbol{a}, \quad \nabla \times \boldsymbol{a} \equiv \operatorname{curl} \boldsymbol{a}.$$

The use of the del symbol has a number of advantages. We shall therefore use such symbols in the following. One must accustom oneself to identify the symbol $\nabla\varphi$ with the words "gradient of phi" (*i.e.*, to say not "del phi", but "gradient of phi"), the symbol $\nabla \cdot \boldsymbol{a}$ with the words "divergence of a" and, finally, the symbol $\nabla \times \boldsymbol{a}$ with the words "curl of a".

When using the vector ∇, one must remember that it is a differential operator acting on all the functions to the right of it. Consequently, in transforming expressions including ∇, one must take into consideration both the rules of vector algebra and those of differential calculus. For example, the derivative of the product of the functions φ and ψ is

$$(\varphi\psi)' = \varphi'\psi + \varphi\psi'.$$

Accordingly,

$$\operatorname{grad}(\varphi\psi) = \nabla(\varphi\psi) = \psi\nabla\varphi + \varphi\nabla\psi = \psi \operatorname{grad} \varphi + \varphi \operatorname{grad} \psi. \tag{1.102}$$

Similarly,

$$\operatorname{div}(\varphi\boldsymbol{a}) = \nabla \cdot (\varphi\boldsymbol{a}) = \boldsymbol{a} \cdot (\nabla\varphi) + \varphi(\nabla \cdot \boldsymbol{a}). \tag{1.103}$$

The gradient of a function φ is a vector function. Therefore, the divergence and curl operations can be performed with it:

$$\begin{aligned} \operatorname{div} \operatorname{grad} \varphi = \nabla \cdot \nabla\varphi = (\nabla \cdot \nabla)\varphi = \left(\nabla_x^2 + \nabla_y^2 + \nabla_z^2\right)\varphi \\ = \frac{\partial^2\varphi}{\partial x^2} + \frac{\partial^2\varphi}{\partial y^2} + \frac{\partial^2\varphi}{\partial z^2} = \Delta\varphi \end{aligned} \tag{1.104}$$

(Δ is the Laplacian operator)

$$\operatorname{curl} \operatorname{grad} \varphi = \nabla \times (\nabla\varphi) = (\nabla \times \nabla)\varphi \tag{1.105}$$

(we remind our reader that the vector product of a vector and itself is zero).

Let us apply the divergence and curl operations to the function curl $\boldsymbol{a}$:

$$\operatorname{div} \operatorname{curl} \boldsymbol{a} = \nabla \cdot \nabla \times \boldsymbol{a} = 0 \tag{1.106}$$

(a scalar triple product equals the volume of a parallelepiped constructed on the vectors being multiplied (see Vol. I, p. 22); if two of these vectors coincide, the

volume of the parallelepiped equals zero):

$$\text{curl curl } \boldsymbol{a} = \nabla \times (\nabla \times \boldsymbol{a}) = \nabla(\nabla \cdot \boldsymbol{a}) - (\nabla \cdot \nabla)\boldsymbol{a} = \text{grad div } \boldsymbol{a} - \Delta \boldsymbol{a} \quad (1.107)$$

[we have used Eq. (1.35) of Vol. I, namely, $\boldsymbol{a} \times \boldsymbol{b} \times \boldsymbol{c} = \boldsymbol{b}(\boldsymbol{a} \cdot \boldsymbol{c}) - \boldsymbol{c}(\boldsymbol{a} \cdot \boldsymbol{b})$].

Equation (1.106) signifies that the field of a curl has no sources. Hence, the lines of the vector curl $\boldsymbol{a}$ have neither a beginning nor an end. It is exactly for this reason that the flux of a curl through any surface S resting on the given contour Γ is the same [see Eq. (1.97)].

We shall note in concluding that when the del operator is used, Eqs. (1.82) and (1.97) can be given the form

$$\oint_S \boldsymbol{a} \cdot d\boldsymbol{S} = \oint_V \nabla \cdot \boldsymbol{a}\, dV, \qquad \text{(the Ostrogradsky-Gauss theorem)} \qquad (1.108)$$

$$\oint_\Gamma \boldsymbol{a} \cdot d\boldsymbol{l} = \int_S (\nabla \times \boldsymbol{a}) \cdot d\boldsymbol{S}. \qquad \text{(Stokes' theorem)} \qquad (1.109)$$

1.12. Circulation and Curl of an Electrostatic Field

We established in Sec. 1.6 that the forces acting on the charge q in an electrostatic field are conservative. Hence, the work of these forces on any closed path Γ is zero:

$$A = \oint_\Gamma q\boldsymbol{E} \cdot d\boldsymbol{l} = 0.$$

Cancelling q, we get

$$\oint_\Gamma \boldsymbol{E} \cdot d\boldsymbol{l} = 0 \qquad (1.110)$$

(compare with Eq. (1.46)].

The integral in the left-hand side of Eq. (1.110) is the circulation of the vector $\boldsymbol{E}$ around contour Γ [see expression (1.80)]. Thus, *an electrostatic field is characterized by the fact that the circulation of the strength (intensity) vector of this field around any closed contour equals zero*.

Let us take an arbitrary surface S resting on contour Γ for which the circulation is calculated (Fig. 1.32). According to Stokes's theorem [see Eq. (1.109)], the integral of curl $\boldsymbol{E}$ taken over this surface equals the circulation of the vector $\boldsymbol{E}$ around contour Γ:

$$\int_S (\nabla \times \boldsymbol{E}) \cdot d\boldsymbol{S} = \oint_\Gamma \boldsymbol{E} \cdot d\boldsymbol{l}. \qquad (1.111)$$

Since the circulation equals zero, we arrive at the conclusion that

$$\int_S (\nabla \times \boldsymbol{E}) \cdot d\boldsymbol{S} = 0.$$

This condition must be observed for any surface S resting on arbitrary contour Γ.

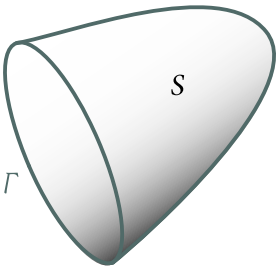


Fig. 1.32

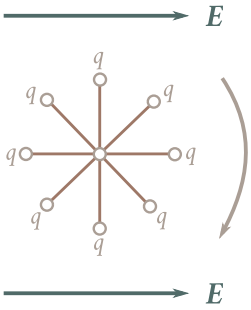


Fig. 1.33

This is possible only if the curl of the vector $\boldsymbol{E}$ at every point of the field equals zero:

$$\nabla \times \boldsymbol{E} = 0. \tag{1.112}$$

By analogy with the fan impeller shown in Fig. 1.25, let us imagine an electrical "impeller" in the form of a light hub with spokes whose ends carry identical positive charges q (Fig. 1.33; the entire arrangement must be small in size). At the points of an electric field where curl $\boldsymbol{E}$ differs from zero, such an impeller would rotate with an acceleration that is the greater, the larger is the projection of the curl onto the impeller axis. For an electrostatic field, such an imaginary arrangement would not rotate with any orientation of its axis.

Thus, a feature of an electrostatic field is that it is a non-circuital one. We established in the preceding section that the curl of the gradient of a scalar function equals zero [see expression (1.96)]. Therefore, the equality to zero of curl $\boldsymbol{E}$ at every point of a field makes it possible to represent $\boldsymbol{E}$ in the form of the gradient of a scalar function φ called the potential. We have already considered this representation in Sec. 1.8 [see Eq. (1.41); the minus sign in this equation was taken from physical considerations].

We can immediately conclude from the need to observe condition (1.110) that the existence of an electrostatic field of the kind shown in Fig. 1.34 is impossible. Indeed, for such a field, the circulation around the contour shown by the dash line would differ from zero, which contradicts condition (1.110). It is also impossible for a field differing from zero in a restricted volume to be homogeneous throughout this volume (Fig. 1.35). In this case, the circulation around the contour shown by the dash line would differ from zero.

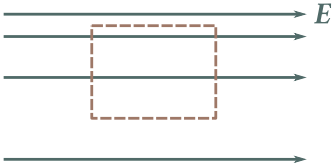


Fig. 1.34

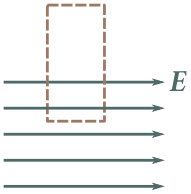


Fig. 1.35

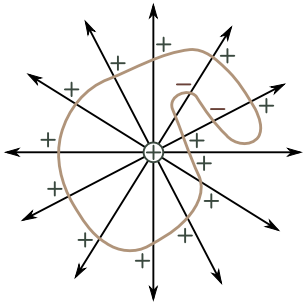


Fig. 1.36

1.13. Gauss's Theorem

We established in the preceding section what the curl of an electrostatic field equals. Now let us find the divergence of a field. For this purpose, we shall consider the field of a point charge q and calculate the flux of the vector $\boldsymbol{E}$ through closed surface S surrounding the charge (Fig. 1.36). We showed in Sec. 1.5 that the number of lines of the vector $\boldsymbol{E}$ beginning at a point charge $+q$ or terminating at a charge $-q$ numerically equals q/ε_0.

By Eq. (1.77), the flux of the vector $\boldsymbol{E}$ through any closed surface equals the number of lines coming out, *i.e.*, beginning on the charge, if it is positive, and the number of lines entering the surface, *i.e.*, terminating on the charge, if it is negative. Taking into account that the number of lines beginning or terminating at a point charge numerically equals q/ε_0 (see Sec. 1.5), we can write that

$$\Phi_E = \frac{q}{\varepsilon_0}. \tag{1.113}$$

The sign of the flux coincides with that of the charge q. The dimensions of both sides of Eq. (1.113) are identical.

Now let us assume that a closed surface surrounds N point charges $q_1, q_2, \ldots, q_N$. On the basis of the superposition principle, the strength $\boldsymbol{E}$ of the field set up by all the charges equals the sum of the strengths $\boldsymbol{E}_i$ set up by each charge separately: $\boldsymbol{E} = \sum_i \boldsymbol{E}_i$. Hence,

$$\Phi_E = \oint_S \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{S} = \oint_S \left(\sum_i \boldsymbol{E}_i \right) \cdot \mathrm{d}\boldsymbol{S} = \sum_i \oint_S \boldsymbol{E}_i \cdot \mathrm{d}\boldsymbol{S}.$$

Each of the integrals inside the sum sign equals q_i/ε_0. Therefore,

$$\Phi_E = \oint_S \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{S} = \frac{1}{\varepsilon_0} \sum_{i=1}^{N} q_i. \tag{1.114}$$

The statement we have proved is called **Gauss's theorem**. According to it, *the flux*

of an electric field strength vector through a closed surface equals the algebraic sum of the charges enclosed by this surface divided by ε_0.

When considering fields set up by macroscopic charges (*i.e.*, charges formed by an enormous number of elementary charges), the discrete structure of these charges is disregarded, and they are considered to be distributed in space continuously with a finite density everywhere. The **volume density of a charge** ρ is determined by analogy with the density of a mass as the ratio of the charge $\mathrm{d}q$ to the infinitely small (physically) volume $\mathrm{d}V$ containing this charge:

$$\rho = \frac{\mathrm{d}q}{\mathrm{d}V}. \tag{1.115}$$

In the given case by an infinitely small (physically) volume, we must understand a volume which on the one hand is sufficiently small for the density within its limits to be considered identical, and on the other is sufficiently great for the discreteness of the charge not to manifest itself.

Knowing the charge density at every point of space, we can find the total charge surrounded by closed surface S. For this purpose, we must calculate the integral of ρ with respect to the volume enclosed by the surface:

$$\sum_i q_i = \int_V \rho \, \mathrm{d}V.$$

Thus, Eq. (1.114) can be written in the form

$$\oint_S \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{S} = \frac{1}{\varepsilon_0} \int_V \rho \, \mathrm{d}V. \tag{1.116}$$

Replacing the surface integral with a volume one in accordance with Eq. (1.108), we have

$$\int_V \nabla \cdot \boldsymbol{E} \, \mathrm{d}V = \frac{1}{\varepsilon_0} \int_V \rho \, \mathrm{d}V.$$

The relation which we have arrived at must be observed for any arbitrarily chosen volume V. This is possible only if the values of the integrands for every point of space are the same. Hence, the divergence of the vector $\boldsymbol{E}$ is associated with the density of the charge at the same point by the equation

$$\nabla \cdot \boldsymbol{E} = \frac{1}{\varepsilon_0} \rho. \tag{1.117}$$

This equation expresses Gauss's theorem in the differential form.

For a flowing liquid, $\nabla \cdot \boldsymbol{v}$ gives the unit power of the sources of the liquid at a given point. By analogy, charges are said to be sources of an electric field.

1.14. Calculating Fields with the Aid of Gauss's Theorem

Gauss's theorem permits us in a number of cases to find the strength of a field in a much simpler way than by using Eq. (1.15) for the field strength of a point charge and the field superposition principle. We shall demonstrate the possibilities of Gauss's theorem by employing a few examples that will be useful for our further exposition. Before starting on our way, we shall introduce the concepts of surface and linear charge densities.

If a charge is concentrated in a thin surface layer of the body carrying the charge, the distribution of the charge in space can be characterized by the surface density σ, which is determined by the expression

$$\sigma = \frac{\mathrm{d}q}{\mathrm{d}S}. \tag{1.118}$$

Here $\mathrm{d}q$ is the charge contained in the layer of area $\mathrm{d}S$. By $\mathrm{d}S$ is meant an infinitely small (physically) section of the surface.

If a charge is distributed over the volume or surface of a cylindrical body (uniformly in each section), the linear charge density is used, *i.e.*,

$$\lambda = \frac{\mathrm{d}q}{\mathrm{d}l} \tag{1.119}$$

where $\mathrm{d}l$ is the length of an infinitely small (physically) segment of the cylinder, and $\mathrm{d}q$ is the charge concentrated on this segment.

Field of an Infinite Homogeneously Charged Plane. Assume that the surface charge density at all points of a plane is identical and equal to σ; for definiteness we shall consider the charge to be positive. It follows from considerations of symmetry that the field strength at any point is directed at right angles to the plane. Indeed, since the plane is infinite and charged homogeneously, there is no reason why the vector $\boldsymbol{E}$ should deflect to a side from a normal to the plane. It is further evident that at points symmetrical relative to the plane, the field strength is identical in magnitude and opposite in direction.

Let us imagine mentally a cylindrical surface with generatrices perpendicular to the plane and bases of a size ΔS arranged symmetrically relative to the plane (Fig. 1.37). Owing to symmetry, we have $E' = E'' = E$. We shall apply Gauss's theorem to the surface. The flux through the side part of the surface will be absent because E_n at each point of it is zero. For the bases, E_n coincides with E. Hence, the total flux through the surface is $2E\Delta S$. The surface encloses the charge $\sigma\Delta S$. According to Gauss's theorem, the condition must be observed that

$$2E\Delta S = \frac{\sigma\Delta S}{\varepsilon_0}$$

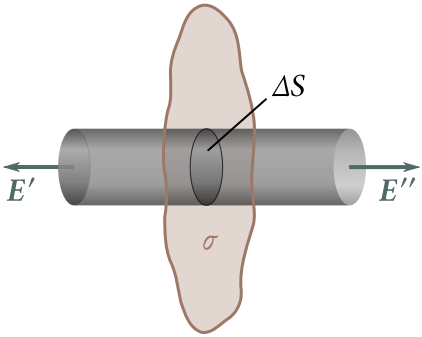


Fig. 1.37

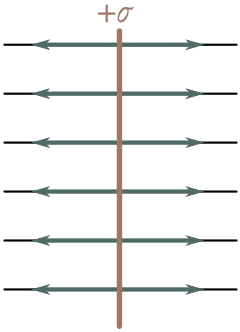


Fig. 1.38

whence

$$E = \frac{\sigma}{2\varepsilon_0}. \tag{1.120}$$

The result we have obtained does not depend on the length of the cylinder. This signifies that at any distances from the plane, the field strength is identical in magnitude. The field lines are shown in Fig. 1.38. For a negatively charged plane, the result will be the same except for the reversal of the direction of the vector $\boldsymbol{E}$ and the field lines.

If we take a plane of finite dimensions, for instance a charged thin plate[12], then the result obtained above will hold only for points, the distance to which from the edge of the plate considerably exceeds the distance from the plate itself (in Fig. 1.39 the region containing such points is outlined by a dash line). At points at an increasing distance from the plane or approaching its edges, the field will differ more and more from that of an infinitely charged plane. It is easy to imagine the nature of the field at great distances if we take into account that at distances considerably exceeding the dimensions of the plate, the field it sets up can be treated as that of a point charge.

Field of Two Uniformly Charged Planes. The field of two parallel infinite planes carrying opposite charges with a constant surface density σ identical in magnitude can be found by superposition of the fields produced by each plane separately (Fig. 1.40). In the region between the planes, the fields being added have the same direction, so that the resultant field strength is

$$E = \frac{\sigma}{\varepsilon_0}. \tag{1.121}$$

[12] For a plate, by σ in Eq. (1.120) should be understood the charge concentrated on 1 m^2 of the plate over its entire thickness. In metal bodies, the charge is distributed over the external surface. Therefore by σ we should understand the double value of the charge density on the surfaces surrounding the metal plate.

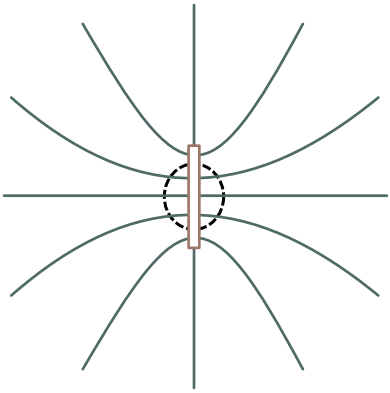


Fig. 1.39

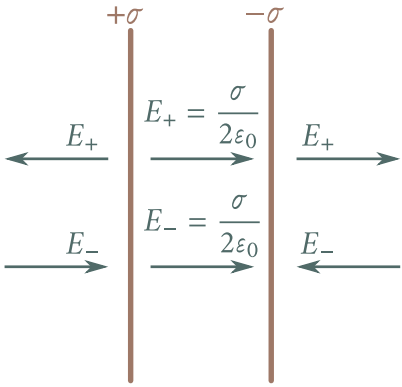


Fig. 1.40

Outside the volume bounded by the planes, the fields being added have opposite directions so that the resultant field strength equals zero.

Thus, the field is concentrated between the planes. The field strength at all points of this region is identical in value and in direction; consequently, the field is homogeneous. The field lines are a collection of parallel equispaced straight lines.

The result we have obtained also holds approximately for planes of finite dimensions if the distance between them is much smaller than their linear dimensions (a parallel-plate capacitor). In this case, appreciable deviations of the field from homogeneity are observed only near the edges of the plates (Fig. 1.41).

Field of an Infinite Charged Cylinder. Assume that the field is produced by an infinite cylindrical surface of radius R whose charge has a constant surface density σ. Considerations of symmetry show that the field strength at any point must be directed along a radial line perpendicular to the cylinder axis, and that the magnitude of the strength can depend only on the distance r from the cylinder axis. Let us mentally imagine a coaxial closed cylindrical surface of radius r and height h with a charged surface (Fig. 1.42). For the bases of the cylinder, we have $E_n = 0$, for the side surface $E_n = E(r)$ (the charge is assumed to be positive). Hence, the flux of the vector $\boldsymbol{E}$ through the surface being considered is $E(r) \times 2\pi rh$. If $r > R$, the charge $q = \lambda h$ (where λ is the linear charge density) will get into the surface. Applying Gauss's theorem, we find that

$$E(r) \times 2\pi rh = \frac{2\lambda}{\varepsilon_0}.$$

Hence,

$$E(r) = \frac{1}{2\pi\varepsilon_0}\frac{\lambda}{r} \quad (r \geqslant R). \tag{1.122}$$

If $r < R$, the closed surface being considered contains no charges inside, owing to

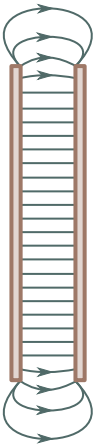


Fig. 1.41

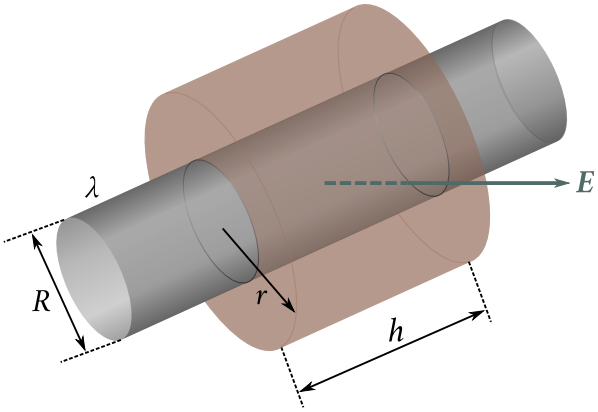


Fig. 1.42

which $E(r) = 0$.

Thus, there is no field inside a uniformly charged cylindrical surface of infinite length. The field strength outside the surface is determined by the linear charge density λ and the distance r from the cylinder axis.

The field of a negatively charged cylinder differs from that of a positively charged one only in the direction of the vector $\boldsymbol{E}$. A glance at Eq. (1.122) shows that by reducing the cylinder radius R (with a constant linear charge density λ), we can obtain a field with a very great strength near the surface of the cylinder.

Introducing $\lambda = 2\pi R\sigma$ into Eq. (1.122) and assuming that $r = R$, we get the following value for the field strength in direct proximity to the surface of a cylinder:

$$E(R) = \frac{\sigma}{\varepsilon_0}. \tag{1.123}$$

The superposition principle makes it simple to find the field of two coaxial cylindrical surfaces carrying a linear charge density λ of the same magnitude, but of opposite signs (Fig. 1.43). There is no field inside the smaller and outside the larger cylinders. The field strength in the gap between the cylinders is determined by Eq. (1.122). This also holds for cylindrical surfaces of a finite length if the gap between the surfaces is much smaller than their length (a cylindrical capacitor). Appreciable deviations from the field of surfaces of an infinite length will be observed only near the edges of the cylinders.

Field of a Charged Spherical Surface. The field produced by a spherical surface of radius R whose charge has a constant surface density σ will obviously be a centrally symmetrical one. This signifies that the direction of the vector $\boldsymbol{E}$ at any point passes through the centre of the sphere, while the magnitude of the field

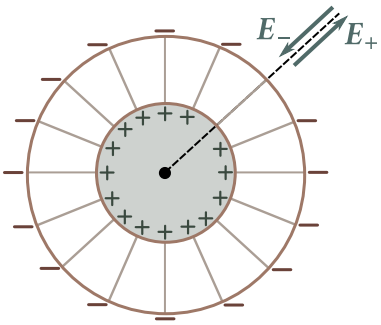


Fig. 1.43

strength is a function of the distance r from the centre of the sphere. Let us imagine a surface of radius r that is concentric with the charged sphere. For all points of this surface, $E_n = E(r)$. If $r > R$, the entire charge q distributed over the sphere will be inside the surface. Hence,

$$E(r) \times 4\pi r^2 = \frac{q}{\varepsilon_0}$$

whence

$$E(r) = \frac{1}{4\pi\varepsilon_0}\frac{q}{r^2}. \quad (r \geqslant R) \tag{1.124}$$

A spherical surface of radius r less than R will contain no charges, owing to which for $r < R$ we get $E(r) = 0$.

Thus, there is no field inside a spherical surface whose charge has a constant surface density σ. Outside this surface, the field is identical with that of a point charge of the same magnitude at the centre of the sphere.

Using the superposition principle, it is easy to show that the field of two concentric spherical surfaces (a spherical capacitor) carrying charges $+q$ and $-q$ that are identical in magnitude and opposite in sign is concentrated in the gap between the surfaces, the magnitude of the field strength in the gap being determined by Eq. (1.124).

Field of a Volume-Charged Sphere. Assume that a sphere of radius R has a charge with a constant volume density ρ. The field in this case has central symmetry. It is easy to see that the same result is obtained for the field outside the sphere [see Eq. (1.124)] as for a sphere with a surface charge. The result will be different for points inside the sphere, however. A spherical surface of radius r ($r < R$) contains a charge equal to $\rho \times 4\pi r^3/3$. Therefore, Gauss's theorem for such a surface will be written as follows:

$$E(r) \times 4\pi r^2 = \frac{1}{\varepsilon_0}\rho\frac{4}{3}\pi r^3.$$

Hence, substituting $q/(4\pi R^3/3)$ for ρ, we get

$$E(r) = \frac{1}{4\pi\varepsilon_0}\frac{q}{R^3}r. \quad (r \leqslant R) \tag{1.125}$$

Thus, the field strength inside a sphere grows linearly with the distance r from the centre of the sphere. Outside the sphere, the field strength diminishes according to the same law as for the field of a point charge.

Chapter 2

ELECTRIC FIELD IN DIELECTRICS

2.1. Polar and Non-Polar Molecules

Dielectrics (or insulators) are defined as substances not capable of conducting an electric current. Ideal insulators do not exist in nature. All substances, even if to a negligible extent, conduct an electric current. But substances called conductors conduct a current from 10^{15} to 10^{20} times better than substances called dielectrics.

If a dielectric is introduced into an electric field, then the field and the dielectric itself undergo appreciable changes. To understand why this happens, we must take into account that atoms and molecules contain positively charged nuclei and negatively charged electrons.

A molecule is a system with a total charge of zero. The linear dimensions of this system are very small, of the order of a few angstroms (the angstrom—Å—is a unit of length equal to 10^{-10} m that is very convenient in atomic physics). We established in Sec. 1.10 that the field set up by such a system is determined by the magnitude and orientation of the dipole electric moment

$$\boldsymbol{p} = \sum_i q_i \boldsymbol{r}_i \tag{2.1}$$

(summation is performed both over the electrons and over the nuclei). True, the electrons in a molecule are in motion, so that this moment constantly changes. The velocities of the electrons are so high, however, that the mean value of the moment (2.1) is detected in practice. For this reason in the following by the dipole moment of a molecule, we shall mean the quantity

$$\boldsymbol{p} = \sum_i q_i \langle \boldsymbol{r}_i \rangle \tag{2.2}$$

(for nuclei, $\boldsymbol{r}_i$ is simply taken as $\langle \boldsymbol{r}_i \rangle$ in this sum). In other words, we shall consider that the electrons are at rest relative to the nuclei at certain points obtained by

averaging the positions of the electrons in time.

The behaviour of a molecule in an external electric field is also determined by its dipole moment. We can verify this by calculating the potential energy of a molecule in an external electric field. Selecting the origin of coordinates inside the molecule and taking advantage of the smallness of $\langle \boldsymbol{r}_i \rangle$, let us write the potential at the point where the i-th charge is in the form

$$\varphi_i = \varphi + \nabla\varphi \cdot \langle \boldsymbol{r}_i \rangle$$

where φ is the potential at the origin of coordinates [see Eq. (1.69)]. Hence,

$$W_\mathrm{p} = \sum_i q_i \varphi_i = \sum_i q_i \left(\varphi + \nabla\varphi \cdot \langle \boldsymbol{r}_i \rangle\right) = \varphi \sum_i q_i + \nabla\varphi \sum_i q_i \langle \boldsymbol{r}_i \rangle .$$

Taking into account that $\sum_i q_i = 0$ and substituting $-\boldsymbol{E}$ for $\nabla\varphi$, we get

$$W_\mathrm{p} = -\boldsymbol{E} \sum_i q_i \langle \boldsymbol{r}_i \rangle = -\boldsymbol{p} \cdot \boldsymbol{E} = -pE \cos\alpha.$$

Differentiating this expression with respect to α, we get Eq. (1.57) for the rotational moment; differentiating with respect to x, we arrive at the force (1.62).

Thus, a molecule is equivalent to a dipole both with respect to the field it sets up and with respect to the forces it experiences in an external field. The positive charge of this dipole equals the total charge of the nuclei and is at the "centre of gravity" of the positive charges; the negative charge equals the total charge of the electrons and is at the "centre of gravity" of the negative charges.

In symmetrical molecules (such as H_2,O_2, N_2), the centres of gravity of the positive and negative charges coincide in the absence of an external electric field. Such molecules have no intrinsic dipole moment and are called **non-polar**. In asymmetrical molecules (such as CO, NH, HCl), the centres of gravity of the charges of opposite signs are displaced relative to each other. In this case, the molecules have an intrinsic dipole moment and are called **polar**.

Under the action of an external electric field, the charges in a non-polar molecule become displaced relative to one another, the positive ones in the direction of the field, the negative ones against the field. As a result, the molecule acquires a dipole moment whose magnitude, as shown by experiments, is proportional to the field strength (intensity). In the rationalized system, the constant of proportionality is written in the form $\varepsilon_0\beta$, where ε_0 is the electric constant, and β is a quantity called the **polarizability of a molecule**. Since the directions of $\boldsymbol{p}$ and $\boldsymbol{E}$ coincide, we can write that

$$\boldsymbol{p} = \beta\varepsilon_0 \boldsymbol{E}. \tag{2.3}$$

The dipole moment has a dimension of $[q]\mathrm{L}$. By Eq. (1.15), the dimension of $\varepsilon_0\boldsymbol{E}$ is $[q]\mathrm{L}^{-2}$. Hence, the polarizability of a molecule β has the dimension L^3.

The process of polarization of a non-polar molecule proceeds as if the positive and negative charges of the molecule were bound to one another by elastic forces. A non-polar molecule is, therefore said, to behave in an external field like an elastic dipole.

The action of an external field on a polar molecule consists mainly in tending to rotate the molecule so that its dipole moment is arranged in the direction of the field. An external field does not virtually affect the magnitude of a dipole moment. Consequently, a polar molecule behaves in an external field like a rigid dipole.

2.2. Polarization of Dielectrics

In the absence of an external electric field, the dipole moments of the molecules of a dielectric usually either equal zero (non-polar molecules) or are distributed in space by directions chaotically (polar molecules). In both cases, the total dipole moment of a dielectric equals zero[1].

A dielectric becomes polarized under the action of an external field. This signifies that the resultant dipole moment of the dielectric becomes other than zero. It is quite natural to take the dipole moment of a unit volume as the quantity characterizing the degree of polarization. If the field or the dielectric (or both) are not homogeneous, the degrees of polarization at different points of the dielectric will differ. To characterize the polarization at a given point, we must separate an infinitely small (physically) volume ΔV containing this point, find the sum $\sum_{\Delta V} \boldsymbol{p}$ of the moments of the molecules confined in this volume, and take the ratio

$$\boldsymbol{P} = \frac{\sum\limits_{\Delta V} \boldsymbol{p}}{\Delta V}. \tag{2.4}$$

The vector quantity $\boldsymbol{P}$ defined by Eq. (2.4) is called the **polarization of a dielectric**.

The dipole moment $\boldsymbol{p}$ has the dimension $[q]\mathrm{L}$. Consequently, the dimension of $\boldsymbol{P}$ is $[q]\mathrm{L}^{-2}$, *i.e.*, it coincides with the dimension of $\varepsilon_0 \boldsymbol{E}$ [see Eq. (1.15)].

The polarization of isotropic dielectrics of any kind is associated with the field strength at the same point by the simple relation

$$\boldsymbol{P} = \chi \varepsilon_0 \boldsymbol{E} \tag{2.5}$$

where χ is a quantity independent of $\boldsymbol{E}$ called the **electric susceptibility of a dielectric**[2]. It was indicated above that the dimensions of $\boldsymbol{P}$ and $\varepsilon_0 \boldsymbol{E}$ are identical. Hence, χ is a dimensionless quantity.

[1] In Sec. 2.9, we shall acquaint ourselves with substances that can have a dipole moment in the absence of an external field.

[2] In anisotropic dielectrics, the directions of $\boldsymbol{P}$ and $\boldsymbol{E}$, generally speaking, do not coincide. In this

In the Gaussian system of units, Eq. (2.5) has the form

$$\boldsymbol{P} = \chi \boldsymbol{E}. \tag{2.6}$$

For dielectrics built of non-polar molecules, Eq. (2.5) issues from the following simple considerations. The volume ΔV contains a number of molecules equal to $n\Delta V$, where n is the number of molecules per unit volume. Each of the moments $\boldsymbol{p}$ is determined in this case by Eq. (2.3). Hence,

$$\sum \Delta V \boldsymbol{p} = n\Delta V \beta \varepsilon_0 \boldsymbol{E}.$$

Dividing this expression by ΔV, we get the polarization $\boldsymbol{P} = n\beta\varepsilon_0 \boldsymbol{E}$. Finally, introducing the symbol $\chi = n\beta$, we arrive at Eq. (2.5).

For dielectrics built of polar molecules, the orienting action of the external field is counteracted by the thermal motion of the molecules tending to scatter their dipole moments in all directions. As a result, a certain preferred orientation of the dipole moments of the molecules sets in in the direction of the field. The relevant statistical calculations, which agree with experimental data, show that the polarization is proportional to the field strength, *i.e.*, leads to Eq. (2.5). The electric susceptibility of such dielectrics varies inversely with the absolute temperature.

In ionic crystals, the separate molecules lose their individuality. An entire crystal is, as it were, a single giant molecule. The lattice of an ionic crystal can be considered as two lattices inserted into each other, one of which is formed by the positive, and the other by the negative ions. When an external field acts on the crystal ions, both lattices are displaced relative to each other, which leads to polarization of the dielectric. The polarization in this case too is associated with the field strength by Eq. (2.5). We must note that the linear relation between $\boldsymbol{E}$ and $\boldsymbol{P}$ described by Eq. (2.5) may be applied only to not too strong fields [a similar remark relates to Eq. (2.3)].

2.3. The Field Inside a Dielectric

The charges in the molecules of a dielectric are called **bound**. The action of a field can only cause bound charges to be displaced slightly from their equilibrium

---

case, the relation between $\boldsymbol{P}$ and $\boldsymbol{E}$ is described by the equations

$$\begin{aligned} P_x &= \varepsilon \left(\chi_{xx}E_x + \chi_{xy}E_y + \chi_{xz}E_z\right), \\ P_y &= \varepsilon \left(\chi_{yx}E_x + \chi_{yy}E_y + \chi_{yz}E_z\right), \\ P_z &= \varepsilon \left(\chi_{zx}E_x + \chi_{zy}E_y + \chi_{zz}E_z\right). \end{aligned}$$

The combination of the nine quantities χ_{ij} forms a symmetrical tensor of rank two called the **tensor of the dielectric susceptibility** [compare with Eqs. (5.30) of Vol. I). This tensor characterizes the electrical properties of an anisotropic dielectric.

positions; they cannot leave the molecule containing them.

Following the example of L. Landau and E. Lifshitz[3], we shall call charges that, although they are within the boundaries of a dielectric, are not inside its molecules, and also charges outside a dielectric, extraneous ones[4].

The field in a dielectric is the superposition of the field $\boldsymbol{E}_{\text{extr}}$ produced by the extraneous charges, and the field $\boldsymbol{E}_{\text{bound}}$ of the bound charges. The resultant field is called **microscopic** (or **true**):

$$\boldsymbol{E}_{\text{micro}} = \boldsymbol{E}_{\text{extr}} + \boldsymbol{E}_{\text{bound}}. \tag{2.7}$$

The microscopic field changes greatly within the limits of the intermolecular distances. Owing to the motion of the bound charges, the field $\boldsymbol{E}_{\text{micro}}$ also changes with time. These changes are not detected in a macroscopic-consideration. Therefore, a field is characterized by the quantity (2.7) averaged over an infinitely small (physically) volume, *i.e.*,

$$\boldsymbol{E} = \langle \boldsymbol{E}_{\text{micro}} \rangle = \langle \boldsymbol{E}_{\text{extr}} \rangle + \langle \boldsymbol{E}_{\text{bound}} \rangle .$$

In the following, we shall designate the averaged field of the extraneous charges by $\boldsymbol{E}_0$, and the averaged field of the bound charges by $\boldsymbol{E}'$. Accordingly, we shall define a macroscopic field as the quantity

$$\boldsymbol{E} = \boldsymbol{E}_0 + \boldsymbol{E}'. \tag{2.8}$$

The polarization $\boldsymbol{P}$ is a macroscopic quantity. Therefore, $\boldsymbol{E}$ in Eq. (2.5) should be understood as the strength determined by Eq. (2.8).

In the absence of dielectrics (*i.e.*, in a "vacuum"), the macroscopic field is

$$\boldsymbol{E} = \boldsymbol{E}_0 = \langle \boldsymbol{E}_{\text{extr}} \rangle .$$

It is exactly this quantity that is understood to be $\boldsymbol{E}$ in Eq. (1.117).

If the extraneous charges are stationary, the field determined by Eq. (2.8) has the same properties as an electrostatic field in a vacuum. In particular, it can be characterized with the aid of the potential φ related to the field strength (2.8) by Eqs. (1.41) and (1.45).

2.4. Space and Surface Bound Charges

When a dielectric is not polarized, the volume density ρ' and the surface density σ' of the bound charges equal zero. Polarization causes the surface density, and in some cases also the volume density of the bound charges to become different from

[3] See L. D. Landau and E. M. Lifshitz. Elektrodinamika sploshnykh sred (Electrodynamics of Continuous Media). Moscow, Gostekhizdat (1957), p. 57.

[4] It is customary practice to call such charges **free**. This name is extremely unsuccessful, however, because in a number of cases extraneous charges are not at all free.

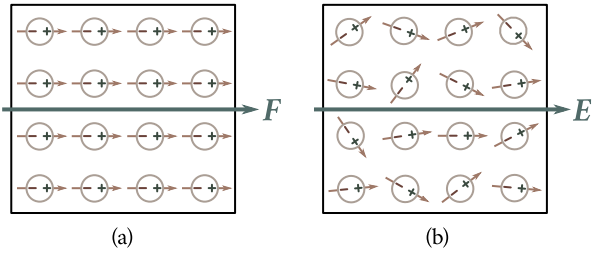


Fig. 2.1

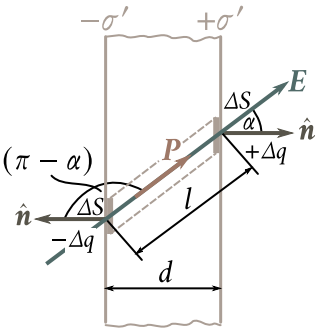


Fig. 2.2

zero.

Figure 2.1 shows schematically a polarized dielectric with nonpolar (a) and polar (b) molecules. Inspection of the figure shows that the polarization is attended by the appearance of a surplus of bound charges of one sign in the thin surface layer of the dielectric. If the normal component of the field strength $\boldsymbol{E}$ for the given section of the surface is other than zero, then under the action of the field, charges of one sign will move away inward, and of the other sign will emerge.

There is a simple relation between the polarization $\boldsymbol{P}$ and the surface density of the bound charges σ'. To find it, let us consider an infinite plane-parallel plate of a homogeneous dielectric placed in a homogeneous electric field (Fig. 2.2). Let us mentally separate an elementary volume in the plate in the form of a very thin cylinder with generatrices parallel to $\boldsymbol{E}$ in the dielectric, and with bases of area ΔS coinciding with the surfaces of the plate. The magnitude of this volume is

$$\Delta V = l \Delta S \cos \alpha$$

where l is the distance between the bases of the cylinder and α the angle between the vector $\boldsymbol{E}$ and an outward normal to the positively charged surface of the dielectric.

The volume ΔV has a dipole electric moment of the magnitude

$$P \Delta V = P l \Delta S \cos \alpha$$

(P is the magnitude of the polarization).

From the macroscopic viewpoint, the volume being considered is equivalent to a dipole formed by the $+\sigma' \Delta S$ and $-\sigma' \Delta S$ with a spacing of l. Therefore, its electric moment can be written in the form $\sigma' \Delta S l$. Equating the two expressions for the electric moment, we get

$$P l \Delta S \cos \alpha = \sigma' \Delta S l.$$

Hence, we get the required relation between σ' and $\boldsymbol{P}$:

$$\sigma' = P \cos \alpha = P_{\mathrm{n}} \tag{2.9}$$

where P_n is the projection of the polarization onto an outward normal to the relevant surface. For the right-hand surface in Fig. 2.2, we have $P_n > 0$, accordingly, σ' for it is positive; for the left-hand surface $P_n < 0$, accordingly, σ' for it is negative.

Expressing $\boldsymbol{P}$ through χ and $\boldsymbol{E}$ by means of Eq. (2.5), we arrive at the formula

$$\sigma' = \chi \varepsilon_0 E_n \tag{2.10}$$

where E_n is the normal component of the field strength inside the dielectric. According to Eq. (2.10), at the places where the field lines emerge from the dielectric ($E_n > 0$), positive bound charges come up to the surface, while where the field lines enter the dielectric ($E_n < 0$), negative surface charges appear.

Equations (2.9) and (2.10) also hold in the most general case when an inhomogeneous dielectric of an arbitrary shape is in an inhomogeneous electric field. By P_n and E_n in this case, we must understand the normal component of the relevant vector taken in direct proximity to the surface element for which σ' is being determined.

Now let us turn to finding the volume density of the bound charges appearing inside an inhomogeneous dielectric. Let us consider an imaginary small area ΔS (Fig. 2.3) in an inhomogeneous isotropic dielectric with non-polar molecules. Assume that a unit volume of the dielectric has n identical particles with a charge of $+e$ and n identical particles with a charge of $-e$. In close proximity to area ΔS, the electric field and the dielectric can be considered homogeneous. Therefore, when the field is switched on, all the positive charges near ΔS will be displaced over the same distance l_1 in the direction of $\boldsymbol{E}$, and all the negative charges will be displaced in the opposite direction over the same distance l_2 (see Fig. 2.3). A certain number of charges of one sign (positive if $\alpha < n/2$ and negative if $\alpha > n/2$) will pass through area ΔS in the direction of a normal to it, and a certain number of charges of the opposite sign (negative if $\alpha < n/2$ and positive if $\alpha < n/2$) in the direction opposite to $\hat{\boldsymbol{n}}$. Area ΔS will be intersected by all the charges $+e$ that were at a distance of not over $l_1 \cos\alpha$ from it before the field was switched on, *i.e.*, by all the $+e$'s in an oblique cylinder of volume $l_1 \Delta S \cos\alpha$. The number of these charges is $nl_1 \Delta S \cos\alpha$, while the charge they carry in the direction of a normal to the area is $enl_1 \Delta S \cos\alpha$ (when $\alpha > \pi/2$, the charge carried in the direction of the normal as a result of displacement of the charges $+e$ will be negative). Similarly, area ΔS will be intersected by all the charges $-e$ in the volume $l_2 \Delta S \cos\alpha$. These charges will carry a charge of $enl_2 \Delta S \cos\alpha$ in the direction of a normal to the area (inspection of Fig. 2.3 shows that when $\alpha < \pi/2$, the charges $-e$ will carry the charge $-enl_2 \Delta S \cos\alpha$ through ΔS in the direction opposite to $\hat{\boldsymbol{n}}$, which is equivalent to carrying the charge $enl_2 \Delta S \cos\alpha$ in the direction of $\hat{\boldsymbol{n}}$).

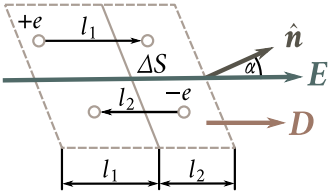


Fig. 2.3

Thus, when the field is switched on, the charge

$$\Delta q' = enl_1 \Delta S \cos\alpha + enl_2 \Delta S \cos\alpha = en(l_1 + l_2)\Delta S \cos\alpha$$

is carried through area ΔS in the direction of a normal to it. The sum $l_1 + l_2$ is the distance l over which the positive and negative bound charges are displaced toward one another in the dielectric. As a result of this displacement, each pair of charges acquires the dipole moment $p = el = e(l_1 + l_2)$. The number of such pairs in a unit volume is n. Consequently, the product $e(l_1 + l_2)n = eln = p$ gives the magnitude of the polarization P. Thus, the charge passing through area ΔS in the direction of a normal to it when the field is switched on is [see Eq. (2.9)]

$$\Delta q' = P\Delta S \cos\alpha.$$

Since the dielectric is isotropic, the directions of the vectors $\boldsymbol{E}$ and $\boldsymbol{P}$ coincide (see Fig. 2.3). Consequently, α is the angle between the vectors $\boldsymbol{P}$ and $\hat{\boldsymbol{n}}$, and in this connection we can write

$$\Delta q' = (\boldsymbol{P} \cdot \hat{\boldsymbol{n}})\, \Delta S.$$

Passing over from deltas to differentials, we get

$$\mathrm{d}q = (\boldsymbol{P} \cdot \hat{\boldsymbol{n}})\, \mathrm{d}S = \boldsymbol{P} \cdot \mathrm{d}\boldsymbol{S}.$$

We have found the bound charge $\mathrm{d}q'$ that passes through elementary area $\mathrm{d}S$ in the direction of a normal to it when the field is switched on; $\boldsymbol{P}$ is the polarization set up under the action of the field at the location of area $\mathrm{d}S$.

Let us imagine closed surface S inside the dielectric. When the field is switched on, a bound charge q' will intersect this surface and emerge from it. This charge is

$$q'_{\mathrm{em}} = \oint_S \mathrm{d}q' = \oint_S \boldsymbol{P} \cdot \mathrm{d}\boldsymbol{S}$$

(we have agreed to take the outward normal to area $\mathrm{d}S$ for closed surfaces). As a result, a surplus bound charge will appear in the volume enclosed by surface S. Its value is

$$q'_{\mathrm{sur}} = -q'_{\mathrm{em}} = -\oint_S \boldsymbol{P} \cdot \mathrm{d}\boldsymbol{S} = -\Phi_{\mathrm{P}} \tag{2.11}$$

(Φ_{P} is the flux of the vector $\boldsymbol{P}$ through surface S).

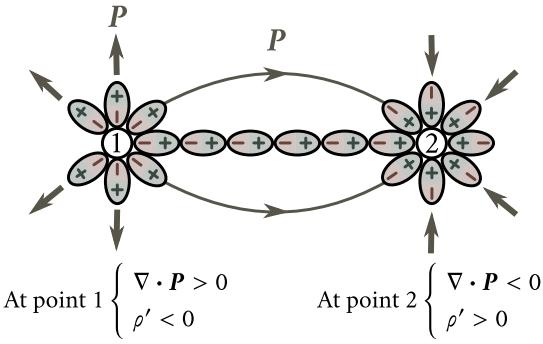


Fig. 2.4

Introducing the volume density of the bound charges ρ', we can write

$$q'_{\text{sur}} = \int_V \rho' \, \mathrm{d}V$$

(the integral is taken over the volume enclosed by surface S). We thus arrive at the formula

$$\int_V \rho' \, \mathrm{d}V = - \oint_S \boldsymbol{P} \cdot \mathrm{d}\boldsymbol{S}.$$

Let us transform the surface integral according to the Ostrogradsky-Gauss theorem [see Eq. (1.108)). The result is

$$\int_V \rho' \, \mathrm{d}V = - \int_V \nabla \cdot \boldsymbol{P} \, \mathrm{d}V.$$

This equation must be observed for any arbitrarily chosen volume V. This is possible only if the following equation is observed at every point of the dielectric:

$$\rho' = -\nabla \cdot \boldsymbol{P}. \tag{2.12}$$

Consequently, the density of bound charges equals the divergence of the polarization $\boldsymbol{P}$ taken with the opposite sign.

We obtained Eq. (2.12) when considering a dielectric with non-polar molecules. This equation also holds, however, for dielectrics with polar molecules.

Equation (2.12) can be given a graphical interpretation. Points with a positive $\nabla \cdot \boldsymbol{P}$ are sources of the field of the vector $\boldsymbol{P}$, and the lines of $\boldsymbol{P}$ diverge from them (Fig. 2.4). Points with a negative $\nabla \cdot \boldsymbol{P}$ are sinks of the field of the vector $\boldsymbol{P}$, and the lines of $\boldsymbol{P}$ converge at them. In polarization of the dielectric, the positive bound charges are displaced in the direction of the vector $\boldsymbol{P}$, *i.e.*, in the direction of the lines $\boldsymbol{P}$; the negative bound charges are displaced in the opposite direction (in the figure the bound charges belonging to separate molecules are encircled by ovals). As a result, a surplus of negative bound charges is formed at places with a positive

$\nabla \cdot \boldsymbol{P}$, and a surplus of positive bound charges at places with a negative $\nabla \cdot \boldsymbol{P}$.

Bound charges differ from extraneous ones only in that they cannot leave the confines of the molecules which they are in. Otherwise, they have the same properties as all other charges. In particular, they are sources of an electric field. Therefore, when the density of the bound charges ρ' differs from zero, Eq. (1.117) must be written in the form

$$\nabla \cdot \boldsymbol{E} = \frac{1}{\varepsilon_0}(\rho + \rho'). \tag{2.13}$$

Here ρ is the density of the extraneous charges.

Let us introduce Eq. (2.5) for $\boldsymbol{P}$ into Eq. (2.12) and use Eq. (1.103). The result is

$$\rho' = -\nabla \cdot (\chi \varepsilon_0 \boldsymbol{E}) = -\varepsilon_0 \nabla \cdot (\chi \boldsymbol{E}) = -\varepsilon_0 [\boldsymbol{E} \cdot \nabla \chi + \chi \nabla \cdot \boldsymbol{E}].$$

Substituting for $\nabla \cdot \boldsymbol{E}$ its value from Eq. (2.13), we arrive at the equation

$$\rho' = -\varepsilon_0 (\boldsymbol{E} \cdot \nabla \chi) - \chi \rho - \chi \rho'.$$

Hence,

$$\rho' = -\left(\frac{1}{1+\chi}\right)[\varepsilon_0 (\boldsymbol{E} \cdot \nabla \chi) + \chi \rho]. \tag{2.14}$$

We can see from Eq. (2.14) that the volume density of bound charges can differ from zero in two cases: (1) if a dielectric is not homogeneous ($\nabla \chi \neq 0$), and (2) if at a given place in a dielectric the density of the extraneous charges is other than zero ($\rho \neq 0$).

When there are no extraneous charges in a dielectric, the volume density of the bound charges is

$$\rho' = -\left(\frac{\varepsilon_0}{1+\chi}\right)(\boldsymbol{E} \cdot \nabla \chi). \tag{2.15}$$

2.5. Electric Displacement Vector

We noted in the preceding section that not only extraneous, but also bound charges are sources of a field. Accordingly,

$$\nabla \cdot \boldsymbol{E} = \frac{1}{\varepsilon_0}(\rho + \rho') \tag{2.16}$$

[see Eq. (2.13)].

Equation (2.16) is of virtually no use for finding the vector $\boldsymbol{E}$ because it expresses the properties of the unknown quantity $\boldsymbol{E}$ through bound charges, which in turn are determined by the unknown quantity $\boldsymbol{E}$ [see Eqs. (2.10) and (2.14)].

Calculation of the fields is often simplified if we introduce an auxiliary quantity whose sources are only extraneous charges ρ. To establish what this quantity looks

like, let us introduce Eq. (2.12) for ρ' into Eq. (2.16):

$$\nabla \cdot \boldsymbol{E} = \frac{1}{\varepsilon_0}(\rho - \nabla \cdot \boldsymbol{P})$$

whence it follows that

$$\nabla \cdot (\varepsilon_0 \boldsymbol{E} + \boldsymbol{P}) = \rho \tag{2.17}$$

(we have put ε_0 inside the del symbol). The expression in parentheses in Eq. (2.17) is the required quantity. It is designated by the symbol $\boldsymbol{D}$ and is called the electric displacement (or electric induction).

Thus, the **electric displacement** is a quantity determined by the relation

$$\boldsymbol{D} = \varepsilon_0 \boldsymbol{E} + \boldsymbol{P}. \tag{2.18}$$

Inserting Eq. (2.5) for $\boldsymbol{P}$, we get

$$\boldsymbol{D} = \varepsilon_0 \boldsymbol{E} + \chi \varepsilon_0 \boldsymbol{E} = \varepsilon_0 (1 + \chi) \boldsymbol{E}. \tag{2.19}$$

The dimensionless quantity

$$\varepsilon = 1 + \chi \tag{2.20}$$

is called the **relative permittivity** or simply the **permittivity** of a medium[5]. Thus, Eq. (2.19) can be written in the form

$$\boldsymbol{D} = \varepsilon_0 \varepsilon \boldsymbol{E}. \tag{2.21}$$

According to Eq. (2.21), the vector $\boldsymbol{D}$ is proportional to the vector $\boldsymbol{E}$. We remind our reader that we are dealing with isotropic dielectrics. In anisotropic dielectrics, the vectors $\boldsymbol{E}$ and $\boldsymbol{D}$, generally speaking, are not collinear.

In accordance with Eqs. Eq. (1.15) and Eq. (2.21), the electric displacement of the field of a point charge in a vacuum is

$$\boldsymbol{D} = \frac{1}{4\pi} \frac{q}{r^2} \hat{\boldsymbol{e}}_r. \tag{2.22}$$

The unit of electric displacement is the coulomb per square metre ($\mathrm{C\,m^{-2}}$).

Equation (2.17) can be written as

$$\nabla \cdot \boldsymbol{D} = \rho. \tag{2.23}$$

Integration of this equation over the arbitrary volume V yields

$$\int_V \nabla \cdot \boldsymbol{D} \, \mathrm{d}V = \int_V \rho \, \mathrm{d}V.$$

Let us transform the left-hand side according to the Ostrogradsky-Gauss theorem [see Eq. (1.108)]:

$$\oint_S \boldsymbol{D} \cdot \mathrm{d}\boldsymbol{S} = \int_V \rho \, \mathrm{d}V. \tag{2.24}$$

[5]The so-called absolute permittivity of a medium $\varepsilon_\mathrm{a} = \varepsilon_0 \varepsilon$ is introduced in electrical engineering. This quantity is deprived of a physical meaning, however, and we shall not use it.

The quantity on the left-hand side is Φ_D—the flux of the vector $\boldsymbol{D}$ through closed surface S, while that on the right-hand side is the sum of the extraneous charges $\sum_i q_i$ enclosed by this surface. Hence, Eq. (2.24) can be written in the form

$$\Phi_D = \sum_i q_i. \tag{2.25}$$

Equations (2.24) and (2.25) express Gauss's theorem for the vector $\boldsymbol{D}$: *the flux of the electric displacement through a closed surface equals the algebraic sum of the extraneous charges enclosed by this surface.*

In a vacuum, $\boldsymbol{P} = 0$, so that the quantity $\boldsymbol{D}$ determined by Eq. (2.18) transforms into $\varepsilon_0 \boldsymbol{E}$, and Eqs. (2.24) and (2.25) transform into Eqs. (1.114) and (1.116).

The unit of the flux of the electric displacement vector is the coulomb. By Eq. (2.25), a charge of 1 C sets up a displacement flux of 1 C through the surface surrounding it.

The field of the vector $\boldsymbol{D}$ can be depicted with the aid of electric displacement lines (we shall call them displacement lines for brevity's sake). Their direction and density are determined in exactly the same way as for the lines of the vector $\boldsymbol{E}$ (see Sec. 1.5). The lines of the vector $\boldsymbol{E}$ can begin and terminate at both extraneous and bound charges. The sources of the field of the vector $\boldsymbol{D}$ are only extraneous charges. Hence, displacement lines can begin or terminate only at extraneous charges. These lines pass without interruption through points at which bound charges are placed.

The electric induction[6] in the Gaussian system is determined by the expression

$$\boldsymbol{D} = \boldsymbol{E} + 4\pi \boldsymbol{P}. \tag{2.26}$$

Substituting for $\boldsymbol{P}$ in this equation its value from Eq. (2.6), we get

$$\boldsymbol{D} = (1 + 4\pi\chi)\boldsymbol{E}. \tag{2.27}$$

The quantity

$$\varepsilon = 1 + 4\pi\chi. \tag{2.28}$$

is called the **permittivity**. Introducing this quantity into Eq. (2.27) we get

$$\boldsymbol{D} = \varepsilon \boldsymbol{E}. \tag{2.29}$$

In the Gaussian system, the electric induction in a vacuum coincides with the field strength $\boldsymbol{E}$. Consequently, the electric induction of the field of a point charge in a vacuum is determined by Eq. (1.16).

By Eq. (2.22) the electric displacement set up by a charge of 1 C at a distance of 1 m is

$$D = \frac{1}{4\pi}\frac{q}{r^2} = \frac{1}{4\pi \times 1^2} = \frac{1}{4\pi}\,\mathrm{C\,m^{-2}}.$$

[6]The term "electric displacement" is not applied to quantity (2.27).

In the Gaussian system, the electric induction in this case is

$$D = \frac{q}{r^2} = \frac{3 \times 10^9}{10^4} = 3 \times 10^5 \ \text{cgse}_D.$$

Thus,

$$1\ \text{C m}^{-2} = 4\pi \times 3 \times 10^5 \ \text{cgse}_D.$$

In the Gaussian system, the expressions of Gauss's theorem have the form

$$\oint_S \boldsymbol{D} \cdot \mathrm{d}\boldsymbol{S} = 4\pi \int_V \rho \, \mathrm{d}V \tag{2.30}$$

$$\Phi_D = 4\pi \sum_i q_i. \tag{2.31}$$

According to Eq. (2.31), a charge of 1 C sets up a flux of the electric induction vector of $4\pi q = 4\pi \times 3 \times 10^9 \ \text{cgse}_{\Phi_D}$. The following relation thus exists between the units of flux of the vector $\boldsymbol{D}$:

$$1\ \text{C} = 4\pi \times 3 \times 10^9 \ \text{cgse}_{\Phi_D}.$$

2.6. Examples of Calculating the Field in Dielectrics

We shall consider several examples of fields in dielectrics to reveal the meaning of the quantities $\boldsymbol{D}$ and ε.

Field Inside a Flat Plate. Let us consider two infinite parallel oppositely charged planes. Let the field they produce in a vacuum be characterized by the strength $\boldsymbol{E}_0$ and the displacement $\boldsymbol{D}_0 = \varepsilon_0 \boldsymbol{E}_0$. Let us introduce into this field a plate of a homogeneous isotropic dielectric and arrange it as shown in Fig. 2.5. The dielectric becomes polarized under the action of the field, and bound charges of density σ' will appear on its surfaces. These charges will set up a homogeneous field inside the plate whose strength by Eq. (1.121) is $E' = \sigma'/\varepsilon_0$. In the given case, $E' = 0$ outside the dielectric.

The field strength E_0 is σ/ε_0. Both fields are directed toward each other, hence, inside the dielectric we have

$$E = E_0 - E' = E_0 - \frac{\sigma'}{\varepsilon_0} = \frac{1}{\varepsilon_0} ((\,) \sigma - \sigma'). \tag{2.32}$$

Outside the dielectric, $E = E_0$.

The polarization of the dielectric is due to field (2.32). The latter is perpendicular to the surfaces of the plate. Hence, $E_\text{n} = E$, and in accordance with Eq. (2.10), $\sigma' = \chi \varepsilon_0 E$. Using this value in Eq. (2.32), we get

$$E = E_0 - \chi E$$

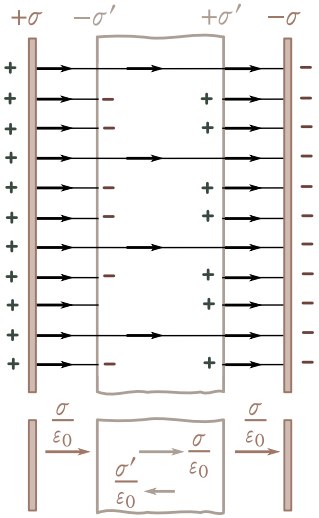


Fig. 2.5

whence

$$E = \frac{E_0}{1+\chi} = \frac{E_0}{\varepsilon}. \tag{2.33}$$

Thus, in the given case, the permittivity ε shows how many times the field in a dielectric weakens.

Multiplying Eq. (2.33) by $\varepsilon_0\varepsilon$, we get the electric displacement inside the plate:

$$D = \varepsilon_0\varepsilon E = \varepsilon_0 E_0 D_0. \tag{2.34}$$

Hence, the electric displacement inside the plate coincides with that of the external field D_0. Substituting σ/ε_0 for E_0 in Eq. (2.34), we find

$$D = \sigma. \tag{2.35}$$

To find σ', let us express E and E_0 in Eq. (2.33) through the charge densities:

$$\frac{1}{\varepsilon_0}\left(\left(\right)\sigma - \sigma'\right) = \frac{\sigma}{\varepsilon_0\varepsilon}$$

whence

$$\sigma' = \frac{\varepsilon - 1}{\varepsilon}\sigma. \tag{2.36}$$

Figure 2.5 has been drawn assuming that $\varepsilon = 3$. Accordingly, the density of the field lines in the dielectric is one-third of that outside the plate. The lines are equally spaced because the field is homogeneous. In the given case, σ' can be found without resorting to Eq. (2.36). Indeed, since the field intensity inside the plate is one-third of that outside it, then of three field lines beginning (or terminating) on

extraneous charges, two must terminate (or begin respectively) on bound charges. It thus follows that the density of the bound charges must be two-thirds that of the extraneous charges.

In the Gaussian system, the field strength E' produced by the bound charges σ' is $4\pi\sigma'$. Therefore, Eq. (2.32) becomes

$$E = E_0 - E' = E_0 - 4\pi\sigma'. \tag{2.37}$$

The surface density σ' is associated with the field strength E by the equation $\sigma' = \chi E_{\mathrm{n}}$. We can thus write that

$$E = E_0 - 4\pi\chi E$$

whence

$$E = \frac{E_0}{1 + 4\pi\chi} = \frac{E_0}{\varepsilon}.$$

Thus, the permittivity ε, like its counterpart ε in the SI, shows how many times the field inside a dielectric weakens. Therefore, the values of ε in the SI and the Gaussian system coincide. Hence, taking into account Eqs. (2.20) and (2.28), we conclude that the susceptibilities in the Gaussian system (χ_{Gs}) and in the SI (χ_{SI}) differ from each other by the factor 4π:

$$\chi_{\mathrm{SI}} = 4\pi\chi_{\mathrm{Gs}}. \tag{2.38}$$

Field Inside a Spherical Layer. Let us surround a charged sphere of radius R with a concentric spherical layer of a homogeneous isotropic dielectric (Fig. 2.6). The bound charge q_1' distributed with the density σ_1' will appear on the internal surface of the layer ($q_1' = 4\pi R_1^2\sigma_1'$), and the charge q_2' distributed with the density σ_2' will appear on its external surface ($q_2' = 4\pi R_2^2\sigma_2'$). The sign of the charge q_2' coincides with that of the charge q of the sphere, while q_1' has the opposite sign. The charges q_1' and q_2' set up a field at a distance r exceeding R_1 and R_2, respectively, that coincides with the field of a point charge of the same magnitude [see Eq. (1.124)]. The charges q_1' and q_2' produce no field inside the surfaces over which they are distributed. Hence, the field strength E' inside a dielectric is

$$E' = \frac{1}{4\pi\varepsilon_0}\frac{q_1'}{r^2} = \frac{1}{4\pi\varepsilon_0}\frac{4\pi R_1^2\sigma_1'}{r^2} = \frac{1}{\varepsilon_0}\frac{R_1^2\sigma_1'}{r^2}$$

and is opposite in direction to the field strength E_0. The resultant field in a dielectric is

$$E(r) = E_0 - E' = \frac{1}{4\pi\varepsilon_0}\frac{q}{r^2} - \frac{1}{\varepsilon_0}\frac{R_1^2\sigma_1'}{r^2}. \tag{2.39}$$

It diminishes in proportion to $1/r^2$. We can, therefore, state that

$$\frac{E(R_1)}{E(r)} = \frac{r^2}{R_1^2} \quad \Rightarrow \quad E(R_1) = E(r)\frac{r^2}{R_1^2},$$

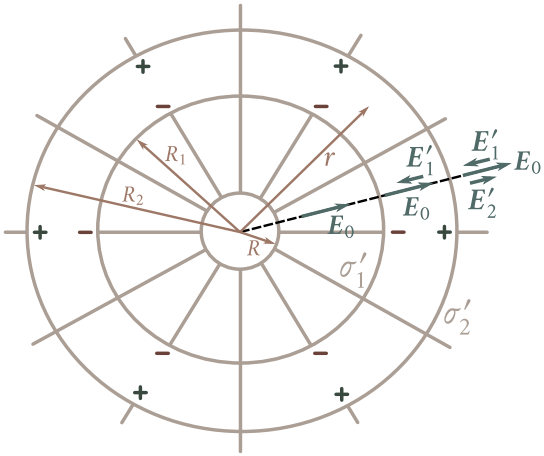


Fig. 2.6

where $E(R_1)$ is the field strength in a dielectric in direct proximity to the internal surface of the layer. It is exactly this strength that determines the quantity σ_1':

$$\sigma_1' = \chi\varepsilon_0 E(R_1) = \chi\varepsilon_0 E(r)\frac{r^2}{R_1^2} \tag{2.40}$$

(at each point of the surface $|E_n| = E$).

Introducing Eq. (2.40) into Eq. (2.39), we get

$$E(r) = \frac{1}{4\pi\varepsilon_0}\frac{q}{r^2} - \frac{1}{\varepsilon_0}\frac{R_1^2\chi\varepsilon_0 E(r)r^2}{r^2R_1^2} = E_0(r) - \chi E(r).$$

From this equation, we find that inside a dielectric $E = E_0/\varepsilon$, and, consequently, $D = \varepsilon_0 E_0$ [compare with Eqs. (2.33) and (2.34)].

The field inside a dielectric changes in proportion to $1/r^2$. Therefore, the relation $\sigma_1' : \sigma_2' = R_1 : R_2$ holds. Hence, it follows that $q_1' = q_2'$. Consequently, the fields set up by these charges at distances exceeding R_2 mutually destroy each other so that outside the spherical layer $E' = 0$ and $E = E_0$.

Assuming that $R_1 = R$ and $R_2 = \infty$, we arrive at the case of a charged sphere immersed in an infinite homogeneous and isotropic dielectric. The field strength outside such a sphere is

$$E = \frac{1}{4\pi\varepsilon_0}\frac{q}{\varepsilon r^2}. \tag{2.41}$$

The strength of the field set up in an infinite dielectric by a point charge will be the same.

Both examples considered above are characterized by the fact that the dielectric was homogeneous and isotropic, and the surfaces enclosing it coincided with the

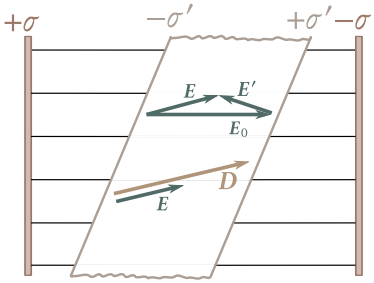


Fig. 2.7

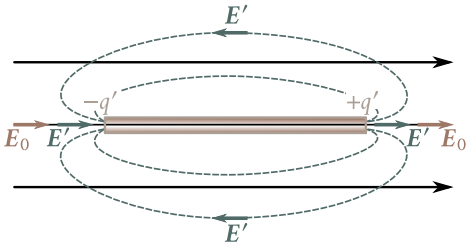


Fig. 2.8

equipotential surfaces of the field of extraneous charges. The result we have obtained in these cases is a general one. *If a homogeneous and isotropic dielectric completely fills the volume enclosed by equipotential surfaces of the field of extraneous charges, then the electric displacement vector coincides with the vector of the field strength of the extraneous charges multiplied by ε_0, and, therefore, the field strength inside the dielectric is $1/\varepsilon$ of that of the field strength of the extraneous charges.*

If the above conditions are not observed, the vectors $\boldsymbol{D}$ and $\varepsilon_0\boldsymbol{E}$ do not coincide. Figure 2.7 shows the field in the plate of a dielectric. The plate is skewed relative to the planes carrying extraneous charges. The vector $\boldsymbol{E}'$ is perpendicular to the faces of the plate, therefore, $\boldsymbol{E}$ and $\boldsymbol{E}_0$ are not collinear. The vector $\boldsymbol{D}$ is directed the same as $\boldsymbol{E}$, consequently, $\boldsymbol{D}$ and $\varepsilon_0\boldsymbol{E}_0$ do not coincide in direction. We can show that they also fail to coincide in magnitude. In the examples considered above owing to the specially selected shape of the dielectric, the field $\boldsymbol{E}'$ differed from zero only inside the dielectric. In the general case, $\boldsymbol{E}'$ may differ from zero outside the dielectric too. Let us place a rod made of a dielectric into an initially homogeneous field (Fig. 2.8). Owing to polarization, bound charges of opposite signs are formed on the ends of the rod. Their field outside the rod is equivalent to the field of a dipole (the lines of $\boldsymbol{E}'$ are dash ones in the figure). It is easy to see that the resultant field $\boldsymbol{E}$ near the ends of the rod is greater than the field $\boldsymbol{E}_0$.

2.7. Conditions on the Interface Between Two Dielectrics

Near the interface between two dielectrics, the vectors $\boldsymbol{E}$ and $\boldsymbol{D}$ must comply with definite boundary conditions following from the relations (1.112) and (2.23):

$$\nabla \times \boldsymbol{E} = 0, \quad \nabla \cdot \boldsymbol{D} = \rho.$$

Let us consider the interface between two dielectrics with the permittivities ε_1 and ε_2 (Fig. 2.9). We choose an arbitrarily directed x-axis on this surface. We take a small rectangular contour of length a and width b that is partly in the first dielectric

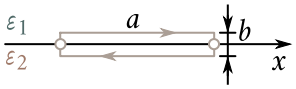


Fig. 2.9

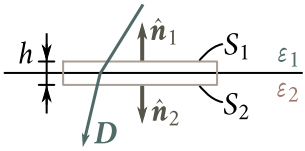


Fig. 2.10

and partly in the second one. The x-axis passes through the middle of the sides b.

Assume that a field has been set up in the first dielectric whose strength is $\boldsymbol{E}_1$, and in the second one whose strength is $\boldsymbol{E}_2$. Since $\nabla \times \boldsymbol{E} = 0$, the circulation of the vector $\boldsymbol{E}$ around the contour we have chosen must equal zero [see Eq. (1.110)]. With small dimensions of the contour and the direction of circumvention shown in Fig. 2.9, the circulation of the vector $\boldsymbol{E}$ can be written in the form

$$\oint E_l \, \mathrm{d}l = E_{1,x} a - E_{2,x} a + \langle E_b \rangle \, 2b \tag{2.42}$$

where $\langle E_b \rangle$ is the mean value of E_l on sections of the contour perpendicular to the interface. Equating this expression to zero, we arrive at the equation

$$\left(E_{1,x} - E_{2,x}\right) a = \langle E_b \rangle \, 2b.$$

In the limit, when the width b of the contour tends to zero, we get

$$E_{1,x} = E_{2,x}. \tag{2.43}$$

The values of the projections of the vectors $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$ onto the x-axis are taken in direct proximity to the interface between the boundary of the dielectrics.

Equation (2.43) is obeyed when the x-axis is selected arbitrarily. It is only essential that this axis be in the plane of the interface between the dielectrics. Inspection of Eq. (2.43) shows that with such a selection of the x-axis when $E_{1,x} = 0$, the projection of $E_{2,x} = 0$ will also equal zero. This signifies that the vectors $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$ at two close points taken at opposite sides of the interface are in the same plane as a normal to the interface. Let us represent each of the vectors $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$ in the form of the sum of the normal and tangential components:

$$\boldsymbol{E}_1 = \boldsymbol{E}_{1,n} + \boldsymbol{E}_{1,\tau}, \quad \boldsymbol{E}_2 = \boldsymbol{E}_{2,n} + \boldsymbol{E}_{2,\tau}.$$

In accordance with Eq. (2.43)

$$E_{1,\tau} = E_{2,\tau}. \tag{2.44}$$

Here $\boldsymbol{E}_{i,\tau}$ is the projection of the vector $\boldsymbol{E}_i$ onto the unit vector $\hat{\boldsymbol{\tau}}$ directed along the line of intersection of the dielectric interface with the plane containing the vectors $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$.

Substituting in accordance with Eq. (2.21) the projections of the vector $\boldsymbol{D}$ divided

by $\varepsilon_0\varepsilon$ for the projections of the vector $\boldsymbol{E}$, we get the proportion

$$\frac{D_{1,\tau}}{\varepsilon_0\varepsilon_1} = \frac{D_{2,\tau}}{\varepsilon_0\varepsilon_2}$$

whence it follows that

$$\frac{D_{1,\tau}}{D_{2,\tau}} = \frac{\varepsilon_1}{\varepsilon_2}. \tag{2.45}$$

Now let us take an imaginary cylindrical surface of height h on the interface between the dielectrics (Fig. 2.10). Base S_1 is in the first dielectric, and base S_2 in the second. Both bases are identical in size ($S_1 = S_2 = S$) and are so small that within the limits of each of them the field may be considered homogeneous. Let us apply Gauss's theorem [see Eq. (2.25)] to this surface. If there are no extraneous charges on the interface between the dielectrics, the right-hand side in Eq. (2.25) equals zero. Hence, $\Phi_D = 0$.

The flux through base S_1 is $D_{1,n}S$, where $D_{1,n}$ is the projection of the vector $\boldsymbol{D}$ in the first dielectric onto the normal $\hat{\boldsymbol{n}}_1$. Similarly, the flux through base S_2 is $D_{2,n}S$, where $D_{2,n}$ is the projection of the vector $\boldsymbol{D}$ in the second dielectric onto the normal $\hat{\boldsymbol{n}}_2$. The flux through the side surface can be written in the form $\langle D\rangle_n S_{\text{side}}$, where $\langle D\rangle_n$ is the value of D_n averaged over the entire side surface, and S_{side} is the magnitude of this surface. We can thus write that

$$\Phi_D = D_{1,n}S + D_{2,n}S + \langle D\rangle_n S_{\text{side}} = 0. \tag{2.46}$$

If the altitude h of the cylinder is made to tend to zero, then S_{side} will also tend to zero. Hence, in the limit, we get

$$D_{1,n} = -D_{2,n}.$$

Here $D_{i,n}$ is the projection onto $\hat{\boldsymbol{n}}_i$ of the vector $\boldsymbol{D}$ the i-th dielectric in direct proximity to its interface with the other dielectric. The signs of the projections are different because the normals $\hat{\boldsymbol{n}}_1$ and $\hat{\boldsymbol{n}}_2$ to the bases of the cylinder have opposite directions. If we project $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ onto the same normal, we get the condition

$$D_{1,n} = D_{2,n}. \tag{2.47}$$

Using Eq. (2.21) to replace the projections of $\boldsymbol{D}$ with the corresponding projections of the vector $\boldsymbol{E}$ multiplied by $\varepsilon_0\varepsilon$, we get the relation

$$\varepsilon_0\varepsilon_1 E_{1,n} = \varepsilon_0\varepsilon_2 E_{2,n}$$

whence

$$\frac{E_{1,n}}{E_{2,n}} = \frac{\varepsilon_2}{\varepsilon_1}. \tag{2.48}$$

The results we have obtained signify that when passing through the interface between two dielectrics, the normal component of the vector $\boldsymbol{D}$ and the tangential component of the vector $\boldsymbol{E}$ change continuously. The tangential component of the

vector $\boldsymbol{D}$ and the normal component of the vector $\boldsymbol{E}$, however, are disrupted when passing through the interface.

Equations (2.44), (2.45), (2.47), and (2.48) determine the conditions which the vectors $\boldsymbol{E}$ and $\boldsymbol{D}$ must comply with on the interface between two dielectrics (if there are no extraneous charges on this interface). We have obtained these equations for an electrostatic field. They also hold, however, for fields varying with time (see Sec. 16.3).

The conditions we have found also hold for the interface between a dielectric and a vacuum. In this case, one of the permittivities must be taken equal to unity.

We must note that condition (2.47) can be obtained on the basis of the fact that the displacement lines pass through the interface between two dielectrics without being interrupted (Fig. 2.11). According to the rule for drawing these lines, the number of lines arriving at area ΔS from the first dielectric is $D_1 \Delta S_1 = D_1 \Delta S \cos \alpha_1$. Similarly, the number of lines emerging from area ΔS into the second dielectric is $D_2 \Delta S_2 = D_2 \Delta S \cos \alpha_2$. If the lines are not interrupted at the interface, both these numbers must be the same:

$$D_1 \Delta S \cos \alpha_1 = D_2 \Delta S \cos \alpha_2.$$

Cancelling ΔS and taking into account that the product $D \cos \alpha$ gives the value of the normal component of the vector $\boldsymbol{D}$, we arrive at condition (2.47).

The displacement lines are bent (refracted) on the interface between dielectrics, owing to which the angle α between a normal to the interface and the line $\boldsymbol{D}$ changes. Inspection of Fig. 2.12 shows that

$$\tan \alpha_1 : \tan \alpha_2 = \frac{D_{1,\tau}}{D_{1,n}} : \frac{D_{2,\tau}}{D_{2,n}}$$

whence with account taken of Eqs. (2.45) and (2.47), we get the law of displacement line refraction:

$$\frac{\tan \alpha_1}{\tan \alpha_2} = \frac{\varepsilon_1}{\varepsilon_2}. \tag{2.49}$$

When displacement lines pass into a dielectric with a lower permittivity ε, the angle made by them with a normal diminishes, hence, the lines are spaced farther apart; when the lines pass into a dielectric with a higher permittivity ε, on the contrary, they become closer together.

2.8. Forces Acting on a Charge in a Dielectric

If we introduce into an electric field in a vacuum a charged body of such small dimensions that the external field within the body can be considered homogeneous,

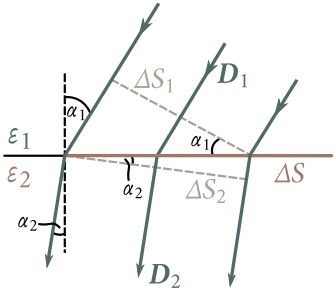


Fig. 2.11

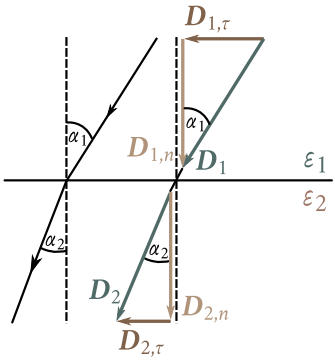


Fig. 2.12

then the body will experience the force

$$\boldsymbol{F} = q\boldsymbol{E}. \tag{2.50}$$

To place a charged body in a field set up in a dielectric, a cavity must be made in the latter. In a fluid dielectric, the body itself forms the cavity by displacing the dielectric from the volume it occupies. The field inside the cavity $\boldsymbol{E}_{\text{cav}}$ will differ from that in a continuous dielectric. Thus, we cannot calculate the force exerted on a charged body placed in a cavity as the product of the charge q and the field strength $\boldsymbol{E}$ in the dielectric before the body was introduced into it.

When calculating the force acting on a charged body in a fluid dielectric, we must take another circumstance into account. Mechanical tension is set up on the boundary with the body in the dielectric. This sets up an additional mechanical force $\boldsymbol{F}_{\text{ten}}$ acting on the body.

Thus, the force acting on a charged body in a dielectric, generally speaking, cannot be determined by Eq. (2.50), and it is usually a very complicated task to calculate it. These calculations give an interesting result for a fluid dielectric. The resultant of the electric force $q\boldsymbol{E}_{\text{cav}}$ and the mechanical force $\boldsymbol{F}_{\text{ten}}$ is found to be exactly equal to $q\boldsymbol{E}$, where $\boldsymbol{E}$ is the field strength in the continuous dielectric

$$\boldsymbol{F} = q\boldsymbol{E}_{\text{cav}} + \boldsymbol{F}_{\text{ten}} = q\boldsymbol{E}. \tag{2.51}$$

The strength of the field produced in a homogeneous infinitely extending dielectric by a point charge is determined by Eq. (2.49). Hence, we get the following expression for the forces of interaction of two point charges immersed in a homogeneous infinitely extending dielectric:

$$F = \frac{1}{4\pi\varepsilon_0}\frac{q_1 q_2}{\varepsilon r^2}. \tag{2.52}$$

This formula expresses Coulomb's law for charges in a dielectric. It holds only for fluid dielectrics.

Some authors characterize Eq. (2.52) as "the most general expression of Coulomb's law". In this connection, we shall cite Richard P. Feynman: "Many older books on electricity start with the 'fundamental' law that the force between two charges is…[Eq. (2.52) is given]…, a point of view which is thoroughly unsatisfactory. For one thing, it is not true in general; it is true only for a world filled with a liquid. Secondly, it depends on the fact that ε is a constant which is only approximately true for most real materials"[7].

We shall not treat questions relating to the forces acting on a charge inside a cavity made in a solid dielectric.

2.9. Ferroelectrics

There is a group of substances that can have the property of spontaneous polarization in the absence of an external field. They are called **ferroelectrics**. This phenomenon was first discovered for Rochelle salt, and the first detailed investigation of the electrical properties of this salt was carried out by the Soviet physicists I. Kurchatov and P. Kobeko.

Ferroelectrics differ from the other dielectrics in a number of features:

1. Whereas the permittivity ε of ordinary dielectrics is only several units, reaching as an exception several scores (for example, for water $\varepsilon = 81$), the permittivity of ferroelectrics may be of the order of several thousands.

2. The dependence of P on E is not linear (see branch 1 of the curve shown in Fig. 2.13). Hence, the permittivity depends on the field strength.

3. When the field changes, the values of the polarization P (and, therefore, of the displacement D too) lag behind the field strength E. As a result, P and D are determined not only by the value of E at the given moment, but also by the preceding values of E, *i.e.*, they depend on the preceding history of the dielectric. This phenomenon is called **hysteresis** (from the Greek word "husterein"—to come late, be behind). Upon cyclic changes of the field, the dependence of P on E follows the curve shown in Fig. 2.13 and called a **hysteresis loop**. When the field is initially switched on, the polarization grows with E according to branch 1 of the curve. Diminishing of P takes place along branch 2. When E vanishes, the substance retains a value of the polarization P_r called the **residual polarization**. The polarization vanishes only under the action of an oppositely directed field E_c. This value of the field strength is called the **coercive force**. Upon a further change in E, branch 3 of the hysteresis loop is obtained, and so on.

[7] R. P. Feynman, R. B. Leighton, M. Sands. The Feynman Lectures on Physics. Vol. II. Reading, Mass., Addison-Wesley (1965), p. 10-8.

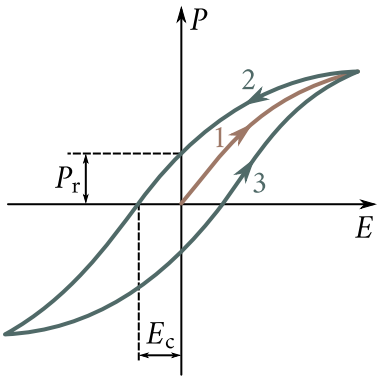


Fig. 2.13

The behaviour of the polarization of ferroelectrics is similar to that of the magnetization of ferromagnetics (see Sec. 7.9), and this is the origin of their name.

Only crystalline substances having no centre of symmetry can be ferroelectrics. For example, the crystals of Rochelle salt belong to the rhombic system (see Sec. 13.2 of Vol. I). The interaction of the particles in a ferroelectric crystal leads to the fact that their dipole moments line up spontaneously parallel to one another. In exclusive cases, the identical orientation of the dipole moments extends to the entire crystal. Ordinarily, however, regions appear in a crystal in whose confines the dipole moments are parallel to one another, but the directions of polarization in different regions are different. Thus, the resultant moment of an entire crystal may equal zero. The regions of spontaneous polarization are also called **domains**. Under the action of an external field, the moments of the domains rotate as a single whole, arranging themselves in the direction of the field.

Every ferroelectric has a temperature at which the substance loses its unusual properties and becomes a normal dielectric. This temperature is called the **Curie point**. Rochelle salt has two Curie points, namely, −15 °C and 22 °C, and it behaves like a ferroelectric only in the interval between these two temperatures. Its electrical properties are conventional at temperatures below −15 °C and above 22 °C.

Chapter 3

CONDUCTORS IN AN ELECTRIC FIELD

3.1. Equilibrium of Charges on a Conductor

The carriers of a charge in a conductor are capable of moving under the action of a vanishingly small force. Therefore, the following conditions must be observed for the equilibrium of charges on a conductor:

1. The strength of the field everywhere inside the conductor must be zero:

$$\boldsymbol{E} = 0. \tag{3.1}$$

   In acccordance with Eq. (1.41), this signifies that the potential inside the conductor must be constant (φ = constant).

2. The strength of the field on the surface of the conductor must be directed along a normal to the surface at every point:

$$\boldsymbol{E} = \boldsymbol{E}_n. \tag{3.2}$$

   Consequently, when the charges are in equilibrium, the surface of the conductor will be an equipotential one.

If a charge q is imparted to a conducting body, the charge will be distributed so as to observe conditions of equilibrium. Let us imagine an arbitrary closed surface completely confined in a body. When the charges are in equilibrium, there is no field at every point inside the conductor; therefore, the flux of the electric displacement vector through the surface vanishes. According to Gauss's theorem, the sum of the charges inside the surface will also equal zero. This holds for a surface of any dimensions arbitrarily arranged inside a conductor. Hence, in equilibrium, there can be no surplus charges anywhere inside a conductor—they will all be distributed over the surface of the conductor with a certain density σ.

Since there are no surplus charges in a conductor in the state of equilibrium,

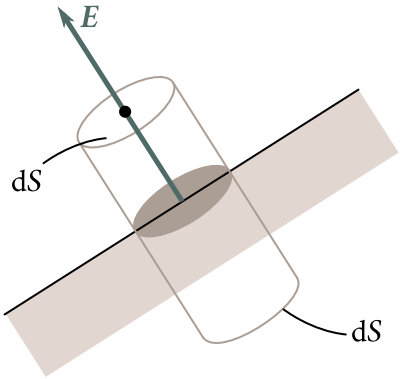


Fig. 3.1

the removal of substance from a volume taken inside the conductor will have no effect whatsoever on the equilibrium arrangement of the charges. Thus, a surplus charge will be distributed on a hollow conductor in the same way as on a solid one, *i.e.*, along its external surface. No surplus charges can be located on the surface of a cavity in the state of equilibrium. This conclusion also follows from the fact that the like elementary charges forming the given charge q mutually repel one another and, consequently, tend to take up positions at the farthest distance apart.

Imagine a small cylindrical surface formed by normals to the surface of a conductor and bases of the magnitude dS, one of which is inside and the other outside the conductor (Fig. 3.1). The flux of the electric displacement vector through the inner part of the surface equals zero because $\boldsymbol{E}$ and, consequently, $\boldsymbol{D}$ vanish inside the conductor. Outside the conductor in direct proximity to it, the field strength $\boldsymbol{E}$ is directed along a normal to the surface. Hence, for the side surface of the cylinder protruding outward, $D_n = 0$, and for the outside base $D_n = D$ (the outside base is assumed to be very close to the surface of the conductor). Hence, the displacement flux through the surface being considered is $D\,\mathrm{d}S$, where D is the value of the displacement in direct proximity to the surface of the conductor. The cylinder contains an extraneous charge $\sigma\,\mathrm{d}S$ (σ is the charge density at the given spot on the surface of the conductor).

Applying Gauss's theorem, we get $D\,\mathrm{d}S = \sigma\,\mathrm{d}S$, *i.e.*, $D = \sigma$. We thus see that the strength of the field near the surface of the conductor is

$$E = \frac{\sigma}{\varepsilon_0 \varepsilon} \tag{3.3}$$

where ε is the permittivity of the medium surrounding the conductor [compare with Eq. (1.123) obtained for the case when $\varepsilon = 1$].

Let us consider the field produced by the charged conductor shown in Fig. 3.2.

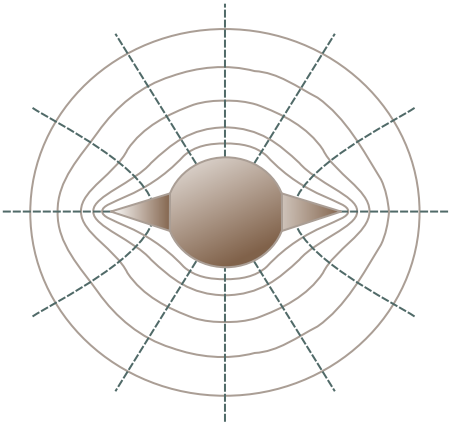


Fig. 3.2

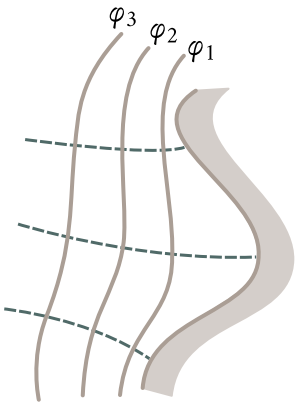


Fig. 3.3

At great distances from the conductor, equipotential surfaces have the shape of a sphere that is characteristic of a point charge (owing to the lack of space, a spherical surface is shown in the figure at a small distance from the conductor; the dash lines are field lines). As we approach the conductor, the equipotential surfaces become more and more similar to the surface of the conductor, which is an equipotential one. Near the projections, the equipotential surfaces are denser, hence, the field strength is also greater here. It thus follows that the density of the charges on the projections is especially great [see Eq. (3.3)]. We can arrive at the same conclusion by taking into account that owing to their mutual repulsion, charges tend to take up positions as far as possible from one another.

Near depressions in a conductor, the equipotential surfaces have a lower density (see Fig. 3.3). Accordingly, the field strength and the density of the charges at these spots will he smaller. In general, the density of charges with a given potential of a conductor is determined by the curvature of the surface—it grows with an increase in the positive curvature (convexity) and diminishes with an increase in the negative curvature (concavity). The density of charges is especially high on sharp points. Consequently, the field strength near such points may be so great that the gas molecules surrounding the conductor become ionized. Ions of the sign opposite to that of q are attracted to the conductor and neutralize its charge. Ions of the same sign as q begin to move away from the conductor, carrying along neutral molecules of the gas. The result is a noticeable motion of the gas called an electric wind. The charge of the conductor diminishes, it flows off the point, as it were, and is carried away by the wind. This phenomenon is therefore called emanation of a charge from a point.

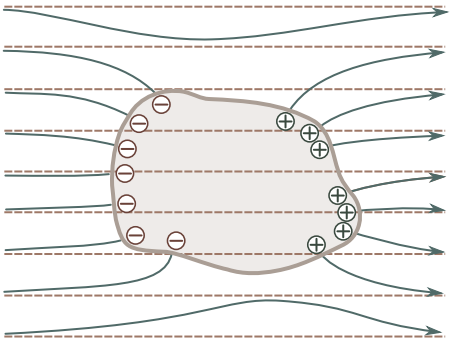


Fig. 3.4

3.2. A Conductor in an External Electric Field

When an uncharged conductor is introduced into an electric field, the charge carriers come into motion: the positive ones in the direction of the vector $\boldsymbol{E}$, the negative ones in the opposite direction. As a result, charges of opposite signs called **induced charges** appear at the ends of the conductor (Fig. 3.4, the dash lines depict the external field lines). The field of these charges is directed oppositely to the external field. Hence, the accumulation of charges at the ends of a conductor leads to weakening of the field in it. The charge carriers will be redistributed until conditions (3.1) and (3.2) are observed, *i.e.*, until the strength of the field inside the conductor vanishes and the field lines outside the conductor are perpendicular to its surface (see Fig. 3.4). Thus, a neutral conductor introduced into an electric field disrupts part of the field lines—they terminate on the negative induced charges and begin again on the positive ones.

The induced charges distribute themselves over the outer surface of a conductor. If a conductor contains a cavity, then upon equilibrium distribution of the induced charges, the field inside it vanishes. Electrostatic shielding is based on this phenomenon. If an instrument is to be protected from the action of external fields, it is surrounded by a conducting screen. The external field is compensated inside the screen by the induced charges appearing on its surface. Such a screen also functions quite well if it is made not solid, but in the form of a dense network.

3.3. Capacitance

A charge q imparted to a conductor distributes itself over its surface so that the strength of the field inside the conductor vanishes. Such a distribution is the only possible one. Therefore, if we impart to a conductor already carrying the charge

q another charge of the same magnitude, then the second charge must distribute itself over the conductor in exactly the same way as the first one. Otherwise, the charge will set up in the conductor a field differing from zero. We must note that this holds only for a conductor remote from other bodies (an isolated conductor). If other bodies are near the conductor, the imparting to the latter of a new portion of charge will produce either a change in the polarization of these bodies or a change in the induced charges on them. As a result, similarity in the distribution of different portions of the charge will be violated.

Thus, charges differing in magnitude distribute themselves on an isolated conductor in a similar way (the ratio of the densities of the charge at two arbitrary points on the surface of the conductor with any magnitude of the charge will be the same). It thus follows that the potential of an isolated conductor is proportional to the charge on it. Indeed, an increase in the charge a certain number of times leads to an increase in the strength of the field at every point of the space surrounding the conductor the same number of times. Accordingly, the work needed for transferring a unit charge from infinity to the surface of a conductor, *i.e.*, the potential of the conductor, grows the same number of times. Thus, for an isolated conductor

$$q = C\varphi. \tag{3.4}$$

The constant of proportionality C between the potential and the charge is called the **capacitance**. From Eq. (3.4), we get

$$C = \frac{q}{\varphi}. \tag{3.5}$$

In accordance with Eq. (3.5), the capacitance numerically equals the charge which when imparted to a conductor increases its potential by unity.

Let us calculate the potential of a charged sphere of radius R. The potential difference and the field strength are related by Eq. (1.45). We can therefore find the potential of the sphere φ by integrating Eq. (2.41) over r from R to ∞ (we assume that the potential at infinity equals zero):

$$\varphi = \frac{1}{4\pi\varepsilon_0}\int_0^\infty \frac{q}{\varepsilon r^2}\,\mathrm{d}r = \frac{1}{4\pi\varepsilon_0}\frac{q}{\varepsilon R}. \tag{3.6}$$

Comparing Eqs. (3.5) and (3.6), we find that the capacitance of an isolated sphere of radius R immersed in a homogeneous infinite dielectric of permittivity ε is

$$C = 4\pi\varepsilon_0\varepsilon R. \tag{3.7}$$

The unit of capacitance is the capacitance of a conductor whose potential changes by 1 V when a charge of 1 C is imparted to it. This unit of capacitance is called the **farad** (F). In the Gaussian system, the formula for the capacitance of an

isolated sphere has the form

$$C = \varepsilon R. \tag{3.8}$$

Since ε is a dimensionless quantity, the capacitance determined by Eq. (3.8) has the dimension of length. The unit of capacitance is the capacitance of an isolated sphere with a radius of 1 cm in a vacuum. This unit of capacitance is called the **centimetre**. According to Eq. (3.5),

$$1\ \text{F} = \frac{1\ \text{C}}{1\ \text{V}} = \frac{3 \times 10^9\ \text{cgse}_C}{1/300} = 9 \times 10^{11}\ \text{cm}. \tag{3.9}$$

An isolated sphere having a radius of 9×10^{11} cm, *i.e.*, a radius 1500 times greater than that of the Earth, would have a capacitance of 1 F. We can thus see that the farad is a very great unit. For this reason, submultiples of a farad are used in practice—the millifarad (mF), the microfarad (μF), the nanofarad (nF), and the picofarad (pF) (see Vol. I, Table 3.1).

3.4. Capacitors

Isolated conductors have a small capacitance. Even a sphere of the Earth's size has a capacitance of only 700 μF. Devices are needed in practice, however, that with a low potential relative to the surrounding bodies would accumulate charges of an appreciable magnitude (*i.e.*, would have a high charge "capacity"). Such devices, called **capacitors**, are based on the fact that the capacitance of a conductor grows when other bodies are brought close to it. This is due to the circumstance that under the action of the field set up by the charged conductor, induced (on a conductor) or bound (on a dielectric) charges appear on the body brought up to it. Charges of the sign opposite to that of the charge q of the conductor will be closer to the conductor than charges of the same sign as q and, consequently, will have a greater influence on its potential. Therefore, when a body is brought close to a charged conductor, the potential of the latter diminishes in absolute value. According to Eq. (3.5), this signifies an increase in the capacitance of the conductor.

Capacitors are made in the form of two conductors placed close to each other. The conductors forming a capacitor are called its **plates**. To prevent external bodies from influencing the capacitance of a capacitor, the plates are shaped and arranged relative to each other so that the field set up by the charges accumulating on them is concentrated inside the capacitor. This condition is satisfied (see Sec. 1.14) by two plates arranged close to each other, two coaxial cylinders, and two concentric spheres. Accordingly, parallel-plate (plane), cylindrical, and spherical capacitors are encountered. Since the field is confined inside a capacitor, the electric displacement lines begin on one plate and terminate on the other. Consequently, the extraneous

charges produced on the plates have the same magnitude and are opposite in sign.

The basic characteristic of a capacitor is its capacitance, by which is meant a quantity proportional to the charge q and inversely proportional to the potential difference between the plates:

$$C = \frac{q}{\varphi_1 - \varphi_2}. \tag{3.10}$$

The potential difference $\varphi_1 - \varphi_2$ is called the **voltage** across the relevant points[1]. We shall use the symbol U to designate the voltage. Hence, Eq. (3.10) can be written as follows:

$$C = \frac{q}{U}. \tag{3.11}$$

Here, U is the voltage across the plates.

The capacitance of capacitors is measured in the same units as that of isolated conductors (see the preceding section).

The magnitude of the capacitance is determined by the geometry of the capacitor (the shape and dimensions of the plates and their separation distance), and also by the dielectric properties of the medium filling the space between the plates. Let us find the equation for the capacitance of a parallel-plate capacitor. If the area of a plate is S and the charge on it is q, then the strength of the field between the plates is

$$E = \frac{\sigma}{\varepsilon_0 \varepsilon} = \frac{q}{\varepsilon_0 \varepsilon S}$$

[see Eqs. (1.121) and (2.33); ε is the permittivity of the medium filling the gap between the plates].

In accordance with Eq. (1.45), the potential difference between the plates is

$$\varphi_1 - \varphi_2 = Ed = \frac{qd}{\varepsilon_0 \varepsilon S}.$$

Hence, for the capacitance of a parallel-plate capacitor, we get

$$C = \frac{\varepsilon_0 \varepsilon S}{d} \tag{3.12}$$

where S is the area of a plate, d is the separation distance of the plates, and ε is the permittivity of the substance filling the gap.

It must be noted that the accuracy of determining the capacitance of a real parallel-plate capacitor by Eq. (3.12) is the greater, the smaller is the separation distance d in comparison with the linear dimensions of the plates.

It can be seen from Eq. (3.12) that the dimension of the electric constant ε_0 equals the dimension of capacitance divided by that of length. Accordingly, ε_0 is measured in farads per metre [see Eq. (1.12)].

If we disregard the dispersion of the field near the plate edges, we can easily

[1] A more general definition of the quantity called voltage will be given in Sec. 5.3 [see Eq. (5.18)].

obtain the following equation for the capacitance of a cylindrical capacitor:

$$C = \frac{2\pi\varepsilon_0\varepsilon l}{\ln\left(\frac{R_2}{R_1}\right)} \tag{3.13}$$

where l is length of the capacitor, R_1 and R_2 the radii of the internal and external plates.

The accuracy of determining the capacitance of a real capacitor by Eq. (3.13) is the greater, the smaller is the separation distance of the plates $d = R_2 - R_1$ in comparison with l and R_1.

The capacitance of a spherical capacitor is

$$C = 4\pi\varepsilon_0\varepsilon\left(\frac{R_1R_2}{R_2 - R_1}\right) \tag{3.14}$$

where R_1 and R_2 are the radii of the internal and external plates.

Apart from the capacitance, every capacitor is characterized by the maximum voltage $U_{\max}$ that may be applied across its plates without the danger of a breakdown. When this voltage is exceeded, a spark jumps across the space between the plates. The result is destruction of the dielectric and failure of the capacitor.

Chapter 4

ENERGY OF AN ELECTRIC FIELD

4.1. Energy of a Charged Conductor

The charge q on a conductor can be considered as a system of point charges Δq. In Sec. 1.7, we obtained the following expression for the energy of interaction of a system of charges [see Eq. (1.39)]:

$$W_{\mathrm{p}} = \frac{1}{2}\sum_i q_i \varphi_i. \tag{4.1}$$

Here, φ_i is the potential set up by all the charges except q_i at the point where the charge q_i is.

The surface of a conductor is equipotential. Therefore, the potentials of the points where the point charges Δq are located are identical and equal the potential φ of the conductor. Using Eq. (4.1), we get the following expression for the energy of a charged conductor

$$W_{\mathrm{p}} = \frac{1}{2}\sum \varphi \Delta q = \frac{1}{2}\varphi \sum \Delta q = \frac{1}{2}\varphi q. \tag{4.2}$$

Taking into account Eq. (3.5), we can write that

$$W_{\mathrm{p}} = \frac{\varphi q}{2} = \frac{q^2}{2C} = \frac{C\varphi^2}{2}. \tag{4.3}$$

Any of these expressions gives the energy of a charged conductor.

4.2. Energy of a Charged Capacitor

Assume that the potential of a capacitor plate carrying the charge $+q$ is φ_1 and that of a plate carrying the charge $-q$ is φ_2. Consequently, each of the elementary charges

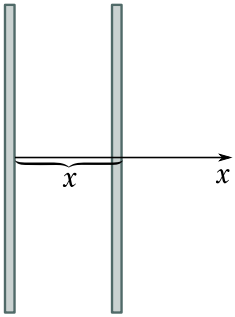


Fig. 4.1

Δq into which the charge $+q$ can be divided is at a point with the potential φ_1, and each of the charges into which the charge $-q$ can be divided is at a point with the potential φ_2. By Eq. (4.1), the energy of such a system of charges is

$$W_\mathrm{p} = \frac{1}{2}\left[(+q)\varphi_1 + (-q)\varphi_2\right] = \frac{1}{2}q\left(\varphi_1 - \varphi_2\right) = \frac{1}{2}qU. \tag{4.4}$$

Using Eq. (3.11), we can write three expressions for the energy of a charged capacitor:

$$W_\mathrm{p} = \frac{qU}{2} = \frac{q^2}{2C} = \frac{CU^2}{2}. \tag{4.5}$$

Equation (4.5) differs from (4.3) only in containing U instead of φ.

The expression for the potential energy permits us to find the force with which the plates of a parallel-plate capacitor attract each other. Let us assume that the separation distance of the plates can be changed. We shall associate the origin of the x-axis with the left-hand plate (Fig. 4.1). The coordinate x of the second plate will, therefore, determine the separation distanced of the plates. According to Eqs. (3.12) and (4.5), we have

$$W_\mathrm{p} = \frac{q^2}{2C} = \frac{q^2}{2\varepsilon_0\varepsilon S}x.$$

Let us differentiate this expression with respect to x, assuming that the charge on the plates is constant (the capacitor is disconnected from a voltage source). As a result, we obtain the projection of the force exerted on the right-hand plate onto the x-axis:

$$F_x = -\frac{\partial W_\mathrm{p}}{\partial x} = -\frac{q^2}{2\varepsilon_0\varepsilon S}.$$

The magnitude of this expression gives the force with which the plates attract each other:

$$F = \frac{q^2}{2\varepsilon_0\varepsilon S}. \tag{4.6}$$

Now, let us try to calculate the force of attraction between the plates of a parallel-plate capacitor as the product of the strength of the field produced by one of the plates and the charge concentrated on the other one. By Eq. (1.120), the strength of the field set up by one plate is

$$E = \frac{\sigma}{2\varepsilon_0} = \frac{q}{2\varepsilon_0 S}. \tag{4.7}$$

A dielectric weakens the field in the space between the plates e times, but this occurs only inside the dielectric [see Eq. (2.33) and the related text]. The charges on the plates are outside the dielectric and are, therefore, acted upon by the field strength given by Eq. (4.7). Multiplying the charge of a plate q by this strength, we get the following expression for the force:

$$F' = \frac{q^2}{2\varepsilon_0 S}. \tag{4.8}$$

Equations (4.6) and (4.8) do not coincide. The value of the force given by Eq. (4.6) obtained from the expression for the energy agrees with experimental data. The explanation is that apart from the "electric" force given by Eq. (4.8), the plates experience mechanical forces from the side of the dielectric that tend to spread them apart (see Sec. 2.8; we must note that we have in mind a fluid dielectric). There is a dispersed field at the edges of the plates whose magnitude diminishes with an increasing distance from the edges (Fig. 4.2). The molecules of the dielectric have a dipole moment and experience the action of a force pulling them into the region with the stronger field [see Eq. (1.62)]. The result is an increase in the pressure between the plates and the appearance of a force that weakens the force given by Eq. (4.8) e times.

If a charged capacitor with an air gap is partially immersed in a liquid dielectric, the latter will be drawn into the space between the plates (Fig. 4.3). This phenomenon is explained as follows. The permittivity of air virtually equals unity. Consequently, before the plates are immersed in the dielectric, we can consider that the capacitance of the capacitor is $C_0 = \varepsilon_0 S/d$, and its energy is $W_0 = q^2/2C_0$. When the space between the plates is partially filled with the dielectric, the capacitor can be considered as two capacitors connected in parallel, one of which has a plate area of xS (x is the relative part of the space filled with the liquid) and is filled with a dielectric for which $\varepsilon > 1$, and the other has a plate area equal to $(1 - x)S$. In the parallel connection of capacitors, their capacitances are summated:

$$C = C_1 + C_2 = \frac{\varepsilon_0 S(1-x)}{d} + \frac{\varepsilon_0 \varepsilon S x}{d} = C_0 + \frac{\varepsilon_0(\varepsilon - 1)S}{d} x > C_0.$$

Since $C > C_0$, the energy $W = q^2/2C$ will be smaller than W_0 (the charge q is assumed to be constant—the capacitor was disconnected from the voltage source

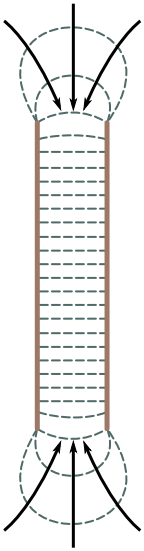


Fig. 4.2

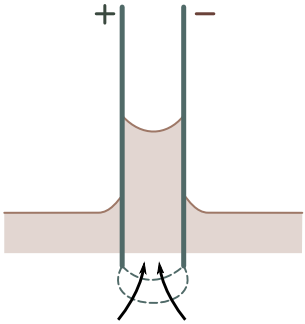


Fig. 4.3

before being immersed in the liquid). Hence, the filling of the space between the plates with the dielectric is profitable from the energy viewpoint. This is why the dielectric is drawn into the capacitor and its level in the space separating the plates rises. This, in turn, results in an increase in the potential energy of gravity. In the long run, the level of the dielectric in the space will establish itself at a certain height corresponding to the minimum total energy (electrical and gravitational). The above phenomenon is similar to the capillary rise of a liquid in the narrow space between plates (see Sec. 14.5 of Vol. I).

The drawing of the dielectric into the space between plates can also be explained from the microscopic viewpoint. There is a nonuniform field at the edges of the capacitor plates. The molecules of the dielectric have an intrinsic dipole moment or acquire it under the action of the field; therefore, they experience forces that tend to transfer them to the region of the strong field, *i.e.*, into the capacitor. These forces cause the liquid to be drawn into the space between the plates until the electric forces exerted on the liquid at the plate edges will be balanced by the weight of the liquid column.

4.3. Energy of an Electric Field

The energy of a charged capacitor can be expressed through quantities characterizing the electric field in the space between the plates. Let us do this for a

parallel-plate capacitor. Introducing expression (3.12) for the capacitance into the equation $W_\mathrm{p} = CU^2/2$ [see Eq. (4.5)], we get

$$W_\mathrm{p} = \frac{CU^2}{2} = \frac{\varepsilon_0 \varepsilon S U^2}{2d} = \frac{\varepsilon_0 \varepsilon}{2}\left(\frac{U}{d}\right)^2 Sd.$$

The ratio U/d equals the strength of the field between the plates; the product Sd is the volume occupied by the field. Hence,

$$W_\mathrm{p} = \frac{\varepsilon_0 \varepsilon E^2}{2} V. \tag{4.9}$$

The equation $W_\mathrm{p} = q^2/(2C)$ relates the energy of a capacitor to the charge on its plates, while Eq. (4.9) relates this energy to the field strength. It is logical to ask the question: where, after all, is the energy localized (*i.e.*, concentrated), what is the carrier of the energy—charges or a field? This question cannot be answered within the scope of electrostatics, which studies the fields of fixed charges that are constant in time. Constant fields and the charges producing them cannot exist separately from each other. Fields varying in time, however, can exist independently of the charges producing them and propagate in space in the form of electromagnetic waves. Experiments show that electromagnetic waves transfer energy. In particular, the energy due to which life exists on the Earth is supplied from the Sun by electromagnetic waves; the energy that causes a radio receiver to sound is carried from the transmitting station by electromagnetic waves, etc. These facts make us acknowledge the circumstance that the carrier of energy is a field.

If a field is homogeneous (which is the case in a parallel-plate capacitor), the energy confined in it is distributed in space with a constant density w equal to the energy of the field divided by the volume it occupies. Inspection of Eq. (4.9) shows that the density of the energy of a field of strength E set up in a medium with the permittivity ε is

$$w = \frac{\varepsilon_0 \varepsilon E^2}{2}. \tag{4.10}$$

With account taken of Eq. (2.21), we can write Eq. (4.10) as follows:

$$w = \frac{\varepsilon_0 \varepsilon E^2}{2} = \frac{ED}{2} = \frac{D^2}{2\varepsilon_0 \varepsilon}. \tag{4.11}$$

In an isotropic dielectric, the directions of the vectors $\boldsymbol{E}$ and $\boldsymbol{D}$ coincide. We can, therefore, write the equation for the energy density in the form

$$w = \frac{\boldsymbol{E} \cdot \boldsymbol{D}}{2}.$$

Substituting for $\boldsymbol{D}$ in this equation its value from Eq. (2.18), we get the following

expression for w:

$$w = \frac{\boldsymbol{E}(\varepsilon_0 \boldsymbol{E} + \boldsymbol{P})}{2} = \frac{\varepsilon_0 \boldsymbol{E}^2}{2} + \frac{\boldsymbol{E} \cdot \boldsymbol{P}}{2}. \tag{4.12}$$

The first addend in this expression coincides with the energy density of the field $\boldsymbol{E}$ in a vacuum. The second addend, as we shall proceed to prove, is the energy spent for polarization of the dielectric.

The polarization of a dielectric consists in that the charges contained in the molecules are displaced from their positions under the action of the electric field $\boldsymbol{E}$. The work done to displace the charges q, over the distance $\mathrm{d}\boldsymbol{r}_i$ per unit volume of the dielectric is

$$\mathrm{d}A = \sum_{V=i} q_i \boldsymbol{E}\, \mathrm{d}\boldsymbol{r}_i = \boldsymbol{E}\, \mathrm{d}\left(\sum_{V=i} q_i \boldsymbol{r}_i\right)$$

(we consider for simplicity's sake that the field is homogeneous). According to Eq. (2.1), $\sum_{V=i} q_i \boldsymbol{r}_i$ equals the dipole moment of a unit volume, *i.e.*, the polarization of the dielectric $\boldsymbol{P}$. Hence,

$$\mathrm{d}A = \boldsymbol{E}\, \mathrm{d}\boldsymbol{P}. \tag{4.13}$$

The vector $\boldsymbol{P}$ is related to the vector $\boldsymbol{E}$ by the expression $\boldsymbol{P} = \chi \varepsilon_0 \boldsymbol{E}$ [see Eq. (2.5)]. Hence, $\mathrm{d}\boldsymbol{P} = \chi \varepsilon_0\, \mathrm{d}\boldsymbol{E}$. Using this value of $\mathrm{d}\boldsymbol{P}$ in Eq. (4.13), we get the expression

$$\mathrm{d}A = \chi \varepsilon_0 \boldsymbol{E}\, \mathrm{d}\boldsymbol{E} = \mathrm{d}\left(\frac{\chi \varepsilon_0 \boldsymbol{E}^2}{2}\right) = \mathrm{d}\left(\frac{\boldsymbol{E} \cdot \boldsymbol{P}}{2}\right).$$

Finally, integration gives us the following expression for the work done to polarize a unit volume of the dielectric:

$$A = \frac{\boldsymbol{E} \cdot \boldsymbol{P}}{2}, \tag{4.14}$$

which coincides with the second addend in Eq. (4.12). Thus, expressions (4.11), apart from the intrinsic energy of a field $\varepsilon_0 E^2/2$, include the energy $(\boldsymbol{E} \cdot \boldsymbol{P})/2$ spent for the polarization of the dielectric when the field is set up.

Knowing the density of the field energy at every point, we can find the energy of the field confined in any volume V. For this purpose, we must calculate the integral

$$W = \int_V w\, \mathrm{d}V = \int_V \frac{\varepsilon_0 \varepsilon E^2}{2}\, \mathrm{d}V. \tag{4.15}$$

Let us calculate, as an example, the energy of the field of a charged conducting sphere of radius R placed in a homogeneous infinite dielectric. The field strength here is a function only of r:

$$E = \frac{1}{4\pi\varepsilon_0} \frac{q}{\varepsilon r^2}.$$

Let us divide the space surrounding our sphere into concentric spherical layers of

thickness dr. The volume of a layer is $\mathrm{d}V = 4\pi r^2\,\mathrm{d}r$. It contains the energy

$$\mathrm{d}W = w\,\mathrm{d}V = \frac{\varepsilon_0\varepsilon}{2}\left(\frac{1}{4\pi\varepsilon_0}\frac{q}{\varepsilon r^2}\right)4\pi r^2\,\mathrm{d}r = \frac{1}{2}\frac{q^2}{4\pi\varepsilon_0\varepsilon}\frac{\mathrm{d}r}{r^2}.$$

The energy of the field is

$$W = \int \mathrm{d}W = \frac{1}{2}\frac{q^2}{4\pi\varepsilon_0\varepsilon}\int_R^\infty \frac{\mathrm{d}r}{r^2} = \frac{1}{2}\frac{q^2}{4\pi\varepsilon_0\varepsilon R} = \frac{q^2}{2C}$$

[according to Eq. (3.7), $4\pi\varepsilon_0\varepsilon R$ is the capacitance of a sphere].

The expression we have obtained coincides with that for the energy of a conductor having the capacitance C and carrying the charge q [see Eq. (4.3)].

Chapter 5
STEADY ELECTRIC CURRENT

5.1. Electric Current

If a total charge other than zero is carried through an imaginary surface, an **electric current** (or simply a **current**) is said to flow through this surface. A current can flow in solids (metals, semiconductors), liquids (electrolytes), and in gases (the flow of a current through a gas is called a gas discharge).

For a current to flow, the given body (or given medium) must contain charged particles that can move within the limits of the entire body. Such particles are called **current carriers**. The latter may be electrons, or ions, or, finally, macroscopic particles carrying a surplus charge (for example, charged dust particles and droplets).

A current is produced if there is an electric field inside a body. The charge carriers participate in the molecular thermal motion and, consequently, travel with a certain velocity $\boldsymbol{v}$ even in the absence of a field. But in this case, an identical number of carriers of either sign pass on the average in both directions through an arbitrary area mentally drawn in the body, so that the current is zero. When a field is switched on, ordered motion with the velocity $\boldsymbol{u}$ is superposed onto the chaotic motion of the carriers with the velocity $\boldsymbol{v}$[1]. The velocity of the carriers will thus be $\boldsymbol{v} + \boldsymbol{u}$. Since the mean value of $\boldsymbol{v}$ (but not of v) equals zero, then the mean velocity of the carriers is $\langle \boldsymbol{u} \rangle$:

$$\langle \boldsymbol{v} + \boldsymbol{u} \rangle = \langle \boldsymbol{v} \rangle + \langle \boldsymbol{u} \rangle = \langle \boldsymbol{u} \rangle .$$

It follows from what has been said above that an electric current can be defined as the ordered motion of electric charges.

A quantitative characteristic of an electric current is the magnitude of the charge carried through the surface being considered in unit time. It is called the **current**

[1]Similarly, in a gas flow, ordered motion is superposed onto the chaotic thermal motion of the molecules.

strength, or more often simply the **current**. We must note that a current is in essence a flow of a charge through a surface (compare with the flow of a fluid, energy flux, etc.).

If the charge $\mathrm{d}q$ is carried through a surface during the time $\mathrm{d}t$, then the current is

$$I = \frac{\mathrm{d}q}{\mathrm{d}t}. \tag{5.1}$$

An electric current may be produced by the motion of either positive or negative charges. The transfer of a negative charge in one direction is equivalent to the transfer of a positive charge of the same magnitude in the opposite direction. If a current is produced by carriers of both signs, the positive carriers transferring the charge $\mathrm{d}q^+$ in one direction through the given surface during the time $\mathrm{d}t$, and the negative carriers the charge $\mathrm{d}q^-$ in the opposite direction during the same time, then

$$I = \frac{\mathrm{d}q^+}{\mathrm{d}t} + \frac{|\mathrm{d}q^-|}{\mathrm{d}t}.$$

The direction of motion of the positive carriers has been conventionally assumed to be the direction of a current.

A current may be distributed non-uniformly over the surface through which it is flowing. A current can be characterized in greater detail by means of the current density vector $\boldsymbol{j}$. This vector numerically equals the current $\mathrm{d}I$ through the area $\mathrm{d}S_\perp$ arranged at the given point perpendicular to the direction of motion of the carriers divided by the magnitude of this area:

$$j = \frac{\mathrm{d}I}{\mathrm{d}S_\perp}. \tag{5.2}$$

The direction of $\boldsymbol{j}$ is taken as that of the velocity vector $\boldsymbol{u}^+$ of the ordered motion of the positive carriers (or as the direction opposite to that of the vector $\boldsymbol{u}^-$).

The field of the current density vector can be depicted by means of current lines that are constructed in the same way as the streamlines in a flowing liquid, the lines of the vector $\boldsymbol{E}$, etc.

Knowing the current density vector at every point of space, we can find the current I through any surface S:

$$I = \int_S \boldsymbol{j} \cdot \mathrm{d}\boldsymbol{S}. \tag{5.3}$$

It can be seen from Eq. (5.3) that the current is the flux of the current density vector through a surface [see Eq. (1.74)].

Assume that a unit volume contains n^+ positive carriers and n^- negative ones. The algebraic value of the carrier charges is e^+ and e^-, respectively. If the carriers acquire the average velocities u^+ and u^- under the action of the field, then n^+u^+

positive carriers will pass in unit time through unit area[2], and they will transfer the charge $e^+n^+u^+$. Similarly, the negative carriers will transfer the charge $e^-n^-u^-$ in the opposite direction. We, thus, get the following expression for the current density:

$$j = e^+n^+u^+ + e^-n^-u^-. \tag{5.4}$$

This expression can be given a vector form:

$$\boldsymbol{j} = e^+n^+\boldsymbol{u}^+ + e^-n^-\boldsymbol{u}^- \tag{5.5}$$

(both addends have the same direction: the vector $\boldsymbol{u}^-$ is directed oppositely to the vector $\boldsymbol{j}$; when it is multiplied by the negative scalar e^-, we get a vector of the same direction as $\boldsymbol{j}$).

The product e^+n^+ gives the charge density of the positive carriers ρ^+. Similarly, e^-n^- gives the charge density of the negative carriers ρ^-. Hence, Eq. (5.5) can be written in the form

$$\boldsymbol{j} = \rho^+\boldsymbol{u}^+ + \rho^-\boldsymbol{u}^- \tag{5.6}$$

A current that does not change with time is called **steady** (do not confuse with a direct current whose direction is constant, but whose magnitude may vary). For a steady current, we have

$$I = \frac{\mathrm{d}q}{\mathrm{d}t}, \tag{5.7}$$

where q is the charge carried through the surface being considered during the finite time t.

In the SI, the unit of current, the **ampere** (A), is a basic one. Its definition will be given on a later page (see Sec. 6.1). The unit of charge, the **coulomb** (C), is defined as the charge carried in one second through the cross section of a conductor at a current of one ampere.

The unit of current in the cgse system is the current at which one cgse unit of charge (1 cgse_q) is carried through a given surface in one second. From Eqs. (1.8) and (5.7) we find that

$$1\ \mathrm{A} = 3 \times 10^9\ \mathrm{cgse}_I. \tag{5.8}$$

[2]The expression for the number of molecules flying in unit time through unit area contains, in addition, the factor 1/4 due to the fact that the molecules move chaotically [see Eq. (11.23) of Vol. I]. This factor is not present in the given case because all the carriers of a given sign have ordered motion in one direction.

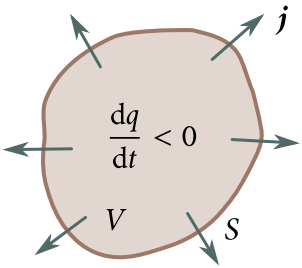


Fig. 5.1

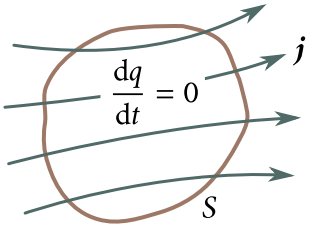


Fig. 5.2

5.2. Continuity Equation

Let us consider an imaginary closed surface S (Fig. 5.1) in a medium in which a current is flowing. The expression $\oint_S \boldsymbol{j} \cdot \mathrm{d}\boldsymbol{S}$ gives the charge emerging in a unit time from the volume V enclosed by surface S. Owing to charge conservation, this quantity must equal the rate of diminishing of the charge q contained in the given volume:

$$\oint_S \boldsymbol{j} \cdot \mathrm{d}\boldsymbol{S} = -\frac{\mathrm{d}q}{\mathrm{d}t}.$$

Substituting for q its value $\int_V \rho \, \mathrm{d}V$, we get the expression

$$\oint_S \boldsymbol{j} \cdot \mathrm{d}\boldsymbol{S} = -\frac{\mathrm{d}}{\mathrm{d}t} \int_V \rho \, \mathrm{d}V = -\int_V \frac{\partial \rho}{\partial t} \, \mathrm{d}V. \tag{5.9}$$

We have written the partial derivative of ρ with respect to t inside the integral because the charge density may depend not only on time, but also on the coordinates (the integral $\int_V \rho \, \mathrm{d}V$ is a function only of time). Let us transform the left-hand side of Eq. (5.9) in accordance with the Ostrogradsky-Gauss theorem. As a result, we get

$$\int_V (\nabla \cdot \boldsymbol{j}) \, \mathrm{d}V = -\int_V \frac{\partial \rho}{\partial t} \, \mathrm{d}V. \tag{5.10}$$

Equation (5.10) must be observed upon an arbitrary choice of the volume V over which the integrals are taken. This is possible only if at every point of space the condition is observed that

$$\nabla \cdot \boldsymbol{j} = -\frac{\partial \rho}{\partial t}. \tag{5.11}$$

Equation (5.11) is known as the **continuity equation**. It [like Eq. (5.9)] expresses the law of charge conservation. According to Eq. (5.11), the charge diminishes at points that are sources of the vector $\boldsymbol{j}$.

For a steady current, the potential at different points, the charge density, and other quantities are constant. Hence, for a steady current, Eq. (5.11) has the form

$$\nabla \cdot \boldsymbol{j} = 0. \tag{5.12}$$

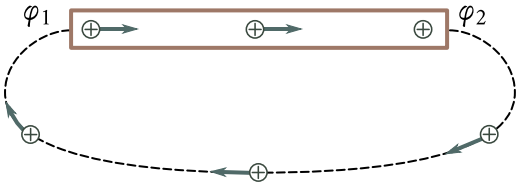


Fig. 5.3

Thus, for a steady current, the vector $\boldsymbol{j}$ has no sources. This signifies that the current lines begin nowhere and terminate nowhere. Hence, the lines of a steady current are always closed. Accordingly, $\oint_S \boldsymbol{j} \cdot \mathrm{d}\boldsymbol{S}$ equals zero. Therefore, for a steady current, the picture similar to that shown in Fig. 5.1 has the form shown in Fig. 5.2.

5.3. Electromotive Force

If an electric field is set up in a conductor and no measures are taken to maintain it, then motion of the current carriers will lead very rapidly to vanishing of the field inside the conductor and stopping of the current. To maintain a current for a sufficiently long time, it is necessary to continuously remove from the end of the conductor with the lower potential (the current carriers are assumed to be positive) the charges carried to it by the current, and continuously supply them to the end with the higher potential (Fig. 5.3). In other words, it is necessary to circulate the charges along a closed path. This agrees with the fact that the lines of a steady current are closed (see the preceding section).

The circulation of the strength vector of an electrostatic field equals zero. Therefore, in a closed circuit, in addition to sections on which the positive carriers travel in the direction of a decrease in the potential φ, there must be sections on which the positive charges are carried in the direction of a growth in φ, *i.e.*, against the forces of the electrostatic field (see the part of the circuit in Fig. 5.3 shown by the dash line). Motion of the carriers on these sections is possible only with the aid of forces of a non-electrostatic origin, called **extraneous forces**. Thus, to maintain a current, extraneous forces are needed that act either over the entire length of the circuit or on separate sections of it. These forces may be due to chemical processes, the diffusion of the current carriers in a non-uniform medium or through the interface between two different substances, to electric (but not electrostatic) fields set up by magnetic fields varying with time (see Sec. 9.1), etc.

Extraneous forces can be characterized by the work they do on charges travelling along a circuit. The quantity equal to the work done by the extraneous forces on a unit positive charge is called the **electromotive force** (**e.m.f.**) $\mathscr{E}$ acting in

a circuit or on a section of it. Hence, if the work of the extraneous forces on the charge q is A, then

$$\mathscr{E} = \frac{A}{q}. \tag{5.13}$$

A comparison of Eqs. (1.31) and (5.13) shows that the dimension of the e.m.f. coincides with that of the potential. Therefore, $\mathscr{E}$ is measured in the same units as φ.

The extraneous force $\boldsymbol{F}_{\text{extr}}$ acting on the charge q can be represented in the form

$$\boldsymbol{F}_{\text{extr}} = \boldsymbol{E}^* q. \tag{5.14}$$

The vector quantity $\boldsymbol{E}^*$ is called the **strength of the extraneous force field**. The work (of the extraneous forces on the charge q on circuit section 1-2 is

$$A_{12} = \int_1^2 \boldsymbol{F}_{\text{extr}}\, \mathrm{d}\boldsymbol{l} = q \int_1^2 \boldsymbol{E}^* \cdot \mathrm{d}\boldsymbol{l}.$$

Dividing this work by q, we get the e.m.f. acting on the given section:

$$\mathscr{E}_{12} = \int_1^2 \boldsymbol{E}^* \cdot \mathrm{d}\boldsymbol{l}. \tag{5.15}$$

A similar integral calculated for a closed circuit gives the e.m.f. acting in this circuit:

$$\mathscr{E} = \oint \boldsymbol{E}^* \cdot \mathrm{d}\boldsymbol{l}. \tag{5.16}$$

Thus, the e.m.f. acting in a closed circuit can be determined as the circulation of the strength vector of the extraneous forces.

In addition to extraneous forces, a charge experiences the forces of an electrostatic field $\boldsymbol{F}_E = q\boldsymbol{E}$. Hence, the resultant force acting at each point of a circuit on the charge q is

$$\boldsymbol{F} = \boldsymbol{F}_E + \boldsymbol{F}_{\text{extr}} = q(\boldsymbol{E} + \boldsymbol{E}^*).$$

The work done by this force on the charge q on circuit section 1-2 is determined by the expression

$$A_{12} = q \int_1^2 \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{l} + q \int_1^2 \boldsymbol{E}^* \cdot \mathrm{d}\boldsymbol{l} = q(\varphi_1 - \varphi_2) + q\mathscr{E}_{12}. \tag{5.17}$$

The quantity numerically equal to the work done by the electrostatic and extraneous forces in moving a unit positive charge is defined as the **voltage drop** or simply the **voltage** U on the given section of the circuit. According to Eq. (5.17),

$$U_{12} = \varphi_1 - \varphi_2 + \mathscr{E}_{12}. \tag{5.18}$$

A section of a circuit on which no extraneous forces act is called **homogeneous**.

A section on which the current carriers experience extraneous forces is called **inhomogeneous**. For a homogeneous section of a circuit

$$U_{12} = \varphi_1 - \varphi_2, \tag{5.19}$$

i.e., the voltage coincides with the potential difference across the ends of the section.

5.4. Ohm's Law. Resistance of Conductors

The German physicist Georg Ohm (1789-1854) experimentally established a law according to which *the current flowing in a homogeneous* (in the meaning that no extraneous forces are present) *metal conductor is proportional to the voltage drop U in the conductor*:

$$I = \frac{1}{R}U. \tag{5.20}$$

We remind our reader that for a homogeneous conductor the voltage U coincides with the potential difference $\varphi_1 - \varphi_2$ [see Eq. (5.18)].

The quantity designated by the symbol R in Eq. (5.20) is called the **electrical resistance** of a conductor. The unit of resistance is the ohm (Ω) equal to the resistance of a conductor in which a current of 1 A flows at a voltage of 1 V.

The value of the resistance depends on the shape and dimensions of a conductor and also on the properties of the material it is made of. For a homogeneous cylindrical conductor

$$R = \rho\frac{1}{S}, \tag{5.21}$$

where l, is the length of the conductor, S its cross-sectional area, and ρ the coefficient depending on the properties of the material and called the **resistivity** of the substance.

If $l = 1$ and $S = 1$, then R numerically equals ρ. In the SI, ρ is measured in ohm-metres (Ω m).

Let us find the relation between the vectors $\boldsymbol{j}$ and $\boldsymbol{E}$ at the same point of a conductor. In an isotropic conductor, the ordered motion of the current carriers takes place in the direction of the vector $\boldsymbol{E}$. Therefore, the directions of the vectors $\boldsymbol{j}$ and $\boldsymbol{E}$ coincide[3]. Let us mentally separate an elementary cylindrical volume with generatrices parallel to the vectors $\boldsymbol{j}$ and $\boldsymbol{E}$ in the vicinity of a certain point (Fig. 5.4). A current equal to $j\,\mathrm{d}S$ flows through the cross section of the cylinder. The voltage across the cylinder is $E\,\mathrm{d}l$, where E is the field-strength at the given point. Finally, the resistance of the cylinder, according to Eq. (5.21), is $\rho(\mathrm{d}l/\mathrm{d}S)$. Using these values

[3] In anisotropic bodies, the directions of the vectors $\boldsymbol{j}$ and $\boldsymbol{E}$, generally speaking, do not coincide. The relation between $\boldsymbol{j}$ and $\boldsymbol{E}$ for such bodies is achieved with the aid of the conductance tensor.

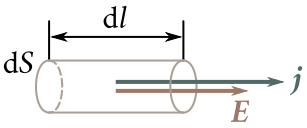


Fig. 5.4

in Eq. (5.21), we arrive at the equation

$$j\,\mathrm{d}S = \frac{1}{\rho}\frac{\mathrm{d}S}{\mathrm{d}l}E\,\mathrm{d}l \quad \text{or} \quad j = \frac{1}{\rho}E.$$

Taking advantage of the fact that the vectors $\boldsymbol{j}$ and $\boldsymbol{E}$ have the same direction, we can write

$$\boldsymbol{j} = \frac{1}{\rho}\boldsymbol{E} = \sigma\boldsymbol{E}. \tag{5.22}$$

This equation expresses Ohm's law in the differential form.

The quantity σ in Eq. (5.22) that is the reciprocal of ρ is called the **conductivity** of a material. The unit that is the reciprocal of the ohm is called the **siemens** (S). The unit of σ is accordingly the siemens per metre ($\mathrm{S\,m^{-1}}$).

Let us assume for simplicity's sake that a conductor contains carriers of only one sign. According to Eq. (5.5), the current density in this case is

$$\boldsymbol{j} = en\boldsymbol{u}. \tag{5.23}$$

A comparison of this expression with Eq. (5.22) leads us to the conclusion that the velocity of ordered motion of current carriers is proportional to the field strength $\boldsymbol{E}$, *i.e.*, to the force imparting ordered motion to the carriers. Proportionality of the velocity to the force applied to a body is observed when apart from the force producing the motion, the body experiences the force of resistance of the medium. This force is due to the interaction of the current carriers with the particles which the substance of the conductor is built of. The presence of the force of resistance to ordered motion of the current carriers results in the electrical resistance of a conductor.

The ability of a substance to conduct an electric current is characterized by its resistivity ρ or conductivity σ. Their magnitude is determined by the chemical nature of the substance and the surrounding conditions, in particular the ambient temperature.

The resistivity ρ varies directly with the absolute temperature T for most metals at temperatures close to room one:

$$\rho \propto T. \tag{5.24}$$

Deviations from this proportion are observed at low temperatures (Fig. 5.5). The dependence of ρ on T usually follows curve 1. The magnitude of the residual

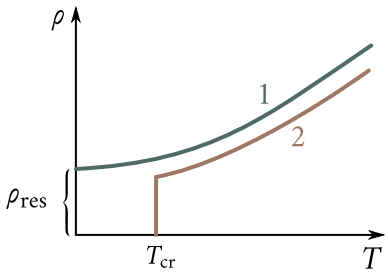


Fig. 5.5

resistivity ρ_{res} depends very greatly on the purity of the material and the presence of residual mechanical stresses in the specimen. This is why ρ_{res} appreciably diminishes after annealing. The resistivity ρ of a perfectly pure metal with an ideal regular crystal lattice vanishes at absolute zero.

The resistance of a large group of metals and alloys at a temperature of the order of several kelvins vanishes in a jump (curve 2 in Fig. 5.5). This phenomenon, called **superconductivity**, was first discovered in 1911 by the Dutch scientist Heike Kamerlingh Onnes (1853-1926) for mercury. Superconductivity was later discovered in lead, tin, zinc, aluminium, and other metals, as well as in a number of alloys. Every superconductor has its own critical temperature T_{cr} at which it passes over into a superconducting state. The superconducting state is violated when a magnetic field acts on a superconductor. The magnitude of the critical field B_{cr} (the symbol B stands for the magnetic induction—see Sec. 6.2) destroying superconductivity equals zero when $T = T_{cr}$ and grows with lowering of the temperature.

A complete theoretical substantiation of superconductivity was given in 1957 by J. Bardeen, L. Cooper, and J. Schrieffer (see Vol. III, Sec. 8.2).

The temperature dependence of resistance underlies the design of resistance thermometers. Such a thermometer is a metal (usually platinum) wire wound onto a porcelain or mica body. A resistance thermometer graduated according to constant temperature points makes it possible to measure both low and high temperatures with an accuracy of the order of several hundredths of a kelvin. Recent times have seen semiconductor resistance thermometers coming into greater and greater favour.

5.5. Ohm's Law for an Inhomogeneous Circuit Section

The extraneous forces $e\boldsymbol{E}^*$ act on current carriers on an inhomogeneous section of a circuit in addition to the electrostatic forces $e\boldsymbol{E}$. Extraneous forces are capable of producing ordered motion of current carriers to the same extent as electrostatic

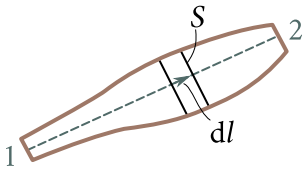


Fig. 5.6

forces are. We found in the preceding section that in a homogeneous conductor, the average velocity of ordered motion of the current carriers is proportional to the electrostatic force $e\boldsymbol{E}$. It is quite obvious that, where extraneous forces are exerted on the carriers in addition to the electrostatic forces, the average velocity of ordered motion of the carriers will be proportional to the total force $e\boldsymbol{E} + e\boldsymbol{E}^*$. Accordingly, the current density at these points is proportional to the sum of the strengths $\boldsymbol{E} + \boldsymbol{E}^*$:

$$\boldsymbol{j} = \sigma\,(\boldsymbol{E} + \boldsymbol{E}^*)\,. \tag{5.25}$$

Equation (5.25) summarizes Eq. (5.22) for an inhomogeneous conductor. It expresses Ohm's law for an inhomogeneous section of a circuit in the differential form.

We can pass over from Ohm's law in the differential form to its integral form. Let us consider an inhomogeneous section of a circuit. Assume that there is a line inside this section (we shall call it the current path) complying with the following conditions: (1) in every cross section perpendicular to the path, the quantities $\boldsymbol{j}$, σ, $\boldsymbol{E}$, $\boldsymbol{E}^*$ have the same values with sufficient accuracy, and (2) the vectors $\boldsymbol{j}$, $\boldsymbol{E}$, and $\boldsymbol{E}^*$ at every point are directed along a tangent to the path. The cross section of the conductor may vary (Fig. 5.6).

Let us choose an arbitrary direction of motion along the path. Assume that the chosen Fig. 5.6 direction corresponds to motion from end 1 to end 2 of the circuit section (direction 1-2). Let us project the vectors in Eq. (5.25) onto the path element $\mathrm{d}\boldsymbol{l}$. The result is

$$j_l = \sigma(E_l + E_l^*). \tag{5.26}$$

Owing to our assumption, the projection of each of the vectors equals the magnitude of the vector taken with the sign plus or minus depending on the direction of the vector relative to $\mathrm{d}\boldsymbol{l}$. For example, $j_l = j$ if the current flows in direction 1-2, and $j_l = -j$ if it flows in direction 2-1.

Owing to charge conservation, the steady current in each section must be the same. Therefore, the quantity $I = j_l S$ is constant along the path. The current in this case should be treated as an algebraic quantity. We remind our reader that we have chosen direction 1-2 arbitrarily. Hence, if the current flows in the chosen direction, it should be considered positive, and if it flows in the opposite direction (*i.e.*, from

end 2 to end 1), it should be considered negative.

Let us substitute the ratio I/S for j_l and the resistivity ρ for the conductivity σ in Eq. (5.26). We get

$$I\frac{\rho}{S} = E_l + E_l^*.$$

Multiplication of the above equation by $\mathrm{d}l$ and integration along the path yield

$$I\int_1^2 \rho\frac{\mathrm{d}l}{S} = \int_1^2 E_l\,\mathrm{d}l + \int_1^2 E_l^*\,\mathrm{d}l.$$

The quantity $\rho(\mathrm{d}l/S)$ is the resistance of the path section of length $\mathrm{d}l$, and the integral of this quantity is the resistance R of the circuit section. The first integral in the right-hand side gives $\varphi_1 - \varphi_2$, and the second integral gives the e.m.f. $\mathscr{E}_{12}$ acting on the section. We, thus, arrive at the equation

$$IR = \varphi_1 - \varphi_2 + \mathscr{E}_{12}. \tag{5.27}$$

The e.m.f. $\mathscr{E}_{12}$, like the current I, is an algebraic quantity. When the e.m.f. facilitates the motion of the positive current carriers in the selected direction (in direction 1-2), we have $\mathscr{E}_{12} > 0$. If the e.m.f. prevents the motion of the positive carriers in the given direction, $\mathscr{E}_{12} < 0$.

Let us write Eq. (5.27) in the form

$$I = \frac{\varphi_1 - \varphi_2 + \mathscr{E}_{12}}{R}. \tag{5.28}$$

This equation expresses Ohm's law for an inhomogeneous circuit section. Assuming that $\varphi_1 = \varphi_2$, we get the equation of Ohm's law for a closed circuit:

$$I = \frac{\mathscr{E}}{R}. \tag{5.29}$$

Here, $\mathscr{E}$ is the e.m.f. acting in the circuit, and R is the total resistance of the entire circuit.

5.6. Multiloop Circuits. Kirchhoff's Rules

The calculation of multiloop circuits or networks is considerably simplified if we use two rules formulated by the German physicist Gustav Kirchhoff (1824-1887). The first of them relates to the junctions of a circuit. A **junction** is defined as a point where three or more conductors meet (Fig. 5.7). A current flowing toward a junction is considered to have one sign (plus or minus), and a current flowing out of a junction is considered to have the opposite sign (minus or plus).

Kirchhoff's first rule, also called the **junction rule**, states that *the algebraic sum*

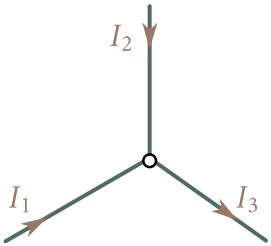


Fig. 5.7

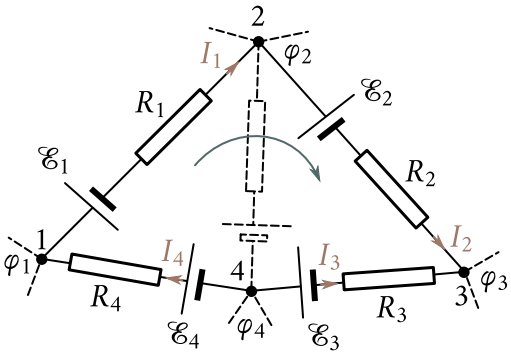


Fig. 5.8

of all the currents coming into a junction must be zero:

$$\sum_k I_k = 0. \tag{5.30}$$

This rule follows from the continuity equation, *i.e.*, in the long run from the law of charge conservation. For a steady current, $\nabla \cdot \boldsymbol{j}$ equals zero everywhere [see Eq. (5.21)]. Hence, the flux of the vector $\boldsymbol{j}$, *i.e.*, the algebraic sum of the currents flowing through an imaginary closed surface surrounding a junction, must be zero.

Equation (5.30) can be written for each of the N junctions of circuit. Only $N-1$ equations will be independent, however, whereas the N-th one will be a corollary of them.

The second rule relates to any closed loop separated from a multiloop circuit (see, for example, loop 1-2-3-4-1 in Fig. 5.8). Let us choose a direction of circumvention (for example, clockwise as in the figure) and apply Ohm's law to each unbranched loop section:

$$\begin{aligned} I_1 R_1 &= \varphi_1 - \varphi_2 + \mathscr{E}_1 \\ I_2 R_2 &= \varphi_2 - \varphi_3 + \mathscr{E}_2 \\ I_3 R_3 &= \varphi_3 - \varphi_4 + \mathscr{E}_3 \\ I_4 R_4 &= \varphi_4 - \varphi_1 + \mathscr{E}_4. \end{aligned}$$

When these expressions are summated, the potentials can be cancelled, and we get the equation

$$\sum_k I_k R_k = \sum_k \mathscr{E}_k, \tag{5.31}$$

that expresses **Kirchhoff's second rule**, also called the **loop rule**.

Equation (5.31) can be written for all the closed loops that can be separated mentally in a given multiloop circuit. Only the equations for the loops that cannot

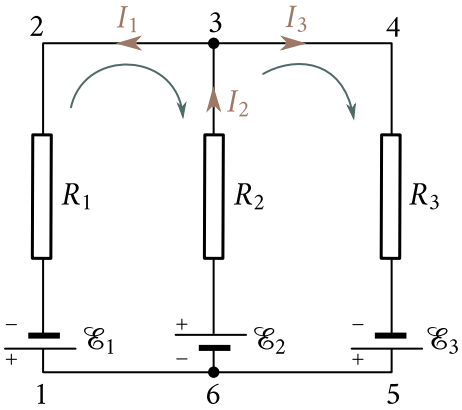


Fig. 5.9

be obtained by the superposition of other loops on one another will be independent, however. For example, for the circuit depicted in Fig. 5.9, we can write three equations:

(1) for loop 1-2-3-6-1,
(2) for loop 3-4-5-6-3, and
(3) for loop 1-2-3-4-5-6-1.

The last loop is obtained by superposition of the first two. The equations will, therefore, not be independent. We can take any two equations of the three as independent ones.

In writing the equations of the loop rule, we must appoint the signs of the currents and e.m.f.'s in accordance with the chosen direction of circumvention. For example, the current I_1 in Fig. 5.9 must be considered negative because it flows oppositely to the chosen direction of circumvention. The e.m.f. must also be considered negative because it acts in the direction opposite to that of circumvention, and so on.

We may choose the direction of circumvention in each loop absolutely arbitrarily and independently of the choice of the directions in the other loops. It may happen, here, that the same current or the same e.m.f. may be included in different equations with opposite signs (this happens with the current I_2 in Fig. 5.9 for the indicated directions of circumvention in the loops). This is of no significance, however, because a change in the direction around a loop results only in a reversal of all the signs in Eq. (5.31).

In compiling the equations, remember that the same current flows in any cross section of an unbranched part of a circuit. For example, the same current I_2 flows from junction 6 to the current source $\mathscr{E}_2$ as from the source $\mathscr{E}_2$ to junction 3.

The number of independent equations compiled in accordance with the junction and loop rules equals the number of different currents flowing in a multiloop circuit. Therefore, if we know the e.m.f.'s and resistances for all the unbranched sections, we can calculate all the currents. We can also solve other problems, for instance find the e.m.f.'s that must be connected in each of the sections of a circuit to obtain the required currents with the given resistances.

5.7. Power of a Current

Let us consider an arbitrary section of a steady current circuit across whose ends the voltage U is applied. The charge $q = It$ will flow during the time t through every cross section of the conductor. This is equivalent to the fact that the charge It is carried during the time t from one end of the conductor to the other. The forces of the electrostatic field and the extraneous forces acting on the given section do the work

$$A = Uq = UIt \tag{5.32}$$

[we remind our reader that the voltage U is determined as the work done by the electrostatic and extraneous forces in moving a unit positive charge; see Eq. (5.18)].

Dividing the work A by the time t during which it is done, we get the power developed by the current on the circuit section being considered:

$$P = UI = (\varphi_1 - \varphi_2)I + \mathscr{E}_{12} + I. \tag{5.33}$$

This power may be spent for the work done by the circuit section being considered on external bodies (for this purpose the section must move in space), for the proceeding of chemical reactions, and, finally, for heating the given circuit section.

The ratio of the power ΔP developed by a current in the volume ΔV of a conductor to the magnitude of this volume is called the **unit power of the current** P_{u}, corresponding to the given point of the conductor. By definition, the unit power is

$$P_{\text{u}} = \frac{\Delta P}{\Delta V}. \tag{5.34}$$

Speaking conditionally, the unit power is the power developed in unit volume of a conductor.

An expression for the unit power can be obtained proceeding from the following considerations. The force $e(\boldsymbol{E} + \boldsymbol{E}^*)$ develops a power of

$$P' = e(\boldsymbol{E} + \boldsymbol{E}^*)(\boldsymbol{v} + \boldsymbol{u})$$

upon the motion of a current carrier. Let us average this expression for the carriers confined in the volume ΔV within which $\boldsymbol{E}$ and $\boldsymbol{E}^*$ may be considered constant.

The result is

$$\langle P' \rangle = e(\boldsymbol{E} + \boldsymbol{E}^*) \langle \boldsymbol{v} + \boldsymbol{u} \rangle = e(\boldsymbol{E} + \boldsymbol{E}^*) \langle \boldsymbol{v} \rangle + e(\boldsymbol{E} + \boldsymbol{E}^*) \langle \boldsymbol{u} \rangle = e(\boldsymbol{E} + \boldsymbol{E}^*) \langle \boldsymbol{u} \rangle$$

(remember that $\langle \boldsymbol{v} \rangle = 0$).

We can find the power ΔP developed in the volume ΔV by multiplying $\langle P' \rangle$ by the number of current carriers in this volume, *i.e.*, by $n\Delta V$ (n is the number of carriers in unit volume). Thus,

$$\Delta P = \langle P' \rangle \, n\Delta V = e(\boldsymbol{E} + \boldsymbol{E}^*) \cdot \langle \boldsymbol{u} \rangle \, n\Delta V = \boldsymbol{j} \cdot (\boldsymbol{E} + \boldsymbol{E}^*) \Delta V$$

[see Eq. (5.23)]. Hence,

$$P_{\text{u}} = \boldsymbol{j} \cdot (\boldsymbol{E} + \boldsymbol{E}^*) \tag{5.35}$$

This expression is a differential form of the integral equation (5.33).

5.8. The Joule-Lenz Law

When a conductor is stationary and no chemical transformations occur in it, the work of a current given by Eq. (5.32) goes to increase the internal energy of the conductor, and as a result the latter gets heated. It is customary to say that when a current flows in a conductor, the heat

$$Q = UIt$$

is liberated. Substituting RI for U in accordance with Ohm's law, we get the formula

$$Q = RI^2 t. \tag{5.36}$$

Equation (5.36) was established experimentally by the British physicist James Joule (1818-1889) and independently of him by the Russian physicist Emil Lenz (1804-1865), and is called the **Joule-Lenz law**.

If the current varies with time, then the amount of heat liberated during the time t is calculated by the equation

$$Q = \int_0^t RI^2 \, \mathrm{d}t. \tag{5.37}$$

We can pass over from Eq. (5.36) determining the heat liberated in an entire conductor to an expression characterizing the liberation of heat at different spots of the conductor. Let us separate in a conductor, in the same way as we did in deriving Eq. (5.22), an elementary volume in the form of a cylinder (see Fig. 5.4). According to the Joule-Lenz law, the following amount of heat will be liberated in this volume during the time $\mathrm{d}t$:

$$\mathrm{d}Q = RI^2 \, \mathrm{d}t = \frac{\rho \, \mathrm{d}l}{\mathrm{d}S} (j \, \mathrm{d}S)^2 \, \mathrm{d}t = \rho j^2 \, \mathrm{d}V \, \mathrm{d}t \tag{5.38}$$

($\mathrm{d}V = \mathrm{d}S \, \mathrm{d}l$ is the magnitude of the elementary volume).

Dividing Eq. (5.38) by dV and dt, we shall find the amount of heat liberated in unit volume per unit time:

$$Q_u = \rho j^2. \tag{5.39}$$

By analogy with the name of quantity Eq. (5.34), the quantity Q_u can be called the **unit thermal power of a current**.

Equation (5.39) is a differential form of the Joule-Lenz law. It can be obtained from Eq. (5.35). Substituting $\boldsymbol{j}/\sigma = \rho \boldsymbol{j}$ for $\boldsymbol{E} + \boldsymbol{E}^*$ in Eq. (5.35) [see Eq. (5.25)], we arrive at the expression

$$P_u = \rho \boldsymbol{j}^2,$$

that coincides with Eq. (5.39).

It must be noted that Joule and Lenz established their law for a homogeneous circuit section. As follows from what has been said in the present section, however, Eqs. (5.36) and (5.39) also hold for an inhomogeneous section provided that the extraneous forces acting in it have a non-chemical origin.

Chapter 6

MAGNETIC FIELD IN A VACUUM

6.1. Interaction of Currents

Experiments show that electric currents exert a force on one another. For example, two thin straight parallel conductors carrying a current (we shall call them line currents) attract each other if the currents in them flow in the same direction, and repel each other if the currents flow in opposite directions. The force of interaction per unit length of each of the parallel conductors is proportional to the magnitudes of the currents I_1 and I_2 in them and inversely proportional to the distance b between them:

$$F_{\mathrm{u}} = k\frac{2I_1 I_2}{b}. \tag{6.1}$$

We have designated the proportionality constant $2k$ for reasons that will become clear on a later page.

The law of interaction of currents was established in 1820 by the French physicist Andre Ampere (1775-1836). A general expression of this law suitable for conductors of any shape will be given in Sec. 6.6. Equation (6.1) is used to establish the unit of current in the SI and in the absolute electromagnetic system (cgsm) of units. The SI unit of current—the **ampere**—is defined as the constant current which, if maintained in two straight parallel conductors of infinite length, of negligible cross section, and placed 1 metre apart in vacuum, would produce between these conductors a force equal to 2×10^{-7} newton per metre of length.

The unit of charge, called the **coulomb**, is defined as the charge passing in 1 second through the cross section of a conductor in which a constant current of 1 ampere is flowing. Accordingly, the coulomb is also called the **ampere-second** (A s).

Equation (6.1) is written in the rationalized form as follows:

$$F_{\mathrm{u}} = \frac{\mu_0}{4\pi}\frac{2I_1 I_2}{b}, \tag{6.2}$$

where μ_0 is the so-called **magnetic constant** [compare with Eq. (1.11)]. To find the numerical value of μ_0, we shall take advantage of the fact that according to the definition of the ampere, when $I_1 = I_2 = 1$ A and $b = 1$ m, the force F_{u} is obtained equal to 2×10^{-7} N m^{-1}.

Let us use these values in Eq. (6.2):

$$2 \times 10^{-7} = \frac{\mu_0}{4\pi}\frac{2 \times 1 \times 1}{1}.$$

Hence,

$$\mu_0 = 4\pi \times 10^{-7} = 1.26 \times 10^{-6}\ \mathrm{H\,m^{-1}} \tag{6.3}$$

(the symbol H m^{-1} stands for henry per metre—see Sec. 8.5).

The constant k in Eq. (6.1) can be made equal to unity by choosing an appropriate unit of current. This is how the absolute electromagnetic unit of current (cgsm$_I$) is established. It is defined as the current which, if maintained in a thin straight conductor of infinite length, would act on an equal and parallel line current at a distance of 1 cm from it with a force equal to 2 dyn per centimetre of length.

In the cgse system, the constant k is a dimension quantity other than unity. According to Eq. (6.1), the dimension of k is determined as follows:

$$[k] = \frac{[F_{\mathrm{u}}b]}{[I]^2} = \frac{[F]}{[I]^2}. \tag{6.4}$$

We have taken into account that the dimension of F_{u} is the dimension of force divided by the dimension of length; hence, the dimension of the product $F_{\mathrm{u}}b$ is that of force. According to Eqs. (1.7) and (5.7):

$$[F] = \frac{[q]^2}{\mathrm{L}^2}; \quad [I] = \frac{[q]}{\mathrm{T}}.$$

Using these values in Eq. (6.4), we find that

$$[k] = \frac{\mathrm{T}^2}{\mathrm{L}^2}.$$

Consequently, in the cgse system, k can be written in the form

$$k = \frac{1}{c^2}, \tag{6.5}$$

where c is a quantity having the dimension of velocity and called the **electromagnetic constant**. To find its value, let us use relation (1.8) between the coulomb and the cgse unit of charge, which was established experimentally. A force of 2×10^{-7} N m^{-1} is equivalent to 2×10^{-4} dyn cm^{-1}. According to Eq. (6.1), this is the force with which currents of 3×10^9 cgse$_I$ (*i.e.*, 1 A) each interact when $b = 100$ cm.

Thus,

$$2 \times 10^{-4} = \frac{1}{c^2} \frac{2 \times 3 \times 10^9 \times 3 \times 10^9}{100},$$

whence

$$c = 3 \times 10^{10} \text{ cm s}^{-1} = 3 \times 10^8 \text{ m s}^{-1}. \tag{6.6}$$

The value of the electromagnetic constant coincides with that of the speed of light in a vacuum. From J. Maxwell's theory, there follows the existence of electromagnetic waves whose speed in a vacuum equals the electromagnetic constant c. The coincidence of c with the speed of light in a vacuum gave Maxwell the grounds to assume that light is an electromagnetic wave.

The value of k in Eq. (6.1) is 1 in the cgsm system and $1/c^2 = 1/\left(3 \times 10^{10}\right)^2$ $\text{s}^2\,\text{cm}^{-2}$ in the cgse system. Hence, it follows that a current of 1 cgsm_I is equivalent to a current of 3×10^{10} cgse_I:

$$1 \text{ cgsm}_I = 3 \times 10^{10} \text{ cgse}_I = 10 \text{ A}. \tag{6.7}$$

Multiplying this relation by 1 s, we get

$$1 \text{ cgsm}_q = 3 \times 10^{10} \text{ cgse}_q = 10 \text{ C}. \tag{6.8}$$

Thus,

$$I_{\text{cgsm}} = \frac{1}{c} I_{\text{cgse}}. \tag{6.9}$$

Accordingly,

$$q_{\text{cgsm}} = \frac{1}{c} q_{\text{cgse}}. \tag{6.10}$$

There is a definite relation between the constants ε_0, μ_0, and c. To establish it, let us find the dimension and numerical value of the product $\varepsilon_0 \mu_0$. In accordance with Eq. (1.11), the dimension of ε_0 is

$$[\varepsilon_0] = \frac{[q]^2}{\text{L}^2 [F]}. \tag{6.11}$$

According to Eq. (6.2)

$$[\mu_0] = \frac{[F_{\text{u}} b]}{[I]^2} = \frac{[F]\text{T}^2}{[q]^2}. \tag{6.12}$$

Multiplication of Eqs. (6.11) and (6.12) yields

$$[\varepsilon_0 \mu_0] = \frac{\text{T}^2}{\text{L}^2} = \frac{1}{[v]^2} \tag{6.13}$$

(v is the speed).

With account taken of Eqs. (1.11) and (6.3), the numerical value of the product

$\varepsilon_0\mu_0$ is

$$\varepsilon_0\mu_0 = \frac{1}{4\pi \times 9 \times 10^9} \times 4\pi \times 10^{-7} = \frac{1}{\left(3 \times 10^8\right)^2} \text{ s}^2 \text{ cm}^{-2}. \tag{6.14}$$

Finally, taking into account Eqs. (6.6), (6.13), and (6.14), we get the relation interesting us:

$$\varepsilon_0\mu_0 = \frac{1}{c^2}. \tag{6.15}$$

6.2. Magnetic Field

Currents interact through a field called **magnetic**. This name originated from the fact that, as the Danish physicist Hans Oersted (1777-1851) discovered in 1820, the field set up by a current has an orienting action on a magnetic pointer. Oersted stretched a wire carrying a current over a magnetic pointer rotating on a needle. When the current was switched on, the pointer aligned itself at right angles to the wire. Reversing of the current caused the pointer to rotate in the opposite direction.

Oersted's experiment shows that a magnetic field has a sense of direction and must be characterized by a vector quantity. The latter is designated by the symbol $\boldsymbol{B}$. It would be logical to call $\boldsymbol{B}$ the magnetic field strength, by analogy with the electric field strength $\boldsymbol{E}$. For historical reasons, however, the basic force characteristic of a magnetic field was called the **magnetic induction**. The name magnetic field strength was given to an auxiliary quantity $\boldsymbol{H}$ similar to the auxiliary characteristic $\boldsymbol{D}$ of an electric field.

A magnetic field, unlike its electric counterpart, does not act on a charge at rest. A force appears only when a charge is moving.

A current-carrying conductor is an electrically neutral system of charges in which the charges of one sign are moving in one direction, and the charges of the other sign in the opposite direction (or are at rest). It thus follows that a magnetic field is set up by moving charges.

Thus, moving charges (currents) change the properties of the space surrounding them—they set up a magnetic field in it. This field manifests itself in that forces are exerted on charges moving in it (currents).

Experiments show that the superposition principle holds for a magnetic field, the same as for an electric field: the *field $\boldsymbol{B}$ set up by several moving charges (currents) equals the vector sum of the fields $\boldsymbol{B}_i$ set up by each charge (current) separately:*

$$\boldsymbol{B} = \sum_i \boldsymbol{B}_i \tag{6.16}$$

[compare with Eq. (1.19)].

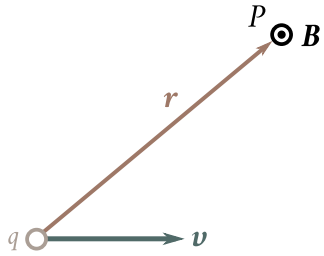


Fig. 6.1

6.3. Field of a Moving Charge

Space is isotropic, consequently, if a charge is stationary, then all directions have equal rights. This underlies the fact that the electrostatic field set up by a point charge is spherically symmetrical.

If a charge travels with the velocity $\boldsymbol{v}$, a preferred direction (that of the vector $\boldsymbol{v}$) appears in space. We can, therefore, expect the magnetic field produced by a moving charge to have axial symmetry. We must note that we have in mind free motion of a charge, *i.e.*, motion with a constant velocity. For an acceleration to appear, the charge must experience the action of a field (electric or magnetic). This field by its very existence would violate the isotropy of space.

Let us consider the magnetic field set up at point P by the point charge q travelling with the constant velocity $\boldsymbol{v}$ (Fig. 6.1). The disturbances of the field are transmitted from point to point with the finite velocity c. For this reason, the induction $\boldsymbol{B}$ at point P at the moment of time t is determined not by the position of the charge at the same moment t, but by its position at an earlier moment of time $t - \tau$:

$$\boldsymbol{B}(P, t) = f\{q, \boldsymbol{v}, \boldsymbol{r}(t - \tau)\}.$$

Here, P signifies the collection of the coordinates of point P determined in a stationary reference frame, and $\boldsymbol{r}(t - \tau)$ is the position vector drawn to point P from the point where the charge was at the moment $t - \tau$.

If the velocity of the charge is much smaller than c ($v \ll c$), then the retardation time τ will be negligibly small. In this case, we can consider that the value of $\boldsymbol{B}$ at the moment t is determined by the position of the charge at the same moment t. If this condition is observed, then

$$\boldsymbol{B}(P, t) = f\{q, \boldsymbol{v}, \boldsymbol{r}(t)\} \tag{6.17}$$

[we remind our reader that $\boldsymbol{v}$ = constant, therefore, $\boldsymbol{v}(t - \tau) = \boldsymbol{v}(t)$].

The form of function (6.17) can be established only experimentally. But before giving the results of experiments, let us try to find the logical form of this relation.

The simplest assumption is that the magnitude of the vector $\boldsymbol{B}$ is proportional to the charge q and the velocity v (when $\boldsymbol{v} \to 0$, a magnetic field is absent). We have to "construct" the vector $\boldsymbol{B}$ we are interested in from the scalar q and the two given vectors $\boldsymbol{v}$ and $\boldsymbol{r}$. This can be done by vector multiplication of the given vectors and then by multiplying their product by the scalar. The result is the expression

$$q(\boldsymbol{v} \times \boldsymbol{r}). \tag{6.18}$$

The magnitude of this expression grows with an increasing distance from the charge (with increasing r). It is improbable that the characteristic of a field will behave in this way—for the fields that we know (electrostatic, gravitational), the field does not grow with an increasing distance from the source, but, on the contrary, weakens, varying in proportion to $1/r^2$. Let us assume that the magnetic field of a moving charge behaves in the same way when r changes. We can obtain an inverse proportion to the square of r by dividing Eq. (6.18) by r^3. The result is

$$\frac{q(\boldsymbol{v} \times \boldsymbol{r})}{r^3}. \tag{6.19}$$

Experiments show that when $v \ll c$, the magnetic induction of the field of a moving charge is determined by the formula

$$\boldsymbol{B} = k' \frac{q(\boldsymbol{v} \times \boldsymbol{r})}{r^3}, \tag{6.20}$$

where k' is a proportionality constant.

We must stress once more that the reasoning which led us to expression (6.19) must by no means be considered as the derivation of Eq. (6.20). This reasoning does not have conclusive force. Its aim is to help us understand and memorize Eq. (6.20). This equation itself can be obtained only experimentally.

It can be seen from Eq. (6.20) that the vector $\boldsymbol{B}$ at every point P is directed at right angles to the plane passing through the direction of the vector $\boldsymbol{v}$ and point P, so that rotation in the direction of $\boldsymbol{B}$ forms a right-handed system with the direction of $\boldsymbol{v}$ (see the circle with the dot in Eq. (6.1)). We must note that $\boldsymbol{B}$ is a pseudo vector. The value of the proportionality constant k' depends on our choice of the units of the quantities in Eq. (6.20). This equation is written in the rationalized form as follows:

$$\boldsymbol{B} = \frac{\mu_0}{4\pi} \frac{q(\boldsymbol{v} \times \boldsymbol{r})}{r^3}. \tag{6.21}$$

This equation can be written in the form

$$\boldsymbol{B} = \frac{\mu_0}{4\pi} \frac{q(\boldsymbol{v} \times \hat{\boldsymbol{e}}_r)}{r^3} \tag{6.22}$$

[compare with Eq. (1.15)]. It must be noted that in similar equations when ε_0 is in the denominator, μ_0 is in the numerator, and vice versa.

The SI unit of magnetic induction is called the **tesla** (T) in honour of the Croatian electrician and inventor Nikola Tesla (1856-1943).

The units of the magnetic induction B are chosen in the cgse and cgsm systems so that the constant k' in Eq. (6.20) equals unity. Hence, the same relation holds between the units of B in these systems as between the units of charge:

$$1 \text{ cgsm}_B = 3 \times 10^{10} \text{ cgse}_B \tag{6.23}$$

[see Eq. (6.8)].

The cgsm unit of magnetic induction has a special name—the **gauss** (Gs).

The German mathematician Karl Gauss (1777-1855) proposed a system of units in which all the electrical quantities (charge, current, electric field strength, etc.) are measured in cgse units, and all the magnetic quantities (magnetic induction, magnetic moment, etc.) in cgsm units. This system of units was named the **Gaussian** one, in honour of its author.

In the Gaussian system, owing to Eqs. (6.9) and (6.10), all the equations containing the current or charge in addition to magnetic quantities include one multiplier $1/c$ for each quantity I or q in the relevant equation. This multiplier converts the value of the pertinent quantity (I or q) expressed in cgse units to a value expressed in cgsm units (the cgsm system of units is constructed so that the proportionality constants in all the equations equal 1). For example, in the Gaussian system, Eq. (6.20) has the form

$$\boldsymbol{B} = \frac{1}{c}\frac{q(\boldsymbol{v} \times \boldsymbol{r})}{r^3}. \tag{6.24}$$

We must note that the appearance of a preferred direction in space (the direction of the vector $\boldsymbol{v}$) when a charge moves leads to the electric field of the moving charge also losing its spherical symmetry and becoming axially symmetrical. The relevant calculations show that the $\boldsymbol{E}$ lines of the field of a freely moving charge have the form shown in Fig. 6.2. The vector $\boldsymbol{E}$ at point P is directed along the position vector $\boldsymbol{r}$ drawn from the point where the charge is at the given moment to point P. The magnitude of the field strength is determined by the equation

$$E = \frac{1}{4\pi\varepsilon_0}\frac{q}{r^2}\frac{1 - (v^2/c^2)}{\left[1 - (v^2/c^2)\sin^2\theta\right]^{3/2}}, \tag{6.25}$$

where θ is the angle between the direction of the velocity $\boldsymbol{v}$ and the position vector $\boldsymbol{r}$.

When $v \ll c$, the electric field of a freely moving charge at each moment of time does not virtually differ from the electrostatic field set up by a stationary charge at the point where the moving charge is at the given moment. It must be remembered, however, that this "electrostatic" field moves together with the charge. Hence, the field at each point of space changes with time.

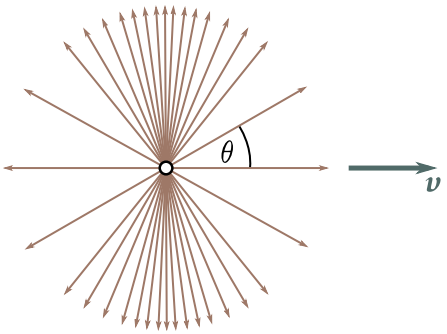


Fig. 6.2

At values of v comparable with c, the field in directions at right angles to $\boldsymbol{v}$ is appreciably stronger than in the direction of motion at the same distance from the charge (see Fig. 6.2 drawn for $v/c = 0.8$). The field "flattens out" in the direction of motion and is concentrated mainly near a plane passing through the charge and perpendicular to the vector $\boldsymbol{v}$.

6.4. The Biot-Savart Law

Let us determine the nature of the magnetic field set up by an arbitrary thin wire through which a current flows. We shall consider a small element of the wire of length $\mathrm{d}l$. This element contains $nS\,\mathrm{d}l$ current carriers (n is the number of carriers in a unit volume, and S is the cross-sectional area of the wire where the element $\mathrm{d}l$ has been taken). At the point whose position relative to the element $\mathrm{d}l$ is determined by the position vector $\boldsymbol{r}$ (Fig. 6.3), a separate carrier of current e sets up a field with the induction

$$\boldsymbol{B} = \frac{\mu_0}{4\pi}\frac{e[(\boldsymbol{v}+\boldsymbol{u})\times\boldsymbol{r}]}{r^3}$$

[see Eq. (6.21)]. Here, $\boldsymbol{v}$ is the velocity of chaotic motion, and $\boldsymbol{u}$ is the velocity of ordered motion of the carrier.

The value of the magnetic induction averaged over the current carriers in the element $\mathrm{d}l$ is

$$\langle\boldsymbol{B}\rangle = \frac{\mu_0}{4\pi}\frac{e[(\langle\boldsymbol{v}\rangle+\langle\boldsymbol{u}\rangle)\times\boldsymbol{r}]}{r^3} = \frac{\mu_0}{4\pi}\frac{e(\langle\boldsymbol{u}\rangle\times\boldsymbol{r})}{r^3}$$

($\langle\boldsymbol{v}\rangle = 0$). Multiplying this expression by the number of carriers in an element of the wire (equal to $nS\,\mathrm{d}l$), we get the contribution to the field introduced by the element $\mathrm{d}l$:

$$\mathrm{d}\boldsymbol{B} = \langle\boldsymbol{B}\rangle\, nS\,\mathrm{d}l = \frac{\mu_0}{4\pi}\frac{S[(ne\langle\boldsymbol{u}\rangle)\times\boldsymbol{r}]\,\mathrm{d}l}{r^3}$$

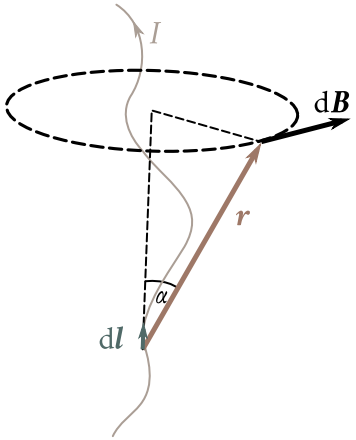


Fig. 6.3

(we have put the scalar multipliers n and e inside the sign of the vector product). Taking into account that $ne\langle \boldsymbol{u} \rangle = \boldsymbol{j}$, we can write

$$\mathrm{d}\boldsymbol{B} = \frac{\mu_0}{4\pi} \frac{S(\boldsymbol{j} \times \boldsymbol{r})\,\mathrm{d}l}{r^3}. \tag{6.26}$$

Let us introduce the vector $\mathrm{d}\boldsymbol{l}$ directed along the axis of the current element $\mathrm{d}l$ in the same direction as the current. The magnitude of this vector is $\mathrm{d}l$. Since the directions of the vectors $\boldsymbol{j}$ and $\mathrm{d}\boldsymbol{l}$ coincide, we can write the equation

$$\boldsymbol{j}\,\mathrm{d}l = j\,\mathrm{d}\boldsymbol{l}. \tag{6.27}$$

Performing such a substitution in Eq. (6.26), we get

$$\mathrm{d}\boldsymbol{B} = \frac{\mu_0}{4\pi} \frac{Sj(\mathrm{d}\boldsymbol{l} \times \boldsymbol{r})}{r^3}.$$

Finally, taking into account that the product Sj gives the current I in the wire, we arrive at the final expression determining the magnetic induction of the field set up by a current element of length $\mathrm{d}l$:

$$\mathrm{d}\boldsymbol{B} = \frac{\mu_0}{4\pi} \frac{I(\mathrm{d}\boldsymbol{l} \times \boldsymbol{r})}{r^3}. \tag{6.28}$$

We have derived Eq. (6.28) from Eq. (6.21). Equation (6.28) was actually established experimentally before Eq. (6.21) was known. Moreover, the latter equation was derived Eq. (6.28).

In 1820, the French physicists Jean Biot (1774-1862) and Felix Savart (1791-1841) studied the magnetic fields flowing along thin wires of various shape. The French astronomer and mathematician Pierre Laplace (1749-1827) analysed the experimental data obtained and found that the magnetic field of any current can be calculated as the vector sum (superposition) of the fields set up by the separate elementary

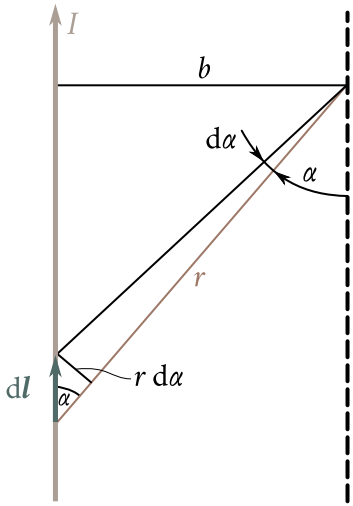


Fig. 6.4

sections of the currents. Laplace obtained Eq. (6.28) for the magnetic induction of the field set up by a current element of length dl. In this connection, Eq. (6.28) is called the **Biot-Savart-Laplace law**, or more briefly the **Biot-Savart law**. A glance at Fig. 6.3 shows that the vector d$\boldsymbol{B}$ is directed at right angles to the plane passing through d$\boldsymbol{l}$ and the point for which the field is being calculated so that rotation about d$\boldsymbol{l}$ in the direction of d$\boldsymbol{B}$ is associated with d$\boldsymbol{l}$ by the right-hand screw rule. The magnitude of d$\boldsymbol{B}$ is determined by the expression

$$\mathrm{d}B = \frac{\mu_0}{4\pi}\frac{I\,\mathrm{d}l\,\sin\theta}{r^3}, \tag{6.29}$$

where α is the angle between the vectors d$\boldsymbol{l}$ and $\boldsymbol{r}$.

Let us use Eq. (6.28) to calculate the field of a line current, *i.e.*, the field set up by a current flowing through a thin straight wire of infinite length (Fig. 6.4). All the vectors d$\boldsymbol{B}$ at a given point have the same direction (in our case beyond the drawing). Therefore, addition of the vectors d$\boldsymbol{B}$ may be replaced with addition of their magnitudes. The point for which we are calculating the magnetic induction is at the distance b from the wire.

Inspection of Fig. 6.4 shows that

$$r = \frac{b}{\sin\alpha}, \quad \mathrm{d}l = \frac{r\,\mathrm{d}\alpha}{\sin\alpha} = \frac{b\,\mathrm{d}\alpha}{\sin^2\alpha}.$$

Let us introduce these values into Eq. (6.29):

$$\mathrm{d}B = \frac{\mu_0}{4\pi}\frac{Ib\,\mathrm{d}\alpha\sin\alpha\sin^2\alpha}{b^2\sin^2\alpha} = \frac{\mu_0}{4\pi}\frac{I}{b}\sin\alpha\,\mathrm{d}\alpha.$$

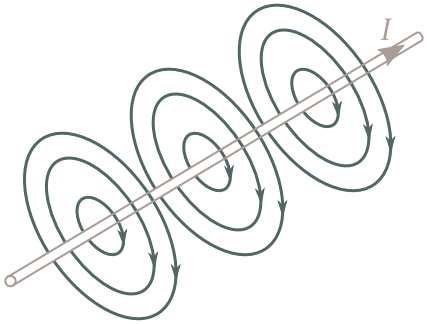


Fig. 6.5

The angle α varies within the limits from 0 to π for all the elements of an infinite line current. Hence,

$$B = \int \mathrm{d}B = \frac{\mu_0}{4\pi}\frac{I}{b}\int_0^{\pi} \sin\alpha \,\mathrm{d}\alpha = \frac{\mu_0}{4\pi}\frac{2I}{b}.$$

Thus, the magnetic induction of the field of a line current is determined by the formula

$$B = \frac{\mu_0}{4\pi}\frac{2I}{b}. \tag{6.30}$$

The magnetic induction lines of the field of a line current are a system of concentric circles surrounding the wire (Fig. 6.5).

6.5. The Lorentz Force

A charge moving in a magnetic field experiences a force which we shall call **magnetic**. The force is determined by the charge q, its velocity $\boldsymbol{v}$, and the magnetic induction $\boldsymbol{B}$ at the point where the charge is at the moment of time being considered. The simplest assumption is that the magnitude of the force F is proportional to each of the three quantities q, v, and B. In addition, F can be expected to depend on the mutual orientation of the vectors $\boldsymbol{v}$ and $\boldsymbol{B}$. The direction of the vector $\boldsymbol{F}$ should be determined by those of the vectors $\boldsymbol{v}$ and $\boldsymbol{B}$.

To "construct" the vector $\boldsymbol{F}$ from the scalar q and the vectors $\boldsymbol{v}$ and $\boldsymbol{B}$, let us find the vector product of $\boldsymbol{v}$ and $\boldsymbol{B}$ and then multiply the result obtained by the scalar q. The result is the expression

$$q(\boldsymbol{v} \times \boldsymbol{B}). \tag{6.31}$$

It has been established experimentally that the force $\boldsymbol{F}$ acting on a charge moving in a magnetic field is determined by the formula

$$\boldsymbol{F} = kq(\boldsymbol{v} \times \boldsymbol{B}), \tag{6.32}$$

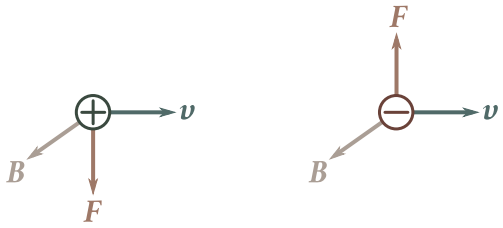


Fig. 6.6

where k is a proportionality constant depending on the choice of the units for the quantities in the formula.

It must be borne in mind that the reasoning which led us to expression (6.31) must by no means be considered as the derivation of Eq. (6.32). This reasoning does not have conclusive force. Its aim is to help us memorize Eq. (6.32). The correctness of this equation can be established only experimentally.

We must note that Eq. (6.32) can be considered as a definition of the magnetic induction $\boldsymbol{B}$.

The unit of magnetic induction $\boldsymbol{B}$—the tesla—-is determined so that the proportionality constant k in Eq. (6.32) equals unity. Hence, in SI units, this equation becomes

$$\boldsymbol{F} = q(\boldsymbol{v} \times \boldsymbol{B}). \tag{6.33}$$

The magnitude of the magnetic force is

$$F = qvB \sin \alpha, \tag{6.34}$$

where α is the angle between the vectors $\boldsymbol{v}$ and $\boldsymbol{B}$. It can be seen from Eq. (6.34) that a charge moving along the lines of a magnetic field does not experience the action of a magnetic force.

The magnetic force is directed at right angles to the plane containing the vectors $\boldsymbol{v}$ and $\boldsymbol{B}$. If the charge q is positive, then the direction of the force coincides with that of the vector $\boldsymbol{v} \times \boldsymbol{B}$. When q is negative, the directions of the vectors $\boldsymbol{F}$ and $\boldsymbol{v} \times \boldsymbol{B}$ are opposite (Fig. 6.6).

Since the magnetic force is always directed at right angles to the velocity of a charged particle, it does no work on the particle. Hence, we cannot change the energy of a charged particle by acting on it with a constant magnetic field.

The force exerted on a charged particle that is simultaneously in an electric and a magnetic field is

$$\boldsymbol{F} = q\boldsymbol{E} + q(\boldsymbol{v} \times \boldsymbol{B}). \tag{6.35}$$

This expression was obtained from the results of experiments by the Dutch physicist Hendrik Lorentz (1853-1928) and is called the **Lorentz force**.

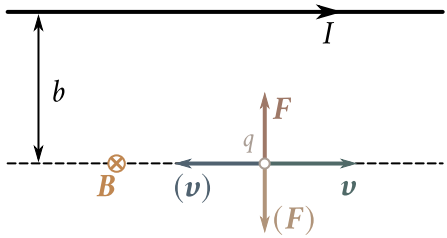


Fig. 6.7

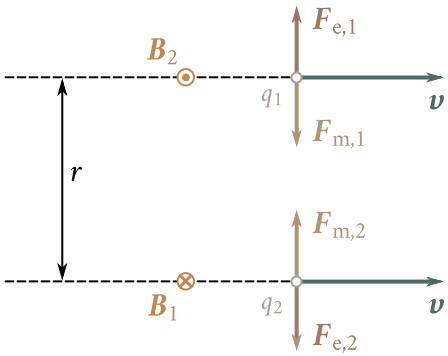


Fig. 6.8

Assume that the charge q is moving with the velocity $\boldsymbol{v}$ parallel to a straight infinite wire along which the current I flows (Fig. 6.7).

According to Eqs. (6.30) and (6.34), the charge in this case experiences a magnetic force whose magnitude is

$$F = qvB = qv\frac{\mu_0}{4\pi}\frac{2I}{b}, \tag{6.36}$$

where b is the distance from the charge to the wire. The force is directed toward the wire when the charge is positive if the directions of the current and motion of the charge are the same, and away from the wire if these directions are opposite (see Fig. 6.7). When the charge is negative, the direction of the force is reversed, the other conditions being equal.

Let us consider two like point charges q_1 and q_2 moving along parallel straight lines with the same velocity v that is much smaller than c (Fig. 6.8). When $v \ll c$, the electric field does not virtually differ from the field of stationary charges (see Sec. 6.3). Therefore, the magnitude of the electric force F_e exerted on the charges can be considered equal to

$$F_{\mathrm{e},1} = F_{\mathrm{e},2} = F_\mathrm{e} = \frac{1}{4\pi\varepsilon_0}\frac{q_1 q_2}{r^2}. \tag{6.37}$$

Equations (6.21) and (6.3) give us the following expression for the magnetic force F_m exerted on the charges:

$$F_{\mathrm{m},1} = F_{\mathrm{m},2} = F_\mathrm{m} = \frac{\mu_0}{4\pi}\frac{q_1 q_2 v^2}{r^2} \tag{6.38}$$

(the position vector $\boldsymbol{r}$ is perpendicular to $\boldsymbol{v}$).

Let us find the ratio between the magnetic and electric forces. It follows from Eqs. (6.37) and (6.38) that

$$\frac{F_\mathrm{m}}{F_\mathrm{e}} = \varepsilon_0\mu_0 v^2 = \frac{v^2}{c^2} \tag{6.39}$$

[see Eq. (6.15)]. We have obtained Eq. (6.39) on the assumption that $v \ll c$. This ratio holds, however, with any v's.

The forces $\boldsymbol{F}_e$ and $\boldsymbol{F}_m$ are directed oppositely. Figure 6.8 has been drawn for like and positive charges. For like negative charges, the directions of the forces will remain the same, while the directions of the vectors $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ will be reversed. For unlike charges, the directions of the electric and magnetic forces will be the reverse of those shown in the figure.

Inspection of Eq. (6.39) shows that the magnetic force is weaker than the Coulomb one by a factor equal to the square of the ratio of the speed of the charge to that of light. The explanation is that the magnetic interaction between moving charges is a relativistic effect (see Sec. 6.7). Magnetism would disappear if the speed of light were infinitely great.

6.6. Ampere's Law

If a wire carrying a current is in a magnetic field, then each of the current carriers experiences the force

$$\boldsymbol{F} = e[(\boldsymbol{v} + \boldsymbol{u}) \times \boldsymbol{B}] \tag{6.40}$$

[see Eq. (6.33)]. Here, $\boldsymbol{v}$ is the velocity of chaotic motion of a carrier, and $\boldsymbol{u}$ is the velocity of ordered motion. The action of this force is transferred from a current carrier to the conductor along which it is moving. As a result, a force acts on a wire with current in a magnetic field.

Let us find the value of the force $\mathrm{d}\boldsymbol{F}$ exerted on an element of a wire of length $\mathrm{d}l$. We shall average Eq. (6.40) over the current carriers contained in the element $\mathrm{d}l$:

$$\langle \boldsymbol{F} \rangle = e[(\langle \boldsymbol{v} \rangle + \langle \boldsymbol{u} \rangle) \times \boldsymbol{B}] = e(\langle \boldsymbol{u} \rangle \times \boldsymbol{B}) \tag{6.41}$$

($\boldsymbol{B}$ is the magnetic induction at the place where the element $\mathrm{d}l$ is). The wire element contains $nS\,\mathrm{d}l$ carriers (n is the number of carriers in unit volume, and S is the cross-sectional area of the wire at the given place). Multiplying Eq. (6.41) by the number of carriers, we find the force we are interested in:

$$\mathrm{d}\boldsymbol{F} = \langle \boldsymbol{F} \rangle\, nS\,\mathrm{d}l = [(ne\,\langle \boldsymbol{u} \rangle) \times \boldsymbol{B}] S\,\mathrm{d}l.$$

Taking into account that $ne\,\langle \boldsymbol{u} \rangle$ is the current density $\boldsymbol{j}$, and $S\,\mathrm{d}l$ gives the volume of a wire element $\mathrm{d}V$, we can write

$$\mathrm{d}\boldsymbol{F} = (\boldsymbol{j} \times \boldsymbol{B})\,\mathrm{d}V. \tag{6.42}$$

Hence, we can obtain an expression for the density of the force, *i.e.*, for the force acting on unit volume of the conductor

$$\boldsymbol{F}_{u.v} = \boldsymbol{j} \times \boldsymbol{B}. \tag{6.43}$$

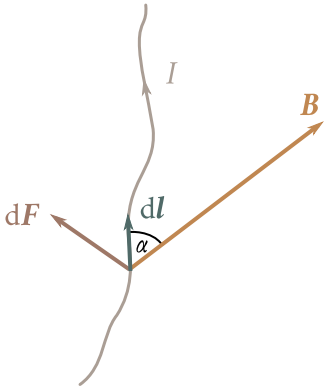


Fig. 6.9

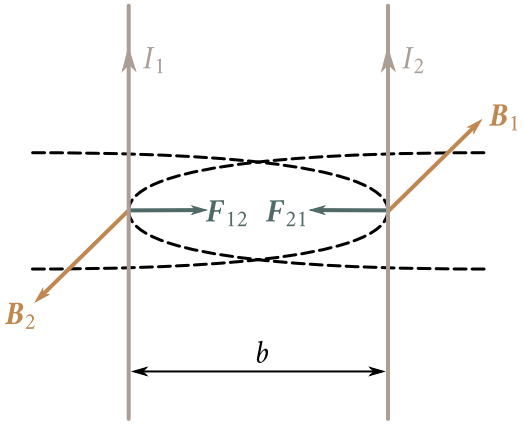


Fig. 6.10

Let us write Eq. (6.42) in the form

$$d\boldsymbol{F} = (\boldsymbol{j} \times \boldsymbol{B}) S\, dl.$$

Replacing in accordance with Eq. (6.27) $\boldsymbol{j}S\, dl$ with $jS\, d\boldsymbol{l} = I\, d\boldsymbol{l}$, we arrive at the equation

$$d\boldsymbol{F} = I(d\boldsymbol{l} \times \boldsymbol{B}). \tag{6.44}$$

This equation determines the force exerted on a current element $d\boldsymbol{l}$ in a magnetic field. Equation (6.44) was established experimentally by Ampere and is called **Ampere's law**.

We have obtained Ampere's law on the basis of Eq. (6.33) for the magnetic force. The expression for the magnetic force was actually obtained from the experimentally established equation (6.44).

The magnitude of the force (6.44) is calculated by the equation

$$dF = IB\, dl\, \sin\alpha, \tag{6.45}$$

where α is the angle between the vectors $d\boldsymbol{l}$ and $\boldsymbol{B}$ (Fig. 6.9). The force is normal to the plane containing the vectors $d\boldsymbol{l}$ and $\boldsymbol{B}$.

Let us use Ampere's law to calculate the force of interaction between two parallel infinitely long line currents in a vacuum. If the distance between the currents is b (Fig. 6.10), then each element of the current I_2 will be in a magnetic field whose induction is $B_1 = (\mu_0/4\pi)(2I_1/b)$ [see Eq. (6.30)]. The angle α between the elements of the current I_2 and the vector $\boldsymbol{B}_1$ is a right one. Hence, according to Eq. (6.45), the force acting on unit length of the current I_2 is

$$F_{21,u} = I_2 B_1 = \frac{\mu_0}{4\pi}\frac{2I_1 I_2}{b}. \tag{6.46}$$

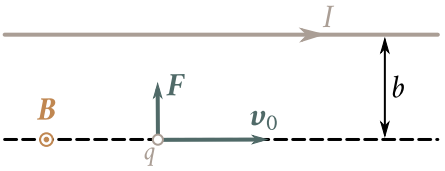


Fig. 6.11

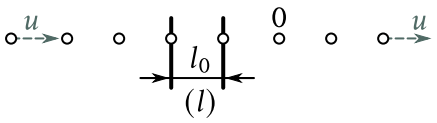


Fig. 6.12

Equation (6.46) coincides with Eq. (6.2).

We get a similar equation for the force $F_{21,u}$ exerted on unit length of the current I_1. It is easy to see that when the currents flow in the same direction they attract each other, and in the opposite direction repel each other.

6.7. Magnetism as a Relativistic Effect

There is a deep relation between electricity and magnetism. On the basis of the postulates of the theory of relativity and of the invariance of an electric charge, we can show that the magnetic interaction of charges and currents is a corollary of Coulomb's law. We shall show this on the example of a charge moving parallel to an infinite line current with the velocity v_0[1] (Fig. 6.11).

According to Eq. (6.36), the magnetic force acting on a charge in the case being considered is

$$F = qv_0 \frac{\mu_0}{4\pi} \frac{2I}{b} \tag{6.47}$$

(the meaning of the symbols is clear from Fig. 6.11). The force is directed toward the conductor carrying the current ($q > 0$). Before commencing to derive Eq. (6.47) for the force on the basis of Coulomb's law and relativistic relations, let us consider the following effect. Assume that we have an infinite linear train of point charges of an identical magnitude e spaced a very small distance l_0 apart (Fig. 6.12). Owing to the smallness of l_0, we can speak of the linear density of the charges λ_0 which obviously is

$$\lambda_0 = \frac{e}{l_0}. \tag{6.48}$$

Let us bring the charges into motion along the train with the identical velocity u. The distance between the charges will therefore diminish and become equal to

$$l = l_0 \left[1 - \frac{u^2}{c^2}\right]^{1/2}$$

[1]We have used the symbol v_0 for the velocity of a charge to make the notation similar to that in Chap. 8 of Vol. I.

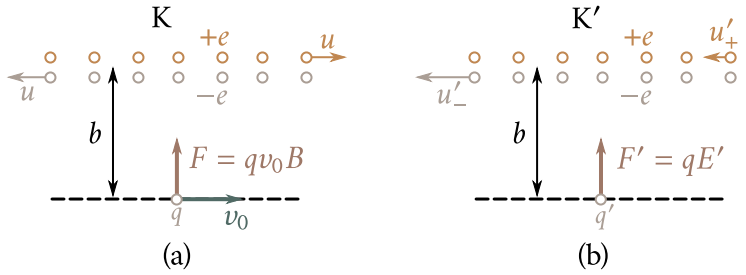


Fig. 6.13

[see Eq. 8.19 of Vol. I]. The magnitude of the charges owing to their invariance, however, remains the same. As a result, the linear density of the charges observed in the reference frame relative to which the charges are moving will change and become equal to

$$\lambda = \frac{e}{l} = \frac{\lambda_0}{\sqrt{1 - (u^2/v^2)}}. \tag{6.49}$$

Now let us consider in the reference frame K two infinite trains formed by charges of the same magnitude, but of opposite signs, moving in opposite directions with the same velocity u and virtually coinciding with each other (Fig. 6.13a). The combination of these trains is equivalent to an infinite line current having the value

$$I = 2\lambda u = \frac{2\lambda_0 u}{\sqrt{1 - (u^2/v^2)}}, \tag{6.50}$$

where λ is the quantity determined by Eq. (6.49). The total linear density of the charges of a train equals zero, therefore an electric field is absent. The charge q experiences a magnetic force whose magnitude according to Eqs. (6.47) and (6.50) is

$$F = qv_0 \frac{\mu_0}{4\pi} \frac{4\lambda_0 u}{b\sqrt{1 - (u^2/v^2)}}. \tag{6.51}$$

Let us pass over to the reference frame K′ relative to which the charge q is at rest (Fig. 6.13b). In this frame, the charge q also experiences a force (let us denote it by F'). This force cannot be of a magnetic origin, however, because the charge q is stationary. The force F' has a purely electrical origin. It appears because the linear densities of the positive and negative charges in the trains are now different (we shall see below that the density of the negative charges is greater). The surplus negative charge distributed over a train sets up an electric field that acts on the positive charge q with the force F' directed toward the train (see Fig. 6.13b).

Let us calculate the force F' and convince ourselves that it "equals" the force F determined by Eq. (6.51). We have taken the word "equals" in quotation marks because force is not an invariant quantity. Upon transition from one inertial refer-

ence frame to another, the force transforms according to a quite complicated law. In a particular case, when the force $\boldsymbol{F}'$ is perpendicular to the relative velocity of the frames K and K' ($\boldsymbol{F}' \perp \boldsymbol{v}_0$), the transformation has the form

$$\boldsymbol{F} = \frac{\boldsymbol{F}'\sqrt{1-\left(v_0^2/c^2\right)} + \boldsymbol{v}_0(\boldsymbol{F}'\cdot\boldsymbol{v}')/c^2}{1+(\boldsymbol{v}_0\cdot\boldsymbol{v}')/c^2}$$

($\boldsymbol{v}'$ is the velocity of a particle experiencing the force $\boldsymbol{F}'$ and measured in the frame K'). If $\boldsymbol{v}' = 0$ (which occurs in the problem we are considering), the formula for transformation of the force is as follows:

$$\boldsymbol{F} = \boldsymbol{F}'\left[1-\left(\frac{v_0^2}{c^2}\right)\right]^{1/2}.$$

A glance at this formula shows that the force perpendicular to $\boldsymbol{v}_0$ exerted on a particle at rest in the frame K' is also perpendicular to the vector $\boldsymbol{v}_0$ in the frame K. The magnitude of the force in this case, however, is transformed by the formula

$$F = F'\left[1-\left(\frac{v_0^2}{c^2}\right)\right]^{1/2}. \tag{6.52}$$

The densities of the charges in the positive and negative trains measured in the frame K' have the values [see Eq. (6.49)]

$$\lambda'_+ = \frac{\lambda_0}{\sqrt{1-\left(u'^2_+/c^2\right)}}, \quad \lambda'_- = -\frac{\lambda_0}{\sqrt{1-\left(u'^2_-/c^2\right)}}, \tag{6.53}$$

where u'_+ and u'_- are the velocities of the charges $+e$ and $-e$ measured in the frame K'. Upon a transition from the frame K to the frame K', the projection of the velocity of a particle onto the direction x coinciding with the direction of $\boldsymbol{v}_0$ is transformed by the equation

$$u'_x = \frac{u_x - v_0}{1-(u_x v_0/c^2)}$$

[see Eqs. (8.28) of Vol. I; we have substituted u and u' for v and v']. For the charges $+e$, the component u_x equals u, for the charges $-e$ it equals $-u$ (see Fig. 6.13a). Hence,

$$\left(u'_x\right)_+ = \frac{u-v_0}{1-(uv_0/c^2)}, \quad \left(u'_x\right)_- = \frac{-u-v_0}{1+(uv_0/c^2)}.$$

Since the remaining projections equal zero, we get

$$u'_+ = \frac{|u-v_0|}{1-(uv_0/c^2)}, \quad u'_- = \frac{u+v_0}{1+(uv_0/c^2)}. \tag{6.54}$$

To simplify our calculations, let us pass over to relative velocities:

$$\beta_0 = \frac{v_0}{c}, \quad \beta = \frac{u}{c}, \quad \beta'_+ = \frac{u'_+}{c}, \quad \beta_- = \frac{u'_-}{c}.$$

Equations (6.53) and (6.54) therefore acquire the form

$$\lambda'_+ = \frac{\lambda_0}{\sqrt{1-\beta'^2_+}}, \quad \lambda_- = \frac{\lambda_0}{\sqrt{1-\beta'^2_-}} \tag{6.55}$$

$$\beta'_+ = \frac{|\beta - \beta_0|}{1-\beta\beta_0}, \quad \beta'_- = \frac{\beta+\beta_0}{1+\beta\beta_0}. \tag{6.56}$$

With account taken of these equations, we get the following expression for the total density of the charges:

$$\begin{aligned}\lambda' &= \lambda'_+ + \lambda'_- \\ &= \frac{\lambda_0}{\left[1-\left(\dfrac{\beta-\beta_0}{1-\beta\beta_0}\right)^2\right]^{1/2}} - \frac{\lambda_0}{\left[1-\left(\dfrac{\beta+\beta_0}{1+\beta\beta_0}\right)^2\right]^{1/2}} \\ &= \frac{\lambda_0(1-\beta\beta_0)}{\sqrt{(1-\beta\beta_0)^2-(\beta-\beta_0)^2}} - \frac{\lambda_0(1+\beta\beta_0)}{\sqrt{(1+\beta\beta_0)^2-(\beta+\beta_0)^2}}.\end{aligned}$$

It is easy to see that

$$(1-\beta\beta_0)^2-(\beta-\beta_0)^2 = (1+\beta\beta_0)^2-(\beta+\beta_0)^2 = \left(1-\beta_0^2\right)\left(1+\beta^2\right).$$

Consequently,

$$\lambda' = \frac{-2\lambda_0\beta\beta_0}{\sqrt{\left(1-\beta_0^2\right)\left(1+\beta^2\right)}} = \frac{-2\lambda_0 u v_0}{c^2\sqrt{1-\left(v_0^2/c^2\right)}\sqrt{1-\left(u^2/c^2\right)}}. \tag{6.57}$$

In accordance with Eq. (1.122), an infinitely long filament carrying a charge of density λ' sets up a field whose strength at the distance b from the filament is

$$E' = \frac{1}{2\pi\varepsilon_0}\frac{\lambda'}{b}.$$

In this field, the charge q experiences the force

$$F' = qE' = \frac{q\lambda'}{2\pi\varepsilon_0 b}.$$

Introduction of Eq. (6.57) yields (we have omitted the minus sign)

$$\begin{aligned}F' &= \frac{q\lambda_0 u v_0}{\pi\varepsilon_0 v c^2\sqrt{1-\left(v_0^2/c^2\right)}\sqrt{1-\left(u^2/c^2\right)}} \\ &= q v_0 \frac{\mu_0}{4\pi}\frac{4\lambda_0 u}{\sqrt{1-\left(u^2/c^2\right)}}\frac{1}{\sqrt{1-\left(v_0^2/c^2\right)}}\end{aligned} \tag{6.58}$$

[we remind our reader that $\mu_0 = 1/(\varepsilon_0 c^2)$; see Eq. (6.15)].

The expression obtained differs from Eq. (6.51) only in the factor $\sqrt{1-\left(v_0^2/c^2\right)}$.

We can, therefore, write that

$$F = F' \left[1 - \left(\frac{v_0^2}{c^2}\right)\right]^{1/2},$$

where F is the force determined by Eq. (6.51), and F' is the force determined by Eq. (6.58). A comparison with Eq. (6.52) shows that F and F' are the values of the same force determined in the frames K and K'.

We must note that in the frame K', which would move relative to the frame K with a velocity differing from that of the charge v_0, the force exerted on the charge would consist of both electric and magnetic forces.

The results we have obtained signify that an electric and a magnetic field are inseparably linked with each other and form a single electromagnetic field. Upon a special choice of the reference frame, a field may be either purely electric or purely magnetic. Relative to other reference frames, however, the same field is a combination of an electric and a magnetic field.

In different inertial reference frames, the electric and magnetic fields of the same collection of charges are different. A derivation beyond the scope of a general course in physics leads to the following equations for the transformation of fields when passing over from a reference frame K to a reference frame K' moving relative to it with the velocity $\boldsymbol{v}_0$:

$$\begin{cases} E'_x = E_x, \quad E_y = \dfrac{E_y - v_0 B_z}{\sqrt{1-\beta^2}}, \quad E'_z = \dfrac{E_z + v_0 B_y}{\sqrt{1-\beta^2}}, \\ B'_x = B_x, \quad B_y = \dfrac{B_y + v_0 E_z}{\sqrt{1-\beta^2}}, \quad B'_z = \dfrac{B_z - v_0 E_y}{\sqrt{1-\beta^2}}. \end{cases} \tag{6.59}$$

Here, E_x, E_y, E_z, B_x, B_y, B_z are the components of the vectors $\boldsymbol{E}$ and $\boldsymbol{B}$ characterizing an electromagnetic field in the frame K, similar primed symbols are the components of the vectors $\boldsymbol{E}'$ and $\boldsymbol{B}'$ characterizing the field in the frame K'. The Greek letter β stands for the ratio v_0/c.

Resolving the vectors $\boldsymbol{E}$ and $\boldsymbol{B}$, and also $\boldsymbol{E}'$ and $\boldsymbol{B}'$, into their components parallel to the vector $\boldsymbol{v}_0$ (and, consequently, to the axes x and x') and perpendicular to this vector (*i.e.*, representing, for example, $\boldsymbol{E}$ in the form $\boldsymbol{E} = \boldsymbol{E}_\parallel + \boldsymbol{E}_\perp$, etc.), we can write Eqs. (6.59) in the vector form:

$$\begin{cases} \boldsymbol{E}'_\parallel = \boldsymbol{E}_\parallel, \quad \boldsymbol{E}'_\perp = \dfrac{\boldsymbol{E}_\perp + (\boldsymbol{v}_0 \times \boldsymbol{B}_\perp)}{\sqrt{1-\beta^2}}, \\ \boldsymbol{B}'_\parallel = \boldsymbol{B}_\parallel, \quad \boldsymbol{B}'_\perp = \dfrac{\boldsymbol{B}_\perp - \left(1/c^2\right)(\boldsymbol{v}_0 \times \boldsymbol{E}_\perp)}{\sqrt{1-\beta^2}}. \end{cases} \tag{6.60}$$

In the Gaussian system of units, Eqs. (6.60) have the form

$$\begin{cases} \boldsymbol{E}'_{\parallel} = \boldsymbol{E}_{\parallel}, \quad \boldsymbol{E}'_{\perp} = \dfrac{\boldsymbol{E}_{\perp} + (1/c)\,(\boldsymbol{v}_0 \times \boldsymbol{B}_{\perp})}{\sqrt{1-\beta^2}}, \\ \boldsymbol{B}'_{\parallel} = \boldsymbol{B}_{\parallel}, \quad \boldsymbol{B}'_{\perp} = \dfrac{\boldsymbol{B}_{\perp} - (1/c)\,(\boldsymbol{v}_0 \times \boldsymbol{E}_{\perp})}{\sqrt{1-\beta^2}}. \end{cases} \tag{6.61}$$

When $\beta \ll 1$ (*i.e.*, $v_0 \ll c$), Eqs. (6.60) are simplified as follows:

$$\boldsymbol{E}'_{\parallel} = \boldsymbol{E}_{\parallel}, \quad \boldsymbol{E}'_{\perp} = \boldsymbol{E}_{\perp} + \boldsymbol{v}_0 \times \boldsymbol{B}_{\perp},$$
$$\boldsymbol{B}'_{\parallel} = \boldsymbol{B}_{\parallel}, \quad \boldsymbol{B}'_{\perp} = \boldsymbol{B}_{\perp} - \left(1/c^2\right)(\boldsymbol{v}_0 \times \boldsymbol{E}_{\perp}).$$

Adding these equations in pairs, we get

$$\begin{cases} \boldsymbol{E}' = \boldsymbol{E}'_{\parallel} + \boldsymbol{E}'_{\perp} = \boldsymbol{E}_{\parallel} + \boldsymbol{E}_{\perp} + (\boldsymbol{v}_0 \times \boldsymbol{B}_{\perp}) = \boldsymbol{E} + (\boldsymbol{v}_0 \times \boldsymbol{B}_{\perp}), \\ \boldsymbol{B}' = \boldsymbol{B}'_{\parallel} + \boldsymbol{B}'_{\perp} = \boldsymbol{B}_{\parallel} + \boldsymbol{B}_{\perp} - \dfrac{1}{c^2}(\boldsymbol{v}_0 \times \boldsymbol{E}_{\perp}) = \boldsymbol{B} + \dfrac{1}{c^2}(\boldsymbol{v}_0 \times \boldsymbol{E}_{\perp}). \end{cases} \tag{6.62}$$

Since the vectors $\boldsymbol{v}_0$ and $\boldsymbol{B}_{\parallel}$ are collinear, their vector product equals zero. Hence, $\boldsymbol{v}_0 \times \boldsymbol{B} = \boldsymbol{v}_0 \times \boldsymbol{B}_{\parallel} + \boldsymbol{v}_0 \times \boldsymbol{B}_{\perp} = \boldsymbol{v}_0 \times \boldsymbol{B}_{\perp}$. Similarly, $\boldsymbol{v}_0 \times \boldsymbol{E} = \boldsymbol{v}_0 \times \boldsymbol{E}_{\perp}$. With this taken into account, Eqs. (6.62) can be given the form

$$\boldsymbol{E}' = \boldsymbol{E} + \boldsymbol{v}_0 \times \boldsymbol{B}, \quad \boldsymbol{B}' = \boldsymbol{B} - \frac{1}{c^2}(\boldsymbol{v}_0 \times \boldsymbol{E}). \tag{6.63}$$

Fields are transformed by means of these equations if the relative velocity of the reference frames v_0 is much smaller than the speed of light in a vacuum c ($v_0 \ll c$).

Equations (6.63) acquire the following form in the Gaussian system of units:

$$\boldsymbol{E}' = \boldsymbol{E} + \frac{1}{c}(\boldsymbol{v}_0 \times \boldsymbol{B}), \quad \boldsymbol{B}' = \boldsymbol{B} - \frac{1}{c}(\boldsymbol{v}_0 \times \boldsymbol{E}). \tag{6.64}$$

In the example in the frame K considered at the beginning of this section, in which the charge q travelled with the velocity $\boldsymbol{v}_0$ parallel to a current-carrying wire, there was only the magnetic field $\boldsymbol{B}_{\perp}$ perpendicular to $\boldsymbol{v}_0$; the components $\boldsymbol{B}_{\parallel}$, $\boldsymbol{E}_{\perp}$. and $\boldsymbol{E}_{\parallel}$ equalled zero. According to Eqs. (6.60) in the frame K', in which the charge q is at rest (this frame travels relative to K with the velocity $\boldsymbol{v}_0$), the component $\boldsymbol{B}'_{\perp}$ equal to $\boldsymbol{B}_{\perp}/\sqrt{1-\beta^2}$ is observed and, in addition, the perpendicular component of the electric field $\boldsymbol{E}'_{\perp} = (\boldsymbol{v}_0 \times \boldsymbol{B}_{\perp})/\sqrt{1-\beta^2}$.

In the frame K, the charge experiences the force

$$\boldsymbol{F} = q(\boldsymbol{v}_0 \times \boldsymbol{B}_{\perp}). \tag{6.65}$$

Since the charge q is at rest in the frame K', it experiences in this frame only the electric force

$$\boldsymbol{F}' = q\boldsymbol{E}'_{\perp} = \frac{q(\boldsymbol{v}_0 \times \boldsymbol{B}_{\perp})}{\sqrt{1-\beta^2}}. \tag{6.66}$$

A comparison of Eqs. (6.65) and (6.66) yields $\boldsymbol{F} = \boldsymbol{F}'\sqrt{1-\beta^2}$, which coincides with

Eq. (6.52).

6.8. Current Loop in a Magnetic Field

Let us see how a loop carrying a current behaves in a magnetic field. We shall begin with a homogeneous field ($\boldsymbol{B}$ = constant). According to Eq. (6.44), a loop element $d\boldsymbol{l}$ experiences the force

$$d\boldsymbol{F} = I(d\boldsymbol{l} \times \boldsymbol{B}). \tag{6.67}$$

The resultant of such forces is

$$\boldsymbol{F} = \oint I(d\boldsymbol{l} \times \boldsymbol{B}). \tag{6.68}$$

Putting the constant quantities I and $\boldsymbol{B}$ outside the integral, we get

$$\boldsymbol{F} = I\left[\left(\oint d\boldsymbol{l}\right) \times \boldsymbol{B}\right].$$

The integral $\oint d\boldsymbol{l}$ equals zero, therefore, $\boldsymbol{F} = 0$. Thus, the resultant force exerted on a current loop in a homogeneous magnetic field equals zero. This holds for loops of any shape (including non-planar ones) with an arbitrary arrangement of the loop relative to the direction of the field. Only homogeneity of the field is essential for the resultant force to equal zero.

In the following, we shall limit ourselves to a consideration of plane loops. Let us calculate the resultant torque set up by the forces (6.67) applied to a loop. Since the sum of these forces equals zero in a homogeneous field, the resultant torque relative to any point will be the same. Indeed, the resultant torque relative to point 0 is determined by the expression

$$\boldsymbol{T} = \int (\boldsymbol{r} \times d\boldsymbol{F}),$$

where $\boldsymbol{r}$ is the position vector drawn from point 0 to the point of application of the force $d\boldsymbol{F}$. Let us take point 0′ displaced relative to 0 by the distance b. Hence, $\boldsymbol{r} = \boldsymbol{b} + \boldsymbol{r}'$, and accordingly $\boldsymbol{r}' = \boldsymbol{r} - \boldsymbol{b}$. Therefore, the resultant torque relative to point 0′ is

$$\boldsymbol{T}' = \int (\boldsymbol{r}' \times d\boldsymbol{F}) = \int ([\boldsymbol{r} - \boldsymbol{b}] \times d\boldsymbol{F}) = \int (\boldsymbol{r} \times d\boldsymbol{F}) - \int (\boldsymbol{b} \times d\boldsymbol{F})$$

$$= \boldsymbol{T} - \left[\boldsymbol{b} \times \int d\boldsymbol{F}\right] = \boldsymbol{T},$$

$\int d\boldsymbol{F} = 0$. The torques calculated relative to two arbitrarily taken points 0 and 0′ were found to coincide. We, thus, conclude that the torque does not depend on the selection of the point relative to which it is taken (compare with a couple of forces).

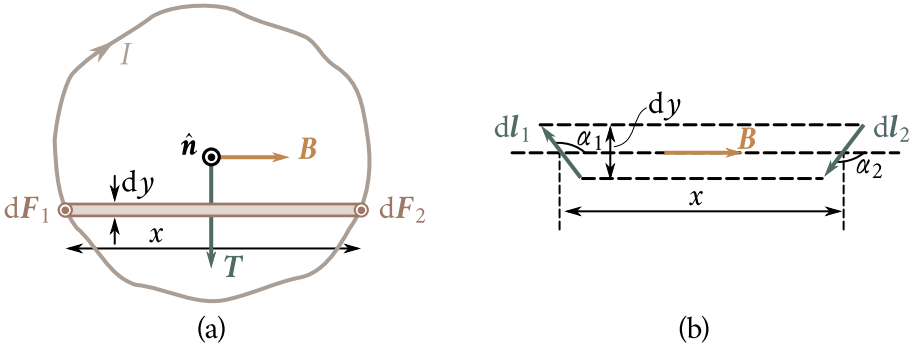


Fig. 6.14

Let us consider an arbitrary plane current loop in a homogeneous magnetic field $\boldsymbol{B}$. Assume that the loop is oriented so that a positive normal to the loop $\hat{\boldsymbol{n}}$ is at right angles to the vector $\boldsymbol{B}$ (Fig. 6.14). A normal is called positive if its direction is associated with that of the current in the loop by the right-hand screw rule.

Let us divide the area of the loop into narrow strips of width $\mathrm{d}y$ parallel to the direction of the vector $\boldsymbol{B}$ (see Fig. 6.14a; Fig. 6.14b is an enlarged view of one of these strips). The force $\mathrm{d}\boldsymbol{F}_1$ directed beyond the drawing is exerted on the loop element $\mathrm{d}\boldsymbol{l}_1$ enclosing the strip at the left. The magnitude of this force is $\mathrm{d}F_1 = IB\,\mathrm{d}l_1\,\sin\alpha_1 = IB\,\mathrm{d}y$ (see Fig. 6.14b). The force $\mathrm{d}\boldsymbol{F}_2$ directed toward us is exerted on the loop element $\mathrm{d}\boldsymbol{l}_2$ enclosing the strip at the right. The magnitude of this force is $\mathrm{d}F_2 = IB\,\mathrm{d}l_2\,\sin\alpha_2 = IB\,\mathrm{d}y$.

The result we have obtained signifies that the forces applied to opposite loop elements $\mathrm{d}\boldsymbol{l}_1$ and $\mathrm{d}\boldsymbol{l}_2$ form a couple whose torque is

$$\mathrm{d}T = IBx\,\mathrm{d}y = IB\,\mathrm{d}S$$

($\mathrm{d}S$ is the area of a strip). A glance at Fig. 6.14 shows that the vector $\mathrm{d}\boldsymbol{T}$ is perpendicular to the vectors $\hat{\boldsymbol{n}}$ and $\boldsymbol{B}$ and, consequently, can be written in the form

$$\mathrm{d}\boldsymbol{T} = I(\hat{\boldsymbol{n}} \times \boldsymbol{B})\,\mathrm{d}S.$$

Summation of this equation over all the strips yields the torque acting on the loop:

$$\boldsymbol{T} = \int I(\hat{\boldsymbol{n}} \times \boldsymbol{B})\,\mathrm{d}S = I(\hat{\boldsymbol{n}} \times \boldsymbol{B}) \int \mathrm{d}S = I(\hat{\boldsymbol{n}} \times \boldsymbol{B})\,\mathrm{d}S \tag{6.69}$$

(the field is assumed to be homogeneous, therefore, the product $\hat{\boldsymbol{n}} \times \boldsymbol{B}$ is the same for all the strips and can be put outside the integral). The quantity S in Eq. (6.69) is the area of the loop.

Equation (6.69) can be written in the form

$$\boldsymbol{T} = (IS\hat{\boldsymbol{n}}) \times \boldsymbol{B}. \tag{6.70}$$

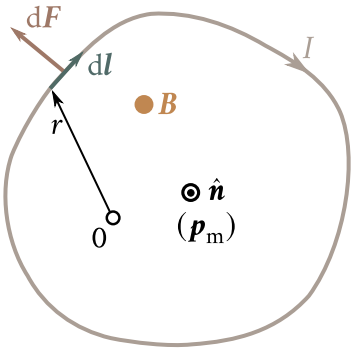


Fig. 6.15

This equation is similar to Eq. (1.58) determining the torque exerted on an electric dipole in an electric field. The analogue of $\boldsymbol{E}$ in Eq. (6.70) is the vector $\boldsymbol{B}$, and that of the electric dipole moment $\boldsymbol{p}$ is the expression $IS\hat{\boldsymbol{n}}$. This served as the grounds to call the quantity

$$\boldsymbol{p}_{\mathrm{m}} = IS\hat{\boldsymbol{n}} \tag{6.71}$$

the **magnetic dipole moment** of a current loop. The direction of the vector $\boldsymbol{p}_{\mathrm{m}}$ coincides with that of a positive normal to the loop.

Using the notation of Eq. (6.71), we can write Eq. (6.70) as follows:

$$\boldsymbol{T} = \boldsymbol{p}_{\mathrm{m}} \times \boldsymbol{B} \quad (\boldsymbol{p}_{\mathrm{m}} \perp \boldsymbol{B}). \tag{6.72}$$

Now, let us assume that the direction of the vector $\boldsymbol{B}$ coincides with that of a positive normal to the loop $\hat{\boldsymbol{n}}$ and, therefore, with that of the vector $\boldsymbol{p}_{\mathrm{m}}$ too (Fig. 6.15). In this case, the forces exerted on different elements of the loop are in one plane—that of the loop. The force exerted on the loop element $\mathrm{d}\boldsymbol{l}$ is determined by Eq. (6.67). Let us calculate the resultant torque produced by such forces relative to point 0 in the plane of the loop:

$$\boldsymbol{T} = \int \mathrm{d}\boldsymbol{T} = \int (\boldsymbol{r} \times \mathrm{d}\boldsymbol{F}) = I \oint [\boldsymbol{r} \times (\mathrm{d}\boldsymbol{l} \times \boldsymbol{B})]$$

($\boldsymbol{r}$ is the position vector drawn from point 0 to the element $\mathrm{d}\boldsymbol{l}$). Let us transform the integrand by means of Eq. (1.35) of Vol. I. The result is

$$\boldsymbol{T} = I \left[\oint (\boldsymbol{r} \cdot \boldsymbol{B})\, \mathrm{d}\boldsymbol{l} - \oint \boldsymbol{B}(\boldsymbol{r} \cdot \mathrm{d}\boldsymbol{l}) \right].$$

The first integral equals zero because the vectors $\boldsymbol{r}$ and $\boldsymbol{B}$ are mutually perpendicular. The scalar product inside the second integral is $r\,\mathrm{d}r = \mathrm{d}\left(r^2\right)/2$. The second integral can, therefore, be written in the form

$$\frac{1}{2}\boldsymbol{B} \oint \mathrm{d}\left(r^2\right).$$

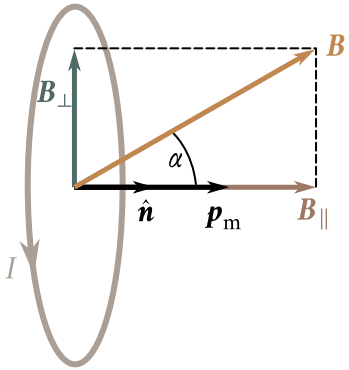


Fig. 6.16

The total differential of the function r^2 is inside the integral. The sum of the increments of a function along a closed path is zero. Hence, the second addend in the expression for $\boldsymbol{T}$ is zero too. We have, thus, proved that the resultant torque $\boldsymbol{T}$ relative to any point 0 in the plane of the loop is zero. The resultant torque relative to all other points has the same value (see above).

Thus, when the vectors $\boldsymbol{p}_\mathrm{m}$ and $\boldsymbol{B}$ have the same direction, the magnetic forces exerted on separate portions of a loop do not tend to turn the loop nor shift it from its position. They only tend to stretch the loop in its plane. If the vectors $\boldsymbol{p}_\mathrm{m}$ and $\boldsymbol{B}$ have opposite directions, the magnetic forces tend to compress the loop.

Assume that the directions of the vectors $\boldsymbol{p}_\mathrm{m}$ and $\boldsymbol{B}$ form an arbitrary angle α (Fig. 6.16). Let us resolve the magnetic induction $\boldsymbol{B}$ into two components: $\boldsymbol{B}_\parallel$ parallel to the vector $\boldsymbol{p}_\mathrm{m}$ and $\boldsymbol{B}_\perp$ perpendicular to it, and consider the action of each component separately. The component $\boldsymbol{B}_\parallel$ will set up forces stretching or compressing the loop. The component $\boldsymbol{B}_\perp$ whose magnitude is $B \sin \alpha$ will lead to the appearance of a torque that can be calculated by Eq. (6.72):

$$\boldsymbol{T} = \boldsymbol{p}_\mathrm{m} \times \boldsymbol{B}_\perp.$$

Inspection of Fig. 6.16 shows that

$$\boldsymbol{p}_\mathrm{m} \times \boldsymbol{B}_\perp = \boldsymbol{p}_\mathrm{m} \times \boldsymbol{B}.$$

Consequently, in the most general case, the torque exerted on a plane current loop in a homogeneous magnetic field is determined by the equation

$$\boldsymbol{T} = \boldsymbol{p}_\mathrm{m} \times \boldsymbol{B}. \tag{6.73}$$

The magnitude of the vector $\boldsymbol{T}$ is

$$T = p_\mathrm{m} B \sin \alpha. \tag{6.74}$$

To increase the angle a between the vectors $\boldsymbol{p}_\mathrm{m}$ and $\boldsymbol{B}$ by $\mathrm{d}\alpha$, the following work

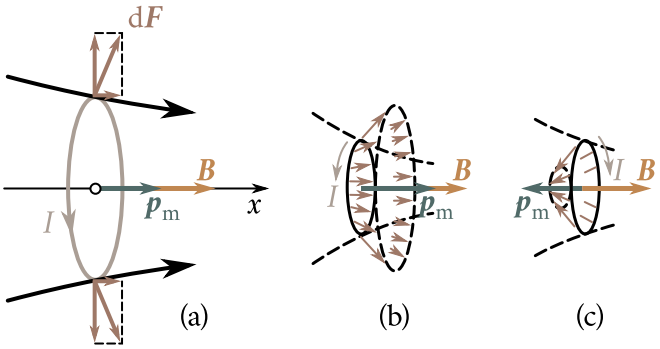


Fig. 6.17

must be done against the forces exerted on a loop in a magnetic field:

$$\mathrm{d}A = T\,\mathrm{d}\alpha = p_{\mathrm{m}} B \sin\alpha\,\mathrm{d}\alpha. \tag{6.75}$$

Upon turning to its initial position, a loop can return the work spent for its rotation by doing it on some other body. Hence, the work (6.75) goes to increase the potential energy $W_{\mathrm{p,mech}}$ which a current loop has in a magnetic field, by the magnitude

$$\mathrm{d}W_{\mathrm{p,mech}} = p_{\mathrm{m}} B \sin\alpha\,\mathrm{d}\alpha.$$

Integration yields

$$W_{\mathrm{p,mech}} = -p_{\mathrm{m}} B \cos\alpha + \text{constant}.$$

Assuming that constant = 0, we get the following expression:

$$W_{\mathrm{p,mech}} = -p_{\mathrm{m}} B \cos\alpha = -\boldsymbol{p}_{\mathrm{m}} \cdot \boldsymbol{B} \tag{6.76}$$

[compare with Eq. (1.61)].

Parallel orientation of the vectors $\boldsymbol{p}_{\mathrm{m}}$ and $\boldsymbol{B}$ corresponds to the minimum energy (6.76) and, consequently, to the position of stable equilibrium of a loop.

The quantity expressed by Eq. (6.76) is not the total potential energy of a current loop, but only the part of it that is due to the existence of the torque (6.73). To stress this, we have provided the symbol of the potential energy expressed by Eq. (6.76) with the subscript "mech". Apart from $W_{\mathrm{p,mech}}$, the total potential energy of a loop includes other addends.

Now let us consider a plane current loop in an inhomogeneous magnetic field. For simplicity, we shall first consider the loop to be circular. Assume that the field changes the fastest in the direction x coinciding with that of $\boldsymbol{B}$ where the centre of the loop is, and that the magnetic moment of the loop is oriented along $\boldsymbol{B}$ (Fig. 6.17a).

Here, $\boldsymbol{B} \neq$ constant, and Eq. (6.68) does not have to be zero. The force $\mathrm{d}\boldsymbol{F}$ exerted on a loop element is perpendicular to $\boldsymbol{B}$, *i.e.*, to the magnetic field line where it intersects $\mathrm{d}\boldsymbol{l}$. Therefore, the forces applied to different loop elements form a

symmetrical conical fan (Fig. 6.17b). Their resultant $\boldsymbol{F}$ is directed toward a growth in $\boldsymbol{B}$ and, therefore, pulls the loop into the region with a stronger field. It is quite obvious that the greater the field changes (the greater is $\partial B/\partial x$), the smaller is the apex angle of the cone and the greater, other conditions being equal, is the resultant force $\boldsymbol{F}$. If we reverse the direction of the current (now $\boldsymbol{p}_{\mathrm{m}}$ is antiparallel to $\boldsymbol{B}$), the directions of all the forces $\mathrm{d}\boldsymbol{F}$ and of their resultant $\boldsymbol{F}$ will be reversed (Fig. 6.17c). Hence, with such a mutual orientation of the vectors $\boldsymbol{p}_{\mathrm{m}}$ and $\boldsymbol{B}$, the loop will be pushed out of the field.

It is a simple matter to find a quantitative expression for the force $\boldsymbol{F}$ by using Eq. (6.76) for the energy of a loop in a magnetic field. If the orientation of the magnetic moment relative to the field remains constant (α = constant), then $W_{\mathrm{p,mech}}$ will depend only on x (through B). Differentiating $W_{\mathrm{p,mech}}$ with respect to x and changing the sign of the result, we get the projection of the force onto the x-axis:

$$F_x = -\frac{\partial W_{\mathrm{p,mech}}}{\partial x} = p_{\mathrm{m}} \frac{\partial B}{\partial x} \cos \alpha.$$

We assume that the field changes only slightly in the other directions. Hence, we may disregard the projections of the force onto the other axes and assume that $F = F_x$. Thus,

$$F = p_{\mathrm{m}} \frac{\partial B}{\partial x} \cos \alpha. \tag{6.77}$$

According to the equation we have obtained, the force exerted on a current loop in an inhomogeneous magnetic field depends on the orientation of the magnetic moment of the loop relative to the direction of the field. If the vectors $\boldsymbol{p}_{\mathrm{m}}$ and $\boldsymbol{B}$ coincide in direction ($\alpha = 0$), then the force is positive, *i.e.*, is directed toward a growth in $vecB$ ($\partial B/\partial x = 0$ is assumed to be positive; otherwise, the sign and the direction of the force will be reversed, but the force will pull the loop into the region of a strong field as before). If $\boldsymbol{p}_{\mathrm{m}}$ and $\boldsymbol{B}$ are antiparallel ($\alpha = \pi$), the force is negative, *i.e.*, directed toward diminishing of $\boldsymbol{B}$. We have already obtained this result qualitatively with the aid of Fig. 6.17.

It is quite evident that apart from the force (6.77), a current loop in an inhomogeneous magnetic field will also experience the torque (6.73).

6.9. Magnetic Field of a Current Loop

Let us consider the field set up by a current flowing in a thin wire having the shape of a circle of radius R (a ring current). We shall determine the magnetic induction at the centre of the ring current (Fig. 6.18). Every current element produces at the centre an induction directed along a positive normal to the loop. Therefore, vector

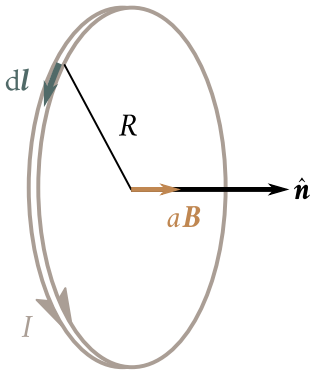


Fig. 6.18

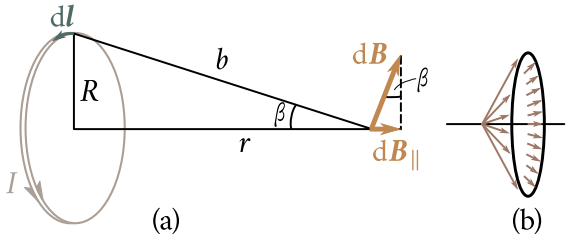


Fig. 6.19

summation of the d$\boldsymbol{B}$'s consists in summation of their magnitudes. By Eq. (6.29),

$$\mathrm{d}B = \frac{\mu_0}{4\pi}\frac{I\,\mathrm{d}l}{R^2}$$

($\alpha = \pi/2$). Let us integrate this expression over the entire loop:

$$B = \int \mathrm{d}B = \frac{\mu_0}{4\pi}\frac{I}{R^2}\oint \mathrm{d}l = \frac{\mu_0}{4\pi}\frac{I}{R^2}2\pi R = \frac{\mu_0}{4\pi}\frac{2I\left(\pi R^2\right)}{R^3}.$$

The expression in parentheses is the magnitude of the magnetic dipole moment p_m [see Eq. (6.71)]. Hence, the magnetic induction at the centre of a ring current has the value

$$B = \frac{\mu_0}{4\pi}\frac{2p_\mathrm{m}}{R^3}. \tag{6.78}$$

Inspection of Fig. 6.18 shows that the direction of the vector $\boldsymbol{B}$ coincides with that of a positive normal to the loop, *i.e.*, with that of the vector $\boldsymbol{p}_\mathrm{m}$. Therefore, Eq. (6.78) can be written in the vector form:

$$\boldsymbol{B} = \frac{\mu_0}{4\pi}\frac{2\boldsymbol{p}_\mathrm{m}}{R^3}. \tag{6.79}$$

Now let us find $\boldsymbol{B}$ on the axis of the ring current at the distance of r from the centre of the loop (Fig. 6.19). The vectors d$\boldsymbol{B}$ are perpendicular to the planes passing through the relevant element d$\boldsymbol{l}$ and the point where we are seeking the field. Hence, they form a symmetrical conical fan (Fig. 6.19b). We can conclude from considerations of symmetry that the resultant vector $\boldsymbol{B}$ is directed along the axis of the loop. Each of the component vectors d$\boldsymbol{B}$ contributes $\mathrm{d}\boldsymbol{B}_{parallel}$ equal in magnitude to $\mathrm{d}B\sin\beta = \mathrm{d}B(R/b)$ to the resultant vector. The angle α between d$\boldsymbol{l}$ and $\boldsymbol{b}$ is a right one, hence,

$$\mathrm{d}B_\parallel = \mathrm{d}B\frac{R}{b} = \frac{\mu_0}{4\pi}\frac{I\,\mathrm{d}l}{b^2}\frac{R}{b} = \frac{\mu_0}{4\pi}\frac{IR\,\mathrm{d}l}{b^3}.$$

Integrating over the entire loop and substituting $\sqrt{R^2 + r^2}$ for b, we obtain

$$B = \int \mathrm{d}B_{\parallel} = \frac{\mu_0}{4\pi}\frac{IR}{b^3}\oint \mathrm{d}l = \frac{\mu_0}{4\pi}\frac{IR}{b^3}2\pi R = \frac{\mu_0}{4\pi}\frac{2\left(I\pi R^2\right)}{\left(R^2 + r^2\right)^{3/2}}$$
$$= \frac{\mu_0}{4\pi}\frac{2p_{\mathrm{m}}}{\left(R^2 + r^2\right)^{3/2}}. \tag{6.80}$$

This equation determines the magnitude of the magnetic induction on the axis of a ring current. With a view to the vectors $\boldsymbol{B}$ and $\boldsymbol{p}_{\mathrm{m}}$ having the same direction, we can write Eq. (6.80) in the vector form:

$$\boldsymbol{B} = \frac{\mu_0}{4\pi}\frac{2\boldsymbol{p}_{\mathrm{m}}}{\left(R^2 + r^2\right)^{3/2}}. \tag{6.81}$$

This expression does not depend on the sign of r. Hence, at points on the axis symmetrical relative to the centre of the current, $\boldsymbol{B}$ has the same magnitude and direction.

When $r = 0$, Eq. (6.81) transforms, as should be expected, into Eq. (6.79) for the magnetic induction at the centre of a ring current.

For great distances from a loop, we may disregard R^2 in the denominator in comparison with r^2. Equation (6.81) now becomes

$$\boldsymbol{B} = \frac{\mu_0}{4\pi}\frac{2\boldsymbol{p}_{\mathrm{m}}}{r^3} \quad \text{(along the current axis)}, \tag{6.82}$$

which is similar to Eq. (1.55) for the electric field strength along the axis of a dipole.

Calculations beyond the scope of the present book show that a magnetic dipole moment $\boldsymbol{p}_{\mathrm{m}}$ can be ascribed to any system of currents or moving charges localized in a restricted portion of space (compare with the electric dipole moment of a system of charges). The magnetic field of such a system at distances that are great in comparison with its dimensions is determined through $\boldsymbol{p}_{\mathrm{m}}$ using the same equations as those used to determine the field of a system of charges at great distances through the electric dipole moment (see Sec. 1.10). In particular, the field of a plane loop of any shape at great distances from it is

$$B = \frac{\mu_0}{4\pi}\frac{2p_{\mathrm{m}}}{r^3}\sqrt{1 + 3\cos^2\theta}, \tag{6.83}$$

where r is the distance from the loop to the given point, and θ is the angle between the direction of the vector $\boldsymbol{p}_{\mathrm{m}}$ and the direction from the loop to the given point of the field [compare with Eq. (1.53)]. When $\theta = 0$, Eq. (6.83) gives the same value as Eq. (6.82) for the magnitude of the vector $\boldsymbol{B}$.

Figure 6.20 shows the magnetic field lines of a ring current. It shows only the lines in one of the planes passing through the current axis. A similar picture will be observed in any of these planes.

It follows from everything said in the preceding and this sections that the

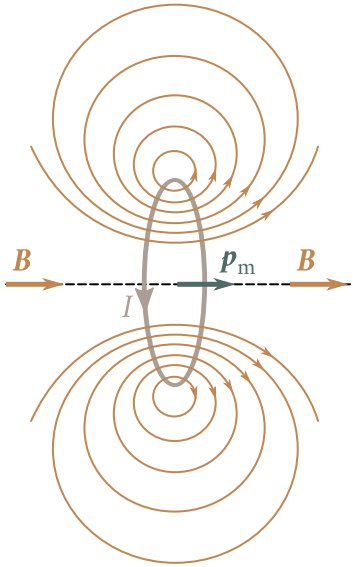


Fig. 6.20

magnetic dipole moment is a very important characteristic of a current loop. It determines both the field set up by a loop and the behaviour of the loop in an external magnetic field.

6.10. Work Done When a Current Moves in a Magnetic Field

Let us consider a current loop formed by stationary wires and a movable rod of length l sliding along them (Fig. 6.21). Let the loop be in an external magnetic field which we shall assume to be homogeneous and at right angles to the plane of the loop. With the directions of the current and field shown in the figure, the force $\boldsymbol{F}$ exerted on the rod will be directed to the right and will equal

$$F = IBl.$$

When the rod moves to the right by $\mathrm{d}h$, this force does the positive work

$$\mathrm{d}A = F\,\mathrm{d}h = IBl\,\mathrm{d}h = IB\,\mathrm{d}S, \tag{6.84}$$

where $\mathrm{d}S$ is the shaded area (see Fig. 6.21a).

Let us see how the magnetic induction flux Φ through the area of the loop will change when the rod moves. We shall agree, when calculating the flux through the area of a current loop, that the quantity $\hat{\boldsymbol{n}}$ in the equation

$$\Phi = \int \boldsymbol{B} \cdot \hat{\boldsymbol{n}}\,\mathrm{d}S,$$

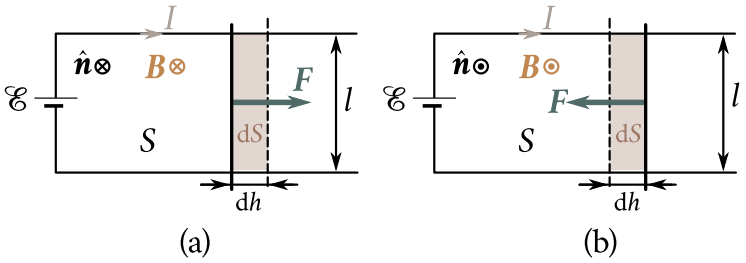


Fig. 6.21

is a positive normal, *i.e.*, one that forms a right-handed system with the direction of the current in the loop (see Sec. 6.8). Hence, in the case shown in Fig. 6.21a, the flux will be positive and equal to BS (S is the area of the loop). When the rod moves to the right, the area of the loop receives the positive increment $\mathrm{d}S$. As a result, the flux also receives the positive increment $\mathrm{d}\Phi = B\,\mathrm{d}S$. Equation (6.84) can, therefore, be written in the form

$$\mathrm{d}A = I\,\mathrm{d}\Phi. \tag{6.85}$$

When the field is directed toward us (Fig. 6.21b), the force exerted on the rod is directed to the left. Therefore when the rod moves to the right through the distance $\mathrm{d}h$, the magnetic force does the negative work

$$\mathrm{d}A = -IBl\,\mathrm{d}h = -IB\,\mathrm{d}S. \tag{6.86}$$

In this case, the flux through the loop is $-BS$. When the area of the loop grows by $\mathrm{d}S$, the flux receives the increment $\mathrm{d}\Phi = -B\,\mathrm{d}S$. Hence, Eq. (6.86) can also be written in the form of Eq. (6.85).

The quantity $\mathrm{d}\Phi$ in Eq. (6.85) can be interpreted as the flux through the area covered by the rod when it moves. We can say accordingly that the work done by the magnetic force on a portion of a current loop equals the product of the current and the magnitude of the magnetic flux through the surface covered by this portion during its motion.

Equations (6.84) and (6.85) can be combined into a single vector expression. For this purpose, we shall compare the vector $\boldsymbol{l}$ having the direction of the current with the rod (Fig. 6.22). Regardless of the direction of the vector $\boldsymbol{B}$ (toward us or away from us), the force exerted on the rod can be represented in the form

$$\boldsymbol{F} = I\boldsymbol{l} \times \boldsymbol{B}.$$

When the rod moves through the distance $\mathrm{d}\boldsymbol{h}$, the force does the work

$$\mathrm{d}A = \boldsymbol{F}\,\mathrm{d}\boldsymbol{h} = I\boldsymbol{l} \times \boldsymbol{B}\,\mathrm{d}\boldsymbol{h}.$$

Let us perform a cyclic transposition of the multipliers in this triple scalar product

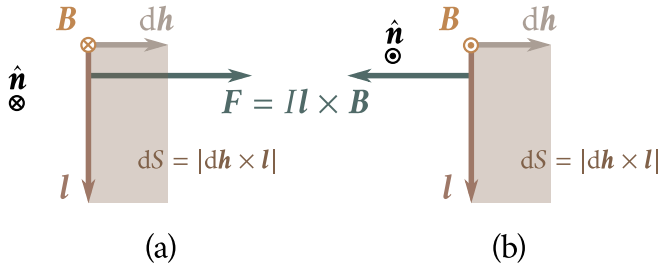


Fig. 6.22

[see Eq. (1.34) of Vol. I]. The result is

$$dA = I\boldsymbol{B}(d\boldsymbol{h} \times \boldsymbol{l}). \tag{6.87}$$

A glance at Fig. 6.22 shows that the vector product $(d\boldsymbol{h} \times \boldsymbol{l})$ equals in magnitude the area dS described by the rod during its motion and has the direction of a positive normal $\hat{\boldsymbol{n}}$. Hence,

$$dA = I(\boldsymbol{B} \cdot \hat{\boldsymbol{n}})\, dS. \tag{6.88}$$

In the case shown in Fig. 6.22a, we have $\boldsymbol{B} \cdot \hat{\boldsymbol{n}} = B$, and we arrive at Eq. (6.84). In the case shown in Fig. 6.22b, we have $\boldsymbol{B} \cdot \hat{\boldsymbol{n}} = -B$. and we arrive at Eq. (6.86).

The expression $\boldsymbol{B} \cdot \hat{\boldsymbol{n}}\, dS$ determines the increment of the magnetic flux through the loop due to motion of the rod. Thus, Eq. (6.88) can be written in the form of (6.85). But Eq. (6.88) has an advantage over (6.85) because we "automatically" get the sign of $d\Phi$ from it and, consequently, the sign of dA too.

Let us consider a rigid current loop of any shape in an arbitrary magnetic field. We shall find the work done upon an arbitrary infinitely small displacement of the loop. Assume that the loop element $d\boldsymbol{l}$ was displaced by $d\boldsymbol{h}$ (Fig. 6.23). The magnetic force does the following work on it:

$$dA_{\text{el}} = I(d\boldsymbol{l} \times \boldsymbol{B}) \cdot d\boldsymbol{h}. \tag{6.89}$$

Here, $\boldsymbol{B}$ is the magnetic induction at the place where the loop element $d\boldsymbol{l}$ is.

Performing a cyclic transposition of the multipliers in Eq. (6.89), we get

$$dA_{\text{el}} = I\boldsymbol{B} \cdot (d\boldsymbol{h} \times d\boldsymbol{l}). \tag{6.90}$$

The magnitude of the vector product $d\boldsymbol{h} \times d\boldsymbol{l}$ equals the area of a parallelogram constructed on the vectors $d\boldsymbol{h}$ and $d\boldsymbol{l}$, *i.e.*, the area dS described by the element $d\boldsymbol{l}$ during its motion. The direction of the vector product coincides with that of a positive normal to the area dS. Consequently,

$$\boldsymbol{B} \cdot (d\boldsymbol{h} \times d\boldsymbol{l}) = (\boldsymbol{B} \cdot \hat{\boldsymbol{n}})\, dS = d\Phi_{\text{el}}, \tag{6.91}$$

where $d\Phi_{\text{el}}$ is the increment of the magnetic flux through the loop due to the displacement of the loop element $d\boldsymbol{l}$.

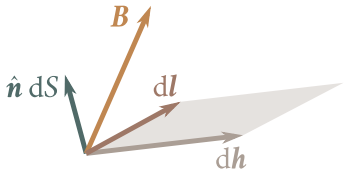


Fig. 6.23

With a view to Eq. (6.91), we can write Eq. (6.90) in the form

$$\mathrm{d}A_{\mathrm{el}} = I\,\mathrm{d}\Phi_{\mathrm{el}}. \tag{6.92}$$

Summation of Eq. (6.92) over all the loop elements yields an expression for the work of the magnetic forces upon an arbitrary infinitely small displacement of the loop:

$$\mathrm{d}A = \int \mathrm{d}A_{\mathrm{el}} = \int I\,\mathrm{d}\Phi_{\mathrm{el}} = I\int \mathrm{d}\Phi_{\mathrm{el}} = I\,\mathrm{d}\Phi \tag{6.93}$$

($\mathrm{d}\Phi$ is the total increment of the flux through the loop).

To find the work done upon a finite arbitrary displacement of a loop, let us integrate Eq. (6.93) over the entire loop:

$$A_{12} = \int \mathrm{d}A = I\int \mathrm{d}\Phi_{\mathrm{el}} = I\,(\Phi_2 - \Phi_1)\,. \tag{6.94}$$

Here, Φ_1 and Φ_2 are the values of the magnetic flux through the loop in its initial and final positions. The work done by the magnetic forces on the loop thus equals the product of the current and the increment of the magnetic flux through the loop.

In particular, when a plane loop rotates in a homogeneous field from a position in which the vectors $\boldsymbol{P}_{\mathrm{m}}$ and $\boldsymbol{B}$ are directed oppositely (in this position $\Phi = -BS$) to a position in which these vectors have the same direction (in this position $\Phi = BS$, the magnetic forces do the following work on the loop:

$$A = I[BS - (BS)] = 2IBS.$$

The same result is obtained with the aid of Eq. (6.91) for the potential energy of a loop in a magnetic field:

$$A = W_{\mathrm{init}} - W_{\mathrm{fin}} = p_{\mathrm{m}}B - (-p_{\mathrm{m}}B) = 2p_{\mathrm{m}}B = 2ISB$$

($p_{\mathrm{m}} = IS$).

We must note that the work expressed by Eq. (6.94) is done not at the expense of the energy of the external magnetic field, but at the expense of the source maintaining a constant current in the loop. We shall show in Sec. 8.2 that when the magnetic flux through a loop changes, an induced e.m.f. $\mathscr{E}_{\mathrm{i}} = -(\mathrm{d}\Phi/\mathrm{d}t)$ is set up in the loop. Hence, the source in addition to the work done to liberate the Joule heat must also do work against the induced e.m.f. determined by the expression

$$A = \int \mathrm{d}A = -\int \mathscr{E}_{\mathrm{i}} I\,\mathrm{d}t = \int \mathrm{d}\Phi/\mathrm{d}t I\,\mathrm{d}t = \int I\,\mathrm{d}\Phi = I(\Phi_2 - \Phi_1),$$

that coincides with Eq. (6.94).

6.11. Divergence and Curl of a Magnetic Field

The absence of magnetic charges in nature[2] results in the fact that the lines of the vector $\boldsymbol{B}$ have neither a beginning nor an end. Therefore, in accordance with Eq. (1.77), the flux of the vector $\boldsymbol{B}$ through a closed surface must equal zero. Thus, for any magnetic field and an arbitrary closed surface, the condition

$$\Phi_B = \oint_S \boldsymbol{B}\, \mathrm{d}\boldsymbol{S} = 0, \tag{6.95}$$

is observed. This equation expresses Gauss's theorem for the vector $\boldsymbol{B}$: the flux of the magnetic induction vector through any closed surface equals zero.

Substituting a volume integral for the surface one in Eq. (6.95) in accordance with Eq. (1.108), we find that

$$\int_V \nabla \cdot \boldsymbol{B}\, \mathrm{d}V = 0.$$

The condition which we have arrived at must be observed for any arbitrarily chosen volume V. This is possible only if the integrand at each point of the field is zero. Thus, a magnetic field has the property that its divergence is zero everywhere:

$$\nabla \cdot \boldsymbol{B} = 0. \tag{6.96}$$

Let us now turn to the circulation of the vector $\boldsymbol{B}$. By definition, the circulation equals the integral

$$\oint \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{l}. \tag{6.97}$$

It is the simplest to calculate this integral for the field of a line current. Assume that a closed loop is in a plane perpendicular to the current (Fig. 6.24; the current is perpendicular to the plane of the drawing and is directed beyond the drawing). At each point of the loop, the vector $\boldsymbol{B}$ is directed along a tangent to the circumference passing through this point. Let us substitute $B\, \mathrm{d}l_B$ for $\boldsymbol{B} \cdot \mathrm{d}\boldsymbol{l}$ in the expression for the circulation ($\mathrm{d}l_B$ is the projection of a loop element onto the direction of the vector $\boldsymbol{B}$). Inspection of the figure shows that $\mathrm{d}l_B$ equals $b\, \mathrm{d}\alpha$, where b is the distance from the wire carrying the current to $\mathrm{d}\boldsymbol{l}$, and $\mathrm{d}\alpha$ is the angle through which a radial straight line turns when it moves along the loop over the element $\mathrm{d}\boldsymbol{l}$. Thus,

[2]The British physicist Paul Dirac made the assumption that magnetic charges (called Dirac's monopoles) should exist in nature. Searches for these charges have meanwhile given no results and the question of the existence of Dirac's...

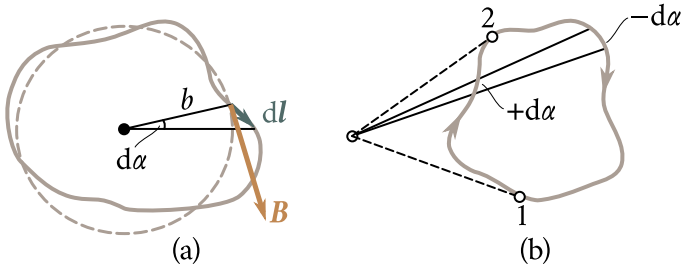


Fig. 6.24

introducing Eq. (6.30) for B, we get

$$\boldsymbol{B} \cdot \mathrm{d}\boldsymbol{l} = B\,\mathrm{d}l_B = \frac{\mu_0}{4\pi}\frac{2I}{b} b\,\mathrm{d}\alpha = \frac{\mu_0 I}{2\pi}\,\mathrm{d}\alpha. \tag{6.98}$$

With a view to Eq. (6.98), we have

$$\oint \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{l} = \frac{\mu_0 I}{2\pi} \oint \mathrm{d}\alpha. \tag{6.99}$$

Upon circumvention of the loop enclosing the current, the radial straight line constantly turns in one direction, therefore, $\oint \mathrm{d}\alpha = 2\pi$. Matters are different if the current is not enclosed by the loop (Fig. 6.24b). Here, upon circumvention of the loop, the radial straight line first turns in one direction (segment 1-2), and then in the opposite one (2-1), owing to which $\oint \mathrm{d}\alpha$ equals zero. With a view to this result. we can write that

$$\oint \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{l} = \mu_0 I, \tag{6.100}$$

where I must be understood as the current enclosed by the loop. If the loop does not enclose the current, the circulation of the vector $\boldsymbol{B}$ is zero.

The sign of expression (6.100) depends on the direction of circumvention of the loop (the angle α is measured in the same direction). If the direction of circumvention forms a right-handed system with the direction of the current, quantity (6.100) is positive, in the opposite case it is negative. The sign can be taken into consideration by assuming I to be an algebraic quantity. A current whose direction is associated with that of circumvention of a loop by the right-hand screw rule must be considered positive; a current of the opposite direction will be negative.

Equation (6.100) will allow us to easily recall Eq. (6.30) for B of the field of a line current. Imagine a plane loop in the form of a circle of radius b (Fig. 6.25). At each point of this loop, the vector $\boldsymbol{B}$ has the same magnitude and is directed along a tangent to the circle. Hence, the circulation equals the product of B and the length of the circumference $2\pi b$, and Eq. (6.100) has the form

$$B \times 2\pi b = \mu_0 I.$$

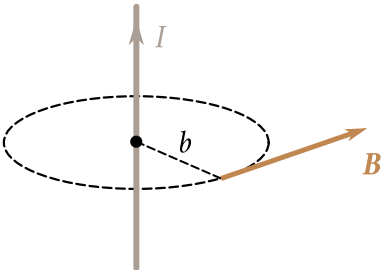


Fig. 6.25

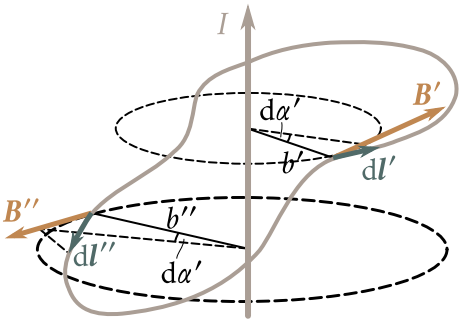


Fig. 6.26

Thus, $B = \mu_0 I/(2\pi b)$ [compare with Eq. (6.30)].

The case of a non-planar loop (Fig. 6.26) differs from that of a plane one considered above, only in that upon motion along the loop the radial straight line not only turns about the wire, but also moves along it. All our reasoning which led us to Eq. (6.100) remains true if we understand $d\alpha$ to be the angle through which the projection of the radial straight line onto a plane perpendicular to the current turns. The total angle of rotation of this projection is 2π if the loop encloses the current, and zero otherwise. We thus again arrive at Eq. (6.100).

We have obtained Eq. (6.100) for a line current. We can show that it also holds for a current flowing in a wire of an arbitrary shape, for example, for a ring current.

Assume that a loop encloses several wires carrying currents. Owing to the superposition principle [see Eq. (6.16)]:

$$\oint \boldsymbol{B} \cdot d\boldsymbol{l} = \oint \left(\sum_k \boldsymbol{B}_k \right) d\boldsymbol{l} = \sum_k \oint \boldsymbol{B}_k \, d\boldsymbol{l}.$$

Each of the integrals in this sum equals $\mu_0 I_k$. Hence,

$$\oint \boldsymbol{B} \cdot d\boldsymbol{l} = \mu_0 \sum_k I_k \tag{6.101}$$

(remember that I_k is an algebraic quantity).

If currents flow in the entire space where a loop is, the algebraic sum of the currents enclosed by the loop can be represented in the form

$$\sum_k I_k = \int_S \boldsymbol{j} \cdot d\boldsymbol{S} = \int_S \boldsymbol{j} \cdot \hat{\boldsymbol{n}} \, dS. \tag{6.102}$$

The integral is taken over the arbitrary surface S enclosing the loop. The vector $\boldsymbol{j}$ is the current density at the point where area element dS is; $\hat{\boldsymbol{n}}$ is a positive normal to this element (*i.e.*, a normal forming a right-handed system with the direction of circumvention of the loop in calculating the circulation).

Substituting Eq. (6.102) for the sum of the currents in Eq. (6.101), we obtain

$$\oint \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{l} = \mu_0 \int_S \boldsymbol{j} \cdot \mathrm{d}\boldsymbol{S}.$$

Transforming the left-band side according to Stokes's theorem, we arrive at the equation

$$\int_S (\nabla \times \boldsymbol{B}) \cdot \mathrm{d}\boldsymbol{S} = \mu_0 \int_S \boldsymbol{j} \cdot \mathrm{d}\boldsymbol{S}.$$

This equation must be obeyed with an arbitrary choice of the surface S over which the integrals are taken. This is possible only if the integrands have identical values at every point. We, thus, arrive at the conclusion that the curl of the magnetic induction vector is proportional to the current density vector at the given point:

$$\nabla \times \boldsymbol{B} = \mu_0 \boldsymbol{j}. \tag{6.103}$$

The proportionality constant in the SI system is μ_0.

We must note that Eqs. (6.101) and (6.103) hold only for the field in a vacuum in the absence of time-varying electric fields.

Thus, we have found the divergence and curl of a magnetic field in a vacuum. Let us compare the equations obtained with the similar equations for an electrostatic field in a vacuum. According to Eqs. (1.112), (1.117), (6.96) and (6.103):

$$\nabla \cdot \boldsymbol{E} = \frac{1}{\varepsilon_0}\rho \quad \text{(the divergence of } \boldsymbol{E} \text{ equals } \rho \text{ divided by } \varepsilon_0)$$

$$\nabla \times \boldsymbol{E} = 0 \quad \text{(the curl of } \boldsymbol{E} \text{ equals zero)}$$

$$\nabla \cdot \boldsymbol{B} = 0 \quad \text{(the divergence of } \boldsymbol{B} \text{ equals zero)}$$

$$\nabla \times \boldsymbol{B} = \mu_0 \boldsymbol{j} \quad \text{(the curl of } \boldsymbol{B} \text{ equals } \mu_0 \text{ multiplied by } \boldsymbol{j}).$$

A comparison of these equations shows that an electrostatic and a magnetic field are of an appreciably different nature. The curl of an electrostatic field equals zero; consequently, an electrostatic field is potential and can be characterized by the scalar potential φ. The curl of a magnetic field at points where there is a current differs from zero. Accordingly, the circulation of the vector $\boldsymbol{B}$ is proportional to the current enclosed by a loop. This is why we cannot ascribe to a magnetic field a scalar potential that would be related to $\boldsymbol{B}$ by an equation similar to Eq. (1.41). This potential would not be unique—upon each circumvention of the loop and return to the initial point it would receive an increment equal to $\mu_0 I$. A field whose curl differs from zero is called a **vortex** or a **solenoidal** one.

Since the divergence of the vector $\boldsymbol{B}$ is zero everywhere, this vector can be represented as the curl of a function $\boldsymbol{A}$:

$$\boldsymbol{B} = \nabla \times \boldsymbol{A}, \tag{6.104}$$

the divergence of a curl always equals zero [see Eq. (1.106)]. The function $\boldsymbol{A}$ is called the **vector potential**. A treatment of the vector potential is beyond the scope of the present book.

6.12. Field of a Solenoid and Toroid

A solenoid is a wire wound in the form of a spiral onto a round cylindrical body. The magnetic field lines of a solenoid are arranged approximately as shown in Fig. 6.27. The direction of these lines inside the solenoid forms a right-handed system with the direction of the current in the turns.

A real solenoid has a current component along its axis. In addition, the linear density of the current j_{lin} (equal to the ratio of the current $\mathrm{d}I$ to an element of solenoid length $\mathrm{d}l$) changes periodically along the solenoid. The average value of this density is

$$\langle j_{\text{lin}} \rangle = \left\langle \frac{\mathrm{d}I}{\mathrm{d}l} \right\rangle = nI, \tag{6.105}$$

where n is the number of solenoid turns per unit length and I the current in the solenoid.

In the science of electromagnetism, a great part is played by an imaginary infinitely long solenoid having no axial current component and, in addition, having a constant linear current density j_{lin} along its entire length. The reason for this is that the field of such a solenoid is homogeneous and is bounded by the volume of the solenoid (similarly, the electric field of an infinite parallel-plate capacitor is homogeneous and is bounded by the volume of the capacitor).

In accordance with what has been said above, let us imagine a solenoid in the form of an infinite thin-walled cylinder around which flows a current of constant linear density

$$j_{\text{lin}} = nI. \tag{6.106}$$

Let us divide the cylinder into identical ring currents—"turns". Examination of Fig. 6.28 shows that each pair of turns arranged symmetrically relative to a plane perpendicular to the solenoid axis sets up a magnetic induction parallel to the axis at any point of this plane. Hence, the resultant of the field at any point inside and outside an infinite solenoid can only have a direction parallel to the axis.

It can be seen from Fig. 6.27 that the directions of the field inside and outside a finite solenoid are opposite. The directions of the fields do not change when the length of a solenoid is increased, and in the limit, when $l \to \infty$, they remain opposite. In an infinite solenoid, as in a finite one, the direction of the field inside the solenoid forms a right-handed system with the direction in which the current

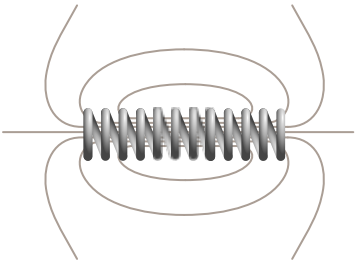


Fig. 6.27

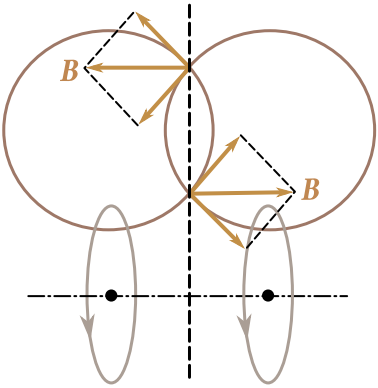


Fig. 6.28

flows around the cylinder.

It follows from the vector $\boldsymbol{B}$ and the axis being parallel that the field both inside and outside an infinite solenoid must be homogeneous. To prove this, let us take an imaginary rectangular loop 1-2-3-4 inside a solenoid (Fig. 6.29; 4-1 is along the axis of the solenoid).

Passing clockwise around the loop, we get the value $(B_2 - B_1)a$ for the circulation of the vector $\boldsymbol{B}$. The loop does not enclose the currents, therefore the circulation must be zero [see Eq. (6.101)]. Hence, it follows that $\boldsymbol{B}_1 = \boldsymbol{B}_2$. Arranging section 2-3 of the loop at any distance from the axis, we shall always find that the magnetic induction $\boldsymbol{B}_2$ at this distance equals the induction $\boldsymbol{B}_1$ on the solenoid axis. Thus, the homogeneity of the field inside the solenoid has been proved.

Now let us turn to loop 1′-2′-3′-4′. We have depicted the vectors $\boldsymbol{B}'_1$ and $\boldsymbol{B}'_2$ by a dash line since, as we shall find out in the following, the field outside an infinite solenoid is zero. Meanwhile, all that we know is that the possible direction of the field outside the solenoid is opposite to that of the field inside it. Loop 1′-2′-3′-4′ does not enclose the currents; therefore, the circulation of the vector $\boldsymbol{B}'$ around this loop, equal to $(B'_1 - B'_2)a$, must be zero. It thus follows that $\boldsymbol{B}'_1 = \boldsymbol{B}'_2$. The distances from the solenoid axis to sections 1′-4′ and 2′-3′ were taken arbitrarily. Consequently, the value of $\boldsymbol{B}'$ at any distance from the axis will be the same outside the solenoid. Thus, the homogeneity of the field outside the solenoid has been proved too.

The circulation around the loop shown in Fig. 6.30 is $a(B + B')$ (for clockwise circumvention). This loop encloses a positive current of magnitude $j_{\text{lin}}a$. In accordance with Eq. (6.101), the following equation must be observed:

$$a(B + B') = \mu_0 j_{\text{lin}} a,$$

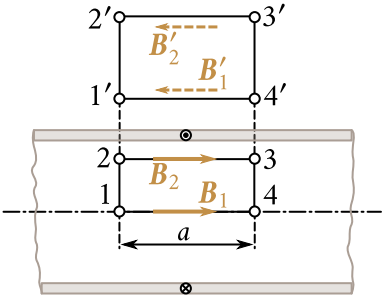


Fig. 6.29

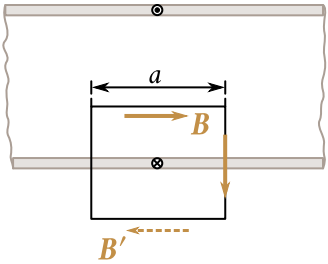


Fig. 6.30

or after cancelling a and replacing j_{lin} with nI [see Eq. (6.106)]

$$(B + B') = \mu_0 nI. \tag{6.107}$$

This equation shows that the field both inside and outside an infinite solenoid is finite.

Let us take a plane at right angles to the solenoid axis (Fig. 6.31). Since the field lines $\boldsymbol{B}$ are closed, the magnetic fluxes through the inner part S of this plane and through its outer part S' must be the same. Since the fields are homogeneous and normal to the plane, each of the fluxes equals the product of the relevant value of the magnetic induction and the area penetrated by the flux. We, thus, get the expression

$$BS = B'S'.$$

The left-hand side of this equation is finite, the factor S' in the right-hand side is infinitely great. Hence, it follows that $B' = 0$.

Thus, we have proved that the magnetic induction outside an infinitely long solenoid is zero. The field inside the solenoid is homogeneous. Assuming in Eq. (6.107) that $B' = 0$, we arrive at an equation for the magnetic induction inside a solenoid:

$$B = \mu_0 nI. \tag{6.108}$$

The product nI is called the number of ampere-turns per metre. At $n = 1000$ turns per metre and a current of 1 A, the magnetic induction inside a solenoid is $4\pi \times 10^{-4}$ T $= 4\pi$Gs.

The symmetrically arranged turns make an identical contribution to the magnetic induction on the axis of a solenoid [see Eq. (6.81)]. Therefore, at the end of a semi-infinite solenoid, the magnetic induction on its axis equals half the value given by Eq. (6.108):

$$B = \frac{1}{2}\mu_0 nI. \tag{6.109}$$

Practically, if the length of a solenoid is considerably greater than its diameter,

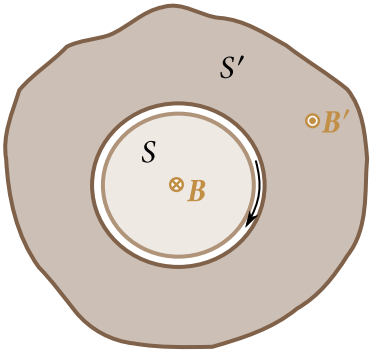


Fig. 6.31

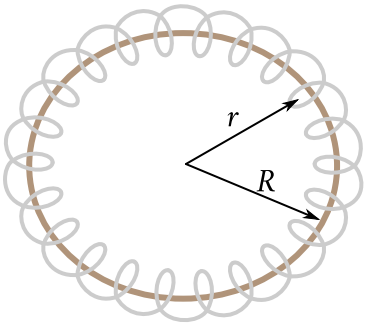


Fig. 6.32

Eq. (6.108) will hold for points in the central part of the solenoid, and Eq. (6.109) for points on its axis near its ends.

A toroid is a wire wound onto a body having the shape of a torus (Fig. 6.32). Let us take a loop in the form of a circle of radius r whose centre coincides with that of a toroid. Owing to symmetry, the vector $\boldsymbol{B}$ at every point must be directed along a tangent to the loop. Hence, the circulation of $\boldsymbol{B}$ is

$$\oint \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{l} = B \times 2\pi r$$

($\boldsymbol{B}$ is the magnetic induction at the points through which the loop passes).

If a loop passes inside a toroid, it encloses the current $2\pi RnI$ (R is the radius of the toroid, and n is the number of turns per unit of its length). In this case,

$$B \times 2\pi r = \mu_0 2\pi RnI,$$

whence

$$B = \mu_0 nI \frac{R}{r}. \tag{6.110}$$

A loop passing outside a toroid encloses no currents, hence, we have $B \times 2\pi r = 0$ for it. Thus, the magnetic induction outside a toroid is zero.

For a toroid whose radius R considerably exceeds the radius of a turn, the ratio R/r for all the points inside the toroid differs only slightly from unity, and instead of Eq. (6.110) we get an equation coinciding with Eq. (6.108) for an infinitely long solenoid. In this case, the field may be considered homogeneous in each of the toroid sections. The field is directed differently in different sections. We can, therefore, speak of the homogeneity of the field within the entire toroid only conditionally, bearing in mind the identical magnitude of $\boldsymbol{B}$.

A real toroid has a current component along its axis. This component sets up a field similar to that of a ring current in addition to the field given by Eq. (6.110).

Chapter 7

MAGNETIC FIELD IN A SUBSTANCE

7.1. Magnetization of a Magnetic

We assumed in the preceding chapter that the conductors carrying a current are in a vacuum. If the conductors carrying a current are in a medium, the magnetic field changes. The explanation is that any substance is a magnetic, *i.e.*, is capable of acquiring a magnetic moment under the action of a magnetic field (of becoming magnetized). The magnetized substance sets up the magnetic field $\boldsymbol{B}'$ that is superposed onto the field $\boldsymbol{B}_0$ produced by the currents. Both fields produce the resultant field

$$\boldsymbol{B} = \boldsymbol{B}_0 + \boldsymbol{B}' \tag{7.1}$$

[compare with Eq. (2.8)].

The true (microscopic) field in a magnetic varies greatly within the limits of intermolecular distances. By $\boldsymbol{B}$ is meant the averaged (macroscopic) field (see Sec. 2.3). To explain the magnetization of bodies, Ampere assumed that ring currents (molecular currents) circulate in the molecules of a substance. Every such current has a magnetic moment and sets up a magnetic field in the surrounding space. In the absence of an external field, the molecular currents are oriented chaotically, owing to which the resultant field set up by them equals zero. The total magnetic moment of a body also equals zero because of the chaotic orientation of the magnetic moments of its separate molecules. The action of a field causes the magnetic moments of the molecules to acquire a predominating orientation in one direction, owing to which the magnetic becomes magnetized–its total magnetic moment becomes other than zero. The magnetic fields of individual molecular currents in this case no longer compensate one another, and the field $\boldsymbol{B}'$ appears.

It is quite natural to characterize the magnetization of a magnetic by the magnetic moment of unit volume. This quantity is called the **magnetization** and is denoted by the symbol $\boldsymbol{M}$. If a magnetic is magnetized inhomogeneously, the magnetization at a given point is determined by the following expression:

$$\boldsymbol{M} = \frac{1}{\Delta V} \sum_{\Delta V} \boldsymbol{p}_{\mathrm{m}}\,, \tag{7.2}$$

where ΔV is an infinitely small volume (from the physical viewpoint) taken in the vicinity of the point being considered, and $\boldsymbol{p}_{\mathrm{m}}$ is the magnetic moment of a separate molecule. Summation is performed over all the molecules confined in the volume ΔV [compare with Eq. (2.4)].

The field $\boldsymbol{B}'$, like the field $\boldsymbol{B}_0$, has no sources. Therefore, the divergence of the resultant field given by Eq. (7.1) is zero:

$$\nabla \cdot \boldsymbol{B} = \nabla \cdot \boldsymbol{B}_0 + \nabla \cdot \boldsymbol{B}' = 0. \tag{7.3}$$

Thus, Eq. (6.96) and, consequently, Eq. (6.95), hold not only for a field in a vacuum, but also for a field in a substance.

7.2. Magnetic Field Strength

Let us write an expression for the curl of the resultant field (7.1):

$$\nabla \times \boldsymbol{B} = \nabla \times \boldsymbol{B}_0 + \nabla \times \boldsymbol{B}'.$$

According to Eq. (6.103), $\nabla \times \boldsymbol{B}_0 = \mu_0 \boldsymbol{j}$, where $\boldsymbol{j}$ is the density of the macroscopic current. Similarly, the curl of the vector $\boldsymbol{B}'$ must be proportional to the density of the molecular currents:

$$\nabla \times \boldsymbol{B}' = \mu_0 \boldsymbol{j}_{\mathrm{mol}}.$$

Consequently, the curl of the resultant field is determined by the equation

$$\nabla \times \boldsymbol{B} = \mu_0 \left(\boldsymbol{j} + \boldsymbol{j}_{\mathrm{mol}}\right). \tag{7.4}$$

Inspection of Eq. (7.4) shows that when calculating the curl of a field in a magnetic, we encounter a difficulty similar to that which we encountered when dealing with an electric field in a dielectric [see Eq. (2.16)]: to determine the curl of $\boldsymbol{B}$, we must know the density not only of the macroscopic, but also of the molecular currents. But the density of the molecular currents, in turn, depends on the value of the vector $\boldsymbol{B}$. The way of circumventing this difficulty is also similar to the one we took advantage of in Sec. 2.5. We are able to find such an auxiliary quantity whose curl is determined only by the density of the macroscopic currents.

To find the form of this auxiliary quantity, let us attempt to express the density of the molecular currents $\boldsymbol{j}_{\mathrm{mol}}$ through the magnetization of a magnetic $\boldsymbol{M}$ (in

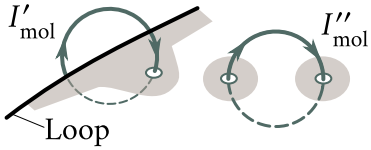


Fig. 7.1

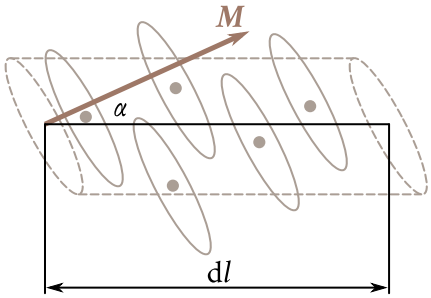


Fig. 7.2

Sec. 2.5 we expressed the density of the bound charges through the polarization of a dielectric $\boldsymbol{P}$). For this purpose, let us calculate the algebraic sum of the molecular currents enclosed by a loop Γ. This sum is

$$\int_S \boldsymbol{j}_{\text{mol}} \cdot \mathrm{d}\boldsymbol{S}, \tag{7.5}$$

where S is the surface enclosing the loop.

The algebraic sum of the molecular currents includes only the molecular currents that are "threaded" onto the loop (see the current I'_{mol} in Fig. 7.1). The currents that are not "threaded" onto the loop either do not intersect the surface enclosing the loop at all, or intersect it twice—once in one direction and once in the opposite one (see the current I''_{mol} in Fig. 7.1). As a result, their contribution to the algebraic sum of the currents enclosed by the loop equals zero.

A glance at Fig. 7.2 shows that the contour element $\mathrm{d}l$ making the angle α with the direction of magnetization $\boldsymbol{M}$ threads onto itself those molecular currents whose centres are inside an oblique cylinder of volume $S_{\text{mol}} \cos\alpha \, \mathrm{d}l$ (where S_{mol} is the area enclosed by a separate molecular current). If n is the number of molecules in unit volume, then the total current enclosed by the element $\mathrm{d}l$ is $I_{\text{mol}} S_{\text{mol}} n \cos\alpha \, \mathrm{d}l$. The product $I_{\text{mol}} S_{\text{mol}}$ equals the magnetic moment p_{m} of an individual molecular current. Hence, the expression $I_{\text{mol}} S_{\text{mol}} n$ is the magnetic moment of unit volume, *i.e.*, it gives the magnitude of the vector $\boldsymbol{M}$, while $I_{\text{mol}} S_{\text{mol}} \cos\alpha$ gives the projection of the vector $\boldsymbol{M}$ onto the direction of the element $\mathrm{d}l$. Thus, the total molecular current enclosed by the element $\mathrm{d}l$ is $\boldsymbol{M} \cdot \mathrm{d}\boldsymbol{l}$, while the sum of the molecular currents enclosed by the entire loop [see Eq. (7.5)] is

$$\int_S \boldsymbol{j}_{\text{mol}} \cdot \mathrm{d}\boldsymbol{S} = \oint \boldsymbol{M} \cdot \mathrm{d}\boldsymbol{l}.$$

Transforming the right-hand side according to Stokes's theorem, we get

$$\int_S \boldsymbol{j}_{\text{mol}} \cdot \mathrm{d}\boldsymbol{S} = \int_S (\nabla \times \boldsymbol{M}) \cdot \mathrm{d}\boldsymbol{S}.$$

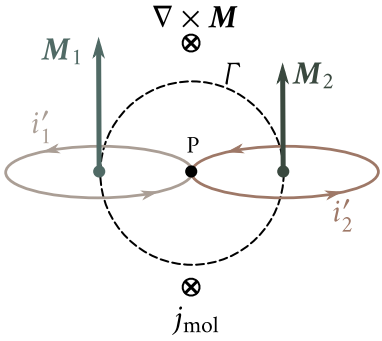


Fig. 7.3

The equation which we have arrived at must be obeyed when the surface S has been chosen arbitrarily. This is possible only if the integrands are equal at every point of a magnetic:

$$\boldsymbol{j}_{\text{mol}} = \nabla \times \boldsymbol{M}. \tag{7.6}$$

Thus, the density of the molecular currents is determined by the value of the curl of the magnetization. When $\nabla \times \boldsymbol{M} = 0$, the molecular currents of individual molecules are oriented so that their sum on an average is zero.

Equation (7.6) allows us to make the following illustrative interpretation. Figure 7.3 shows the magnetization vectors $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ in direct proximity to a certain point P. This point and both vectors are in the plane of the drawing. Loop Γ depicted by a dash line is also in the plane of the drawing. If the nature of the magnetization is such that the vectors $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ are identical in magnitude, then the circulation of $\boldsymbol{M}$ around loop Γ will be zero. Accordingly, $\nabla \times \boldsymbol{M}$ at point P will also be zero.

The molecular currents i'_1 and i'_2 flowing in the loops depicted in Fig. 7.3 by solid lines can be compared with the magnetizations $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$. These loops are in a plane normal to the plane of the drawing. With an identical direction of the vectors $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$, the directions of the currents i'_1 and i'_2 at point P will be opposite. Since $M_1 = M_2$, the currents i'_1 and i'_2 are identical in magnitude, owing to which the resultant molecular current at point P, like $\nabla \times \boldsymbol{M}$, will be zero: $\boldsymbol{j}_{\text{mol}} = 0$.

Now let us assume that $M_1 > M_2$. Therefore, the circulation of $\boldsymbol{M}$ around loop Γ will differ frow zero. Accordingly, the field of the vector $\boldsymbol{M}$ at point P will be characterized by the vector $\nabla \times \boldsymbol{M}$ directed beyond the drawing. A greater molecular current corresponds to a greater magnetization; hence, $i'_1 > i'_2$. Consequently, at point P, there will be observed a resultant current other than zero characterized by the density $\boldsymbol{j}_{\text{mol}}$. The latter, like $\nabla \times \boldsymbol{M}$, is directed beyond the drawing. When $M_1 < M_2$, the vectors $\nabla \times \boldsymbol{M}$ and $\boldsymbol{j}_{\text{mol}}$ will be directed toward us instead of beyond the drawing.

Thus, at points where the curl of the magnetization is other than zero, the density of the molecular currents also differs from zero, the vectors $\nabla \times \boldsymbol{M}$ and $\boldsymbol{j}_{\text{mol}}$ having the same direction [see Eq. (7.6)].

Let us introduce Eq. (7.6) for the density of the molecular currents into Eq. (7.4):

$$\nabla \times \boldsymbol{B} = \mu_0 \boldsymbol{j} + \mu_0 \nabla \times \boldsymbol{M}.$$

Dividing this equation by μ_0 and combining the curls, we get

$$\nabla \times \left(\frac{\boldsymbol{B}}{\mu_0} - \boldsymbol{M} \right) = \boldsymbol{j}. \tag{7.7}$$

Whence it follows that

$$\boldsymbol{H} = \frac{\boldsymbol{B}}{\mu_0} - \boldsymbol{M}, \tag{7.8}$$

is our required auxiliary quantity whose curl is determined only by the macroscopic currents. This quantity is called the **magnetic tield strength.**

In accordance with Eq. (7.7),

$$\nabla \times \boldsymbol{H} = \boldsymbol{j} \tag{7.9}$$

(the curl of the vector $\boldsymbol{H}$ equals the vector of the density of the macroscopic currents).

Let us take an arbitrary loop Γ enclosed by surface S and form the expression

$$\int_S \nabla \times \boldsymbol{H} \cdot \mathrm{d}\boldsymbol{S} = \int_S \boldsymbol{j} \cdot \mathrm{d}\boldsymbol{S}.$$

According to Stokes's theorem, the left-hand side of this equation is equivalent to the circulation of the vector $\boldsymbol{H}$ around loop Γ. Hence,

$$\oint_\Gamma \boldsymbol{H} \cdot \mathrm{d}\boldsymbol{l} = \int_S \boldsymbol{j} \cdot \mathrm{d}\boldsymbol{S}. \tag{7.10}$$

If macroscopic currents flow through wires enclosed by a loop, Eq. (7.10) can be written in the form

$$\oint_\Gamma \boldsymbol{H} \cdot \mathrm{d}\boldsymbol{l} = \sum_k I_k. \tag{7.11}$$

Equations (7.10) and (7.11) express the theorem on the circulation of the vector $\boldsymbol{H}$: *the circulation of the magnetic field strength vector around a loop equals the algebraic sum of the macroscopic currents enclosed by this loop.*

The magnetic field strength $\boldsymbol{H}$ is the analogue of the electric displacement $\boldsymbol{D}$. It was originally assumed that magnetic masses similar to electric charges exist in nature, and the science of magnetism developed along the lines of that of electricity. Back in those times, the relevant names were introduced: the "magnetic induction" for $\boldsymbol{B}$ and the "field strength" (formerly "field intensity") for $\boldsymbol{H}$. It was later established that no magnetic masses exist in nature and that the quantity called the magnetic induction is actually the analogue not of the electric displacement $\boldsymbol{D}$, but of the

electric field strength $\boldsymbol{E}$ (accordingly, $\boldsymbol{H}$ is the analogue of $\boldsymbol{D}$ instead of $\boldsymbol{E}$). It was decided not to change the established terminology, however, moreover because owing to the different nature of an electric and a magnetic field (an electrostatic field is potential, a magnetic one is solenoidal[1]) the quantities $\boldsymbol{B}$ and $\boldsymbol{D}$ display many similarities in their behaviour (for example, the $\boldsymbol{B}$ lines, like the $\boldsymbol{D}$ lines, are not disrupted at the interface between two media).

In a vacuum, $\boldsymbol{M} = 0$, therefore, $\boldsymbol{H}$ transforms into $\boldsymbol{B}/\mu_0$ and Eqs. (7.10) and (7.11) transform into Eqs. (6.103) and (6.101).

In accordance with Eq. (6.30), the strength of the field of a line current in a vacuum is determined by the expression

$$H = \frac{1}{4\pi}\frac{2I}{b}, \tag{7.12}$$

whence it can be seen that the magnetic field strength has a dimension equal to that of current divided by that of length. In this connection, the SI unit of magnetic field strength is called the ampere per metre ($\mathrm{A\,m^{-1}}$).

In the Gaussian system, the magnetic field strength is defined as the quantity

$$\boldsymbol{H} = \boldsymbol{B} - 4\pi\boldsymbol{M}. \tag{7.13}$$

It follows from this definition that in a vacuum $\boldsymbol{H}$ coincides with $\boldsymbol{B}$. Accordingly, the unit of $\boldsymbol{H}$ in the Gaussian system, called the **oersted** (Oe), has the same value and dimension as the unit of magnetic induction—the gauss (Gs). In essence, the oersted and gauss are different names of the same unit. If the latter measures $\boldsymbol{H}$, it is called the oersted, and if it measures $\boldsymbol{B}$, the gauss.

It is customary practice to associate the magnetization not with the magnetic induction, but with the field strength. It is assumed that at every point of a magnetic

$$\boldsymbol{M} = \chi_{\mathrm{m}}\boldsymbol{H}, \tag{7.14}$$

where χ_{m} is a quantity characteristic of a given magnetic and called the **magnetic susceptibility**[2]. Experiments show that for weakly magnetic (non-ferromagnetic) substances in not too strong fields χ_{m} is independent of $\boldsymbol{H}$. According to Eq. (7.8), the dimension of $\boldsymbol{H}$ coincides with that of $\boldsymbol{M}$. Hence, χ_{m} is a dimensionless quantity.

Using Eq. (7.14) for $\boldsymbol{M}$ in Eq. (7.8), we get

$$\boldsymbol{H} = \frac{\boldsymbol{B}}{\mu_0} - \chi_{\mathrm{m}}\boldsymbol{H},$$

[1] A solenoidal field is one having no sources. At each point of such a field, the divergence is zero.

[2] In anisotropic media, the directions of the vectors $\boldsymbol{M}$ and $\boldsymbol{H}$, generally speaking, do not coincide. For such media, the relatiom between the vectors $\boldsymbol{M}$ and $\boldsymbol{H}$ is achieved by means of the **magnetic susceptibility tensor** (see the footnote number 2 on page 55).

whence

$$\boldsymbol{H} = \frac{\boldsymbol{B}}{\mu_0(1+\chi_\mathrm{m})}. \tag{7.15}$$

The dimensionless quantity

$$\mu = 1 + \chi_\mathrm{m}, \tag{7.16}$$

is called the **relative permeability** or simply the **permeability** of a substance[3].

Unlike the dielectric susceptibility χ that can have only positive values (the polarization $\boldsymbol{P}$ in an isotropic dielectric is always directed along the $\boldsymbol{E}$ field), the magnetic susceptibility χ_m may be either positive or negative. Hence, the permeability may be either greater or smaller than unity.

With account taken of Eq. (7.16), Eq. (7.15) can be written as follows:

$$\boldsymbol{H} = \frac{\boldsymbol{B}}{\mu_0\mu}. \tag{7.17}$$

Thus, the magnetic field strength $\boldsymbol{H}$ is a vector having the same direction as the vector $\boldsymbol{B}$, but whose magnitude is $\mu_0\mu$ times smaller (in anisotropic media the vectors $\boldsymbol{H}$ and $\boldsymbol{B}$, generally speaking, do not coincide in direction).

Equation (7.14) relating the vectors $\boldsymbol{M}$ and $\boldsymbol{H}$ has exactly the same form in the Gaussian system too. Using this equation in Eq. (7.13), we get

$$\boldsymbol{H} = \boldsymbol{B} - 4\pi\chi_\mathrm{m}\boldsymbol{H},$$

whence

$$\boldsymbol{H} = \frac{\boldsymbol{B}}{1+4\pi\chi_\mathrm{m}}. \tag{7.18}$$

The dimensionless quantity

$$\mu = 1 + 4\pi\chi_\mathrm{m}, \tag{7.19}$$

is called the **permeability** of a substance. Introducing this quantity into Eq. (7.18), we get

$$\boldsymbol{H} = \frac{\boldsymbol{B}}{\mu}. \tag{7.20}$$

The value of μ in the Gaussian system of units coincides with its value in the SI. A comparison of Eqs. (7.16) and (7.19) shows that the value of the magnetic susceptibility in the SI is 4π times that of χ_m in the Gaussian system:

$$\chi_\mathrm{m,SI} = 4\pi\chi_\mathrm{m,Gs}. \tag{7.21}$$

[3]The so-called absolute permeability $\mu_\mathrm{a} = \mu_0\mu$ is introduced in electrical engineering. This quantity is deprived of a physical meaning, however, and we shall not use it.

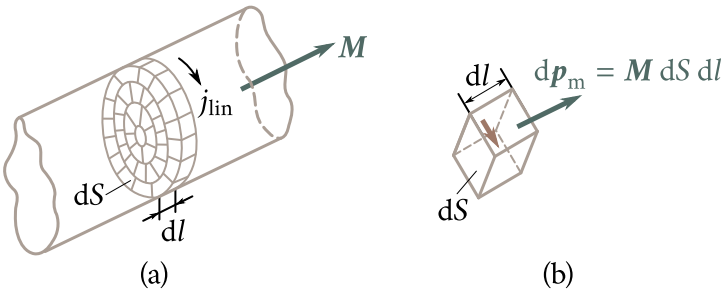


Fig. 7.4

7.3. Calculation of the Field in Magnetics

Let us consider the field produced by an infinitely long round magnetized rod. We shall consider the magnetization $\boldsymbol{M}$ to be the same everywhere and directed along the axis of the rod. Let us divide the rod mentally into layers of thickness dl at right angles to the axis. We shall divide each layer in turn into small cylindrical elements with bases of an arbitrary shape and of area dS (Fig. 7.4a). Each such element has the magnetic moment

$$dp_m = M\,dS\,dl. \tag{7.22}$$

The field $d\boldsymbol{B}'$ set up by an element at distances that are great in comparison with its dimensions is equivalent to the field that would produce the current $I = M\,dl$ flowing around the element along its side surface (see Fig. 7.4b). Indeed, the magnetic moment of such a current is $dp_m = I\,dS = M\,dl\,dS$ [compare with Eq. (7.22)], while the magnetic field at great distances is determined only by the magnitude and direction of the magnetic moment (see Sec. 6.9).

The imaginary currents flowing in the section of the surface common for two adjacent elements are identical in magnitude and opposite in direction, therefore their sum is zero. Thus, when summating the currents flowing around the side surfaces of the elements of one layer, only the currents flowing along the side surface of the layer will remain uncompensated.

It follows from the above that a rod layer of thickness dl sets up a field equivalent to the one which would be produced by the current $M\,dl$ flowing around the layer along its side surface (the linear density of this current is $j_{lin} = M$). The entire infinite magnetized rod sets up a field equivalent to the field of a cylinder around which flows a current having the linear density $j_{lin} = M$. We established in Sec. 6.2 that outside such a cylinder the field vanihes, while inside it the field is homogeneous and equals $\mu_0 j_{lin}$ in magnitude.

We have, thus, determined the nature of the field $\boldsymbol{B}'$ set up by a homogeneously

magnetized infinitely long round rod. Outside the rod, the field vanishes. Inside it, the field is homogeneous and equals

$$\boldsymbol{B}' = \mu_0 \boldsymbol{M}. \tag{7.23}$$

Assume that we have a homogeneous field $\boldsymbol{B}_0$ set up by macrocurrents in a vacuum. According to Eq. (7.17), the strength of this field is

$$\boldsymbol{H}_0 = \frac{\boldsymbol{B}_0}{\mu_0}. \tag{7.24}$$

Let us introduce into this field (we shall call it an external one) an infinitely long round rod of a homogeneous and isotropic magnetic, arranging it along the direction of $\boldsymbol{B}_0$. It follows from considerations of symmetry that the magnetization $\boldsymbol{M}$ set up in the rod is collinear with the vector $\boldsymbol{B}_0$.

The magnetized rod produces inside itself the field $\boldsymbol{B}'$ determined by Eq. (7.23). The field inside the rod, as a result, becomes equal to

$$\boldsymbol{B} = \boldsymbol{B}_0 + \boldsymbol{B}' = \boldsymbol{B}_0 + \mu_0 \boldsymbol{M}. \tag{7.25}$$

Using this value of $\boldsymbol{B}$ in Eq. (7.8), we get the strength of the field inside the rod

$$\boldsymbol{H} = \frac{\boldsymbol{B}}{\mu_0} - \boldsymbol{M} = \frac{\boldsymbol{B}_0}{\mu_0} = \boldsymbol{H}_0$$

[see Eq. (7.24)]. Thus, the strength of the field in the rod coincides with that of the external field.

Multiplying $\boldsymbol{H}$ by $\mu_0\mu$ we get the magnetic induction inside the rod:

$$\boldsymbol{B} = \mu_0 \mu \boldsymbol{H} = \mu_0 \mu \frac{\boldsymbol{B}_0}{\mu_0} = \mu \boldsymbol{B}_0. \tag{7.26}$$

Hence, it follows that the permeability μ shows how many times the field increases in a magnetic [compare with Eq. (7.26)].

It must be noted that since the field $\boldsymbol{B}'$ is other than zero only inside the rod, the magnetic field outside the rod remains unchanged.

The result we have obtained is correct when a homogeneous and isotropic magnetic fills the volume bounded by surfaces formed by the strength lines of the external field[4]. Otherwise, the field strength determined by Eq. (7.8) does not coincide with $\boldsymbol{H}_0 = \boldsymbol{B}_0/\mu_0$.

It is conditionally assumed that the field strength in a magnetic is

$$\boldsymbol{H} = \boldsymbol{H}_0 - \boldsymbol{H}_\mathrm{d}, \tag{7.27}$$

where $\boldsymbol{H}_0$ is the external field, and $\boldsymbol{H}_\mathrm{d}$ is the so-called **demagnetizing field**. The

[4]We remind our reader that for an electric field $\boldsymbol{D} = \boldsymbol{D}_0$ provided that a homogeneous and isotropic dielectric fills the volume bounded by equipotential surfaces, *i.e.*, surfaces orthogonal to the strength lines of the external field.

latter is assumed to be proportional to the magnetization

$$\boldsymbol{H}_{\mathrm{d}} = N\boldsymbol{M}. \tag{7.28}$$

The proportionality constant N is known as the **demagnetization factor**. It depends on the shape of a magnetic. We have seen that $\boldsymbol{H} = \boldsymbol{H}_0$ for a body whose surface is not intersected by strength lines of the external field, *i.e.*, the demagnetization factor is zero. For a thin disk perpendicular to the external field, $N = 1$, and for a sphere, $N = 1/3$.

The relevant calculations show that when a homogeneous and isotropic magnetic having the shape of an ellipsoid is placed in a homogeneous external field, the magnetic field in it is also homogeneous, although it differs from the external one. This also holds for a sphere, which is a particular case of an ellipsoid, and for a long rod and a thin disk, which can be considered as the extreme cases of an ellipsoid.

In concluding, let us find the field strength of an infinitely long solenoid filled with a homogeneous and isotropic magnetic (or submerged in an infinite homogeneous and isotropic magnetic.). Applying the theorem on circulation [see Eq. (7.11)] to the loop shown in Fig. 6.30, we get the equation $Ha = naI$. Hence,

$$H = nI. \tag{7.29}$$

Thus, the field strength inside an infinitely long solenoid equals the product of the current and the number of turns per unit length. Outside the solenoid, the field strength vanishes.

7.4. Conditions at the Interface of Two Magnetics

Near the interface of two magnetics, the vectors $\boldsymbol{B}$ and $\boldsymbol{H}$ must comply with definite boundary conditions that follow from the relations

$$\nabla \cdot \boldsymbol{B} = 0, \quad \nabla \times \boldsymbol{H} = \boldsymbol{j} \tag{7.30}$$

[see Eqs. (7.3) and (7.9)]. We are considering stationary fields, *i.e.*, ones that do not vary with time.

Let us take on the interface of two magnetics of permeabilities μ_1 and μ_2 an imaginary cylindrical surface of height h with bases S_1 and S_2 at different sides of the interface (Fig. 7.5). The flux of the vector $\boldsymbol{B}$ through this interface is

$$\Phi_B = B_{1,\mathrm{n}}S + B_{2,\mathrm{n}}S + \langle B_{\mathrm{n}} \rangle S_{\mathrm{side}} \tag{7.31}$$

[compare with Eq. (2.46)].

Since $\nabla \cdot \boldsymbol{B} = 0$, the flux of the vector $\boldsymbol{B}$ through any closed surface is zero. Equating expression (7.31) to zero and making the transition $h \to 0$, we arrive at the equation $B_{1,\mathrm{n}} = -B_{2,\mathrm{n}}$. If we project $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ onto the same normal, we get

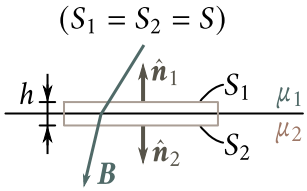


Fig. 7.5

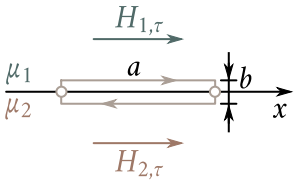


Fig. 7.6

the condition

$$B_{1,\mathrm{n}} = B_{2,\mathrm{n}} \tag{7.32}$$

[compare with Eq. (2.47)].

Replacing in accordance with Eq. (7.17) the components of $\boldsymbol{B}$ with the corresponding components of $\boldsymbol{H}$ multiplied by $\mu_0\mu$, we get the equation

$$\mu_0\mu_1 H_{1,\mathrm{n}} = \mu_0\mu_2 H_{2,\mathrm{n}},$$

whence

$$\frac{H_{1,\mathrm{n}}}{H_{2,\mathrm{n}}} = \frac{\mu_2}{\mu_1}. \tag{7.33}$$

Now let us take a rectangular loop on the interface of the magnetics (Fig. 7.6) and calculate the circulation of $\boldsymbol{H}$ for it. With small dimensions of the loop, the circulation can be written in the form

$$\oint H_l \, \mathrm{d}l = H_{1,\tau}a - H_{2,\tau}a + \langle H_l \rangle \, 2b, \tag{7.34}$$

where $\langle H_l \rangle$ is the average value of $\boldsymbol{H}_l$ on the parts of the loop at right angles to the interface. If no macroscopic currents flow along the interface of the magnetics, $\nabla \times \boldsymbol{H}$ within the limits of the loop will equal zero. Consequently, the circulation will also be zero. Assuming that Eq. (7.34) is zero and performing the limit transition $b \to 0$, we arrive at the expression

$$H_{1,\tau} = H_{2,\tau} \tag{7.35}$$

[compare with Eq. (2.44)].

Replacing the components of $\boldsymbol{H}$ with the corresponding components of $\boldsymbol{B}$ divided by $\mu_0\mu$, we get the relation

$$\frac{B_{1,\tau}}{\mu_0\mu_1} = \frac{B_{2,\tau}}{\mu_0\mu_2},$$

whence it follows that

$$\frac{B_{1,\tau}}{B_{2,\tau}} = \frac{\mu_1}{\mu_2}. \tag{7.36}$$

Summarizing, we can say that in passing through the interface between two magnetics, the normal component of the vector $\boldsymbol{B}$ and the tangential component of

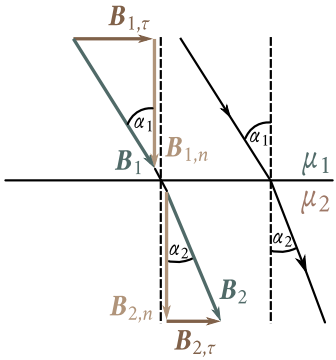


Fig. 7.7

the vector $\boldsymbol{H}$ change continuously.

The tangential component of the vector $\boldsymbol{B}$ and the normal component of the vector $\boldsymbol{H}$ in passing through the interface of the magnetics, however, experience a discontinuity. Thus, when passing through the interface of two media, the vector $\boldsymbol{B}$ behaves similar to the vector $\boldsymbol{D}$, and the vector $\boldsymbol{H}$ similar to the vector $\boldsymbol{E}$.

Figure 7.7 shows the behaviour of the $\boldsymbol{B}$ lines when intersecting the surface between two magnetics. Let the angles between the $\boldsymbol{B}$ lines and a normal to the interface be α_1 and α_2, respectively. The ratio of the tangents of these angles is

$$\frac{\tan\alpha_1}{\tan\alpha_2} = \frac{B_{1,\tau}/B_{1,\mathrm{n}}}{B_{2,\tau}/B_{2,\mathrm{n}}},$$

whence with a view to Eqs. (7.32) and (7.36) we get a law of refraction of the magnetic field lines similar to Eq. (2.49):

$$\frac{\tan\alpha_1}{\tan\alpha_2} = \frac{\mu_1}{\mu_2}. \tag{7.37}$$

Upon passing into a magnetic with a greater value of μ, the magnetic field lines deviate from a normal to the surface. This leads to crowding of the lines. The crowding of the $\boldsymbol{B}$ lines in a substance with a great permeability makes it possible to form magnetic beams, *i.e.*, impart the required shape and direction to them. In particular, for magnetic shielding of a space, it is surrounded with an iron screen. A glance at Fig. 7.8 shows that the crowding of the magnetic field lines in the body of the screen results in weakening of the field inside it.

Figure 7.9 is a schematic view of a laboratory electromagnet. It consists of an iron core onto which coils supplied with a current are fitted. The magnetic field lines are mainly concentrated inside the core. Only in the narrow air gap do they pass in a medium with a low value of μ. The vector $\boldsymbol{B}$ intersects the boundaries between the air gap and the core along a normal to the interface. It thus follows,

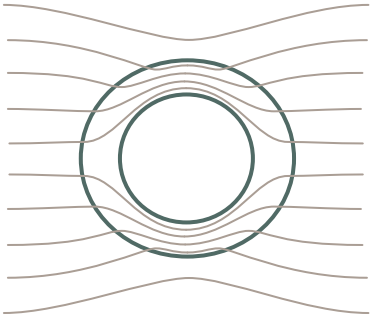


Fig. 7.8

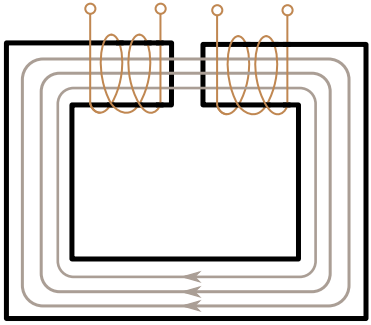


Fig. 7.9

in accordance with Eq. (7.32), that the magnetic induction in the gap and in the core is identical in value. Let us apply the theorem on the circulation of $\boldsymbol{H}$ to the loop along the axis of the core. We can assume that the field strength is identical everywhere in the iron and is $H_{\text{iron}} = B/(\mu_0\mu_{\text{iron}})$. In the air, $H_{\text{air}} = B/(\mu_0\mu_{\text{air}})$. Let us denote the length of the loop section in the iron by l_{iron} and in the gap by l_{air}. The Circulation can, thus, be written in the form $H_{\text{iron}}l_{\text{iron}} + H_{\text{air}}l_{\text{air}}$. According to Eq. (7.11), this circulation must equal NI, where N is the total number of turns of the electromagnet coils, and I is the current. Thus,

$$\frac{B}{\mu_0\mu_{\text{iron}}}l_{\text{iron}} + \frac{B}{\mu_0\mu_{\text{air}}}l_{\text{air}} = NI.$$

Hence,

$$B = \mu_0 I \frac{N}{\left(\dfrac{l_{\text{air}}}{\mu_{\text{air}}} + \dfrac{l_{\text{iron}}}{\mu_{\text{iron}}}\right)} \approx \mu_0 I \frac{N}{\left(l_{\text{air}} + \dfrac{l_{\text{iron}}}{\mu_{\text{iron}}}\right)}$$

(μ_{air} differs from unity only in the fifth digit after the decimal point).

Usually, l_{air} is of the order of 0.1 m, l_{iron} is of the order of 1 m, while μ_{iron} reaches values of the order of several thousands. We may, therefore, disregard the second addend in the denominator and write that

$$B = \mu_0 I \frac{N}{l_{\text{air}}}. \tag{7.38}$$

Consequently, the magnetic induction in the gap of an electromagnet has the same value as it would have inside a toroid without a core when N/l_{air} turns are wound on the torus per unit length [see Eq. (6.110)]. By increasing the total number of turns and reducing the dimensions of the air gap, we can obtain fields with a high value of B. In practice, fields with B of the order of several teslas (several tens of thousands of gausses) are obtained with the aid of electromagnets having an iron core.

7.5. Kinds of Magnetics

Equation (7.14) determines the magnetic susceptibility χ_m of a unit volume of a substance. This susceptibility is often replaced with the molar (for chemically simple substances—the atomic) susceptibility $\chi_{m,mol}$ ($\chi_{m,at}$) related to one mole of a substance. It is evident that $\chi_{m,mol} = \chi_m V_{mol}$, where V_{mol} is the volume of a mole of a substance. Whereas χ_m is a dimensionless quantity, $\chi_{m,mol}$ is measured in $m^3\ mol^{-1}$.

Depending on the sign and magnitude of the magnetic susceptibility, all magnetics are divided into three groups:

(1) **diamagnetics**, for which χ_m is negative and small in absolute value ($|\chi_{m,mol}|$ is about $10^{-11}\ m^3\ mol^{-1}$ to $10^{-10}\ m^3\ mol^{-1}$);

(2) **paramagnetics**, for which χ_m is also not great, but positive ($\chi_{m,mol}$ is about $10^{-10}\ m^3\ mol^{-1}$ to $10^{-10}\ m^3\ mol^{-1}$);

(3) **ferromagnetics**, for which χ_m is positive and reaches very great values ($\chi_{m,mol}$ is about $1\ m^3\ mol^{-1}$). In addition, unlike diamagnetics and paramagnetics for which χ_m does not depend on H, the susceptibility of ferromagnetics is a function of the magnetic field strength.

Thus, the magnetization $\boldsymbol{M}$ in isotropic substances may either coincide in direction with $\boldsymbol{H}$ (in paramagnetics and ferromagnetics), or be directed oppositely to it (in diamagnetics). We remind our reader that in isotropic dielectrics the polarization is always directed in the same way as $\boldsymbol{E}$.

7.6. Gyromagnetic Phenomena

The nature of molecular currents became clear after the British physicist Ernest Rutherford (1871-1937) established experimentally that the atoms of all substances consist of a positively charged nucleus and negatively charged electrons travelling around it.

The motion of electrons in atoms obeys quantum laws; in particular, the concept of a trajectory cannot be applied to the electrons travelling in an atom. The diamagnetism of a substance can be explained, however, by using the very simple Bohr model of an atom. According to this model, the electrons in atoms travel along stationary circular orbits.

Assume that an electron is moving with the speed v in an orbit of radius r (Fig. 7.10). The charge ev, where e is the charge of an electron and ν is its number of revolutions a second, will be carried through an area at any place along the path of the electron in one second. Hence, an electron travelling in orbit will form the ring current $I = ev$. Since the charge of an electron is negative, the direction of motion

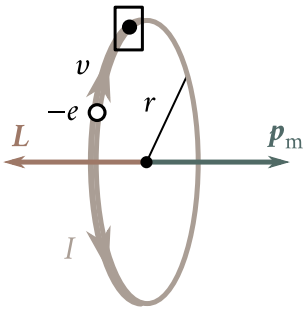


Fig. 7.10

of the electron and the direction of the current will be opposite. The magnetic moment of the current set up by an electron is

$$p_{\mathrm{m}} = IS = e\nu\pi r^3.$$

The product $2\pi r\nu$ gives the speed of the electron v, therefore, we can write that

$$p_{\mathrm{m}} = \frac{evr}{2}. \tag{7.39}$$

The moment (7.39) is due to the motion of an electron in orbit and is, therefore, called the **orbital magnetic moment**. The direction of the vector $\boldsymbol{p}_{\mathrm{m}}$ forms a right-handed system with the direction of the current, and a left-handed one with that of motion of the electron (see Fig. 7.10).

An electron moving in orbit has the angular momentum

$$L = mvr \tag{7.40}$$

(m is the mass of an electron). The vector $\boldsymbol{L}$ is called the **orbital angular momentum** of an electron. It forms a right-handed system with the direction of motion of the electron. Hence, the vectors $\boldsymbol{p}_{\mathrm{m}}$ and $\boldsymbol{L}$ are directed oppositely.

The ratio of the magnetic moment of an elementary particle to its angular momentum is called the **gyromagnetic** (or **magneto mechanical**) ratio. For an electron, it is

$$\frac{p_{\mathrm{m}}}{L} = -\frac{e}{2m} \tag{7.41}$$

(m is the mass of an electron; the minus sign indicates that the magnetic moment and the angular momentum are directed oppositely).

Owing to its rotation about the nucleus, an electron is similar to a spinning top or gyroscope. This circumstance underlies the so-called **gyromagnetic phenomena** consisting in that the magnetization of a magnetic leads to its rotation, and, conversely, the rotation of a magnetic leads to its magnetization. The existence of the first phenomenon was proved experimentally by A. Einstein and W. de Haas, and of the second by S. Barnett.

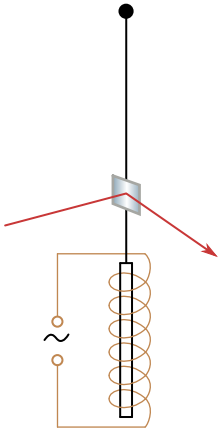


Fig. 7.11

Einstein and de Haas based their experiment on the following reasoning. If we magnetize a rod made of a magnetic, then the magnetic moments of the electrons will be aligned in the direction of the field and the angular momenta in the opposite direction. As a result, the total angular momentum of the electrons $\sum_i \boldsymbol{L}_i$ will become other than zero (initially owing to the chaotic orientation of the individual momenta it equalled zero). The angular momentum of the system rod + electrons must remain unchanged. Therefore, the rod acquires the angular momentum $-\sum_i \boldsymbol{L}_i$ and, consequently, begins to rotate. A change in the direction of magnetization leads to a change in the direction of rotation of the rod.

A mechanical model of this experiment can be carried out by seating a person on a rotatable stool and having him hold a massive rotating wheel in his hands. When he holds the axle of the wheel upward, he begins to rotate in the direction opposite to that of rotation of the wheel. When he turns the axle downward, he begins to rotate in the other direction.

Einstein and de Haas conducted their experiment as follows (Fig. 7.11). A thin iron rod was suspended on an elastic thread and placed inside a solenoid. The thread was twisted very slightly when the rod was magnetized using a constant magnetic field. The resonance method was used to increase the effect—the solenoid was fed with an alternating current whose frequency was chosen equal to the natural frequency of mechanical oscillations of the system. In these conditions, the amplitude of the oscillations reached values that could be measured by watching the displacement of a light spot reflected by a mirror fastened to the thread. The data obtained in the experiment were used to calculate the gyromagnetic ratio, which was found to equal $-(e/m)$. Thus, the sign of the charge of the carriers setting up

the molecular currents coincided with the sign of the charge of an electron. The result obtained, however, was double the expected value of the gyromagnetic ratio (7.41).

To understand Barnett's experiment, we must remember that when an attempt was made to bring a gyroscope into rotation about a certain direction, the gyroscope axis turned so that the directions of the natural and forced rotations of the gyroscope coincided (see Sec. 5.9 of Vol. I). If we place a gyroscope fastened in a universal joint on the disk of a centrifugal machine and begin to rotate it, the gyroscope axis will align itself vertically, and in such a way that the direction of rotation of the gyroscope will coincide with that of the disk. When the direction of rotation of the centrifugal machine is reversed, the gyroscope axis will turn through 180 degrees, *i.e.*, in such a way that the directions of the two rotations will again coincide.

Barnett rotated an iron rod very rapidly about its axis and measured the produced magnetization. Barnett also obtained a value for the gyromagnetic ratio from the results of his experiment double that given by Eq. (7.41).

It was discovered later that apart from the orbital magnetic moment (7.39) and the orbital angular momentum (7.40), an electron has its intrinsic angular momentum L_s and magnetic moment $p_{m,s}$ for which the gyromagnetic ratio is

$$\frac{p_{m,s}}{L_s} = -\frac{e}{m}, \tag{7.42}$$

i.e., coincides with the value obtained in the experiments conducted by Einstein and de Haas and by Barnett. It thus follows that the magnetic properties of iron are due not to the orbital, but to the intrinsic magnetic moment of its electrons.

Attempts were initially made to explain the existence of the intrinsic magnetic moment and angular momentum of an electron by considering it as a charged sphere spinning about its axis. Accordingly, the intrinsic angular momentum of an electron was named its **spin**. It was discovered quite soon, however, that such a notion results in a number of contradictions, and it became necessary to reject the hypothesis of a "spinning" electron. It is assumed at present that the intrinsic angular momentum (spin) and the intrinsic (spin) magnetic moment associated with it are inherent properties of an electron like its mass and charge.

Not only electrons, but also other elementary particles have a spin. The spin[5] of elementary particles is an integral or half-integral multiple of the quantity $\hbar$ equal to Planck's constant h divided by 2π

$$\hbar = \frac{h}{2\pi} = 1.05 \times 10^{-34}\,\text{J s} = 1.05 \times 10^{-2}\,\text{erg s}. \tag{7.43}$$

[5]More exactly, the maximum value of the projection of the spin onto a direction separated in space, for example, onto that of the extenal field.

In particular, for an electron, $L_s = \hbar/2$; in this connection, the spin of an electron is said to equal $1/2$. Thus, $\hbar$ is a natural unit of the angular momentum like the elementary charge e is a natural unit of charge.

In accordance with Eq. (7.42), the intrinsic magnetic moment of an electron is

$$p_m = -\frac{e}{m}L_s = -\frac{e}{m}\frac{\hbar}{2} = -\frac{e\hbar}{2m}. \tag{7.44}$$

The quantity[6]

$$\mu_B = \frac{e\hbar}{2m} = 0.927 \times 10^{-23}\,\mathrm{J\,T^{-1}} = 0.927 \times 10^{-20}\,\mathrm{erg\,Gs^{-1}} \tag{7.45}$$

is called the **Bohr magneton**. Hence, the intrinsic magnetic moment of an electron equals one Bohr magneton.

The magnetic moment of an atom consists of the orbital and intrinsic moments of the electrons in it, and also of the magnetic moment of the nucleus (which is due to the magnetic moments of the elementary particles—protons and neutrons—forming the nucleus). The magnetic moment of a nucleus is much smaller than the moments of the electrons. For this reason, they may be disregarded when considering many questions, and we may consider the magnetic moment of an atom to equal the vector sum of the magnetic moments of its electrons. The magnetic moment of a molecule may also be considered equal to the sum of the magnetic moments of all its electrons.

o. Stern and W. Gerlach determined the magnetic moments of atoms experimentally. They passed a beam of atoms through a greatly inhomogeneous magnetic field. The inhomogeneity of the field was achieved by using a special shape of the electromagnet pole shoes (Fig. 7.12). By Eq. (6.77), the atoms of the beam must experience the force

$$F = p_m \frac{\partial B}{\partial x}\cos\alpha,$$

whose magnitude and sign depend on the angle α made by the vector $\boldsymbol{p}_m$ with the direction of the field. When the moments of the atoms are distributed chaotically by directions, the beam contains particles for which the values of α vary within the limits from 0 to π. It was assumed accordingly that a narrow beam of atoms after passing between the poles would form on a screen a continuous extended trace whose edges would correspond to atoms having orientations at angles of $\alpha = 0$ and $a = \pi$ (Fig. 7.13). The experiment gave unexpected results. Instead of a continuous extended trace, separate lines were obtained that were arranged symmetrically with respect to the trace of the beam obtained in the absence of a field.

[6]According to the equation $W = -\boldsymbol{p}_m \cdot \boldsymbol{B}$, the dimension of magnetic moment equals that of energy (joule or erg) divided by the dimension of magnetic induction (tesla or gauss).

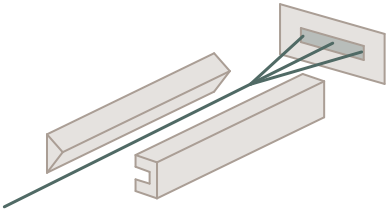


Fig. 7.12

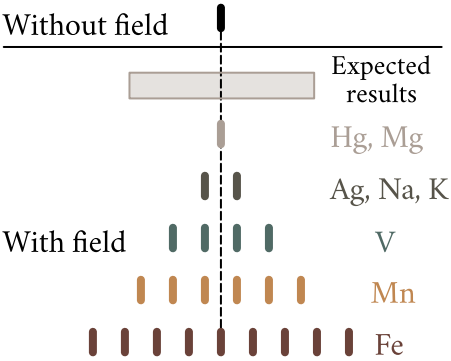


Fig. 7.13

The Stern-Gerlach experiment showed that the angles at which the magnetic moments of atoms are oriented relative to a magnetic field can have only discrete values, *i.e.*, that the projection of a magnetic moment onto the direction of a field is quantized.

The number of possible values of the projection of the magnetic moment onto the direction of the magnetic field for different atoms is different. It is two for silver, aluminium, copper, and the alkali metals, four for vanadium, nitrogen, and the halogens, five for oxygen, six for manganese, nine for iron, ten for cobalt, etc.

Measurements gave values of the order of several Bohr magnetons for the magnetic moments of atoms. Some atoms showed no deflections (see, for example, the trace of mercury and magnesium atoms in Fig. 7.13), which indicates that they have no magnetic moment.

7.7. Diamagnetism

An electron travelling in an orbit is like a spinning top. Therefore, all the features of behaviour of gyroscopes under the action of external forces must be inherent in it, in particular, precession of the electron orbit must appear in the appropriate conditions. The conditions needed for precession appear if an atom is in an external magnetic field $\boldsymbol{B}$ (Fig. 7.14). In this case, the torque $\boldsymbol{T} = \boldsymbol{p}_\mathrm{m} \times \boldsymbol{B}$ is exerted on the orbit. It tends to set up the orbital magnetic moment of an electron $\boldsymbol{p}_\mathrm{m}$ in the direction of the field (the angular momentum $\boldsymbol{L}$ will be set np against the field). The torque $\boldsymbol{T}$ causes the vectors $\boldsymbol{p}_\mathrm{m}$ and $\boldsymbol{L}$ to precess about the direction of the magnetic induction vector $\boldsymbol{B}$ whose velocity is simple to find (see Sec. 5.9 of Vol. I).

During the time dt, the vector $\boldsymbol{L}$ receives the increment d$\boldsymbol{L}$ equal to

$$\mathrm{d}\boldsymbol{L} = \boldsymbol{T}\,\mathrm{d}t.$$

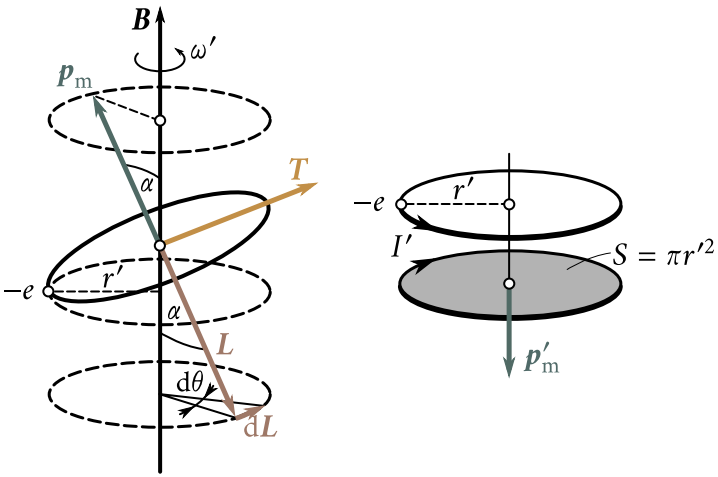


Fig. 7.14

The vector d$\boldsymbol{L}$ like the vector $\boldsymbol{T}$, is perpendicular to the plane passing through the vectors $\boldsymbol{B}$ and $\boldsymbol{L}$; its magnitude is

$$|\mathrm{d}\boldsymbol{L}| = p_\mathrm{m} B \sin\alpha \, \mathrm{d}t,$$

where α is the angle between $\boldsymbol{p}_\mathrm{m}$ and $\boldsymbol{B}$.

During the time dt, the plane containing the vector $\boldsymbol{L}$ will turn about the direction of $\boldsymbol{B}$ through the angle

$$\mathrm{d}\theta = \frac{|\mathrm{d}L|}{L\sin\alpha} = \frac{p_\mathrm{m} B \sin\alpha \, \mathrm{d}t}{L\sin\alpha} = \frac{p_\mathrm{m}}{L} B \, \mathrm{d}t.$$

Dividing this angle by the time dt, we find the angular velocity of precession

$$\omega_\mathrm{L} = \frac{\mathrm{d}\theta}{\mathrm{d}t} = \frac{p_\mathrm{m}}{L} B.$$

Introducing the value of the ratio of the magnetic moment and angular momentum from Eq. (7.41), we get

$$\omega_\mathrm{L} = \frac{eB}{2m}. \tag{7.46}$$

The frequency (7.46) is called the **frequency of Larmor precession** or simply the **Larmor frequency**. It depends neither on the angle of inclination of an orbit with respect to the direction of the magnetic field nor on the radius of the orbit or the speed of the electron, and, consequently, is the same for all the electrons in an atom.

The precession of an orbit causes additional motion of the electron about the direction of the field. If the distance r' from the electron to an axis parallel to $\boldsymbol{B}$ and passing through the centre of the orbit did not change, the additional motion of

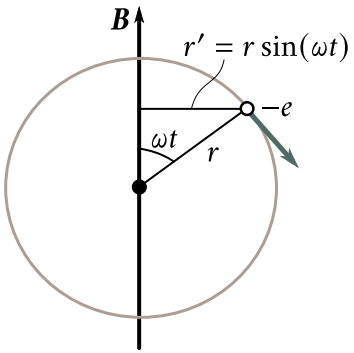


Fig. 7.15

the electron would occur along a circle of radius r' (see the unshaded circle in the right part of Fig. 7.14). The ring current $I' = e(\omega_L/2\pi)$ (see the shaded circle) would correspond to it. The magnetic moment of this current is

$$p'_m = I'S' = e\frac{\omega_L}{2\pi}\pi r'^2 = \frac{e\omega_L}{2}r'^2, \tag{7.47}$$

and is directed oppositely to $\boldsymbol{B}$ (see the figure). It is called the **induced magnetic moment**.

Indeed, owing to the motion of an electron in its orbit, the distance r' constantly changes. Therefore, in Eq. (7.47), we must replace r'^2 with its average value in time $\langle r'^2 \rangle$. The latter depends on the angle α characterizing the orientation of the orbit plane relative to $\boldsymbol{B}$. In particular, for an orbit perpendicular to the vector $\boldsymbol{B}$, the quantity r' is constant and equals the radius of the orbit r. For an orbit whose plane passes through the direction of $\boldsymbol{B}$, the quantity r' varies according to the law $r' = r\sin(\omega t)$, where ω is the angular velocity of revolution of an electron in its orbit (Fig. 7.15; the vector $\boldsymbol{B}$ and the orbit are in the plane of the drawing). Consequently, $\langle r'^2 \rangle = \langle r^2 \sin^2(\omega t) \rangle = r^2/2$ (the quantity $\langle \sin^2(\omega t) \rangle = 1$). Averaging over all possible values of α, considering them to be equally probable, yields

$$\langle r'^2 \rangle = \frac{2}{3}r^2. \tag{7.48}$$

Using in Eq. (7.47) the value (7.46) for ω_L and (7.48) for $\langle r'^2 \rangle$, we get the following expression for the average value of the induced magnetic moment of one electron:

$$\langle p'_m \rangle = -\frac{e^2}{6m}r^2 B \tag{7.49}$$

(the minus sign reflects the circumstance that the vectors $\langle \boldsymbol{p}'_m \rangle$ and $\boldsymbol{B}$ have opposite directions). We assumed the orbit to be circular. In the general case (for example, for an elliptical orbit), we must take $\langle r^2 \rangle$ instead of r^2, *i.e.*, the mean square of the distance from an electron to the nucleus.

Summation of Eq. (7.49) over all the electrons yields the induced magnetic moment of an atom

$$p'_{\mathrm{m,at}} = \sum \left\langle p'_{\mathrm{m}} \right\rangle = -\frac{e^2 B}{6m} \sum_{k=1}^{Z} \left\langle r_k^2 \right\rangle \tag{7.50}$$

(Z is the atomic number of a chemical element; the number of electrons in an atom is Z).

Thus, the action of an external magnetic field sets up precession of the electron orbits with the same angular velocity (7.46) for all the electrons. The additional motion of the electrons due to precession leads to the production of an induced magnetic moment of an atom [Eq. (7.50)] directed against the field. Larmor precession appears in all substances without exception. When atoms by themselves have a magnetic moment, however, a magnetic field not only induces the moment (7.50), but also has an orienting action on the magnetic moments of atoms, aligning them in the direction of the field. The positive (*i.e.*, directed along the field) magnetic moment that appears may be considerably greater than the negative induced moment. The resultant moment is, therefore, positive and the substance behaves like a paramagnetic.

Diamagnetism is found only in substances whose atoms have no magnetic moment (the vector sum of the orbital and spin magnetic moments of the atom electrons is zero). If we multiply Eq. (7.50) by the Avogadro constant N_{A} for such a substance, we get the magnetic moment for a mole of the substance. Dividing it by the field strength H, we find the molar magnetic susceptibility $\chi_{\mathrm{m,mol}}$. The permeability of dielectrics virtually equals unity. We can therefore assume that $B/H = \mu_0$. Thus,

$$\chi_{\mathrm{m,mol}} = \frac{N_{\mathrm{A}} p'_{\mathrm{m,at}}}{H} = -\frac{\mu_0 N_{\mathrm{A}} e^2}{6m} \sum_{k=1}^{Z} \left\langle r_k^2 \right\rangle . \tag{7.51}$$

We must note that the strict quantum-mechanical theory gives exactly the same expression.

Introduction of the numerical values of μ_0, N_{A}, e and m in Eq. (7.51) yields

$$\chi_{\mathrm{m,mol}} = -3.55 \times 10^9 \sum_{k=1}^{Z} \left\langle r_k^2 \right\rangle .$$

The radii of electron orbits have a value of the order of 10^{-10} m. Hence, the molar diamagnetic susceptibility of the order of 10^{-11} to 10^{-10} is obtained, which agrees quite well with experimental data.

7.8. Paramagnetism

If the magnetic moment p_m of the atoms differs from zero, the relevant substance is paramagnetic. A magnetic field tends to align the magnetic moments of the atoms along $\boldsymbol{B}$, while thermal motion tends to scatter them uniformly in all directions. As a result, a certain preferential orientation of the moments is established along the field. Its value grows with increasing $\boldsymbol{B}$ and diminishes with increasing temperature.

The French physicist and chemist Pierre Curie (1859-1906) established experimentally a law (named **Curie's law** in his honour) according to which the susceptibility of a paramagnetic is

$$\chi_{m,mol} = \frac{C}{T}, \tag{7.52}$$

where C is the Curie constant depending on the kind of substance and T the absolute temperature.

The classical theory of paramagnetism was developed by the French physicist Paul Langevin (1872-1946) in 1905. We shall limit ourselves to a treatment of this theory for not too strong fields and not very low temperatures.

According to Eq. (6.76), an atom in a magnetic field has the potential energy $W = -p_m B \cos\theta$ that depends on the angle θ between the vectors $\boldsymbol{p}_m$ and $\boldsymbol{B}$. Therefore, the equilibrium distribution of the moments by directions must obey Boltzmann's law (see Sec. 11.8 of Vol. I). According to this law, the probability of the fact that the magnetic moment of an atom will make with the direction of the vector $\boldsymbol{B}$ an angle within the limits from θ to $\theta + d\theta$ is proportional to

$$\exp\left(-\frac{W}{kT}\right) = \exp\left(\frac{p_m B \cos\theta}{kT}\right).$$

Introducing the notation

$$a = \frac{p_m B}{kT}, \tag{7.53}$$

we can write the expression determining the probability in the form

$$\exp(a\cos\theta). \tag{7.54}$$

In the absence of a field, all the directions of the magnetic moments are equally probable. Consequently, the probability of the fact that the direction of a moment will form with a certain direction z an angle within the limits from θ to $\theta + d\theta$ is

$$(dP_\theta)_{B=0} = \frac{d\Omega_\theta}{d4\pi} = \frac{2\pi \sin\theta\, d\theta}{4\pi} = \frac{1}{2}\sin\theta\, d\theta. \tag{7.55}$$

Here, $d\Omega_\theta = 2\pi \sin\theta\, d\theta$ is the solid angle enclosed between cones having apex angles of θ and $\theta + d\theta$ (Fig. 7.16).

When a field is present, the multiplier (7.54) appears in the expression for the

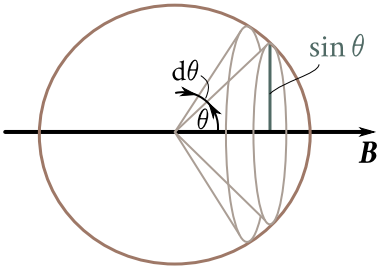


Fig. 7.16

probability:

$$dP_\theta = A\,\exp(a\cos\theta)\frac{1}{2}\sin\theta\,d\theta \tag{7.56}$$

(A is a proportionality constant that is meanwhile unknown).

The magnetic moment of an atom has a magnitude of the order of one Bohr magneton, *i.e.*, about $10^{-23}\,\mathrm{J\,T^{-1}}$ [see Eq. (7.45)]. At the usually achieved fields, the magnetic induction is of the order of 1 T (10^4 Gs). Hence, $p_m B$ is of the order of 10^{-23} J. The quantity kT at room temperature is about 4×10^{-21} J. Thus, $a = p_m B/(kT)$ is much smaller than unity, and $\exp(a\cos\theta)$ may be replaced with the approximate expression $1 + a\cos\theta$. In this approximation, Eq. (7.56) becomes

$$dP_\theta = A(1 + a\cos\theta)\frac{1}{2}\sin\theta\,d\theta.$$

The constant A can be found by proceeding from the fact that the sum of the probabilities of all possible values of the angle θ must equal unity:

$$1 = \int_0^\pi A(1 + a\cos\theta)\frac{1}{2}\sin\theta\,d\theta.$$

Hence, $A = 1$, so that

$$dP_\theta = (1 + a\cos\theta)\frac{1}{2}\sin\theta\,d\theta.$$

Assume that unit volume of a paramagnetic contains n atoms. Consequently, the number of atoms whose magnetic moments form angles from θ to $\theta + d\theta$ with the direction of the field will be

$$dn_\theta = n\,dP_\theta = n(1 + a\cos\theta)\frac{1}{2}\sin\theta\,d\theta.$$

Each of these atoms makes a contribution of $p_m\cos\theta$ to the resultant magnetic moment. Therefore, we get the following expression for the magnetic moment of unit volume (*i.e.*, for the magnetization):

$$M = \int_0^\pi p_m\cos\theta\,dn_\theta = \frac{1}{2}np_m\int_0^\pi(1 + a\cos\theta)\frac{1}{2}\sin\theta\,d\theta = \frac{np_m a}{3}.$$

Substitution for a of its value from Eq. (7.53) yields

$$M = \frac{np_{\mathrm{m}}^2 B}{3kT}.$$

Finally, dividing M by H and assuming that $B/H = \mu_0$ (for a paramagnetic μ is virtually equal to unity), we find the susceptibility

$$\chi_{\mathrm{m}} = \frac{\mu_0 n p_{\mathrm{m}}^2}{3kT}. \tag{7.57}$$

Substituting the Avogadro constant N_{A} for n, we get an expression the molar susceptibility:

$$\chi_{\mathrm{m,mol}} = \frac{\mu_0 N_{\mathrm{A}} p_{\mathrm{m}}^2}{3kT}. \tag{7.58}$$

We have arrived at Curie's law. A comparison of Eqs. (7.52) and (7.58) gives the following expression for the Curie constant:

$$C = \frac{\mu_0 N_{\mathrm{A}} p_{\mathrm{m}}^2}{3k}. \tag{7.59}$$

It must be remembered that Eq. (7.58) has been obtained assuming that $p_{\mathrm{m}} B \ll kT$. In very strong fields and at low temperatures, deviations are observed from proportionality between the magnetization of a paramagnetic M and the field strength H. In particular, a state of magnetic saturation may set in when all the $\boldsymbol{p}_{\mathrm{m}}$'s are lined up along the field, and a further increase in H does not result in a growth in M.

The values of $\chi_{\mathrm{m,mol}}$ calculated by Eq. (7.58) in a number of cases agree quite well with the values obtained experimentally.

The quantum theory of paramagnetism takes account of the fact that only discrete orientations of the magnetic moment of an atom relative to a field are possible. It arrives at an expression for $\chi_{\mathrm{m,mol}}$ similar to Eq. (7.58).

7.9. Ferromagnetism

Substances capable of having magnetization in the absence of an external magnetic field form a special class of magnetics. According to the name of their most widespread representative—ferrum (iron)—they have been called **ferromagnetics**. In addition to iron, they include nickel, cobalt, gadolinium, their alloys and compounds, and also certain alloys and compounds of manganese and chromium with non-ferromagnetic elements. All these substances display ferromagnetism only in the crystalline state.

Ferromagnetics are strongly magnetic substances. Their magnetization exceeds that of diamagnetics and paramagnetics which belong to the category of weakly

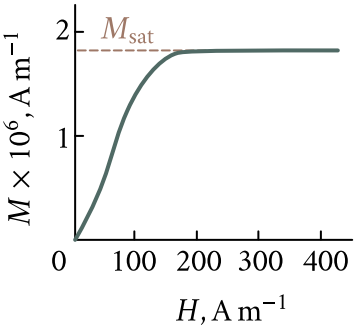


Fig. 7.17

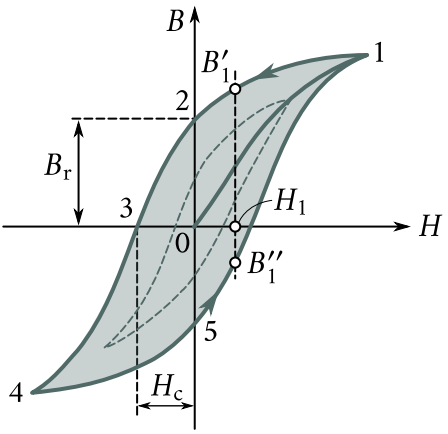


Fig. 7.18

magnetized substances an enormous number of times (up to 10^{10}).

The magnetization of weakly magnetized substances varies linearly with the field strength. The magnetization of ferromagnetics depends on H in an intricate way. Figure 7.17 shows the magnetization curve for a ferromagnetic whose magnetic moment was initially zero (it is called the **initial** or **zero magnetization curve**). Already in fields of the order of several oersteds (about 100 A m^{-1}), the magnetization M reaches saturation. The initial magnetization curve in a B-H diagram is shown in Fig. 7.18 (curve 0-1). We remind our reader that $B = \mu_0(H+M)$. Therefore, when saturation is reached, B continues to grow with increasing H according to a linear law: $B = \mu_0 H$ + constant, where constant = $\mu_0 M_{sat}$.

A magnetization curve for iron was first obtained and investigated in detail by the Russian scientist Aleksandr Stoletov (1839-1896). The ballistic method of measuring the magnetic induction which he developed has been finding wide application (see Sec. 8.3).

Apart from the non-linear relation between H and M (or between H and B), ferromagnetics are characterized by the presence of hysteresis. If we bring magnetization up to saturation (point 1 in Fig. 7.18) and then diminish the magnetic field strength, the induction B will no longer follow the initial curve 0-1, but will change in accordance with curve 1-2. As a result, when the strength of the external field vanishes (point 2), the magnetization does not vanish and is characterized by the quantity B_r called the **residual induction**. The magnetization for this point has the value M_r called the **retentivity** or **remanence**.

The magnetization vanishes only under the action of the field H_c directed oppositely to the field that produced the magnetization. The field strength H_c is called the coercive force.

The existence of remanence makes it possible to manufacture permanent magnets, *i.e.*, bodies that have a magnetic moment and produce a magnetic field in the space surrounding them without the expenditure of energy for maintaining the macroscopic currents. A permanent magnet retains its properties better when the coercive force of the material it is made of is higher.

When an alternating magnetic field acts on a ferromagnetic, the induction changes in accordance with curve 1-2-3-4-5-1 (Fig. 7.18) called a **hysteresis loop** (a similar curve is obtained in an M-H diagram). If the maximum values of H are such that the magnetization reaches saturation, we get the so-called **maximum hysteresis loop** (the solid loop in Fig. 7.18). If saturation is not reached at the amplitude values of H, we get a loop called a **partial cycle** (the dash line in the figure). The number of such partial cycles is infinite, and all of them are within the maximum hysteresis loop.

Hysteresis results in the fact that the magnetization of a ferromagnetic is not a unique function of H. It depends very greatly on the previous history of a specimen—on the fields which it was in previously. For example, in a field of strength H_1 (Fig. 7.18), the induction may have any value ranging from B_1' to B_1''.

It follows from everything said above about ferromagnetics that they are very similar in their properties to ferroelectrics (see Sec. 2.9). In connection with the ambiguity of the dependence of B on H, the concept of permeability is applied only to the initial magnetization curve. The permeability of ferromagnetics μ (and, consequently, their magnetic susceptibility χ_m) is a function of the field strength. Figure 7.19a shows an initial magnetization curve. Let us draw from the origin of coordinates a siraight line that passes through an arbitrary point on the curve. The slope of this line is proportional to the ratio B/H, *i.e.*, to the permeability μ for the relevant value of the field strength. When H grows from zero, the slope (and, consequently, μ) first grows. At point 2 it reaches a maximum (straight line 0-2 is a tangent to the curve) and then diminishes. Figure Fig. 7.19b shows how μ depends on H. A glance at the figure shows that the maximum value of the permeability is reached somewhat earlier than saturation. Upon an unlimited increase in H, the permeability approaches unity asymptotically. This can be seen from the circumstance that M in the expression $\mu = 1 + M/H$ cannot exceed the value M_{sat}.

The quantities B_r (or M_r), H_c and μ_{max} are the basic characteristics of a ferromagnetic. If the coercive force H_c is great, the ferromagnetic is called **hard**. It is characterized by a broad hysteresis loop. A ferromagnetic with a low H_c (and accordingly with a narrow hysteresis loop) is called **soft**. The characteristic of a ferromagnetic is chosen depending on the use it is to be put to. Thus, hard ferromagnetics are used for permanent magnets, and soft ones for the cores of transformers.

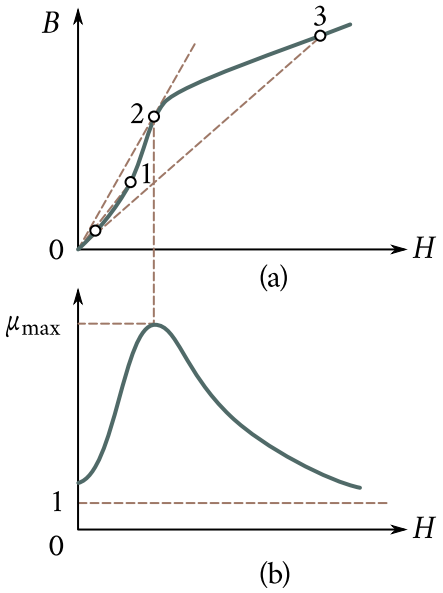


Fig. 7.19

Table 7.1 gives the characteristics of several typical ferromagnetics.

The fundamentals of the theory of ferromagnetism were presented by the Soviet physicist Yakov Frenkel (1894-1952) and the German physicist Werner Heisenberg (1901-1976) in 1928. It follows from experiments involving the studying of gyromagnetic phenomena (see Sec. 7.6) that the intrinsic (spin) magnetic moments of electrons are responsible for the magnetic properties of ferromagnetics. In definite conditions, forces[7] may appear in crystals that make the magnetic moments of the electrons become lined up parallel to one another. The result is the setting up of regions of **spontaneous magnetization**, also called **domains**. Within the confines of each domain, a ferromagnetic is spontaneously magnetized to saturation

[7]These forces are called **exchange** ones. Their explanation is given only by quantum mehcanics.

Table 7.1

Substance	**Composition**	μ_{max}	B_r, T	H_c, A m^{-1}
Iron	99.9% Fe	5000	—	80
Supermalloy	79% Ni, 5%, Mo, 16% Fe	800000	—	0.3
Alniko	10% Al, 19% Ni, 18% Co	—	0.9	52000

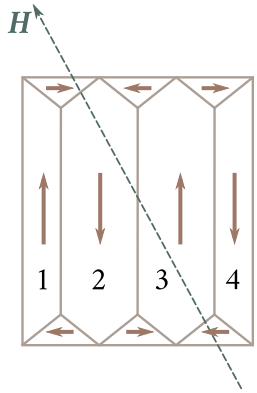


Fig. 7.20

and has a definite magnetic moment. The directions of these moments are different for different domains (Fig. 7.20), so that in the absence of an external field the total moment of an entire body is zero. Domains have dimensions of the order of 1 μm to 10 μm.

The action of a field on domains at different stages of the magnetization process is different. First, with weak fields, displacement of the domain boundaries is observed. As a result, the domains whose moments make a smaller angle with $\boldsymbol{H}$ grow at the expense of the domains for which the angle θ between the vectors $\boldsymbol{p}_\mathrm{m}$ and $\boldsymbol{H}$ is greater. For example, domains 1 and 3 in Fig. 7.20 grow at the expense of domains 2 and 4. With an increase in the field strength, this process goes on further and further until the domains with a smaller θ (which have a smaller energy in a magnetic field) completely absorb the domains that are less advantageous from the energy viewpoint. In the next stage, the magnetic moments of the domains turn in the direction of the field. The moments of the electrons within the confines of a domain turn simultaneously without violating their strict parallelism to one another. *These processes (excluding slight displacements of the boundaries between the domains in very weak fields) are irreversible, and this is exactly what causes hysteresis.*

There is a definite temperature T_C for every ferromagnetic at which the regions of spontaneous magnetization (domains) break up and the substance loses its ferromagnetic properties. This temperature is called the **Curie point**. It is 768 °C for iron and 365 °C for nickel. At a temperature above the Curie point, a ferromagnetic becomes an ordinary paramagnetic whose magnetic susceptibility obeys the **Curie-Weiss law**

$$\chi_{\mathrm{m,mol}} = \frac{C}{T - T_\mathrm{C}} \tag{7.60}$$

[compare with Eq. (7.52)]. When a ferromagnetic is cooled to below its Curie point, domains once more appear in it.

Exchange forces sometimes result in the appearance of so-called **antiferromagnetics** (chromium, manganese, etc.). The existence of antiferromagnetics was predicted by the Soviet physicist Lev Landau (1908-1968) in 1933. In antiferromagnetics, the intrinsic magnetic moments of the electrons are spontaneously oriented antiparallel to one another. Such an orientation involves adjacent atoms in pairs. The result is that antiferromagnetics have an extremely low magnetic susceptibility and behave like very weak paramagnetics. There is also a temperature T_N for antiferromagnetics at which the antiparallel orientation of the spins vanishes. This temperature is known as the **antiferromagnetic Curie point** or the **Neél point**. Some antiferromagnetics (for example, erbium, dysprosium, alloys of manganese and copper) have two such points (an upper and a lower Neel point), the antiferromagnetic properties being observed only at the intermediate temperatures. Above the upper point, the substance behaves like a paramagnetic, and at temperatures below the lower Neél point it becomes a ferromagnetic.

Chapter 8

ELECTROMAGNETIC INDUCTION

8.1. The Phenomenon of Electromagnetic Induction

In 1831, the British physicist and chemist Michael Faraday (1791-1867) discovered that an electric current is produced in a closed conducting loop when the flux of magnetic induction through the surface enclosed by this loop changes. This phenomenon is called **electromagnetic induction**, and the current produced an **induced current**.

The phenomenon of electromagnetic induction shows that when the magnetic flux in a loop changes, an induced electromotive force $\mathscr{E}_i$ is set up. The value of $\mathscr{E}_i$ does not depend on how the magnetic flux Φ is changed and is determined only by the rate of change of Φ, *i.e.*, by the value of $d\Phi/dt$. A change in the sign of $d\Phi/dt$ is attended by a change in the direction of $\mathscr{E}_i$.

Let us consider the following example. Figure 8.1 shows loop 1 whose current I_1 can be varied by means of a rheostat. This current sets up a magnetic field through loop 2. If we increase the current I_1, the magnetic induction flux Φ through loop 2 will grow. This will lead to the appearance in loop 2 of the induced current I_2 registered by a galvanometer. Diminishing of the current I_1 will cause the magnetic flux through the second loop to decrease. This will result in the appearance in it of an induced current of a direction opposite to that in the first case. An induced current I_2 can also be set up by bringing loop 2 closer to loop 1 or moving it away from it. In these two cases, the directions of the induced current are opposite. Finally, electromagnetic induction can be produced without translational motion of loop 2, but by turning it so as to change the angle between a normal to the loop and the direction of the field.

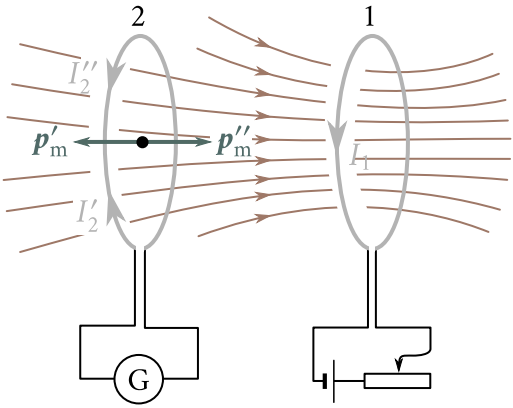


Fig. 8.1

E. Lenz established a rule permitting us to find the direction of an induced current. **Lenz's rule** states that an induced current is always directed so as to oppose the cause producing it. If, for example, a change in Φ is due to motion of loop 2, then an induced current is set up of a direction such that the force of interaction with the first loop opposes the motion of the loop. When loop 2 approaches loop 1 (see Fig. 8.1), a current I_2' is set up whose magnetic moment is directed oppositely to the field of the current I_1 (the angle a between the vectors $\boldsymbol{p}'_m$ and $\boldsymbol{B}$ is πs). Hence, loop 2 will experience a force repelling it from loop 1 [see Eq. (6.77)]. When loop 2 is moved away from loop 1, the current I_2'' is produced whose moment $\boldsymbol{p}''_m$ coincides in direction with the field of the current I_1 ($\alpha = 0$) so that the force exerted on loop 2 is directed toward loop 1.

Assume that both loops are stationary and the current in loop 2 is induced by changing the current I_1 in loop 1. Now a current I_2 is induced of a direction such that the intrinsic magnetic flux it produces tends to weaken the change in the external flux leading to the setting up of the induced current. When I_1 grows, *i.e.*, the external magnetic flux directed to the right is increased, a current I_2' is induced that sets up a flux directed to the left. When I_1 diminishes, the current I_2'' is set up whose intrinsic magnetic flux has the same direction as the external flux and, consequently, tends to keep the external flux unchanged.

8.2. Induced E.M.F.

We have established in the preceding section that changes in the magnetic flux Φ through a loop set up an induced e.m.f. $\mathscr{E}_i$ in it. To find the relation between $\mathscr{E}_i$ and the rate of change of Φ, we shall consider the following example.

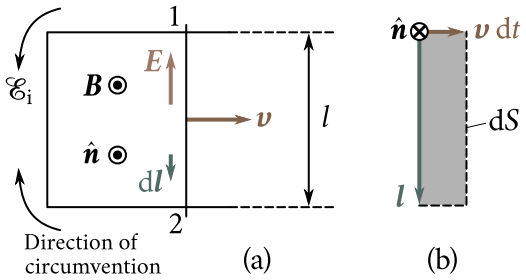


Fig. 8.2

Let us take a loop with a movable rod of length l (Fig. 8.2a). We shall put it in a homogeneous magnetic field at right angles to the plane of the loop and directed beyond the drawing. Let us bring the rod into motion with the velocity $\boldsymbol{v}$. The current carriers in the rod—electrons—will also begin to move relative to the field with the same velocity. As a result, each electron will begin to experience the magnetic force

$$\boldsymbol{F}_{\parallel} = -e(\boldsymbol{v} \times \boldsymbol{B}), \tag{8.1}$$

directed along the rod [see Eq. (6.33); the charge of an electron is $-e$]. The action of this force is equivalent to the action on an electron of an electric field of strength

$$\boldsymbol{E} = \boldsymbol{v} \times \boldsymbol{B}.$$

This field is of a non-electrostatic origin. Its circulation around a loop gives the value of the e.m.f. induced in the loop:

$$\mathscr{E}_i = \oint \boldsymbol{E} \cdot d\boldsymbol{l} = \oint (\boldsymbol{v} \times \boldsymbol{B}) \cdot d\boldsymbol{l} = \int_1^2 (\boldsymbol{v} \times \boldsymbol{B}) \cdot d\boldsymbol{l} \tag{8.2}$$

(the integrand differs from zero only on section 1-2 formed by the rod).

To be able to judge about the direction in which the e.m.f. acts according to the sign of $\mathscr{E}_i$, we shall consider $\mathscr{E}_i$ positive when its direction forms a right-handed system with the direction of a normal to the loop.

Let us choose the normal as shown in Fig. 8.2. Hence, when calculating the circulation, we must circumvent the loop clockwise and choose the direction of the vectors $d\boldsymbol{l}$ accordingly. If we put the constant vector $\boldsymbol{v} \times \boldsymbol{B}$ in Eq. (8.2) outside the integral, we get

$$\mathscr{E}_i = (\boldsymbol{v} \times \boldsymbol{B}) \int_1^2 d\boldsymbol{l} = (\boldsymbol{v} \times \boldsymbol{B}) \cdot \boldsymbol{l},$$

where $\boldsymbol{l}$ is the vector depicted in Fig. 8.2b. Let us perform a cyclic rearrangement of the multipliers in the expression obtained, after which we shall multiply and divide

it by dt:

$$\mathscr{E}_\mathrm{i} = \boldsymbol{B} \cdot (\boldsymbol{l} \times \boldsymbol{v}) = \frac{\boldsymbol{B} \cdot (\boldsymbol{l} \times \boldsymbol{v}\, \mathrm{d}t)}{\mathrm{d}t}. \tag{8.3}$$

A glance at Fig. 8.2b shows that

$$\boldsymbol{l} \times \boldsymbol{v}\, \mathrm{d}t = -\hat{\boldsymbol{n}}\, \mathrm{d}S,$$

where dS is the increment of the loop area during the time dt. By the definition of a flux, $\boldsymbol{B} \cdot \mathrm{d}\boldsymbol{S} = \boldsymbol{B} \cdot \hat{\boldsymbol{n}}\, \mathrm{d}S$ is the flux through the area dS, *i.e.*, the increment of the flux dΦ through the loop. Thus,

$$\boldsymbol{B} \cdot (\boldsymbol{l} \times \boldsymbol{v}\, \mathrm{d}t) = -\boldsymbol{B} \cdot \hat{\boldsymbol{n}}\, \mathrm{d}S = -\mathrm{d}\Phi.$$

With a view to this expression, Eq. (8.3) can be written as

$$\mathscr{E}_\mathrm{i} = -\frac{\mathrm{d}\Phi}{\mathrm{d}t}. \tag{8.4}$$

We have found that dΦ/dt and $\mathscr{E}_\mathrm{i}$ have opposite signs. The sign of the flux and that of $\mathscr{E}_\mathrm{i}$ are associated with the choice of the direction of a normal to the plane of a loop. With our selection of the normal (see Fig. 8.2), the sign of dΦ/dt is positive, and that of $\mathscr{E}_\mathrm{i}$ is negative. If we had chosen a normal directed not beyond the drawing, but toward us, the sign of dΦ/dt would be negative and that of It positive.

The SI unit of magnetic induction flux is the **weber** (Wb), which is the flux through a surface of 1 m^2 intersected by magnetic field lines normal to it with $B = 1$ T. At a rate of change of the flux equal to 1 Wb s^{-1}, an e.m.f. of 1 V is induced in the loop. In the Gaussian system of units, Eq. (8.4) has the form

$$\mathscr{E}_\mathrm{i} = -\frac{1}{c}\frac{\mathrm{d}\Phi}{\mathrm{d}t}. \tag{8.5}$$

The unit of Φ in this system is the **maxwell** (Mx) equal to the flux through a surface of 1 cm^2 at $B = 1$ Gs. Equation (8.5) gives $\mathscr{E}_\mathrm{i}$ in cgse$_U$. To find it in volts, we must multiply the result obtained by 300. Since $300/c = 10^{-8}$, we have

$$\mathscr{E}_\mathrm{i}(V) = -10^{-8}\, \frac{\mathrm{d}\Phi}{\mathrm{d}t}\ \left(\mathrm{Mx\,s^{-1}}\right). \tag{8.6}$$

In the reasoning that led us to Eq. (8.4), the part of the extraneous forces maintaining a current in a loop was played by magnetic forces. The work of these forces on a unit positive charge, equal by definition to the e.m.f., is other than zero. This circumstance apparently contradicts the statement made in Sec. 6.5 that a magnetic force can do no work on a charge. This contradiction is eliminated if we take into account that the force (8.1) is not the total magnetic force exerted on an electron, but only the component of this force parallel to the conductor and due to the velocity $\boldsymbol{v}$ (see the force $\boldsymbol{F}_\parallel$ in Fig. 8.3). This component causes the electron to start moving along the conductor with the velocity $\boldsymbol{u}$, as a result of which a magnetic

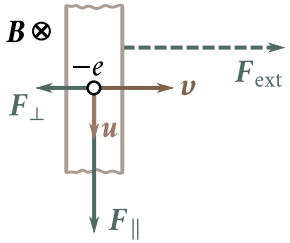


Fig. 8.3

force perpendicular to the wire is set up equal to

$$\boldsymbol{F}_{\parallel} = -e(\boldsymbol{u} \times \boldsymbol{B})$$

(this component makes no contribution to the circulation because it is perpendicular to $\mathrm{d}\boldsymbol{l}$).

The total magnetic force exerted on an electron is

$$\boldsymbol{F} = \boldsymbol{F}_{\parallel} + \boldsymbol{F}_{\perp},$$

and the work of this force on an electron during the time $\mathrm{d}t$ is

$$\mathrm{d}A = \boldsymbol{F}_{\parallel} \cdot \boldsymbol{u}\, \mathrm{d}t + \boldsymbol{F}_{\perp} \cdot \boldsymbol{v}\, \mathrm{d}t = F_{\parallel} u\, \mathrm{d}t + F_{\perp} v\, \mathrm{d}t$$

(the directions of the vectors $\boldsymbol{F}_{\parallel}$ and $\boldsymbol{u}$ are the same, and of the vectors $\boldsymbol{F}_{\perp}$ and $\boldsymbol{v}$ are opposite; see Fig. 8.3). Substituting for the magnitudes of the forces their values $F_{\parallel} = evB$ and $F_{\perp} = euB$, we find that the work of the total magnetic force equals zero.

The force $\boldsymbol{F}_{\perp}$ is directed oppositely to the velocity of the rod $\boldsymbol{v}$. Therefore, for the rod to move with the constant velocity $\boldsymbol{v}$, the external force $\boldsymbol{F}_{\text{ext}}$ must be applied to it that balances the sum of the forces $\boldsymbol{F}_{\perp}$ applied to all the electrons contained in the rod. It is exactly at the expense of the work of this force that the energy liberated in the loop by the induced current will be produced.

Our explanation of the appearance of an induced e.m.f. relates to the case when the magnetic field is constant, while the geometry of the loop changes. The magnetic flux through the loop can also be changed, however, by changing $\boldsymbol{B}$. In this case, the explanation of the appearance of an e.m.f. will differ in principle. The time-varying magnetic field sets up a vortex electric field $\boldsymbol{E}$ (this is treated in detail in Sec. 9.1). The action of the field $\boldsymbol{E}$ causes the current carriers in a conductor to start moving—an induced current is set up. The relation between the induced e.m.f. and the changes in the magnetic flux in this case too is described by Eq. (8.4).

Assume that the loop in which an e.m.f. is induced consists of N turns instead of one, *i.e.*, it is a solenoid, for example. Since the turns are connected in series, $\mathscr{E}_{\text{i}}$

will equal the sum of the e. m.f.'s induced in each of the turns separately:

$$\mathscr{E}_i = -\sum \frac{d\Phi}{dt} = -\frac{d}{dt}\left(\sum \Phi\right).$$

The quantity

$$\Psi = \sum \Phi, \tag{8.7}$$

is called the **flux linkage** or the **total magnetic flux**. It is measured in the same units as Φ. If the flux through each of the turns is the same, then

$$\Psi = N\Phi. \tag{8.8}$$

The e.m.f. induced in an intricate loop is determined by the formula

$$\mathscr{E}_i = -\frac{d\Psi}{dt}. \tag{8.9}$$

8.3. Ways of Measuring the Magnetic Induction

Assume that the total magnetic flux linked to a loop changes from Ψ_1 to Ψ_2. Let us find the charge q that flows through each section of the loop. The instantaneous value of the current in the loop is

$$I = \frac{\mathscr{E}}{R} = -\frac{1}{R}\frac{d\Psi}{dt}.$$

Hence,

$$dq = I\,dt = -\frac{1}{R}\frac{d\Psi}{dt}\,dt = -\frac{1}{R}\,d\Psi.$$

Integration of this expression yields the total charge:

$$q = \int dq = -\frac{1}{R}\int_1^2 d\Psi = \frac{1}{R}(\Psi_1 - \Psi_2). \tag{8.10}$$

Equation (8.10) underlies the ballistic method of measuring the magnetic induction developed by A. Stoletov. It consists in the following. A small coil with N turns is placed in the field being studied. The coil is arranged so that the vector $\boldsymbol{B}$ is perpendicular to the plane of the turns (Fig. 8.4a). Hence, the total magnetic flux linked with the coil will be

$$\Psi_1 = NBS,$$

where S is the area of one turn, which must be so small that the field within its limits may be considered homogeneous.

When the coil is turned through 180 degrees (Fig. 8.4b), the flux linkage becomes equal to $\Psi_2 = -NBS$ ($\hat{\boldsymbol{n}}$ and $\boldsymbol{B}$ are directed oppositely). Hence, the change in the total flux linkage when the coil is turned is $\Psi_1 - \Psi_2 = 2NBS$. If the coil is turned sufficiently quickly, a short current pulse is produced in the loop upon which the

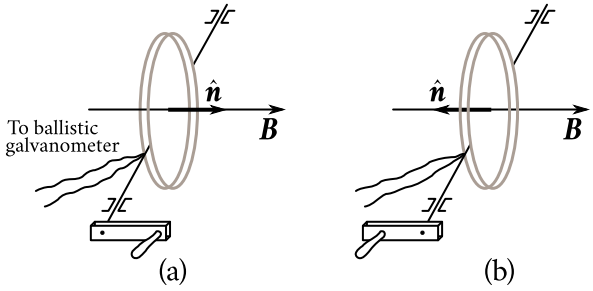


Fig. 8.4

charge

$$q = \frac{1}{R} 2NBS \tag{8.11}$$

flows [see Eq. (8.10)].

The charge flowing in the circuit during the short current pulse can be measured with the aid of a so-called **ballistic galvanometer**. The latter is a galvanometer with a great period of natural oscillations. Having measured q and knowing R, N, and S, we can find B by Eq. (8.11). By R, here, is meant the resistance of the entire circuit including the coil, the connecting wires, and the galvanometer.

Instead of turning the coil, we may switch on (or off) the magnetic field being studied, or reverse its direction.

To measure B, the circumstance is also used that the electric resistance of bismuth grows greatly under the action of a magnetic field—by about five per cent per tenth of a tesla (per 1000 Gs). Consequently, we can determine the magnetic induction of a magnetic field by placing a preliminarily graduated bismuth coil (Fig. 8.5) into the field and measuring the relative change in its resistance.

We must note that the electric resistance of other metals also grows in a magnetic field, but to a much smaller extent. For copper, for example, the increase in the resistance is about one-ten thousandth of that for bismuth.

8.4. Eddy Currents

Induced currents can also be produced in solid massive conductors. In this case, they are known as **eddy currents**. The electric resistance of a massive conductor is small, therefore, the eddy currents may reach a very high value.

In accordance with Lenz's rule, eddy currents choose paths and directions in a conductor such as to resist by their action the reason setting them up as much as possible. This is why good conductors moving in a strong magnetic field experience great retardation due to the interaction of the eddy currents with the magnetic

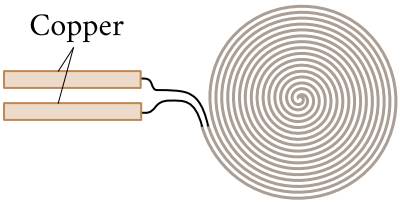


Fig. 8.5

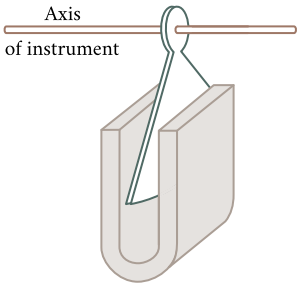


Fig. 8.6

field. This is taken advantage of for damping the movable parts of galvanometers, seismographs, and other instruments. A conducting (for example, aluminium) plate in the form of a sector is fastened to the movable part of an instrument (Fig. 8.6) and is introduced into the gap between the poles of a strong permanent magnet. Movement of the plate causes eddy currents to be produced in it that brake the system. The advantage of such a device is that the braking action appears only when the plate moves and vanishes when the plate is stationary. Therefore, the electromagnetic damper is absolutely no hindrance to the instrument accurately arriving at its equilibrium position.

The thermal action of eddy currents is used in induction furnaces. Such a furnace is a coil supplied with a high-frequency current of a high value. If we place a conducting body inside the coil, intensive eddy currents will be produced in it that can heat the body up to its melting point. This method is used to melt metals in vacuum. The resulting materials have an exceedingly high purity.

Eddy currents are also used to heat the internal metal components of vacuum installations in order to degas them.

Eddy currents are quite often undesirable, and special measures must be taken to eliminate them. For example, to prevent the losses of energy for heating transformer cores by eddy currents, the cores are assembled of thin insulated sheets. The latter are arranged so that the possible directions of the eddy currents will be perpendicular to them. The appearance of ferrites (semiconductor magnetic materials with a high electric resistance) made it possible to manufacture solid cores.

The eddy currents set up in conductors carrying alternating currents are directed so as to weaken the current inside a conductor and increase it near the surface. As a result, the fast-varying current is distributed unevenly over the cross section of the conductor—it is forced out, as it were, to the surface of the conductor. This phenomenon is called the **skin effect**. Owing to this effect, the internal part of conductors in high-frequency circuits is useless. This is why the conductors used

for such circuits have the form of tubes.

8.5. Self-Induction

An electric current flowing in any loop produces the magnetic flux Ψ through this loop. When I changes, Ψ also changes, and the result is the induction of an e.m.f. in the loop. This phenomenon is called **self-induction**. In accordance with the Biot-Savart law, the magnetic induction B is proportional to the current setting up the field. Hence, it follows that the current I in a loop and the total magnetic flux Ψ through the loop it produces are proportional to each other:

$$\Psi = LI. \tag{8.12}$$

The constant of proportionality L between the current and the total magnetic flux is called the **inductance** of a loop.

A linear dependence of Ψ on I is observed only if the permeability μ of the medium surrounding the loop does not depend on the field strength H, *i.e.*, in the absence of ferromagnetics. Otherwise, μ is an intricate function of I (through H, see Fig. 7.19b), and, since $B = \mu_0 \mu H$, the dependence of Ψ on I will also be quite intricate. Equation (8.12), however, is also extended to this case, and the inductance L is considered as a function of I. With a constant current I, the total flux Ψ can change as a result of changes in the shape and dimensions of a loop.

It can be seen from the above that the inductance L depends on the geometry of a loop (*i.e.*, on its shape and dimensions), and also on the magnetic properties (on μ) of the medium surrounding the loop. If the loop is rigid and there are no ferromagnetics near it, the inductance L is a constant quantity. The SI unit of inductance is the inductance of a conductor in which a total flux Ψ of 1 Wb linked with it is set up at a current of 1 A in the conductor. This unit is called the **henry** (H).

In the Gaussian system of units, the inductance has the dimension of length. Accordingly, the unit of inductance in this system is called the **centimetre**. A loop with which a flux of 1 Mx (10^{-8} Wb) is linked at a current of 1 cgsm_I (*i.e.*, 10 A) has an inductance of 1 cm.

Let us calculate the inductance of a solenoid. We shall take a solenoid so long that it can virtually be considered infinite. When a current I flows in it, a homogeneous field is produced inside the solenoid whose induction is $B = \mu_0 \mu n I$ [see Eqs. (6.108) and (7.26)]. The flux through each of the turns is $\Phi = BS$, and the total magnetic flux linked with the solenoid is

$$\Psi = N\Phi = nlBS = \mu_0 \mu n^2 lSI, \tag{8.13}$$

where l is the length of the solenoid (which is assumed to be very great), S is the

cross-sectional area, and n the number of turns per unit length (the product nl gives the total number of turns N).

A comparison of Eqs. (8.12) and (8.13) gives the following expression for the inductance of a very long solenoid:

$$L = \mu_0 \mu n^2 lS = \mu_0 \mu n^2 V, \tag{8.14}$$

where $V = lS$ is the volume of the solenoid.

It follows from Eq. (8.14) that the dimension of μ_0 equals that of inductance divided by the dimension of length. Accordingly, μ_0 is measured in henry per metre [see Eq. (6.3)].

When the current in a loop changes, a self-induced e.m.f. $\mathscr{E}_s$ is set up that equals

$$\mathscr{E}_s = -\frac{d\Psi}{dt} = -\frac{d(LI)}{dt} = -\left(L\frac{dI}{dt} + I\frac{dL}{dt}\right). \tag{8.15}$$

If the inductance remains constant when the current changes (which is possible only in the absence of ferromagnetics), the expression for the self-induced e.m.f. becomes

$$\mathscr{E}_s = -L\frac{dI}{dt}. \tag{8.16}$$

The minus sign in Eq. (8.16) is due to Lenz's rule according to which an induced current is directed so as to oppose the cause producing it. In the case being considered, what sets up $\mathscr{E}_s$ is the change of the current in the circuit. Let us assume clockwise circumvention to be the positive direction. In these conditions, the current will be greater than zero if it flows clockwise in the circuit and less than zero if it flows counterclockwise. Similarly, $\mathscr{E}_s$ will be greater than zero if it is exerted in a clockwise direction, and less than zero if it is exerted in a counterclockwise one.

The derivative dI/dt is positive in two cases—either upon a growth in a positive current or upon a decrease in the absolute value of a negative current. Inspection of Eq. (8.16) shows that in these cases $\mathscr{E}_s < 0$. This signifies that the self-induced e.m.f. is directed counterclockwise and, therefore, is opposed to the above current changes (a growth in a positive or a decrease in a negative current).

The derivative dI/dt is negative also in two cases—either when a positive current diminishes, or when the magnitude of a negative current grows. In these cases, $\mathscr{E}_s > 0$ and, consequently, opposes changes in the current (a decrease in a positive or a growth in the magnitude of a negative current).

Equation (8.16) makes it possible to define the inductance as a constant of proportionality between the rate of change of the current in a loop and the resulting self-induced e.m.f.. Such a definition is lawful, however, only when L = constant. In the presence of ferromagnetics, L of an undeforming loop will be a function of I (through H). Hence, dL/dt can be written as $(dL/dI)(dI/dt)$. Making such a

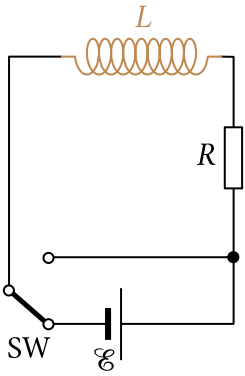


Fig. 8.7

substitution in Eq. (8.15), we get

$$\mathscr{E}_s = -\left(L + I\,\frac{dL}{dI}\right)\frac{dI}{dt}. \tag{8.17}$$

We can, thus, see that in the presence of ferromagnetics the constant of proportionality between dI/dt and $\mathscr{E}_s$ does not at all equal L.

8.6. Current When a Circuit Is Opened or Closed

According to Lenz's rule, the additional currents set up owing to self-induction are always directed so as to prevent any changes in the current in a circuit. The result is that a current grows to its steady value when a circuit is closed or drops to zero when the circuit is opened not instantaneously, but gradually.

Let us first find how a current changes when the switch of a circuit is opened. Assume that a current source of e.m.f. $\mathscr{E}$ is connected in a circuit with an inductance L not depending on I and a resistance R (Fig. 8.7). The steady current flowing in the circuit will be

$$I_0 = \frac{\mathscr{E}}{R} \tag{8.18}$$

(we consider the resistance of the current source to be negligibly small).

At the moment $t = 0$, let us switch off the current source and simultaneously short the circuit by means of switch SW. As soon as the current in the circuit begins to diminish, a self-inductance e.m.f. opposing this decrease appears. The current in the circuit will comply with the equation

$$IR = \mathscr{E}_s = -L\frac{dI}{dt},$$

or

$$\frac{dI}{dt} + \frac{R}{L} I = 0. \tag{8.19}$$

Equation (8.19) is a linear homogeneous differential equation of the first order. Separating variables, we get

$$\frac{dI}{I} = -\frac{R}{L}\, dt,$$

whence

$$\ln I = -\frac{R}{L} t + \ln(\text{constant})$$

(with a view to further transformations, we have written the integration constant in the form "ln(constant)"). Converting this relation to a power yields

$$I = \text{constant} \times \exp\left(-\frac{R}{L} t\right). \tag{8.20}$$

Equation (8.20) is a general solution of Eq. (8.19). We shall find the value of the constant from the initial conditions. When $t = 0$, the current had the value given by Eq. (8.18). Hence, constant $= I_0$. Introducing this value into Eq. (8.20), we arrive at the expression

$$I = I_0 \exp\left(-\frac{R}{L} t\right). \tag{8.21}$$

Thus, after the e.m.f. source had been switched off, the current in the circuit did not vanish instantaneously, but diminished according to the exponential law (8.21). A plot of the diminishing of I is given in Fig. 8.8 (curve 1). The rate of diminishing is determined by the quantity

$$\tau = \frac{L}{R}, \tag{8.22}$$

having the dimension of time and called the **time constant** of the circuit. Substituting $1/\tau$ for R/L in Eq. (8.21), we get

$$I = I_0 \exp\left(-\frac{t}{\tau}\right). \tag{8.23}$$

According to this equation, τ is the time during which the current diminishes to $1/e$-th of its initial value. A glance at Eq. (8.22) shows that the time constant τ grows and the current in the circuit diminishes at a slower rate with an increasing inductance L and a decreasing resistance R of the circuit.

To simplify our calculations, we considered that the circuit is shorted when the current source is switched off. If we simply break a circuit with a high inductance, the high induced voltage set up produces a spark or an arc at the place of breaking of the circuit.

Now let us consider the closing of a circuit. After the e.m.f. source is switched

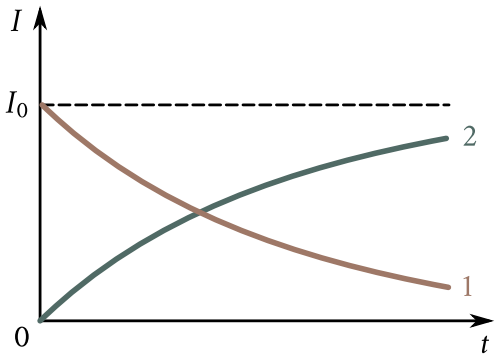


Fig. 8.8

on, a self-induced e.m.f. will act in the circuit apart from the e.m.f. $\mathscr{E}$ until the current reaches its steady value given by Eq. (8.18). Hence, in accordance with Ohm's law

$$IR = \mathscr{E} + \mathscr{E}_s + \mathscr{E} - L\frac{dI}{dt},$$

or

$$\frac{dI}{dt} + \frac{R}{L}I = \frac{\mathscr{E}}{L}. \tag{8.24}$$

We have arrived at a linear inhomogeneous differential equation that differs from Eq. (8.19) only in that the right-hand side contains the constant quantity $\mathscr{E}/L$ instead of zero. It is known from the theory of differential equations that the general solution of a linear inhomogeneous equation can be obtained by adding any partial solution of it to the general solution of the corresponding homogeneous equation (see Sec. 7.4 of Vol. I). The general solution of our homogeneous equation has the form of Eq. (8.20). It is easy to see that $I = \mathscr{E}/R = I_0$ is a partial solution of Eq. (8.24). Hence, the function

$$I = I_0 + \text{constant} \times \exp\left(-\frac{R}{L}t\right),$$

will be the general solution of Eq. (8.24). At the initial moment, the current is zero. Thus, constant $= -I_0$, and

$$I = I_0\left[1 - \exp\left(-\frac{R}{L}t\right)\right]. \tag{8.25}$$

This function describes the growth of the current in a circuit after a source of an e.m.f. has been switched on in it. A plot of function (8.25) is shown in Fig. 8.8 (curve 2).

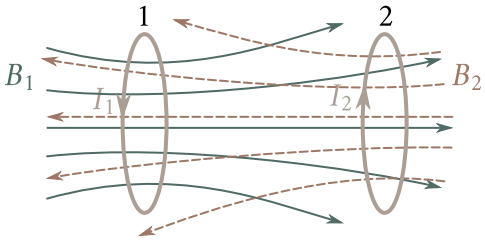


Fig. 8.9

8.7. Mutual Induction

Let us take two loops 1 and 2 close to each other (Fig. 8.9). If the current I_1 flows in loop 1, it sets up through loop 2 a total magnetic flux proportional to I_1, *i.e.*,

$$\Psi_2 = L_{21} I_1 \tag{8.26}$$

(the field producing this flux is depicted in the figure by solid lines). When the current I_1 changes, the e.m.f.

$$\mathscr{E}_{i,2} = -L_{21} \frac{dI_1}{dt}, \tag{8.27}$$

is induced in loop 2 (we assume that there are no ferromagnetics near the loops).

Similarly, when the current I_2 flows in loop 2, the following flux linked with loop 1 appears:

$$\Psi_1 = L_{12} I_2 \tag{8.28}$$

(the field producing this flux is depicted in the figure by dash lines). When the current I_2 changes, the e.m.f.

$$\mathscr{E}_{i,1} = -L_{12} \frac{dI_2}{dt}, \tag{8.29}$$

is induced in loop 1.

Loops 1 and 2 are called **coupled**, while the phenomenon of the setting up of an e.m.f. in one of the loops upon changes in the current in the other is called **mutual induction**.

The coefficients of proportionality L_{12} and L_{21} are called the **mutual inductances** of the loops. The relevant calculations show that in the absence of ferromagnetics these coefficients are always equal to each other:

$$L_{12} = L_{21}. \tag{8.30}$$

Their magnitude depends on the shape, dimensions, and mutual arrangement of the loops, and also on the permeability of the medium surrounding the loops. The quantity L_{12} is measured in the same units as the inductance L.

Let us find the mutual inductance of two coils wound onto a common toroidal

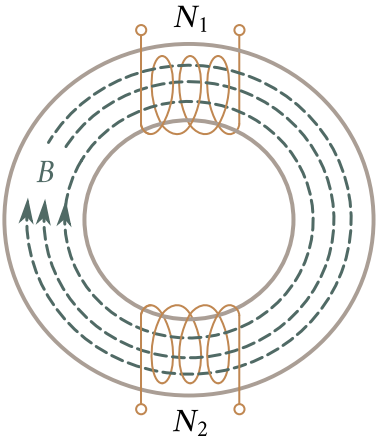


Fig. 8.10

iron core (Fig. 8.10). The magnetic induction lines are concentrated inside the core [see the text following Eq. (7.31)]. We can, therefore, consider that the magnetic field set up by any of the windings will have the same strength throughout the core. If the first winding has N_1 turns and the current I_1 flows through it, then according to the theorem on circulation [see Eq. (7.11)], we have

$$Hl = N_1 I_1 \tag{8.31}$$

(here, l is the length of the core).

The magnetic flux through the cross section of the core is $\Phi = BS = \mu_0 \mu HS$, where S is the cross-sectional area of the core. Introducing the value of H from Eq. (8.31) and multiplying the expression obtained by N_2, we get the total flux linked with the second winding:

$$\Psi_2 = \frac{S}{l} \mu_0 \mu N_1 N_2 I_1.$$

A comparison of this equation with Eq. (8.26) shows that

$$L_{21} = \frac{S}{l} \mu_0 \mu N_1 N_2. \tag{8.32}$$

Calculations of the flux Ψ_1 linked with the first winding when the current I_2 flows through the second winding yields the equation

$$L_{12} = \frac{S}{l} \mu_0 \mu N_1 N_2, \tag{8.33}$$

which coincides in form with L_{21} [see Eq. (8.31)]. In the given case, however, we cannot assert that $L_{12} = L_{21}$. The factor μ in the expressions for these coefficients depends on the field strength H in the core. If $N_1 \neq N_2$, then the same current passed once through the first winding and another time through the second one

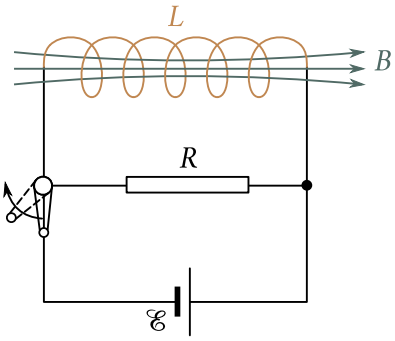


Fig. 8.11

will set up a field of different strength H in the core. Accordingly, the values of μ in both cases will be different so that when $I_1 = I_2$ the numerical values of L_{12} and L_{21} do not coincide.

8.8. Energy of a Magnetic Field

Let us consider the circuit shown in Fig. 8.11. When the switch is closed, the current I will be set up in the solenoid. It will produce a magnetic field linked with the solenoid turns. If the switch is opened, a gradually diminishing current will flow for a certain time through resistor R. This current is maintained by the self-induced e.m.f. produced in the solenoid. The work done by the current during the time dt is

$$\mathrm{d}A = \mathscr{E}_s I \,\mathrm{d}t = -\frac{\mathrm{d}\Psi}{\mathrm{d}t} I \,\mathrm{d}t = -I \,\mathrm{d}\Psi. \tag{8.34}$$

If the inductance of the solenoid does not depend on I (L = constant), then $\mathrm{d}\Psi = L \,\mathrm{d}I$, and Eq. (8.34) becomes

$$\mathrm{d}A = -LI \,\mathrm{d}I. \tag{8.35}$$

Integrating this expression with respect to I within the limits from the initial value of I to zero, we get the work done in the circuit during the entire time needed for vanishing of the magnetic field:

$$A = -\int_I^0 LI \,\mathrm{d}I = \frac{LI^2}{2}. \tag{8.36}$$

The work (8.36) is spent on an increment of the internal energy of the resistor R, the solenoid, and the connecting wires (*i.e.*, on heating them). This work is attended by vanishing of the magnetic field that initially existed in the space surrounding the solenoid. Since no other changes occur in the bodies surrounding the circuit, it remains for us to conclude that the magnetic field is a carrier of energy, and it

is exactly at the expense of the latter that the work given by Eq. (8.36) is done. We, thus, arrive at the conclusion that a conductor of inductance L carrying the current I has the energy

$$W = \frac{LI^2}{2}, \tag{8.37}$$

that is localized in the magnetic field set up by the current [compare this equation with the expression $CU^2/2$ for the energy of a charged capacitor; see Eq. (4.5)].

Equation (8.36) can be interpreted as the work that must be done against the self-induced e.m.f. when the current grows from 0 to I, and that is used to set up a magnetic field having the energy given by Eq. (8.37). Indeed, the work done against the self-induced e.m.f. is

$$A' = \int_0^I (-\mathscr{E}_s) I \, \mathrm{d}t.$$

Performing transformations similar to those which led us to Eq. (8.35), we get

$$A' = \int_0^I LI \, \mathrm{d}I = \frac{LI^2}{2}, \tag{8.38}$$

that coincides with Eq. (8.36). The work according to Eq. (8.38) is done when the current sets in at the expense of the e.m.f. source. It is used completely for producing a magnetic field linked with the solenoid turns. Equation (8.38) takes no account of the work spent by the e.m.f. source for heating the conductors during the time the current reaches its steady value.

Let us express the energy of a magnetic field given by Eq. (8.37) through quantities characterizing the field itself. For a long (virtually infinite) solenoid

$$L = \mu_0 \mu n^2 V, \quad H = nI, \quad \text{or} \quad I = \frac{H}{n}$$

[see Eqs. (7.29) and (8.14)]. Using these values of L and I in Eq. (8.37) and performing the relevant transformations, we obtain

$$W = \frac{\mu_0 \mu H^2}{2} V. \tag{8.39}$$

It was shown in Sec. 6.12 that the magnetic field of an infinitely long solenoid is homogeneous and differs from zero only inside the solenoid. Hence, the energy according to Eq. (8.39) is localized inside the solenoid and is distributed over its volume with a constant density w that can be found by dividing W by V. This division yields

$$w = \frac{\mu_0 \mu H^2}{2}. \tag{8.40}$$

Using Eq. (7.17), we can write the equation for the energy density of a magnetic field

as follows:

$$w = \frac{\mu_0 \mu H^2}{2} = \frac{HB}{2} = \frac{B^2}{2\mu_0 \mu}. \tag{8.41}$$

The expressions we have obtained for the energy density of a magnetic field differ from Eqs. (4.11) for the energy density of an electric field only in that the electrical quantities in them have been replaced with the relevant magnetic ones.

Knowing the density of the field energy at every point, we can find the energy of the field enclosed in any volume V. For this purpose, we must calculate the integral

$$W = \int_V w \, \mathrm{d}V = \int_V \frac{\mu_0 \mu H^2}{2} \, \mathrm{d}V. \tag{8.42}$$

It can be shown that for coupled loops (in the absence of ferromagnetics) the field energy is determined by the equation

$$W = \frac{L_1 I_1^2}{2} + \frac{L_2 I_2^2}{2} + \frac{L_{12} I_1 I_2}{2} + \frac{L_{21} I_2 I_1}{2}. \tag{8.43}$$

A similar expression is obtained for the energy of N loops coupled to one another:

$$W = \frac{1}{2} \sum_{i,k=1}^{N} L_{i,k} I_i I_k, \tag{8.44}$$

where $L_{i,k} = L_{k,i}$ is the mutual inductance of the i-th and k-th loops, and $L_{i,i} = L_i$ is the inductance of the i-th loop.

8.9. Work in Magnetic Reversal of a Ferromagnetic

Changes in a current in a circuit are attended by the performance of work against the self-induced e.m.f.:

$$\mathrm{d}'A = (-\mathscr{E}_s) I \, \mathrm{d}t = \frac{\mathrm{d}\Psi}{\mathrm{d}t} I \, \mathrm{d}t = I \, \mathrm{d}\Psi. \tag{8.45}$$

If the inductance of the circuit L remains constant (which is possible only in the absence of ferromagnetics), this work is used completely for producing the energy of a magnetic field: $\mathrm{d}'A = \mathrm{d}W$. We shall now see that matters are different when ferromagnetics are present.

For a very long ("infinite") solenoid, $H = nI, \Psi = nlBS$. Hence,

$$I = \frac{H}{n}, \quad \mathrm{d}\Psi = nlS \, \mathrm{d}B.$$

Introducing these expressions into Eq. (8.45), we get

$$\mathrm{d}'A = H \, \mathrm{d}B \times V, \tag{8.46}$$

where $V = lS$ is the volume of the solenoid, *i.e.*, the volume in which a homogeneous magnetic feild has been produced.

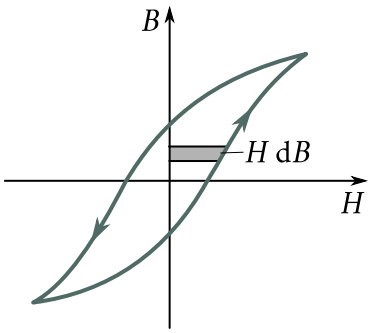


Fig. 8.12

Let us see whether we can identify Eq. (8.46) with the increment of the energy of a magnetic field. We remind our reader that energy is a function of state. Therefore, the sum of its increments for a cyclic process is zero:

$$\oint \mathrm{d}W = 0.$$

If we fill a solenoid with a ferromagnetic, then the relation between B and H is depicted by the curve shown in Fig. 8.12. The expression $H\,\mathrm{d}B$ gives the area of the shaded strip. Consequently, the integral $\oint H\,\mathrm{d}B$ calculated along the hysteresis loop equals the area S_l enclosed by the loop. Thus, the integral of expression (8.46), *i.e.*, $\oint \mathrm{d}'A$, differs from zero. It, therefore, follows that in the presence of ferromagnetics, the work given by Eq. (8.46) cannot be equated to the increment of the energy of a magnetic field. Upon completion of the cycle of magnetic reversal, H and B and, therefore, the magnetic energy will have their initial values. Hence, the work $\oint \mathrm{d}'A$ is not used to produce the energy of a magnetic field. Experiments show that it is used to increase the internal energy of the ferromagnetic, *i.e.*, to heat it.

Thus, the completion of one cycle of magnetic reversal of a ferromagnetic is attended by the expenditure of work per unit volume numerically equal to the area of the hysteresis loop:

$$A_{\mathrm{u.vol}} = \oint H\,\mathrm{d}B = S_l. \tag{8.47}$$

This work goes to heat the ferromagnetic.

In the absence of ferromagnetics, B is an unambiguous function of H ($B = \mu_0\mu H$, where μ = constant). Therefore, the expression $H\,\mathrm{d}B = \mu_0\mu H\,\mathrm{d}H$ is a total differential

$$\mathrm{d}w = H\,\mathrm{d}B, \tag{8.48}$$

determining the increment of the energy of a magnetic field. Integration of Eq. (8.48) within the limits from 0 to H leads to Eq. (8.40) for the density of the field energy

(before performing integration, $H\,\mathrm{d}B$ must be transformed by substituting $\mu_0\mu\,\mathrm{d}H$ for $\mathrm{d}B$).

Chapter 9
MAXWELL'S EQUATIONS

9.1. Vortex Electric Field

Let us consider electromagnetic induction when a wire loop in which a current is induced is stationary, and the changes in the magnetic flux are due to changes in the magnetic field. The setting up of an induced current signifies that the changes in the magnetic field produce extraneous forces in the loop that are exerted on the current carriers. These extraneous forces are associated neither with chemical nor with thermal processes in the wire. They also cannot be magnetic forces because such forces do not work on charges. It remains to conclude that the induced current is due to the electric field set up in the wire. Let us use the symbol $\boldsymbol{E}_B$ to denote the strength of this field (this symbol, like the one $\boldsymbol{E}_q$ used below, is an auxiliary one; we shall omit the subscripts B and q later on). The e.m.f. equals the circulation of the vector $\boldsymbol{E}_B$ around the given loop:

$$\mathscr{E}_\mathrm{i} = \oint \boldsymbol{E}_B \cdot \mathrm{d}\boldsymbol{l}. \tag{9.1}$$

Introducing into Eq. (9.1) $\mathscr{E}_\mathrm{i} = -\mathrm{d}\Phi/\mathrm{d}t$ for $\mathscr{E}_\mathrm{i}$ and the expression $\int \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{S}$ for Φ, we arrive at the equation

$$\oint \boldsymbol{E}_B \cdot \mathrm{d}\boldsymbol{l} = -\frac{\mathrm{d}}{\mathrm{d}t} \int_S \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{S}$$

(the integral in the right-hand side of the equation is taken over an arbitrary surface resting on the loop). Since the loop and the surface are stationary, the operations of time differentiation and integration over the surface can have their places exchanged:

$$\oint \boldsymbol{E}_B \cdot \mathrm{d}\boldsymbol{l} = -\int_S \frac{\partial \boldsymbol{B}}{\partial t} \cdot \mathrm{d}\boldsymbol{S}. \tag{9.2}$$

In connection with the fact that the vector $\boldsymbol{B}$ depends, generally speaking, both on the time and on the coordinates, we have put the symbol of the partial time

derivative inside the integral (the integral $\int \boldsymbol{B} \cdot d\boldsymbol{S}$ is a function only of time).

Let us transform the left-hand side of Eq. (9.2) in accordance with Stokes's theorem. The result is

$$\int_S (\nabla \times \boldsymbol{E}_B) \cdot d\boldsymbol{S} = -\int_S \frac{\partial \boldsymbol{B}}{\partial t} \cdot d\boldsymbol{S}.$$

Owing to the arbitrary nature of choosing the integration surface, the following equation must be obeyed:

$$\nabla \times \boldsymbol{E}_B = -\frac{\partial \boldsymbol{B}}{\partial t}. \tag{9.3}$$

The curl of the field $\boldsymbol{E}_B$ at each point of space equals the time derivative of the vector $\boldsymbol{B}$ taken with the opposite sign.

The British physicist James Maxwell (1831-1879) assumed that a time-varying magnetic field causes the field $\boldsymbol{E}_B$ to appear in space regardless of whether or not there is a wire loop in this space. The presence of a loop only makes it possible to detect the existence of an electric field at the corresponding points of space as a result of a current being induced in the loop.

Thus, according to Maxwell's idea, *a time-varying magnetic field gives birth to an electric field*. This field $\boldsymbol{E}_B$ differs appreciably from the electrostatic field $\boldsymbol{E}_q$ set up by fixed charges. An electrostatic field is a potential one, its strength lines begin and terminate at charges. The curl of the vector $\boldsymbol{E}_B$ is zero at any point:

$$\nabla \times \boldsymbol{E}_q = 0 \tag{9.4}$$

[see Eq. (1.112)]. According to Eq. (9.3), the curl of the vector $\boldsymbol{E}_B$ differs from zero. Hence, the field $\boldsymbol{E}_B$ like a magnetic field, is a vortex one. The strength lines of the field $\boldsymbol{E}_B$ are closed.

Thus, an electric field may be either a potential ($\boldsymbol{E}_q$) or a vortex ($\boldsymbol{E}_B$) one. In the general case, an electric field can consist of the field $\boldsymbol{E}_q$ produced by charges and the field $\boldsymbol{E}_B$ set up by a time-varying magnetic field. Adding Eqs. (9.3) and (9.4), we get the following equation for the curl of the strength of the total field $\boldsymbol{E} = \boldsymbol{E}_B + \boldsymbol{E}_B$:

$$\nabla \times \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t}. \tag{9.5}$$

This equation is one of the fundamental ones in Maxwell's electromagnetic theory.

The existence of a relationship between electric and magnetic fields [expressed in particular by Eq. (9.5)] is a reason why the separate treatment of these fields has only a relative meaning. Indeed, an electric field is set up by a system of fixed charges. If the charges are fixed relative to a certain inertial reference frame, however, they are moving relative to other inertial frames and, consequently, set up not only an electric, but also a magnetic field. A stationary wire carrying a steady current sets up a constant magnetic field at every point of space. This wire is in motion, however,

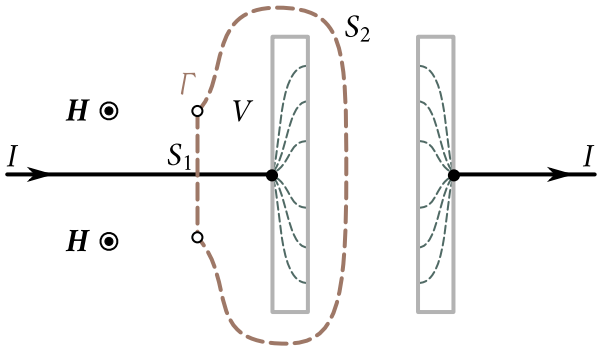


Fig. 9.1

relative to other inertial frames. Consequently, the magnetic field it sets up at any point with the given coordinates x, y, z will change and, therefore, give birth to a vortex electric field. Thus, a field which is "purely" electric or "purely" magnetic relative to a certain reference frame will be a combination of an electric and a magnetic field forming a single electromagnetic field relative to other reference frames.

9.2. Displacement Current

For a stationary (*i.e.*, not varying with time) electromagnetic field, the curl of the vector $\boldsymbol{H}$ by Eq. (7.9) equals the density of the conduction current at each point:

$$\nabla \times \boldsymbol{H} = \boldsymbol{j}.$$

The vector $\boldsymbol{j}$ is associated with the charge density at the same point by continuity equation (5.11):

$$\nabla \cdot \boldsymbol{j} = -\frac{\partial \rho}{\partial t}.$$

An electromagnetic field can be stationary only provided that the charge density ρ and the current density $\boldsymbol{j}$ do not depend on the time. In this case, according to Eq. (5.11), the divergence of $\boldsymbol{j}$ equals zero. Therefore, the current lines (lines of the vector $\boldsymbol{j}$) have no sources and are closed.

Let us see whether Eq. (7.9) holds for time-varying fields. We shall consider the current flowing when a capacitor is charged from a source of constant voltage U. This current varies with time (the current stops flowing when the voltage across the capacitor becomes equal to U). The lines of the conduction current are interrupted in the space between the capacitor plates (Fig. 9.1; the current lines inside the plates are shown by dash lines).

Let us take a circular loop Γ enclosing the wire in which the current flows toward the capacitor and integrate Eq. (7.9) over surface S_1 intersecting the wire and enclosed by the loop:

$$\int_{S_1} \nabla \times \boldsymbol{H} \cdot \mathrm{d}\boldsymbol{S} = \int_{S_1} \boldsymbol{j} \cdot \mathrm{d}\boldsymbol{S}.$$

Transforming the left-hand side according to Stokes's theorem we get the circulation of the vector $\boldsymbol{H}$ over loop Γ:

$$\oint_{\Gamma} \boldsymbol{H} \cdot \mathrm{d}\boldsymbol{l} = \int_{S_1} \boldsymbol{j} \cdot \mathrm{d}\boldsymbol{S} = I \tag{9.6}$$

(I is the current charging the capacitor). After performing similar calculations for surface S_2 that does not intersect the current-carrying wire (see Fig. 9.1), we arrive at the obviously incorrect relation

$$\oint_{\Gamma} \boldsymbol{H} \cdot \mathrm{d}\boldsymbol{l} = \int_{S_2} \boldsymbol{j} \cdot \mathrm{d}\boldsymbol{S} = 0. \tag{9.7}$$

The result we have obtained indicates that for time-varying fields Eq. (7.9) stops being correct. The conclusion suggests itself that this equation lacks an addend depending on the time derivatives of the fields. For stationary fields, this addend vanishes.

That Eq. (7.9) is not correct for non-stationary fields is also indicated by the following reasoning. Let us take the divergence of both sides of Eq. (7.9):

$$\nabla \cdot (\nabla \times \boldsymbol{H}) = \nabla \cdot \boldsymbol{j}.$$

The divergence of a curl must equal zero [see Eq. (1.106)]. We, thus, arrive at the conclusion that the divergence of the vector $\boldsymbol{j}$ must also always equal zero. But this conclusion contradicts the continuity equation (5.11). Indeed, in non-stationary processes, ρ may change with time (this, in particular, is what happens with the charge density on the plates of a capacitor being charged). In this case in accordance with Eq. (5.11), the divergence of $\boldsymbol{j}$ differs from zero.

To bring Eqs. (5.11) and (7.9) into agreement, Maxwell introduced an additional addend into the right-hand side of Eq. (7.9). It is quite natural that this addend should have the dimension of current density. Maxwell called it the density of the displacement current. Thus, according to Maxwell, Eq. (7.9) should have the form

$$\nabla \times \boldsymbol{H} = \boldsymbol{j} + \boldsymbol{j}_{\mathrm{d}}. \tag{9.8}$$

The sum of the conduction current and the displacement current is usually called the **total current**. The density of the total current is

$$\boldsymbol{j}_{\mathrm{tot}} = \boldsymbol{j} + \boldsymbol{j}_{\mathrm{d}}. \tag{9.9}$$

If we assume that the divergence of the displacement current equals that of the

conduction current taken with the opposite sign:

$$\nabla \cdot \boldsymbol{j}_{\mathrm{d}} = -\nabla \cdot \boldsymbol{j}, \tag{9.10}$$

then the divergence of the right-hand side of Eq. (9.8), like that of the left-hand side, will always be zero.

Substituting $\partial\rho/\partial t$ for $\nabla \cdot \boldsymbol{j}$ in Eq. (9.10) in accordance with Eq. (5.11), we get the following expression for the divergence of the displacement current:

$$\nabla \cdot \boldsymbol{j}_{\mathrm{d}} = \partial\rho/\partial t. \tag{9.11}$$

To associate the displacement current with quantities characterizing the change in an electric field with time, let us use Eq. (2.23) according to which the divergence of the electric displacement vector equals the density of the extraneous charges:

$$\nabla \cdot \boldsymbol{D} = 0.$$

Time differentiation of this equation yields

$$\frac{\partial}{\partial t}(\nabla \cdot \boldsymbol{D}) = \frac{\partial\rho}{\partial t}.$$

Now, let us change the sequence of differentiation with respect to time and to the coordinates in the left-hand side. As a result, we get the following. expression for the derivative of ρ with respect to t:

$$\frac{\partial\rho}{\partial t} = \nabla \cdot \left(\frac{\partial \boldsymbol{D}}{\partial t}\right).$$

Introduction of this expression into Eq. (9.11) yields

$$\nabla \cdot \boldsymbol{j}_{\mathrm{d}} = \nabla \cdot \left(\frac{\partial \boldsymbol{D}}{\partial t}\right).$$

Hence,

$$\boldsymbol{j}_{\mathrm{d}} = \frac{\partial \boldsymbol{D}}{\partial t}. \tag{9.12}$$

Using Eq. (9.12) in Eq. (9.8), we arrive at the equation

$$\nabla \times \boldsymbol{H} = \boldsymbol{j} + \frac{\partial \boldsymbol{D}}{\partial t}, \tag{9.13}$$

which, like Eq. (9.5), is one of the fundamental equations in Maxwell's theory.

We must underline the fact that the term "displacement current" is purely conventional. In essence, the displacement current is a time-varying electric field. The only reason for calling the quantity given by Eq. (9.12) a "current" is that the dimension of this quantity coincides with that of current density. Of all the physical properties belonging to a real current, a displacement current has only one—the ability of producing a magnetic field.

The introduction of the displacement current determined by Eq. (9.12) has "given equal rights" to an electric field and a magnetic field. It can be seen from the

phenomenon of electromagnetic induction that a varying magnetic field sets up an electric field. It follows from Eq. (9.13) that a varying electric field sets up a magnetic field.

There is a displacement current wherever there is a time-varying electric field. In particular, it also exists inside conductors carrying an alternating electric current. The displacement current inside conductors, however, is usually negligibly small in comparison with the conduction current.

We must note that Eq. (9.6) is approximate. For it to become quite strict, we must add a term to its right-hand side that takes account of the displacement current due to the weak dispersed electric field in the vicinity of surface S_1.

Let us convince ourselves that the surface integral of the right-hand side of Eq. (9.8) has the same value for surfaces S_1 and S_2 (see Fig. 9.1). Both the conduction current and the displacement current due to the electric field outside the capacitor "flow" through surface S_1. Hence, for the first surface, we have

$$\mathrm{Int}_1 = \int_{S_1} \boldsymbol{j} \cdot \mathrm{d}\boldsymbol{S} + \frac{\mathrm{d}}{\mathrm{d}t} \int_{S_1} \boldsymbol{D} \cdot \mathrm{d}\boldsymbol{S} = I + \frac{\mathrm{d}}{\mathrm{d}t}\Phi_{1,\mathrm{in}}$$

(we have changed the sequence of the operations of differentiation with respect to time and integration over the coordinates in the second addend). The quantity designated by the letter I is the current flowing in the conductor to the left-hand plate of the capacitor, $\Phi_{1,\mathrm{in}}$ is the flux of the vector $\boldsymbol{D}$ flowing into the volume V bounded by surfaces S_1 and S_2 (see Fig. 9.1).

For the second surface, $\boldsymbol{j} = 0$, consequently

$$\mathrm{Int}_2 = \frac{\mathrm{d}}{\mathrm{d}t} \int_{S_2} \boldsymbol{D} \cdot \mathrm{d}\boldsymbol{S} = \frac{\mathrm{d}}{\mathrm{d}t}\Phi_{2,\mathrm{out}}$$

where $\Phi_{2,\mathrm{out}}$ is the flux of the vector $\boldsymbol{D}$ flowing out of volume V through surface S_2.

The difference between the integrals is

$$\mathrm{Int}_2 - \mathrm{Int}_1 = \frac{\mathrm{d}}{\mathrm{d}t}\Phi_{2,\mathrm{out}} - \frac{\mathrm{d}}{\mathrm{d}t}\Phi_{1,\mathrm{in}} - I.$$

The current I can be represented as $\mathrm{d}q/\mathrm{d}t$, where q is the charge on a capacitor plate. The flux passing inward through surface S_1 equals the flux passing outward through the same surface taken with the opposite sign. Substituting $-\Phi_{1,\mathrm{out}}$ for $\Phi_{1,\mathrm{in}}$ and $\mathrm{d}q/\mathrm{d}t$ for I, we get

$$\mathrm{Int}_2 - \mathrm{Int}_1 = \frac{\mathrm{d}}{\mathrm{d}t}\left(\Phi_{2,\mathrm{out}} + \Phi_{1,\mathrm{out}}\right) - \frac{\mathrm{d}q}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}\left(\Phi_D - q\right), \tag{9.14}$$

where Φ_D is the flux of the vector $\boldsymbol{D}$ through the closed surface formed by surfaces S_1 and S_2. According to Eq. (2.25), this flux must equal the charge enclosed by the surface. In the given case, it is the charge q on a capacitor plate. Thus, the right-hand side of Eq. (9.14) equals zero. It follows that the magnitude of the surface integral of

the total current density vector does not depend on the choice of the surface over which the integral is being calculated.

We can construct current lines for the displacement current like those for the conduction current. According to Eq. (2.35), the electric displacement in the space between the capacitor plates equals the surface charge density on a plate: $D = \sigma$. Hence,

$$\dot{D} = \dot{\sigma}.$$

The left-hand side gives the density of the displacement current in the space between the plates, and the right-hand side-the density of the conduction current inside the

current uninterruptedly transform into lines of the displacement current at the boundary of the plates. Consequently, the lines of the total current are closed.

9.3. Maxwell's Equations

The discovery of the displacement current permitted Maxwell to present a single general theory of electrical and magnetic phenomena. This theory explained all the experimental facts known at that time and predicted a number of new phenomena whose existence was confirmed later on. The main corollary of Maxwell's theory was the conclusion on the existence of electromagnetic waves propagating with the speed of light. Theoretical investigation of the properties of these waves led Maxwell to the electromagnetic theory of light.

The theory is based on **Maxwell's equations**. These equations play the same part in the science of electromagnetism as Newton's laws do in mechanics, or the fundamental laws in thermodynamics.

The **first pair of Maxwell's equations** is formed by Eqs. (9.5) and (7.3):

$$\nabla \times \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t}, \tag{9.5}$$

$$\nabla \cdot \boldsymbol{B} = 0. \tag{7.3}$$

The first of them relates the values of $\boldsymbol{E}$ to changes in the vector $\boldsymbol{B}$ in time and is in essence an expression of the law of electromagnetic induction. The second one points to the absence of sources of a magnetic field, *i.e.*, magnetic charges.

The **second pair of Maxwell's equations** is formed by Eqs. (9.13) and (2.23):

$$\nabla \times \boldsymbol{H} = \boldsymbol{j} + \frac{\partial \boldsymbol{D}}{\partial t}, \tag{9.13}$$

$$\nabla \cdot \boldsymbol{D} = \rho. \tag{2.23}$$

The first of them establishes a relation between the conduction and displacement

currents and the magnetic field they produce. The second one shows that extraneous charges are the sources of the vector $\boldsymbol{D}$.

Equations (9.5), (7.3), (9.13) and (2.23) are Maxwell's equations in the differential form. We must note that the first pair of equations includes only the basic characteristics of a field, namely, $\boldsymbol{E}$ and $\boldsymbol{B}$. The second pair includes only the auxiliary quantities $\boldsymbol{D}$ and $\boldsymbol{H}$.

Each of the vector equations (9.5) and (9.13) is equivalent to three scalar equations relating the components of the vectors in the left-hand and right-hand sides of the equations. Using Eqs. (1.81) and (1.92)-(1.91), let us present Maxwell's equation in the scalar form:

$$\begin{cases} \dfrac{\partial E_z}{\partial y} - \dfrac{\partial E_y}{\partial z} = -\dfrac{\partial B_x}{\partial t} \\ \dfrac{\partial E_x}{\partial z} - \dfrac{\partial E_z}{\partial x} = -\dfrac{\partial B_y}{\partial t} \\ \dfrac{\partial E_y}{\partial x} - \dfrac{\partial E_x}{\partial y} = -\dfrac{\partial B_z}{\partial t} \end{cases} \tag{9.15}$$

$$\frac{\partial B_x}{\partial x} + \frac{\partial B_y}{\partial y} + \frac{\partial B_z}{\partial z} = 0, \tag{9.16}$$

(the first pair of equations),

$$\begin{cases} \dfrac{\partial H_z}{\partial y} - \dfrac{\partial H_y}{\partial z} = j_x + \dfrac{\partial D_x}{\partial t} \\ \dfrac{\partial H_x}{\partial z} - \dfrac{\partial H_z}{\partial x} = j_y + \dfrac{\partial D_y}{\partial t} \\ \dfrac{\partial H_y}{\partial x} - \dfrac{\partial H_x}{\partial y} = j_z + \dfrac{\partial D_z}{\partial t} \end{cases} \tag{9.17}$$

$$\frac{\partial D_x}{\partial x} + \frac{\partial D_y}{\partial y} + \frac{\partial D_z}{\partial z} = \rho, \tag{9.18}$$

(the second pair of equations).

We get a total of 8 equations including 12 functions (three components each of the vectors $\boldsymbol{E}$, $\boldsymbol{B}$, $\boldsymbol{D}$, $\boldsymbol{H}$). Since the number of equations is less than the number of unknown functions, (9.5), (7.3), (9.13) and (2.23) are not sufficient for finding the fields according to the given distribution of the charges and currents. To calculate the fields, we must add equations relating $\boldsymbol{D}$ and $\boldsymbol{j}$ to $\boldsymbol{E}$ and also $\boldsymbol{H}$ to $\boldsymbol{B}$ to these equations. They have the form

$$\boldsymbol{D} = \varepsilon_0 \varepsilon \boldsymbol{E}, \tag{2.21}$$

$$\boldsymbol{B} = \mu_0 \mu \boldsymbol{H}, \tag{7.17}$$

$$\boldsymbol{j} = \sigma \boldsymbol{E}. \tag{5.22}$$

The collection of equations (9.5), (7.3), (9.13) and (2.23), and (2.21), (7.17), (5.22) forms the foundation of the electrodynamics of media at rest.

The equations

$$\oint_\Gamma \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{l} = -\frac{\mathrm{d}}{\mathrm{d}} \int_S \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{S}, \tag{9.19}$$

$$\oint_S \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{S} = 0, \tag{9.20}$$

(the first pair) and

$$\oint_\Gamma \boldsymbol{H} \cdot \mathrm{d}\boldsymbol{l} = \int_S \boldsymbol{j} \cdot \mathrm{d}\boldsymbol{S} + \frac{\mathrm{d}}{\mathrm{d}} \int_S \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{S}, \tag{9.21}$$

$$\oint_S \boldsymbol{D} \cdot \mathrm{d}\boldsymbol{S} = \int_V \rho \, \mathrm{d}V, \tag{9.22}$$

(the second pair) are **Maxwell's equations in the integral form**.

Equation (9.19) is obtained by integration of Eq. (9.5) over arbitrary surface S with the following transformation of the left-hand side according to Stokes's theorem into an integral over loop Γ enclosing surface S. Equation (9.21) is obtained in the same way from Eq. (9.13). Equations (9.20) and (9.22) are obtained from Eqs. (7.3) and (2.23) by integration over the arbitrary volume V with the following transformation of the left-hand side according to the Ostrogradsky-Gauss theorem into an integral over closed surface S enclosing volume V.

Chapter 10

MOTION OF CHARGED PARTICLES IN ELECTRIC AND MAGNETIC FIELDS

10.1. Motion of a Charged Particle in a Homogeneous Magnetic Field

Imagine a charge e' moving in a homogeneous magnetic field with the velocity $\boldsymbol{v}$ perpendicular to $\boldsymbol{B}$. The magnetic force imparts to the charge an acceleration perpendicular to the velocity

$$a_{\mathrm{n}} = \frac{F}{m} = \frac{e'}{m} vB \tag{10.1}$$

[see Eq. (6.33); the angle between $\boldsymbol{v}$ and $\boldsymbol{B}$ is a right one]. This acceleration changes only the direction of the velocity, while the magnitude of the latter remains unchanged. Hence, the acceleration given by Eq. (10.1) will be constant in magnitude too. In these conditions, the charged particle moves uniformly around a circle whose radius is determined by means of the equation $a_{\mathrm{n}} = v^2/R$. Substituting for a_{n} in this equation its value from Eq. (10.1) and solving the resulting equation relative to R, we get

$$R = \frac{m}{e'} \frac{v}{B}. \tag{10.2}$$

Thus, when a charged particle moves in a homogeneous magnetic field perpendicular to the plane in which the motion is taking place, the trajectory of the particle is a circle. The radius of the circle depends on the velocity of the particle, the magnetic induction of the field, and the ratio of the charge of the particle e' to its mass m. The ratio e'/m is called the **specific charge**.

Let us find the time T needed for the particle to complete one revolution. For this purpose, we shall divide the length of the circumference $2\pi R$ by the velocity of

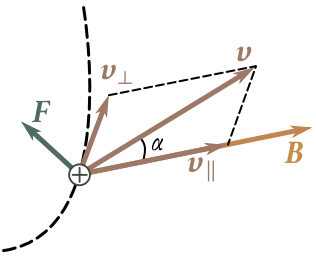


Fig. 10.1

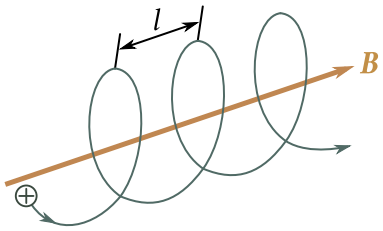


Fig. 10.2

the particle v. The result is

$$T = 2\pi \frac{m}{e'} \frac{1}{B}. \tag{10.3}$$

Inspection of Eq. (10.3) shows that the period of revolution of the particle does not depend on its velocity. It is determined only by the specific charge of the particle and the magnetic induction of the field.

Let us determine the nature of motion of a charged particle when its velocity makes the angle α with the direction of a homogeneous magnetic field, and α is not a right angle. We shall resolve the vector $\boldsymbol{v}$ into two components: $\boldsymbol{v}_\perp$ perpendicular to $\boldsymbol{B}$, and $\boldsymbol{v}_\parallel$ parallel to $\boldsymbol{B}$ (Fig. 10.1). The magnitudes of these components are

$$v_\perp = v \sin \alpha, \quad v_\parallel = v \cos \alpha.$$

The magnetic force has the magnitude

$$F = e' v B \sin \alpha = e' v_\perp B,$$

and is in a plane at right angles to $\boldsymbol{B}$. The acceleration produced by this force is normal for the component $\boldsymbol{v}_\perp$. The component of the magnetic force in the direction of $\boldsymbol{B}$ is zero. Hence, this force cannot affect the magnitude of $\boldsymbol{v}_\parallel$. The motion of the particle can, thus, be considered as the superposition of two motions: (1) translation along the direction of $\boldsymbol{B}$ with a constant velocity $v_\parallel = v \cos \alpha$, and (2) uniform circular motion in a plane at right angles to the vector $\boldsymbol{B}$.

The radius of the circle is determined by Eq. (10.2) with $v_\perp = v \sin \alpha$ substituted for v. The trajectory of motion is a helix (spiral) whose axis coincides with the direction of $\boldsymbol{B}$ (Fig. 10.2). The pitch of the helix l can be found by multiplying $v_\parallel$ by the period of revolution T determined by Eq. (10.3):

$$l = v_\parallel T = 2\pi \frac{m}{e'} \frac{1}{B} v \cos \alpha. \tag{10.4}$$

The direction in which the helix curls depends on the sign of the particle's charge. If the latter is positive, the helix curls counterclockwise. A helix along which a negatively charged particle is moving curls clockwise (it is assumed that we are looking at the helix along the direction of $\boldsymbol{B}$; the particle flies away from us

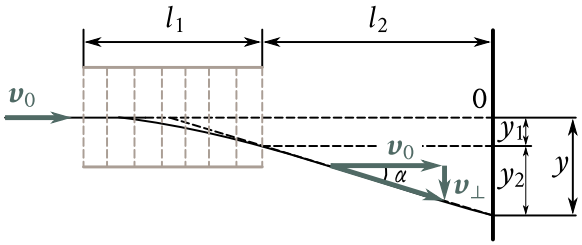


Fig. 10.3

if $\alpha < \pi/2$, and toward us if $\alpha > \pi/2$).

10.2. Deflection of Moving Charged Particles by an Electric and a Magnetic Field

Let us consider a narrow beam of identically charged particles (for example, electrons) that in the absence of fields falls on a screen perpendicular to it at point 0 (Fig. 10.3). Let us find the displacement of the trace of the beam produced by a homogeneous electric field perpendicular to the beam and acting on a path of length l_1. Let the initial velocity of the particles be $\boldsymbol{v}_0$. Upon entering the region of the field, each particle will move with an acceleration $a_\perp = (e'/m)E$ constant in magnitude and in direction and perpendicular to $\boldsymbol{v}_0$ (here, e'/m is the specific charge of a particle). Motion under the action of the field continues during the time $t = l_1/v_0$. During this time, the particles will be displaced over the distance

$$y_1 = \frac{1}{2}a_\perp t^2 = \frac{1}{2}\frac{e'}{m}E\frac{l_1^2}{v_0^2}, \tag{10.5}$$

and will acquire the following velocity component perpendicular to $\boldsymbol{v}_0$:

$$v_\perp = a_\perp t = \frac{e'}{m}E\frac{l_1}{v_0}.$$

The particles now fly in a straight line in a direction that makes with the vector $\boldsymbol{v}_0$ the angle α determined by the expression

$$\tan\alpha = \frac{v_\perp}{v_0} - \frac{e'}{m}E\frac{l_1}{v_0^2}. \tag{10.6}$$

As a result in addition to the displacement given by Eq. (10.5) the beam receives the displacement

$$y_2 = l_2\tan\alpha = \frac{e'}{m}E\frac{l_1 l_2}{v_0^2},$$

where l_2 is the distance to the screen from the boundary of the region which the field is in.

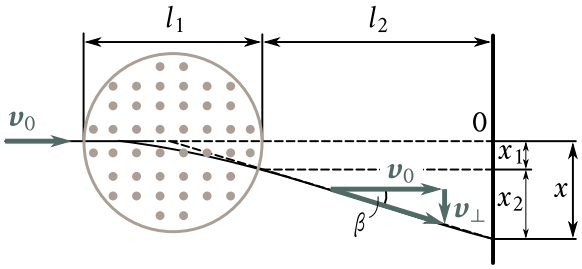


Fig. 10.4

The displacement of the trace of the beam relative to point 0 is thus

$$y = y_1 + y_2 = \frac{e'}{m} E \frac{l_1}{v_0^2} \left(\frac{1}{2} l_1 + l_2 \right). \tag{10.7}$$

Taking into account Eq. (10.6), the expression for the displacement can be written in the form

$$y = \left(\frac{1}{2} l_1 + l_2 \right) \tan \alpha.$$

It thus follows that the particles leaving the field fly as if they were leaving the centre of the capacitor setting up the field at the angle α determined by means of Eq. (10.6).

Now let us assume that on a particle path of l_1 a homogeneous magnetic field is switched on perpendicular to the velocity $\boldsymbol{v}_0$ of the particles (Fig. 10.4; the field is perpendicular to the plane of the drawing, the region of the field is surrounded by a dash circle). Under the action of the field, each particle receives the acceleration $a_\perp = (e'/m) v_0 B$ constant in magnitude. Limiting ourselves to the case when the deflection of the beam by the field is not great, we can consider that the acceleration $a_\perp$ is constant in magnitude and perpendicular to v_0. Hence, we can use the equations we have obtained for calculating the displacement, replacing the acceleration $a_\perp = (e'/m)E$ in them with the value $a_\perp = (e'/m) v_0 B$. As a result, we get the following expression for the displacement, which we shall now denote by x:

$$x = \frac{e'}{m} B \frac{l_1}{v_0} \left(\frac{1}{2} l_1 + l_2 \right). \tag{10.8}$$

The angle through which the beam is deflected by the magnetic field is determined by the expression

$$\tan \beta = \frac{e'}{m} B \frac{l_1}{v_0}. \tag{10.9}$$

With a view to Eq. (10.3), we can write Eq. (10.8) in the form

$$x = \left(\frac{1}{2} l_1 + l_2 \right) \tan \beta.$$

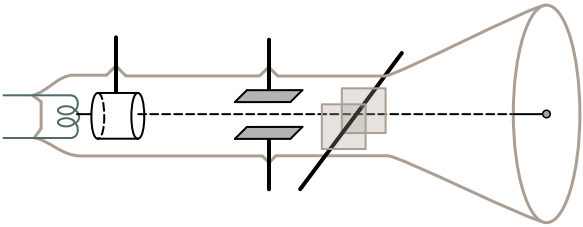


Fig. 10.5

Consequently, upon small deflections, the particles after leaving the magnetic field fly as if they had left the centre of the region containing the deflecting field at the angle β whose magnitude is determined by Eq. (10.9).

Inspection of Eqs. (10.7) and (10.8) shows that both the deflection by an electric field and the deflection by a magnetic one are proportional to the specific charge of the particles.

The deflection of a beam of electrons by an electric or magnetic field is used in cathode-ray tubes. A tube with electrical deflection (Fig. 10.5), apart from the so-called electron gun producing a narrow beam of fast electrons (an electron beam), contains two pairs of mutually perpendicular deflecting plates. By feeding a voltage to any pair of plates, we can produce a proportional displacement of the electron beam in a direction normal to the given plates. The screen of the tube is coated with a fluorescent composition. Therefore, a brightly luminescent spot appears on the screen where the electron beam falls on it.

Cathode-ray tubes are used in oscillographs—instruments making it possible to study rapid processes. A voltage changing linearly with time (the scanning voltage) is fed to one pair of deflecting plates, and the voltage being studied to the other. Owing to the negligibly small inertia of an electron beam, its deflection without virtually any delay follows the changes in the voltages across both pairs of deflecting plates, and the beam draws on the oscillograph screen a plot of time dependence of the voltage being studied. Many nonelectrical quantities can be transformed into electric voltages with the aid of the relevant devices (transducers). Consequently, oscillographs are used to study the most diverse processes.

A cathode-ray tube is an integral part of television equipment. In television, tubes with magnetic control of the electron beam are used most frequently. In these tubes, the deflecting plates are replaced with two external mutually perpendicular systems of coils each of which sets up a magnetic field perpendicular to the beam. Changing of the current in the coils produces motion of the light spot created by the electron beam on the screen.

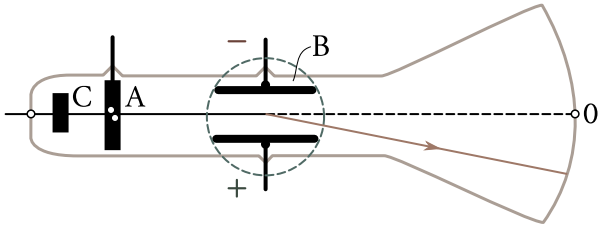


Fig. 10.6

10.3. Determination of the Charge and Mass of an Electron

The specific charge of an electron (*i.e.*, the ratio e/m) was first measured by the British physicist Joseph J. Thomson (1856-1940) in 1897 with the aid of a discharge tube like the one shown in Fig. 10.6. The electron beam (cathode rays; see Sec. 12.6) emerging from the opening in anode A passed between the plates of a parallel-plate capacitor and impinged on a fluorescent screen producing a light spot on it. By feeding a voltage to the capacitor plates, it was possible to act on the beam with a virtually homogeneous electric field.

The tube was placed between the poles of an electromagnet, which could produce a homogeneous magnetic field perpendicular to the electric one on the same portion of the path of the electrons (the region of the magnetic field is shown in Fig. 10.6 by the dash circle). When the fields were switched off, the beam impinged on the screen at point 0. Each of the fields separately caused deflection of the beam in a vertical direction. The magnitudes of the displacements were determined with the aid of Eqs. (10.7) and (10.8) obtained in the preceding section.

After switching on the magnetic field and measuring the displacement of the beam trace

$$x = \frac{e}{m} B \frac{l_1}{v_0} \left(\frac{1}{2} l_1 + l_2 \right), \tag{10.10}$$

which it produced, Thomson also switched on the electric field and selected its value so that the beam would again reach point 0. In this case, the electric and magnetic fields acted on the electrons of the beam simultaneously with forces identical in value, but opposite in direction. The condition was observed that

$$eE = ev_0 B. \tag{10.11}$$

By solving the simultaneous equations (10.10) and (10.11), Thomson calculated e/m and v_0. H. Busch used the method of magnetic focussing to determine the specific charge of electrons. The essence of this method consists in the following. Assume that a slightly diverging beam of electrons having a velocity v identical in magnitude flies out from a certain point of a homogeneous magnetic field. The beam is sym-

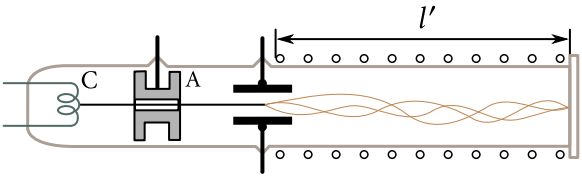


Fig. 10.7

metrical relative to the direction of the field. The directions in which the electrons fly out form small angles α with the direction of $\boldsymbol{B}$. It was shown in Sec. 10.1 that the electrons in this case travel along helical trajectories, performing during the identical time

$$T = 2\pi \frac{m}{e} \frac{1}{B},$$

a complete revolution and being displaced along the direction of the field over the distance l equal to

$$l = v \cos \alpha \times T. \tag{10.12}$$

Owing to the smallness of the angles α, the distances (10.12) for different electrons are virtually the same and equal vT (for small angles $\cos \alpha \approx 1$). Consequently, the slightly diverging beam is focussed at a point that is at the distance

$$l = vT = 2\pi \frac{m}{e} \frac{v}{B} \tag{10.13}$$

from the point of emergence of the electrons.

In Busch's experiment, the electrons emitted by hot cathode C (Fig. 10.7) are accelerated when passing through the potential difference U applied between the cathode and anode A. As a result, they acquire the velocity v whose value can be found from the relation

$$eU = \frac{mv^2}{2}. \tag{10.14}$$

After next flying out through an opening in the anode, the electrons form a narrow beam directed along the axis of the evacuated tube inserted into a solenoid. A capacitor fed with a varying voltage is placed at the inlet of the solenoid. The field set up by the capacitor deflects the electrons of the beam from the axis of the instrument through small angles α changing with time. This leads to "eddying" of the beam—the electrons begin to move along different helical trajectories. A fluorescent screen is placed at the outlet from the solenoid. If the magnetic induction B is selected so that the distance l' from the capacitor to the screen complies with the condition

$$l' = nl \tag{10.15}$$

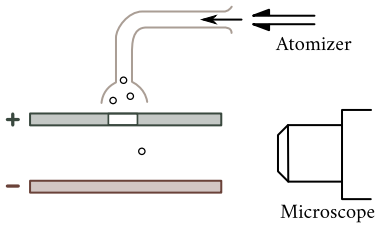


Fig. 10.8

(l is the pitch of the helix, and n is an integer), then the point of intersection of the trajectories of the electrons gets onto the screen the electron beam is focussed at this point and produces a sharp luminescent spot on the screen. If condition (10.15) is not observed, the luminescent spot on the screen will be blurred. We can find e/m and v by solving the system of equations (10.13), (10.14), and (10.15).

The most accurate value of the specific charge of an electron established with account taken of the results obtained by different methods, is

$$\frac{e}{m} = 1.76 \times 10^{11}\ \mathrm{C\,kg^{-1}} = 5.27 \times 10^{17}\ \mathrm{cgse}_q\ \mathrm{g}^{-1}. \tag{10.16}$$

Equation (10.16) gives the ratio of the charge of an electron to its rest mass m. In the experiments conducted by Thomson, Busch, and in other similar experiments, the ratio of the charge to the relativistic mass

$$m_{\mathrm{r}} = \frac{m}{\sqrt{1 - (v^2/c^2)}}, \tag{10.17}$$

was determined. In Thomson's experiments, the speed of the electrons was about $0.1c$. At such a speed, the relativistic mass exceeds the rest mass by 0.5%. In subsequent experiments, the speed of the electrons reached very high values. In all cases, the experimenters discovered a reduction in the measured values of e/m with a growth in v, which occurred in complete accordance with Eq. (10.17).

The charge of an electron was determined with high accuracy by the American scientist Robert Millikan (1886-1953) in 1909. He introduced very minute oil droplets into the closed space between horizontally arranged capacitor plates (Fig. 10.8). When atomized, the droplets became electrolyzed, and they could be suspended in mid air by properly choosing the magnitude and the sign of the voltage across the capacitor. Equilibrium set in when the following condition was observed:

$$P' = e'E. \tag{10.18}$$

Here, e' is the charge of a droplet, and P' is the resultant of the force of gravity and the buoyant force equal to

$$P' = \frac{4}{3}\pi r^2(\rho - \rho_0)g \tag{10.19}$$

(ρ is the density of a droplet, r is its radius, and ρ_0 is the density of air).

Equations (10.18) and (10.19) can be used to find e if we know r. To determine the radius, the speed v_0 of uniform falling of a droplet was measured in the absence of a field. Uniform motion of a droplet sets in provided that the force P' is balanced by the force of resistance $F = 6\pi\eta r v$ [see Eq. (9.24) of Vol. I; η is the viscosity of air]:

$$P' = 6\pi\eta r v_0. \tag{10.20}$$

The motion of a droplet was observed with the aid of a microscope. To measure v_0, the time was determined during which a droplet covered the distance between two threads that could be seen in the field of vision of the microscope.

It is very difficult to accurately suspend a droplet in equilibrium. Therefore, instead of a field complying with condition (10.18), such a field was switched on under whose action a droplet began to move upward with a small speed. The steady speed of rising v_E is determined from the condition that the force P' and the force $6\pi\eta r v$ together balance the force $e'E$:

$$P' + 6\pi\eta r v_E = e'E. \tag{10.21}$$

Excluding P' and r from Eqs. (10.19), (10.20), and (10.21), we get an expression for e':

$$e' = 9\pi \left[\frac{2\eta^3 v_0}{(\rho - \rho_0)g}\right]^{1/2} \left(\frac{v_0 + v_E}{E}\right) \tag{10.22}$$

(Millikan introduced a correction into this equation taking into account that the dimensions of the droplets were comparable with the free path of air molecules).

Thus, by measuring the speed of free fall of a droplet v_0 and the speed of its rise v_E in a known electric field E, one could find the charge of a droplet e'. In measuring the speed v_E at a certain value of the charge e', Millikan ionized the air by radiating X-rays through the space between the plates. Separate ions adhered to a droplet and changed its charge. As a result, the speed v_E also changed. After measuring the new value of the speed, the space between the plates was again irradiated, and so on.

The changes in the charge of a droplet $\Delta e'$ and the charge e' itself measured by Millikan were each time found to be integral multiples of the same quantity e. This was an experimental proof of the discrete nature of an electric charge, *i.e.*, of the fact that any charge consists of elementary charges of the same magnitude.

The value of the elementary charge established with a view to Millikan's measurements and to the data obtained in other ways is

$$e = 1.60 \times 10^{-19}\ \text{C} = 4.80 \times 10^{-10}\ \text{cgse}_q. \tag{10.23}$$

The charge of an electron has the same value.

The rest mass of an electron obtained from Eqs. (10.16) and (10.23) is

$$m = 0.91 \times 10^{-30}\ \text{kg} = 0.91 \times 10^{-23}\ \text{g}. \tag{10.24}$$

It is about 1/1840 of the mass of the lightest of all atoms-the hydrogen atom.

The laws of electrolysis established experimentally by Michael Faraday in 1836 played a great part in discovering the discrete nature of electricity. According to these laws, the mass m of a substance liberated when a current passes through an electrolyte[1] is proportional to the charge q carried by the current:

$$m = \frac{1}{F}\frac{M}{z}q. \tag{10.25}$$

Here, M is the mass of one mole of the liberated substance, z the valence of the substance and F the **Faraday's constant** (**Faraday's number**) equal to

$$F = 96.5 \times 10^3 \text{ C mol}^{-1}. \tag{10.26}$$

Dividing both sides of Eq. (10.25) by the mass of an ion, we get

$$N = \frac{1}{F}\frac{N_A}{z}q,$$

where N_A is the Avogadro's constant and N the number of ions contained in the mass m.

Hence, for the charge of one ion, we have

$$e' = \frac{q}{N} = \frac{F}{N_A}z.$$

Consequently, the charge of an ion is an integral multiple of the quantity

$$e = \frac{F}{N_A}, \tag{10.27}$$

which is the elementary charge.

Thus, the discrete nature of the charges which ions in electrolytes can have follows from an analysis of the laws of electrolysis.

Substituting for F in Eq. (10.27) its value from Eq. (10.26) and for N_A its value found from J. Perrin's experiments (see Sec. 11.9 of Vol. I), we get a value fore that agrees quite well with that found by Millikan.

Since the accuracy with which Faraday's constant is determined and the accuracy of the value of e obtained by Millikan are greatly superior to the accuracy of Perrin's experiments for determining N_A, Eq. (10.27) was used to determine Avogadro's constant. Here, the value of F found from experiments in electrolysis and the value of e obtained by Millikan were used.

[1]Electrolytes are solutions of salts, alkalies or acids in water and some other liquids, and also molten salts that are ionic crystals in the solid state. Chemical transformations occur in electrolytes when a current passes through them. Such substances are called electrolytic conductors (conductors of the second kind) to distinguish them from electronic conductors (conductors of the first kind) in which the passage of a current is not attended by chemical transformations.

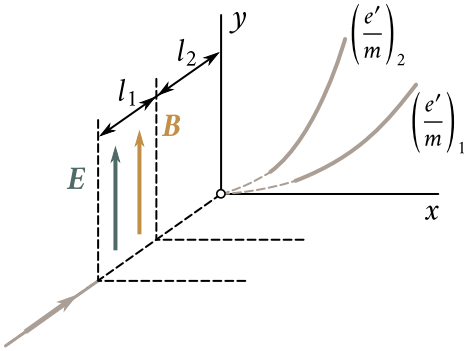


Fig. 10.9

10.4. Determination of the Specific Charge of Ions. Mass Spectrographs

The methods of determining the specific charge described in the preceding section are suitable when all the particles in a beam have the same velocity. All the electrons forming a beam are accelerated by the same potential difference applied between the cathode from which they fly out and the anode. Therefore, the scattering of the values of the velocities of the electrons in a beam is very small. If matters were different, an electron beam would produce a greatly blurred spot on the screen, and measurements would be impossible.

Ions are formed as a result of ionization of molecules of a gas that takes place in a volume having an appreciable length. Appearing in different places of this volume, the ions then pass through different potential differences, and, consequently, their velocities are different. Thus, the methods used to determine the specific charge of electrons cannot be applied to ions. In 1907, J. J. Thomson developed the "method of parabolas", which made it possible to circumvent the difficulty noted above.

In Thomson's experiment, a narrow beam of positive ions passed through a region in which it simultaneously experienced the action of parallel electric and magnetic fields (Fig. 10.9). Both fields were virtually homogeneous and made a right angle with the initial direction of the beam. They produced deflections of the ions: the magnetic field deflected them in the direction of the x-axis, the electric one along they y-axis. According to Eqs. (10.8) and (10.7), these deflections are

$$\begin{aligned} x &= \frac{e'}{m} B \frac{l_1}{v} \left(\frac{1}{2} l_1 + l_2 \right) \\ y &= \frac{e'}{m} E \frac{l_1}{v^2} \left(\frac{1}{2} l_1 + l_2 \right), \end{aligned} \tag{10.28}$$

where v is the velocity of a given ion with the specific charge e'/m, l_1 the length

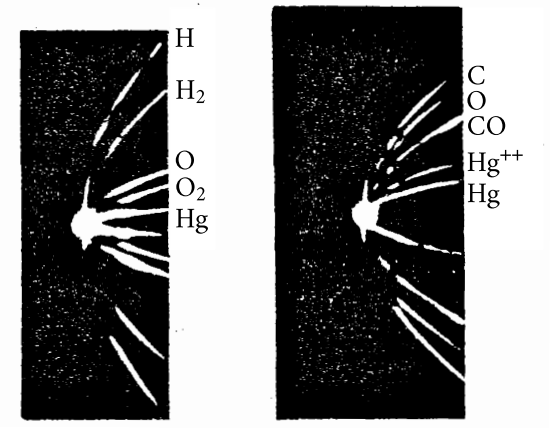


Fig. 10.10

of the region in which the field acts on the beam and l_2 is the distance from the boundary of this region to the photographic plate registering the ions impinging on it.

Equations (10.28) are the coordinates of the point at which an ion having the given values of e'/m and the velocity v impinges on the plate. Ions having the same specific charge, but different velocities, reached different points of the plate. Eliminating the velocity v from Eqs. (10.28), we get the equation of a curve along which the traces of ions having the same value of e'/m are arranged:

$$y = \frac{E}{B^2 l_1 (0.5 l_1 + l_2)} \frac{m}{e'} x^2. \tag{10.29}$$

Inspection of Eq. (10.29) shows that ions having identical values of e'/m and different values of v left a trace in the form of a parabola on the plate. Ions having different values of e'/m occupied different parabolas. Equation (10.29) can be used to find the specific charge of the ions corresponding to each parabola if the parameters of the instrument are known (*i.e.*, E, B, l_1, and l_2), and the displacements x and y are measured. When the direction of one of the fields was reversed, the relevant coordinate reversed its sign, and parabolas symmetrical to the initial ones were obtained. Dividing the distance between similar points of symmetrical parabolas in half made it possible to find x and y. The trace left on the plate by the beam with the fields switched off gave the origin of coordinates. Figure 10.10 shows the first parabolas obtained by Thomson.

When performing experiments with chemically pure neon, Thomson discovered that this gas produced two parabolas corresponding to relative atomic masses of 20 and 22. This result gave rise to the assumption that there are two chemically

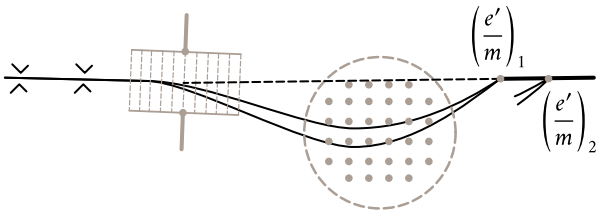


Fig. 10.11

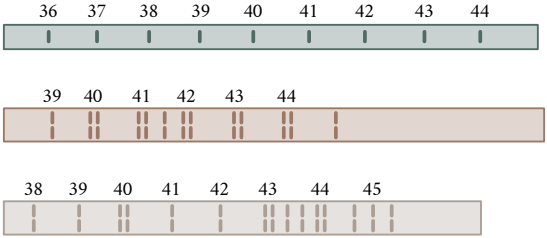


Fig. 10.12

indistinguishable varieties of the neon atoms (today we call them **isotopes** of neon). This assumption was proved by the British scientist Francis Aston (1877-1945), who improved the method of determining the specific charge of ions.

Aston's instrument, which he called a **mass spectrograph**, was designed as follows (Fig. 10.11). A beam of ions separated by a system of slits was consecutively passed through an electric field and a magnetic field. These fields were directed so that they caused the ions to travel to opposite sides. When they passed through the electric field, ions with a given value of e'/m were deflected more when their velocity was lower. Consequently, the ions left the electric field in the form of a diverging beam. The trajectories of the ions were also curved more in the magnetic field when their velocity was lower. Since the ions were deflected to opposite sides by the two fields, after leaving the magnetic field they formed a beam converging at one point.

Ions with other values of the specific charge were focussed at other points (the trajectories of the ions for only one value of e'/m are shown in Fig. 10.11). The relevant calculations show that points at which beams formed by ions having different values of e'/m converge are approximately on a single straight line (shown by a dash line in the figure). Putting a photographic plate along this line, Aston obtained a number of short lines on it, each of which corresponded to a definite value of e'/m. The similarity of the image obtained on the plate to a photograph of an optical line spectrum was the reason why Aston called it a mass spectrogram, and the instrument itself—a mass spectrograph. Figure 10.12 shows mass spectrograms obtained by Aston (the mass numbers of the relevant ions are indicated opposite

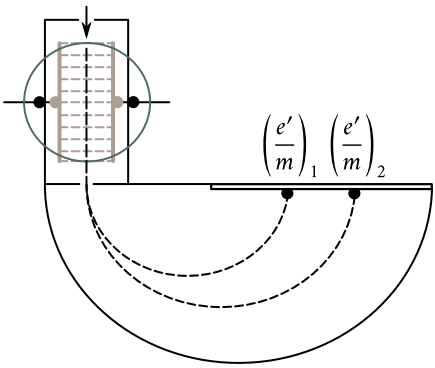


Fig. 10.13

the lines).

K. Bainbridge designed an instrument of a different kind. In the Bainbridge mass spectrograph (Fig. 10.13), a beam of ions first passes through the so-called velocity selector that separates ions having a definite velocity from the beam. In the selector, the ions experience the action of mutually perpendicular electric and magnetic fields that deflect the ions to opposite sides. Only those ions pass through the selector slit for which the actions of the electric and magnetic fields compensate each other. This occurs when $e'E = e'vB$. Hence, the velocities of the ions leaving the selector regardless of their mass and charge have identical values equal to $v = E/B$.

After leaving the selector, the ions get into the region of a homogeneous magnetic field of induction B' at right angles to their velocity. In this field, they move along circles whose radii depend on e'/m:

$$R = \frac{m}{e'}\frac{v}{B'}$$

[see Eq. (10.21)].

After completing a semi-circle, the ions strike a photographic plate at distances of $2R$ from the slit. Hence, the ions of each species (determined by the value of e'/m) leave a trace on the plate in the form of a narrow strip. The specific charges of the ions can be calculated if the parameters of the instrument are known. Since the charges of the ions are integral multiples of the elementary charge e, the masses of the ions can be calculated from the found values of e'/m.

Numerous kinds of mass spectrographs are in use at present. Instruments have also been designed in which the ions are registered by means of an electrical device instead of by a photographic plate. They are called **mass spectrometers**.

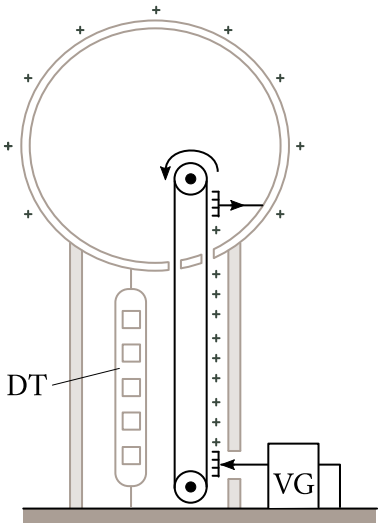


Fig. 10.14

10.5. Charged Particle Accelerators

Experiments using beams of high-energy charged particles play a great part in the physics of atomic nuclei and elementary particles. The devices used for obtaining such beams are called **charged particle accelerators**. There are many types of such devices. We shall acquaint ourselves with the operating principles of some of them.

The Van De Graaff Generator. In 1929, R. van de Graaff proposed an electrostatic generator based on the fact that surplus charges take up a position on the external surface of a conductor. A schematic view of the generator is shown in Fig. 10.14. A hollow metal sphere called a conductor is mounted on an insulating column. An endless moving belt of silk or rubberized fabric mounted on shafts is introduced into the sphere. A comb of sharp points is installed at the base of the column near the belt. The charge produced by a voltage generator (VG) for several scores of kilovolts flows onto the belt from the comb points. The conductor contains a second comb onto whose points the charge flows from the belt. This comb is connected Fig. 10.14 to the conductor so that the charge taken off the belt immediately passes over to its external surface. As charges accumulate on the conductor, its potential grows until the charge that leaks away becomes equal to the newly supplied charge. The leakage is mainly due to ionization of the gas near the surface of the conductor. The resulting passage of a current through the gas is called a corona discharge (see Sec. 12.8). The surface of the conductor is carefully polished

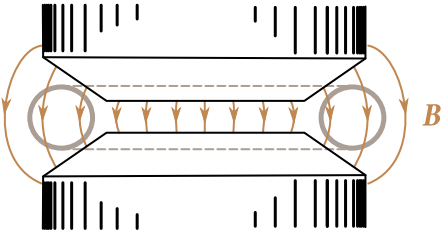


Fig. 10.15

to reduce the corona discharge.

The potential up to which the conductor can be discharged is limited by the circumstance that at a field strength of about 3 kV m^{-1} (30 kV cm^{-1}) a discharge appears in the air at atmospheric pressure. For a sphere, $E = \varphi/r$. Therefore, to obtain higher potential differences, the size of the conductor has to be increased (up to 10 m in diameter). The maximum potential difference that can be obtained in practice with the aid of a van de Graaff generator is about 10 MV (10^7 V).

Particles are accelerated in a discharge tube (DT) to whose electrodes the potential difference obtained in the generator is applied. A van de Graaff generator is sometimes designed in the form of two identical columns near each other whose conductors are charged oppositely. In this case, the discharge tube is connected between the conductors.

It must be noted that the generator belt, conductor, discharge tube, and the earth form a closed direct current circuit. Inside the tube, the charges move under the action of the electrostatic field. Charges are carried to the conductor from the earth by extraneous forces whose part is played by the mechanical forces bringing the generator belt into motion.

Betatron. This is the name given to an induction accelerator of electrons using a vortex electric field. It consists of a toroidal evacuated chamber (a doughnut) placed between the poles of an electromagnet of a special shape (Fig. 10.15). The winding of the magnet is supplied with alternating current having a frequency of about 100 Hz. The varying magnetic field produced performs two functions: first, it sets up a vortex electric field accelerating the electrons, and, second, it retains the electrons in an orbit coinciding with the axis of the doughnut.

To keep an electron in an orbit of constant radius, the magnetic induction of the field must be increased as its velocity grows [according to Eq. (10.2), the radius of the orbit is proportional to v/B]. Consequently, only the second and fourth quarters of the current period can be used for acceleration because at their beginning the current in the magnet winding is zero. A betatron thus operates in pulse conditions. At the beginning of the pulse, an electron gun feeds a beam of electrons into the

doughnut. The beam is caught up by the vortex electric field and begins to travel in a circular orbit with a constantly growing velocity. During the growth of the magnetic field (about 10^{-3} s), the electrons are able to complete up to a million revolutions and acquire an energy that may reach several hundred MeV. With such an energy, the speed of the electrons almost equals the speed of light c.

For an electron being accelerated to travel in a circular orbit of radius r_0, a simple relation, which we shall now proceed to derive, must be observed between the magnetic induction of the field in the orbit and inside it. The vortex electric field is directed along a tangent to the orbit along which the electron is travelling. Hence, the circulation of the vector $\boldsymbol{E}$ along this orbit is $2\pi r_0 E$. At the same time according to Eq. (9.19), the circulation of the vector $\boldsymbol{E}$ is $-(\mathrm{d}\Phi/\mathrm{d}t)$, where Φ is the magnetic flux through the surface enclosed by the orbit. The minus sign indicates the direction of $\boldsymbol{E}$. We shall be interested only in the magnitude of the field strength, therefore, we shall omit the minus sign. Equating the two expressions for the circulation, we find that

$$E = \frac{1}{2\pi r_0}\frac{\mathrm{d}\Phi}{\mathrm{d}t}.$$

The magnetic field is perpendicular to the plane of the orbit. We can, therefore, assume that $\Phi = \pi r_0^2 \langle B\rangle$, where $\langle B\rangle$ is the average value of the magnetic induction over the area of the orbit. Hence,

$$E = \frac{1}{2\pi r_0}\frac{\mathrm{d}}{\mathrm{d}t}\left(\pi r_0^2 \langle B\rangle\right) = \frac{r_0}{2}\frac{\mathrm{d}}{\mathrm{d}t}\langle B\rangle . \tag{10.30}$$

Let us write the relativistic equation of motion of an electron in orbit:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[\frac{m\boldsymbol{v}}{\sqrt{1-(v^2/c^2)}}\right] = e\boldsymbol{E} + e\boldsymbol{v}\times\boldsymbol{B}_{\mathrm{orb}} \tag{10.31}$$

($\boldsymbol{B}_{\mathrm{orb}}$ is the magnetic induction of the field in the orbit).

The velocity of an electron moving along a circle of radius r_0 can be written in the form $\boldsymbol{v} = \omega r_0 \hat{\boldsymbol{\tau}}$, where ω is the angular velocity of the electron, and $\hat{\boldsymbol{\tau}}$ is the unit vector of a tangent to the orbit. The vector $\boldsymbol{E}$ can be represented in the form

$$\boldsymbol{E} = E\hat{\boldsymbol{\tau}} = \frac{r_0}{2}\frac{\mathrm{d}}{\mathrm{d}t}\langle B\rangle\,\hat{\boldsymbol{\tau}}$$

[see Eq. (10.30)]. Finally, the product $\boldsymbol{v}\times\boldsymbol{B}$ can be written in the form $vB\hat{\boldsymbol{n}} = \omega r_0 B\hat{\boldsymbol{n}}$, where $\hat{\boldsymbol{n}}$ is a unit vector of a normal to the orbit. In view of what has been said above, let us write Eq. (10.31) as follows:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[\frac{\omega r_0\hat{\boldsymbol{\tau}}}{\sqrt{1-\left(\omega^2 r_0^2/c^2\right)}}\right] = \frac{er_0}{2}\frac{\mathrm{d}}{\mathrm{d}t}\langle B\rangle\,\hat{\boldsymbol{\tau}} + e\omega r_0 B_{\mathrm{orb}}\hat{\boldsymbol{n}}. \tag{10.32}$$

The time derivative of the unit vector $\hat{\boldsymbol{\tau}}$ is $\dot{\hat{\boldsymbol{\tau}}} = \omega\hat{\boldsymbol{n}}$ [see Eq. (1.56) of Vol. I; the angular velocity of rotation of the unit vector $\hat{\boldsymbol{\tau}}$ coincides with the angular velocity of an electron]. Consequently, performing differentiation in the left-hand side of Eq. (10.32), we arrive at the equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[\frac{\omega r_0}{\sqrt{1-\left(\omega^2 r_0^2/c^2\right)}}\right]\hat{\boldsymbol{\tau}} + \left[\frac{\omega r_0}{\sqrt{1-\left(\omega^2 r_0^2/c^2\right)}}\right]\omega\hat{\boldsymbol{n}} = \frac{er_0}{2}\frac{\mathrm{d}}{\mathrm{d}t}\langle B\rangle\,\hat{\boldsymbol{\tau}} + e\omega r_0 B_{\mathrm{orb}}\hat{\boldsymbol{n}}.$$

Equating the factors of similar unit vectors in the left-hand and righthand sides of the equation, we get

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[\frac{\omega r_0}{\sqrt{1-\left(\omega^2 r_0^2/c^2\right)}}\right] = \frac{er_0}{2}\frac{\mathrm{d}}{\mathrm{d}t}\langle B\rangle\,, \tag{10.33}$$

$$\frac{\omega r_0}{\sqrt{1-\left(\omega^2 r_0^2/c^2\right)}} = er_0 B_{\mathrm{orb}}. \tag{10.34}$$

It follows from Eq. (10.33) that

$$\frac{\omega r_0}{\sqrt{1-\left(\omega^2 r_0^2/c^2\right)}} = \frac{er_0}{2}\langle B\rangle \tag{10.35}$$

(ω and $\langle B\rangle$ at the beginning of a pulse equal zero).

A comparison of Eqs. (10.34) and (10.35) yields:

$$B_{\mathrm{orb}} = \frac{1}{2}\langle B\rangle\,.$$

Thus, for an electron to travel constantly in a circular orbit, the magnetic induction in the orbit must be half of the average value of the magnetic induction inside the orbit. This is achieved by making the pole shoes in the form of truncated cones (see Fig. 10.15).

At the end of an acceleration cycle, an additional magnetic field is switched on that deflects the accelerated electrons from their stationary orbit and directs them onto a special target inside the doughnut. Upon striking the target, the electrons emit hard electromagnetic radiation (gamma rays, X-rays).

Betatrons are mainly used in nuclear investigations. Small accelerators for an energy up to 50 MeV have found use in industry as sources of very hard X-rays employed for flaw detection in massive articles.

Cyclotron. The accelerator bearing this name is based on the period of revolution of a charged particle in a homogeneous magnetic field being independent of its velocity [see Eq. (10.3)]. This apparatus consists of two electrodes in the form

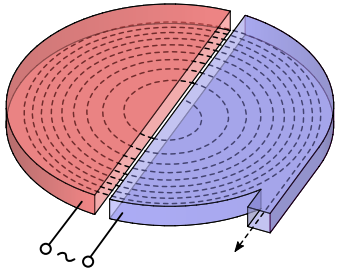


Fig. 10.16

of halves of a low round box (Fig. 10.16[2]) called dees. The latter are confined in an evacuated housing placed between the poles of a large electromagnet. The field produced by the magnet is homogeneous and perpendicular to the plane of the dees. The dees are supplied with an alternating voltage produced by a high-frequency generator.

Let us introduce a charged particle into the slit between the dees at the moment when the voltage reaches its maximum value. The particle will be caught up by the electric field and pulled into one of the dees. The space inside the dee is equipotential, therefore, the particle in it will be under the action of only a magnetic field. In this case, the particle travels along a circle whose radius is proportional to the velocity of the particle [see Eq. (10.2)]. Let us choose the frequency of the change in the voltage between the dees so that by the moment when the particle, after covering half of the circle, approaches the slit between the dees, the potential difference between them will change its sign and reach its amplitude value. The particle will now be accelerated again and fly into the second dee with an energy double that with which it travelled in the first dee. Having a greater velocity, the particle will travel in the second dee along a circle of a greater radius (R is proportional to v), but the time during which it covers half the circle remains the same as previously. Therefore, by the moment when the particle flies into the slit between the dees, the voltage between them will again change its sign and take on the amplitude value.

Thus, the particle travels along a curve close to a spiral, and each time it passes through the slit between the dees it receives an additional portion of energy equal to $e'U_{\mathrm{m}}$ (e' is the charge of the particle, and U_{m} is the amplitude of the voltage produced by the generator). Having a source of alternating voltage of a comparatively small value (U_{m} is about 10^5 V) at our disposal, we can use a cyclotron to accelerate protons up to energies of about 25 MeV. At higher energies, the dependence of the mass of the protons on the velocity begins to tell—the period of revolution increases [according to Eq. (10.3) it is proportional to m], and the synchronism between the

[2]This figure was taken from `https://commons.wikimedia.org/wiki/File:Zyclotron.svg`.

motion of the particles and the changes in the accelerating field is violated.

To prevent this violation of synchronism and to obtain particles having higher energies, either the frequency of the voltage fed to the dees or the magnetic field induction is made to vary. An apparatus in which in the course of accelerating each portion of particles the frequency of the accelerating voltage is diminished as required is called a **phasotron** (or a **synchrocyclotron**). An accelerator in which the frequency remains constant, while the magnetic field induction is changed so that the ratio m/B remains constant is called a **synchrotron** (equipment of this type is used only to accelerate electrons).

In the accelerator called a **synchrophasotron** or a proton synchrotron, both the frequency of the accelerating voltage and the magnetic field induction are changed. The particles being accelerated travel in this machine along a circular path instead of a spiral. An increase in the velocity and mass of the particles is attended by a growth in the magnetic field induction so that the radius determined by Eq. (10.2) remains constant. The period of revolution of the particles changes both owing to the growth in their mass and to the growth in B. For the accelerating voltage to be synchronous with the motion of the particles, the frequency of this voltage is made to change according to the relevant law. A synchrophasotron has no dees, and the particles are accelerated on separate sections of the path by the electric field produced by the varying frequency voltage generator.

The most powerful accelerator at present (in 1979)—a proton synchrotron—was started in 1974 at the Fermi National Accelerator Laboratory at Batavia, Illinois, in the USA. It accelerates protons up to an energy of 400 GeV (4×10^{11} eV). The speed of protons having such an energy differs from that of light in a vacuum by less than 0.0003% ($v = 0.9999972c$).

Chapter 11

THE CLASSICAL THEORY OF ELECTRICAL CONDUCTANCE OF METALS

11.1. The Nature of Current Carriers in Metals

A number of experiments were run to reveal the nature of the current carriers in metals. Let us first of all note the experiment conducted in 1901 by the German physicist Carl Riecke (1845-1915). He took three cylinders—two of copper and one of aluminium—with thoroughly polished ends. After being weighed, the cylinders were put end to end in the sequence copper-aluminium-copper. A current was passed in one direction through this composite conductor during a year. During this time, a total charge of 3.5×10^6 C passed through the cylinders. Weighing showed that the passage of a current had no effect on the weight of the cylinders. When the ends that had been in contact were studied under a microscope, no penetration of one metal into another was detected. The results of the experiment indicate that a charge is carried in metals not by atoms, but by particles encountered in all metals. The electrons discovered by I. J. Thomson in 1897 could be such particles.

To identify the current carriers in metals with electrons, it was necessary to determine the sign and numerical value of the specific charge of the carriers. Experiments run for this purpose were based on the following considerations. If metals contain charged particles capable of moving, then upon the deceleration (braking) of a metal conductor these particles should continue to move by inertia for a certain time, as a result of which a current pulse will appear in the conductor, and a certain charge will be carried in it.

Assume that a conductor initially moves with the velocity $\boldsymbol{v}_0$ (Fig. 11.1). We shall begin to decelerate it with the acceleration $\boldsymbol{a}$. Continuing to move by inertia, the

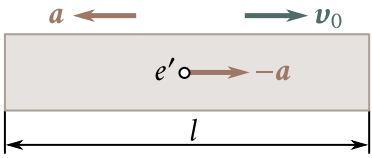


Fig. 11.1

current carriers will acquire the acceleration $-\boldsymbol{a}$ relative to the conductor. The same acceleration can be imparted to the carriers in a stationary conductor if an electric field of strength $\boldsymbol{E} = -m\boldsymbol{a}/e'$ is set up in it, *i.e.*, the potential difference,

$$\varphi_1 - \varphi_2 = \int_1^2 \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{l} = -\int_1^2 \frac{m\boldsymbol{a}}{e'} \cdot \mathrm{d}\boldsymbol{l} = -\frac{mal}{e'},$$

is applied to the ends of the conductor (m and e' are the mass and the charge of a current carrier, l is the length of a conductor). In this case, the current $I = (\varphi_1 - \varphi_2)/R$, where R is the resistance of the conductor, will flow through it (I is considered to be positive if the current flows in the direction of motion of the conductor). Hence, the following charge will pass through each cross section of the conductor during the time $\mathrm{d}t$:

$$\mathrm{d}q = I\,\mathrm{d}t = -\frac{mal}{e'R}\,\mathrm{d}t = -\frac{ml}{e'R}\,\mathrm{d}v.$$

The charge passing during the entire time of deceleration is

$$q = \int_1^2 \mathrm{d}q = -\int_{v_0}^0 \frac{ml}{e'R}\,\mathrm{d}v = \frac{m}{e'}\frac{lv_0}{R} \tag{11.1}$$

(the charge is positive if it is carried in the direction of motion of the conductor).

Thus, by measuring l, v_0, and R, and also the charge q flowing through the circuit when the conductor is decelerated, we can find the specific charge of the carriers. The direction of the current pulse will indicate the sign of the carriers.

The first experiment with conductors moving with acceleration was run in 1913 by the Soviet physicists Leonid Mandelshtam (1879-1944) and Nikolai Papaleksi (1880-1947). They made a wire coil perform rapid torsional oscillations about its axis. A telephone was connected to the ends of the coil, and the sound due to the current pulses was heard in it.

A quantitative result was obtained by the American physicists R. Tolman and T. Stewart in 1916. A coil of a wire 500 m long was made to rotate with a linear velocity of the turns of 300 m s^{-1}. The coil was then sharply braked, and a ballistic galvanometer was used to measure the charge flowing in the circuit during the braking time. The value of the specific charge of the carriers calculated by Eq. (11.1) was obtained very close to e/m for electrons. It was, thus, proved experimentally that electrons are the current carriers in metals.

A current can be produced in metals by an extremely small potential difference. This gives us the grounds to consider that the current carriers—electrons—move without virtually any hindrance in a metal. The result of Tolman's and Stewart's experiment lead to the same conclusion.

The existence of free electrons in metals can be explained by the fact that when a crystal lattice is formed, the most weakly bound (valence) electrons detach themselves from the atoms of the metal. They become the "collective" property of the entire piece of metal. If one electron becomes detached from every atom, then the concentration of the free electrons (*i.e.*, their number n in a unit volume) will equal the number of atoms in a unit volume. The latter number is $(\delta/M)N_A$, where δ is the density of the metal, M is the mass of a mole, N_A is Avogadro's constant. The values of δ/M for metals range from 2×10^4 mol m^{-3} (for potassium) to 2×10^5 mol m^{-3} (for beryllium). Hence, we get values of the following order for the concentration of the free electrons (or conduction electrons, as they are also called):

$$n = 10^{28}\ \text{m}^{-3} \text{ to } 10^{29}\ \text{m}^{-3}\ \left(10^{22}\ \text{cm}^{-3} \text{ to } 10^{23}\ \text{cm}^{-3}\right). \tag{11.2}$$

11.2. The Elementary Classical Theory of Metals

Proceeding from the notions of free electrons, the German physicist Paul Drude (1863-1906) created the classical theory of metals that was later improved by H. Lorentz. Drude assumed that the conduction electrons in a metal behave like the molecules of an ideal gas. In the intervals between collisions, they move absolutely freely, covering on an average a certain path l. True, unlike the molecules of a gas whose free path is determined by collisions of the molecules with one another, the electrons collide chiefly not with one another, but with the ions forming the crystal lattice of the metal. These collisions result in the establishment of thermal equilibrium between the electron gas and the crystal lattice.

Assuming that the results of the kinetic theory of gases may be extended to an electron gas, we can use the following formula to assess the average velocity of thermal motion of the electrons:

$$\langle v \rangle = \left(\frac{8kT}{\pi m}\right)^{1/2} \tag{11.3}$$

[see Eq. (11.65) of Vol. I]. Calculations by this equation for room temperature (about 300 K) give the following result:

$$\langle v \rangle = \left(\frac{8 \times 1.38 \times 10^{-23} \times 300}{3.14 \times 0.91 \times 10^{-30}}\right)^{1/2} \approx 10^5\ \text{m s}^{-1}.$$

When a field is switched on, the ordered motion of the electrons with a certain

average velocity $\langle u \rangle$ is superposed onto the chaotic thermal motion occurring with the velocity $\langle v \rangle$. It is simple to assess the value of $\langle u \rangle$ by the equation

$$j = ne \langle u \rangle \tag{11.4}$$

[see Eq. (5.23)]. The maximum current density for copper wires allowed by the relevant specifications is about 10^7 A m^{-2} (10 A mm^{-2}). Taking the value of 10^{29} m^{-3} for n, we get

$$\langle u \rangle = \frac{j}{en} \approx \frac{10^7}{1.6 \times 10^{-19} \times 10^{29}} \approx 10^{-3} \text{ m s}^{-1}.$$

Thus, even at very high current densities, the average velocity of ordered motion of the charges $\langle u \rangle$ is about $1/10^8$ of the average velocity of thermal motion $\langle v \rangle$. Therefore, in calculations, the magnitude of the resultant velocity $|\boldsymbol{v} + \boldsymbol{u}|$ may be replaced with that of the velocity of thermal motion $|\boldsymbol{v}|$.

Let us find the change in the mean value of the kinetic energy of the electrons produced by a field. The mean square of the resultant velocity is

$$\left\langle (\boldsymbol{v} + \boldsymbol{u})^2 \right\rangle = \left\langle v^2 + 2\boldsymbol{v} \cdot \boldsymbol{u} + u^2 \right\rangle = \left\langle \boldsymbol{v}^2 \right\rangle + 2 \langle \boldsymbol{v} \cdot \boldsymbol{u} \rangle + \left\langle \boldsymbol{u}^2 \right\rangle. \tag{11.5}$$

The two events consisting in that the velocity of thermal motion of the electrons will take on the value $\boldsymbol{v}$, while the velocity of ordered motion—the value $\boldsymbol{u}$, are statistically independent. Therefore, according to the theorem on the multiplication of probabilities [see Eq. (11.4) of Vol. I], we have $\langle \boldsymbol{v} \cdot \boldsymbol{u} \rangle = \langle \boldsymbol{v} \rangle \cdot \langle \boldsymbol{u} \rangle$. But $\langle \boldsymbol{v} \rangle$ equals zero, so that the second addend in Eq. (11.5) vanishes, and it acquires the form

$$\left\langle (\boldsymbol{v} + \boldsymbol{u})^2 \right\rangle = \left\langle \boldsymbol{v}^2 \right\rangle + \left\langle \boldsymbol{u}^2 \right\rangle.$$

Hence, it follows that the ordered motion increases the kinetic energy of the electrons on an average by

$$\langle \Delta \varepsilon_k \rangle = \frac{m \left\langle u^2 \right\rangle}{2}. \tag{11.6}$$

Ohm's Law. Drude considered that when an electron collides with an ion of the crystal lattice, the additional energy (11.6) acquired by the electron is transmitted to the ion and, consequently, the velocity u as a result of the collision vanishes. Let us assume that the field accelerating the electrons is homogeneous. Hence, under the action of the field, the electron receives a constant acceleration equal to eE/m, and toward the end of its path the velocity of ordered motion will reach, on an average, the value

$$u_{\max} = \frac{eE}{m} \tau, \tag{11.7}$$

where τ is the average time elapsing between two consecutive collisions of the electron with ions of the lattice.

Drude did not take into consideration the distribution of the electrons by

velocities and ascribed the same value of the velocity v to all the electrons. In this approximation

$$\tau = \frac{l}{v}$$

(we remind our reader that $|\boldsymbol{v} + \boldsymbol{u}|$ virtually equals $|\boldsymbol{v}|$). Using this value of τ in Eq. (11.7), we get

$$u_{\max} = \frac{eEl}{mv}. \tag{11.8}$$

The velocity u changes linearly during the time it takes to cover the path l. Therefore, its average value over the path equals half the maximum value:

$$\langle u \rangle = \frac{1}{2} u_{\max} = \frac{eEl}{2mv}.$$

Introducing this equation into Eq. (11.4), we get

$$j = \frac{ne^2 l}{2mv} E.$$

The current density is found to be proportional to the field strength. We have, thus, arrived at Ohm's law. According to Eq. (5.22), the constant of proportionality between j and E is the conductivity

$$\sigma = \frac{ne^2 l}{2mv}. \tag{11.9}$$

If the electrons did not collide with the ions of the lattice, their free path and, consequently, the conductivity of the metal would be infinitely great. Thus, *according to the classical notions, the electrical resistance of metals is due to the collisions of their free electrons with the ions at the crystal lattice points of the metal.*

The Joule-Lenz Law. By the end of its free path, an electron acquires additional kinetic energy whose average value is

$$\langle \Delta\varepsilon_{\mathrm{k}} \rangle = \frac{mu_{\max}^2}{2} = \frac{e^2 l^2}{2mv} E^2 \tag{11.10}$$

[see Eqs. (11.6) and (11.8)]. Upon colliding with an ion, the electron, according to the assumption, completely transfers the additional energy it has acquired to the crystal lattice. The energy given up to the lattice goes to increase the internal energy of the metal, which manifests itself in its becoming heated.

Every electron experiences on an average $1/\tau = v/l$ collisions a second, communicating each time the energy expressed by Eq. (11.10) to the lattice. Hence, the following amount of heat should be liberated in unit volume per unit time:

$$Q_{\mathrm{u}} = n \frac{1}{\tau} \langle \Delta\varepsilon_{\mathrm{k}} \rangle = \frac{ne^2 l}{2mv} E^2$$

(n is the number of conduction electrons per unit volume).

The quantity Q_u is the unit thermal power of a current (see Sec. 5.8). The factor of E^2 coincides with the value given by Eq. (11.9) for σ. Passing over in the expression σE^2 from σ and E to ρ and j, we arrive at the formula $Q_u = \rho j^2$ expressing the Joule-Lenz law [see Eq. (5.39)].

The Wiedemann-Franz Law. It is known from experiments that in addition to their high electrical conductivity, metals are distinguished by a high thermal conductivity. The German physicists G. Wiedemann and R. Franz discovered an empirical law according to which the ratio of the thermal conductivity $\varkappa$ to the electrical conductivity σ is about the same for all metals and changes in proportion to the absolute temperature. For example, for aluminium at room temperature, this ratio is $5.8 \times 10^{-6}\,\mathrm{J\,\Omega\,s^{-1}\,K^{-1}}$, for copper it is $6.4 \times 10^{-6}\,\mathrm{J\,\Omega\,S^{-1}\,K^{-1}}$, and for lead it is $7.0 \times 10^{-6}\,\mathrm{J\,\Omega\,s^{-1}\,K^{-1}}$.

Non-metallic crystals are also capable of conducting heat. The thermal conductivity of metals, however, considerably exceeds that of dielectrics. It thus follows, that the free electrons instead of the crystal lattice are responsible for the transfer of heat in metals. Considering these electrons as a monatomic gas, we can adopt an expression from the kinetic theory of gases for the thermal conductivity:

$$\varkappa = \frac{1}{3} nmvlc_V \tag{11.11}$$

[see Eq. (16.26) of Vol. I; the density ρ has been replaced with the product nm, and $\langle v \rangle$ with v]. The specific heat capacity of a monatomic gas is $c_V = 3R/(2M) = 3k/(2m)$. Using this value in Eq. (11.11), we obtain

$$\varkappa = \frac{1}{2} nkvl.$$

Dividing $\varkappa$ by Eq. (11.9) for σ and then substituting $3k/(2T)$ for $mv^2/2$, we arrive at the expression

$$\frac{\varkappa}{\sigma} = \frac{kmv^2}{e^2} = 3\left(\frac{k}{e}\right)^2 T. \tag{11.12}$$

that expresses the Wiedemann-Franz law.

Introduction of the numerical values of k and e into Eq. (11.12) yields

$$\frac{\varkappa}{\sigma} = 2.23 \times 10^{-8}\, T.$$

When $T = 300\,\mathrm{K}$, we get the value $3.7 \times 10^{-6}\,\mathrm{J\,\Omega\,s^{-1}\,K^{-1}}$ for $\varkappa/\sigma$, which agrees quite well with experimental data (see the values of $\varkappa/\sigma$ given above for aluminium, copper, and lead). It was later established, however, that such a good coincidence is accidental, because when H. Lorentz performed the calculations more accurately, taking into account the distribution of the electrons by velocities, the value of $2(k/e)^2 T$ was obtained for the ratio $\varkappa/\sigma$, and it does not agree so well with the data

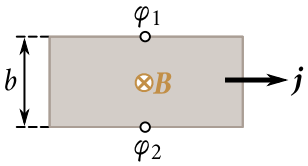


Fig. 11.2

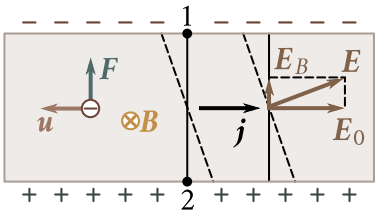


Fig. 11.3

of experiments.

Thus, the classical theory was able to explain Ohm's and the Joule-Lenz laws, and also gave a qualitative explanation of the Wiedemann-Franz law. At the same time, this theory encountered quite appreciable difficulties. They include two basic ones. It can be seen from Eq. (11.9) that the resistance of metals (*i.e.*, the quantity that is the reciprocal of σ) must increase as the square root of T. Indeed, we have no grounds to assume that the quantities n and l depend on the temperature. The velocity of thermal motion, on the other hand, is proportional to the square root of T. This theoretical conclusion contradicts experimental data according to which the electrical resistance of metals grows in proportion to the first power of T, *i.e.*, more rapidly than $T^{1/2}$ [see expression (5.24)].

The second difficulty of the classical theory is that an electron gas must have a molar heat capacity equal to $(3/2)R$. Adding this quantity to the heat capacity of the lattice, which is $3R$ [see Eq. (13.1) of Vol. I], we get the value of $(9/2)R$ for the molar heat capacity of a metal. Thus, in accordance with the classical electron theory, the molar heat capacity of metals ought to be 1.5 times higher than that of dielectrics. Actually, however, the heat capacity of metals does not differ appreciably from that of non-metallic crystals. Only the quantum theory of metals was able to explain this discrepancy.

11.3. The Hall Effect

If a metal plate through which a steady electric current is flowing is placed in a magnetic field perpendicular to it, then a potential difference of $U_{\mathrm{H}} = \varphi_1 - \varphi_2$ (Fig. 11.2) is set up between the plate faces parallel to the directions of the current and field. This phenomenon was discovered by the American physicist E. Hall in 1879 and is called the **Hall effect** or the **galvanomagnetic effect**.

The Hall potential difference is determined by the expression

$$U_{\mathrm{H}} = R_{\mathrm{H}} b j B. \tag{11.13}$$

Here, b is the width of the plate, I the current density, B the magnetic induction of

the field and $abRH$ is a constant of proportionality known as the **Hall coefficient**.

The Hall effect is easily explained by the electron theory. In the absence of a magnetic field, the current in the plate is due to the electric field $\boldsymbol{E}_0$ (Fig. 11.3). The equipotential surfaces of this field form a system of planes perpendicular to the vector $\boldsymbol{E}_0$. Two of them are shown in the figure by solid straight lines. The potential at all the points of each surface and, consequently, at points 1 and 2 too is the same. The current carriers—electrons—have a negative charge, therefore, the velocity of their ordered motion $\boldsymbol{u}$ is directed oppositely to the current density vector $\boldsymbol{j}$.

When the magnetic field is switched on, each carrier experiences the magnetic force $\boldsymbol{F}$ directed along side b of the plate and having a magnitude of

$$F = euB. \tag{11.14}$$

As a result, the electrons acquire a velocity component directed toward the upper (in the figure) face of the plate. A surplus of negative charges is formed at this face and, accordingly, a surplus of positive charges at the lower face. Consequently, an additional transverse electric field $\boldsymbol{E}_B$ is produced. When the strength of this field reaches a value such that its action on the charges balances the force given by Eq. (11.14), a stationary distribution of the charges in a transverse direction will set in. The corresponding value of E_B is determined by the condition $eE_B = euB$. Hence,

$$E_B = uB.$$

The field $\boldsymbol{E}_B$ adds to the field $\boldsymbol{E}_0$ to form the resultant field $\boldsymbol{E}$. The equipotential surfaces are perpendicular to the field strength vector. Consequently, they will turn and occupy the position shown by the dash line in Fig. 11.3. Points 1 and 2 which were formerly on the same equipotential surface now have different potentials. To find the voltage appearing between these points, the distance b between them must be multiplied by the strength E_B:

$$U_{\mathrm{H}} = bE_B = buB.$$

Let us express u through j, n, and e in accordance with the equation $j = neu$. The result is

$$U_{\mathrm{H}} = \frac{1}{ne} bjB. \tag{11.15}$$

Equations (11.13) and (11.15) coincide if we assume that

$$R_{\mathrm{H}} = \frac{1}{ne}. \tag{11.16}$$

Inspection of Eq. (11.16) shows that by measuring the Hall coefficient, we can find the concentration of the current carriers in a given metal (*i.e.*, the number of carriers per unit volume).

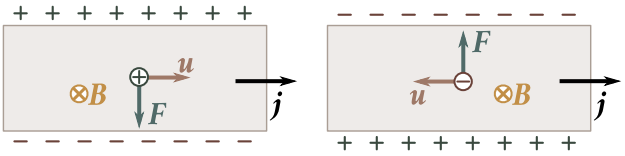


Fig. 11.4

An important characteristic of a substance is the mobility of the current carriers in it. By the mobility of the current carriers is meant the average velocity acquired by the carriers at unit electric field strength. If the carriers acquire the velocity u in a field of strength E, then their mobility u_0 is

$$u_0 = \frac{u}{E}. \tag{11.17}$$

The mobility can be related to the conductivity σ and to the carrier concentration n. For this purpose, let us divide the equation $j = neu$ by the field strength E. Taking into account that $j/E = \sigma$ and $u/E = u_0$, we get

$$\sigma = neu_0. \tag{11.18}$$

Having measured the Hall coefficient R_H and the conductivity σ, we can use Eqs. (11.16) and (11.18) to find the concentration and mobility of the current carriers in the relevant specimen.

The Hall effect is observed not only in metals, but also in semiconductors. The sign of the effect can be used to see whether a semiconductor belongs to the n- or p-type[1]. Figure 11.4 compares the Hall effect for specimens with positive and negative carriers. The direction of the magnetic force is reversed both when the direction of motion of the charge changes and when its sign is reversed. Hence, when the current and field have the same direction, the magnetic force exerted on positive and negative carriers has the same direction. Therefore, with positive carriers, the potential of the upper (in the figure) face is higher than that of the lower one, and with negative carriers the potential is lower. We can thus establish the sign of the current carriers after determining that of the Hall potential difference.

It is of interest to note that in some metals the sign of U_H corresponds to positive current carriers. This anomaly is explained by the quantum theory.

[1]In n-type semiconductors, the current carriers are negative, and in p-type ones they are positive (see Vol. III).

Chapter 12
ELECTRIC CURRENT IN GASES

12.1. Semi-Self-Sustained and Self-Sustained Conduction

The passage of an electric current through gases is called a **gas discharge**. Gases in their normal state are insulators, and current carriers are absent in them. Only when special conditions are created in gases can current carriers appear in them (ions, electrons) and an electric discharge be produced.

Current carriers may appear in gases as a result of external action not associated with the presence of an electric field. In this case, the gas is said to have **semi-self-sustained conduction**. Semi-self-sustained discharge may be due to heating of a gas (thermal ionization), the action of ultraviolet rays or X-rays, and also to the action of radiation of radioactive substances.

If the current carriers appear as a result of processes due to an electric field being produced in a gas, the conduction is called **self-sustained**. The nature of a gas discharge depends on many factors: on the chemical nature of the gas and electrodes, on the temperature and pressure of the gas, on the shape, dimensions, and mutual arrangement of the electrodes, on the voltage applied to them, on the density and power of the current, etc. This is why a gas discharge may have very diverse forms. Some kinds of discharge are attended by a glow and sound effects—hissing, rustling, or crackling.

12.2. Semi-Self-Sustained Gas Discharge

Assume that a gas between electrodes (Fig. 12.1) continuously experiences a constant in intensity action of an ionizing agent (for example, X-rays). The action of the ionizer results in one or more electrons being detached from some of the gas molecules. The latter, thus, become positively charged ions. At not very low pressures, the

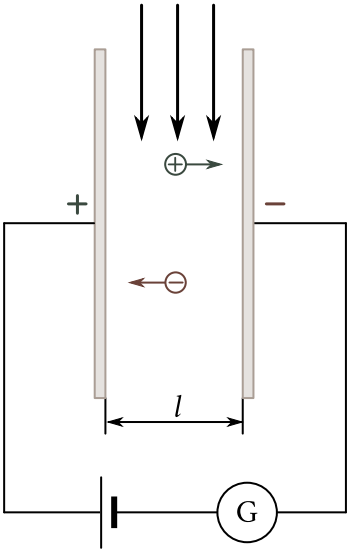


Fig. 12.1

detached electrons are usually captured by neutral molecules, which, thus, become negatively charged ions. Let Δn_{i} stand for the number of pairs of ions appearing under the action of the ionizer in unit volume per second.

The process of ionization in a gas is attended by recombination of the ions, *i.e.*, neutralization of unlike ions when they meet or the formation of a neutral molecule by a positive ion and an electron.

The probability of two ions of opposite signs meeting each other is proportional to the number of both positive and negative ions. Hence, the number of pairs of ions Δn_{r} recombining in unit volume per second is proportional to the square of the number of pairs of ions n per unit volume:

$$\Delta n_{\mathrm{r}} = rn^2 \tag{12.1}$$

(r is a constant of proportionality).

In a state of equilibrium, the number of appearing ions equals the number of recombining ones, hence,

$$\Delta n_{\mathrm{i}} = rn^2. \tag{12.2}$$

We, thus, get the following expression for the equilibrium concentration of ions (the number of pairs of ions in unit volume)·:

$$n = \left(\frac{\Delta n_{\mathrm{i}}}{r}\right)^{1/2}. \tag{12.3}$$

Several pairs of ions appear per second in 1 cm of atmospheric air under the ac-

tion of cosmic radiation and traces of radioactive substances in the Earth's crust. The constant r for air is 1.6×10^{-6} cm^3 s^{-1}. Introduction of these values into Eq. (12.3) gives a value of about 10^3 cm^3 for the equilibrium concentration of ions in the air. This concentration is not adequate for the conduction to be noticeable. Pure dry air is a very good insulator.

If we feed a voltage to electrodes, the ions will decrease in number not only because of recombination, but also because of the ions being drawn off by the field to the electrodes. Assume that Δn_j pairs of ions are drawn off from unit volume every second. If the charge of each ion is e', then the neutralization of one pair of ions on the electrodes is attended by the transfer of the charge e' along the circuit. Every second, $\Delta n_j Sl$ pairs of ions reach the electrodes (here, S is the area of the electrodes, l is the distance between them; the product Sl equals the volume of the space between the electrodes). Consequently, the current in the circuit is

$$I = e' \Delta n_j Sl,$$

whence

$$\Delta n_j = \frac{I}{e' lS} = \frac{j}{e' l}, \tag{12.4}$$

where j is the current density.

When a current is present, the condition of equilibrium is as follows:

$$\Delta n_i = \Delta n_r + \Delta n_j$$

Substituting for Δn_r and Δn_j their values from Eqs. (12.1) and (12.4), we arrive at the equation

$$\Delta n_i = rn^2 + \frac{j}{e' l}. \tag{12.5}$$

The current density is determined by the expression

$$j = e' n \left(u_0^+ + u_0^-\right) E, \tag{12.6}$$

where u_0^+ and u_0^- are the mobilities of the positive and negative ions, respectively [see Eq. (11.17)].

Let us consider two extreme cases—weak and strong fields.

With weak fields, the current density will be very small, and the addend $j/(e'l)$ in Eq. (12.5) may be disregarded in comparison with rn^2 (this signifies that the ions leave the space between the electrodes mainly as a result of recombination). Equation (12.5) thus transforms into Eq. (12.2), and we get Eq. (12.3) for the equilibrium concentration of the ions. Using this value of n in Eq. (12.6), we get

$$j = e' \left(\frac{\Delta n_i}{r}\right)^{1/2} \left(u_0^+ + u_0^-\right) E. \tag{12.7}$$

The multiplier of E in Eq. (12.7) does not depend on the field strength. Hence, with

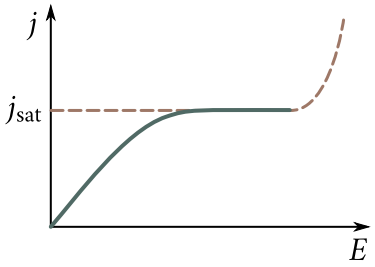


Fig. 12.2

weak fields, a semi-self-sustained gas discharge obeys Ohm's law.

The mobility of ions in gases has a value of the order of $10^{-4}\ m^2\ V^{-1}\ s^{-1}$ [or $1\ cm^2\ V^{-1}\ s^{-1}$]. Hence, at the equilibrium $n = 10^3\ cm^{-3} = 10^9\ m^{-3}$, and the field strength $E = 1\ V\ m^{-1}$, the current density will be

$$j = 1.6 \times 10^{-19} \times 10^9 \left(10^{-4} + 10^{-4}\right) \times 1 \sim 10^{-14}\ A\ m^{-2} = 10^{-18}\ A\ cm^{-2}$$

[see Eq. (12.6); the ions are assumed to be singly charged].

With strong fields, we may disregard the addend rn^2 in Eq. (12.5) in comparison with $j/e'l$. This signifies that virtually all the appearing ions will reach the electrodes without having time to recombine. In these conditions, Eq. (12.5) becomes

$$\Delta n_i = \frac{j}{e'l},$$

whence

$$j = e' \Delta n_i l. \tag{12.8}$$

This current density is produced by all the ions originated by the ionizer in a column of the gas with unit cross-sectional area between the electrodes. Consequently, this current density is the greatest at the given intensity of the ionizer and the given distance l between the electrodes. It is called the saturation current density j_{sat}.

Let us calculate j_{sat} for the following conditions: $\Delta n_i = 10^{-3}\ cm^{-3}$ (this is approximately the rate of ion formation in the atmospheric air in ordinary conditions), $l = 0.1$ m. The introduction of these data into Eq. (12.8) yields

$$j_{sat} = 1.6 \times 10^{-19} \times 10^7 \times 10^1 \sim 10^{-13}\ A\ m^{-2} = 10^{-17}\ A\ cm^{-2}.$$

These calculations show that the conduction of air in ordinary conditions is negligibly small.

At intermediate values of E, there is a smooth transition from a linear dependence of j on E to saturation; when the latter is reached, j stops depending on E (see the solid curve in Fig. 12.2). The region of saturation is followed by a region of a sharp growth in the current (see the portion of the curve depicted by the dash line). The explanation of this growth is that beginning from a certain value of E, the

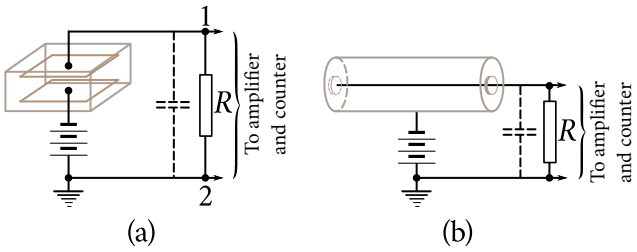


Fig. 12.3

electrons[1] given birth to by the external ionizer manage to acquire a considerable energy while on their free path. This energy is sufficient to ionize the molecules they collide with. The free electrons produced in this ionization, after gaining speed, cause ionization in their turn. Thus, an avalanche-like reproduction of the primary ions produced by the external ionizer occurs, and the discharge current is amplified. The process does not lose its nature of a semi-self- sustained discharge, however, because after the action of the external ionizer stops, the discharge continues only until all the electrons (primary and secondary) reach the anode (the rear boundary of the space containing ionizing particles—electrons—moves toward the anode). For a discharge to become self-sustained, two meeting avalanches of ions are needed. This is possible only if ionization by a collision is capable of giving birth to carriers of both signs.

It is very important that the semi-self-sustained discharge currents amplified as a result of reproduction of the carriers are proportional to the number of primary ions produced by the external ionizer. This property of a discharge is used in proportional counters (see the following section).

12.3. Ionization Chambers and Counters

Ionization chambers and counters are employed for detecting and counting elementary particles, and also for measuring the intensity of X-rays and gamma rays. The functioning of these instruments is based on the use of a semi-self-sustained gas discharge.

The schematic diagram of an ionization chamber and a counter is the same (Fig. 12.3). They differ only in their operating conditions and structural features. A counter (Fig. 12.3b) consists of a cylindrical body along whose axis a thin wire (anode) fastened on insulators is stretched. The body of the counter is the cathode.

[1] Owing to the greater length of their free path, electrons acquire the ability to produce ionization by a collision earlier than gas ions do.

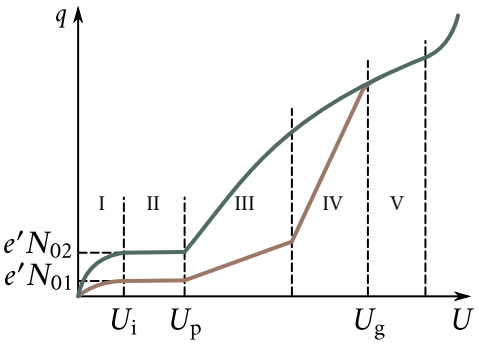


Fig. 12.4

A window of mica or aluminium foil is made in the end of the counter to admit the ionizing particles. Some particles, and also X-rays and gamma rays penetrate into a counter or an ionization chamber directly through their walls. An ionization chamber (Fig. 12.3b) can have electrodes of various shapes. In particular, they may be the same as in a counter, have the shape of plane parallel plates, etc.

Assume that a high-speed charged particle producing N_0 pairs of primary ions (electrons and positive ions) flies into the space between the electrodes. The ions produced are carried along by the field toward the electrodes, and as a result a certain charge q, which we shall call a current pulse, passes through resistor R. Figure 12.4 shows how the current pulse q depends on the voltage U between the electrodes for two different amounts of primary ions N_0 differing by three times ($N_{02} = 3N_{01}$). Six regions can be earmarked on the graph. Regions I and II were considered in the preceding section. In particular, region II is the region of the saturation current—all the ions produced by an ionizing particle reach the electrodes without having time to recombine. It is quite natural that the current pulse does not depend on the voltage in these conditions.

Beginning from the value U_p, the field strength becomes sufficient for the electrons to be able to ionize the molecules by a collision. Therefore, the number of electrons and positive ions grows like an avalanche. As a result, AN_0 ions reach each of the electrodes. The quantity A is called the **gas amplification factor**. In region III, this factor does not depend on the number of primary ions (but does depend on the voltage). Therefore, if we keep the voltage constant, the current pulse will be proportional to the number of primary ions. Region III is called the **proportional region**, and the voltage U_p the **threshold of the proportional region**. The gas amplification factor changes in this region from 1 at its beginning to 10^3-10^4 at its end (the scale along the q-axis has not been observed in Fig. 12.4; only the ratio of 1:3 between the ordinates in regions II and III has been observed).

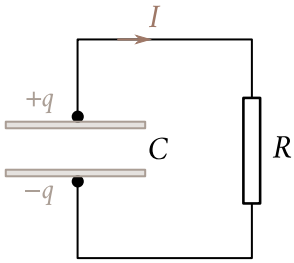


Fig. 12.5

In region IV, called the **region of partial proportionality**, the gas amplification factor A depends to a greater and greater extent on N_0. In this connection, the difference between the current pulses produced by different numbers of primary ions becomes smoothed out more and more.

At voltages corresponding to region V (it is known as the **Geiger region**, and the voltage U_g as the threshold of this region), the process acquires the nature of a self-sustained discharge. The primary ions only produce an impetus for its appearance. The current pulse in this region is absolutely independent of the number of primary ions.

In region VI, the voltage is so high that a discharge, after once being set up, does not stop. It is, therefore, called the **region of continuous discharge**.

Ionization Chambers. An ionization chamber is an instrument operating without gas amplification, *i.e.*, at voltages corresponding to region II. There are two kinds of ionization chambers. Chambers of one kind are used for registering the pulses initiated by individual particles (pulse chambers). A particle flying into the chamber produces a certain number of ions in it, and as a result the current I begins to flow through resistor R. The result is that the potential of point 1 (see Fig. 12.3a) rises and becomes equal to IR (the initial potential of this point was the same as that of earthed point 2). This potential is fed to an amplifier, and after being amplified operates a counting device. After all the charges that have reached the inner electrode pass through resistor R, the current stops and the potential of point 1 again becomes equal to zero. The nature of operation of the chamber depends on the duration of the current pulse set up by one ionizing particle.

To determine what the duration of a pulse depends on, let us consider a circuit consisting of capacitor C and resistor R (Fig. 12.5). If we impart the opposite charges $+q$ and $-q$ to the capacitor plates, a current will flow through resistor R, and the charges on the plates will diminish. The instantaneous value of the voltage applied across the resistor is $U = q/C$. Hence, we get the following expression for the

current:

$$I = \frac{U}{R} = \frac{q}{RC}. \tag{12.9}$$

Let us substitute $-\mathrm{d}q/\mathrm{d}t$ for the current, where $-\mathrm{d}q$ is the decrement of the charge on the plates during the time $\mathrm{d}t$. As a result, we get the differential equation

$$-\frac{\mathrm{d}q}{\mathrm{d}t} = \frac{q}{RC} \quad \text{or} \quad \frac{\mathrm{d}q}{q} = -\frac{q}{RC}\,\mathrm{d}t.$$

According to Eq. (12.9), $\mathrm{d}q/q = \mathrm{d}I/I$. We can, therefore, write

$$\frac{\mathrm{d}I}{I} = -\frac{1}{RC}\,\mathrm{d}t.$$

Integration of this equation yields

$$\ln I = -\frac{1}{RC}\,t + \ln I_0$$

($\ln I_0$ is the integration constant). Finally, raising the expression obtained to a power, we arrive at the equation

$$I = I_0 \exp\left(-\frac{t}{RC}\right). \tag{12.10}$$

It is easy to see that I_0 is the initial value of the current.

It follows from Eq. (12.10) that during the time

$$\tau = RC, \tag{12.11}$$

the current diminishes to $1/e$ of its original value. Accordingly, the quantity τ is called the **time constant** of a circuit. The greater this quantity, the slower is the rate of diminishing of the current in a circuit.

The diagram of an ionization chamber (see Fig. 12.3a) is similar to that shown in Fig. 12.5. The part of C is played by the interelectrode capacitance shown by a dash line on the diagram of the chamber. An increase in the resistance of R is attended by a growth in the voltage across points 1 and 2 at a given current, and this, consequently, facilitates the registration of the pulses. This circumstance induces designers to use the highest possible resistance of R. At the same time, for the chamber to be able to register separately the current pulses set up by particles rapidly following one another, the time constant must not be great. Therefore, designers have to make a compromise when choosing the resistance of R for pulse chambers. It is usually taken of the order of $10^8\ \Omega$. Hence, at $C \sim 10^{-11}$ F, the time constant is 10^{-3} s.

Another kind of ionization chamber is the so-called integrating chamber. The resistance of R in them is of the order of $10^{15}\ \Omega$. At $C \sim 10^{-11}$ F, the time constant is 10^4 s. In this case, the current pulses produced by separate ionizing particles merge and a steady current flows through the resistor. Its magnitude characterizes the

total charge of the ions produced in the chamber in unit time. Thus, the ionization chambers of these two kinds differ only in the value of the time constant RC.

Proportional Counters. The pulses set up by separate particles can be amplified quite considerably (up to 10^3-10^4 times) if the voltage between the electrodes is in region III (see Fig. 12.4). An instrument operating in such conditions is called a **proportional counter**. The anode of the counter is made in the form of a wire of several hundredths of a millimetre in diameter. The field strength near the wire is especially high. With a sufficiently great voltage between the electrodes, the electrons produced near the wire acquire an energy under the action of the field that is adequate for producing ionization of the molecules by a collision. The result is reproduction of the ions. The dimensions of the space in which reproduction occurs increase with the voltage. The gas amplification factor grows accordingly.

The number of primary ions depends on the nature and energy of the particles producing the pulse. Therefore, the magnitude of the pulses at the output of a proportional counter makes it possible to distinguish various particles, and also to sort particles of the same nature by their energies.

Geiger-Müller Counters. A still greater amplification of the pulse (up to 10^8) can be attained by making a counter function in the Geiger region (region V in Fig. 12.4). A counter operating in these conditions is called a **Geiger-Müller counter** (or more briefly a **Geiger counter**). A discharge in the Geiger region, being "launched" by an ionizing particle, subsequently transforms into a self-sustained one. Hence, the magnitude of the pulse does not depend on the initial ionization. To obtain separate pulses from individual particles, the discharge produced must be rapidly interrupted (quenched). This is achieved either with the aid of an external resistance R (in non-self-quenching counters), or at the expense of processes appearing in the counter itself. In the latter case, the counter is called self quenching.

The quenching of a discharge with the aid of an external resistance is due to the fact that when a discharge current flows in the resistance, a great voltage drop is set up in it. Consequently, only part of the applied voltage falls to the lot of the interelectrode space, and it is insufficient for maintaining the discharge.

Stopping of a discharge in self-quenching counters is due to the following reasons. Electrons have a mobility that is about 1000 times greater than the mobility of positive ions. Therefore, during the time it takes the electrons to reach the wire, the positive ions do not virtually move from their places. These ions produce a positive space charge that weakens the field near the wire, and the discharge stops. Quenching of the discharge in this case is prevented by additional processes which we shall not consider. To suppress them, an admixture of a polyatomic organic gas (for example, alcohol vapour) is added to the gas filling the counter (usually argon). Such a counter separates pulses from particles following one another with

an interval of the order of 10^4 s.

12.4. Processes Leading to the Appearance of Current Carriers in a Self-Sustained Discharge

Before commencing to describe the various kinds of self-sustained gas discharge, we shall consider the basic processes leading to the production of current carriers (electrons and ions) in such discharges.

Collisions of Electrons with Molecules. The collisions of electrons (and also ions) with molecules can have an elastic or inelastic nature. The energy of a molecule (like that of an atom) is quantized. This signifies that it can have only discrete (*i.e.*, separated by finite intervals) values called energy levels. The state with the smallest energy is called the **ground** one. To transfer a molecule from its ground state to various excited ones, definite values of the energy W_1, W_2, etc., are needed. A molecule can be ionized by imparting to it a sufficiently great energy W_1.

Upon transition to an excited state, a molecule usually stays in it only $\sim 10^{-8}$ s, after which it passes back to its ground state, emitting its surplus energy in the form of a quantum of light—a **photon**. Molecules can spend a considerably greater time (about 10^{-3} s) in certain excited states called **metastable**.

The laws of energy and momentum conservation must be obeyed when particles collide. Therefore, definite limitations are imposed on the transfer of energy in a collision—not all the energy which a colliding particle has can be transferred to another particle.

If in a collision, an energy sufficient for exciting a molecule cannot be imparted to it, the total kinetic energy of the particles remains unchanged, and the collision will be **elastic**. Let us find the energy imparted to the particle that is struck in an elastic collision. Assume that a particle of mass m_1 having the velocity v_{10} collides with a stationary ($v_{20} = 0$) particle of mass m_2. The following conditions must be observed in a central collision:

$$\frac{m_1 v_{10}^2}{2} = \frac{m_1 v_1^2}{2} + \frac{m_2 v_2^2}{2}$$

$$m_1 v_{10} = m_1 v_1 + m_2 v_2,$$

where v_1 and v_2 are the velocities of the particles after the collision. The velocity of the second particle from these equations will be

$$v_2 = \frac{2m_1}{m_1 + m_2} v_{10}$$

(see Sec. 3.11 of Vol. I).

The energy transmitted to the second particle in an elastic collision is deter-

mined by the expression

$$\Delta W_{\text{el}} = \frac{m_2 v_2^2}{2} = \frac{m_1 v_{10}^2}{2} \frac{4 m_1 m_2}{(m_1 + m_2)^2}.$$

If $m_1 \ll m_2$, this equation is simplified as follows:

$$\Delta W_{\text{el}} = \frac{m_1 v_{10}^2}{2} \frac{4 m_1}{m_2} = W_{10} \frac{4 m_1}{m_2} \tag{12.12}$$

where W_{10} is the initial energy of the incident particle.

It can be seen from Eq. (12.12) that a light particle (electron) in an elastic collision with a heavy particle (molecule) gives up to it only a small fraction of its stock of energy. The light particle "rebounds" from the heavy one like a ball from a wall, and its velocity remains virtually unchanged in magnitude. The relevant calculations show that in a non-central collision the fraction of the energy transferred is still smaller.

With a sufficiently high energy of the incident particle (electron or ion), a molecule may be excited or ionized. In this case, the total kinetic energy of the particles is not conserved—part of the energy goes for excitation or ionization, *i.e.*, for increasing the internal energy of the colliding particles or for splitting one of the particles into two fragments.

Collisions attended by the excitation of particles are called in **elastic collisions of the first kind**. A molecule in an excited state upon colliding with another particle (electron, ion, or neutral molecule) can pass over to the ground state without emitting its surplus energy, but transferring it to this particle. As a result, the total kinetic energy of the particles after the collision will be greater than before it. Such collisions are known as **inelastic collisions of the second kind**. Molecules pass over from a metastable state to the ground one as a result of collisions of the second kind.

In an inelastic collision of the first kind, the equations of energy and momentum conservation have the form

$$\frac{m_1 v_{10}^2}{2} = \frac{m_1 v_1^2}{2} + \frac{m_2 v_2^2}{2} + \Delta W_{\text{int}}, \tag{12.13}$$

$$m_1 v_{10} = m_1 v_1 + m_2 v_2, \tag{12.14}$$

where ΔW_{int} is the increment of the internal energy of a molecule corresponding to its transition to an excited state. Deleting v_1 from these equations, we get

$$\Delta W_{\text{int}} = m_2 v_{10} v_2 - \left(\frac{m_1 + m_2}{m_1} \right) \frac{m_2 v_2^2}{2}. \tag{12.15}$$

At a given velocity of the striking particle (v_{10}), the increment of the internal energy ΔW_{int} depends on the velocity v_2 with which the molecule travels after

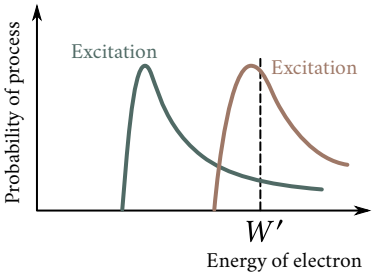


Fig. 12.6

the collision. Let us find the greatest possible value of ΔW_{int}. To do this, we shall differentiate function (12.15) with respect to v_2 and equate the derivative to zero:

$$\frac{\mathrm{d}\,(\Delta W_{\text{int}})}{\mathrm{d}v_2} = m_2 v_{10} - \left(\frac{m_1 + m_2}{m_1}\right) m_2 v_2 = 0.$$

Hence, $_2 = m_1 v_{10}/(m_1 + m_2)$. Substitution of this value for v_2 in Eq. (12.15) yields

$$\Delta W_{\text{int,max}} = \left(\frac{m_2}{m_1 + m_2}\right) \frac{m_1 v_1^2}{2}. \tag{12.16}$$

If the incident particle is considerably lighter than the struck one ($m_1 \ll m_2$), the factor $m_2/(m_1 + m_2)$ in Eq. (12.16) is close to unity. Thus, when a light particle (electron) strikes a heavy one (molecule), almost all the energy of the incident particle can be used to excite or ionize the molecule[2].

Even if the energy of the incident particle (electron) is sufficiently great, however, a collision does not necessarily result in the excitation or ionization of a molecule. These processes have definite probabilities depending on the energy (and, therefore, on the velocity) of the electron. Figure 12.6 shows the approximate path followed by these probabilities. The higher the velocity of the electron, the smaller is the duration of its interaction with the molecule near which it flies. Hence, both probabilities rapidly reach a maximum, and then diminish with an increase in the energy of the electron. Inspection of the figure shows that an electron having, for example, the energy W' will cause ionization of a molecule with greater probability than its excitation.

Photoionization. Electromagnetic radiation consists of elementary particles called **photons**. The energy of a photon is $\hbar\omega$, where $\hbar$ is Planck's constant divided by 2π [see Eq. (7.43)], and ω is the cyclic frequency of the radiation. A photon can be absorbed by a molecule, and its energy goes to excite or ionize the molecule.

[2]When ionization occurs, Eqs. (12.13) become more complicated because there will be three particles instead of two after a collision. The conclusion on the possibility of spending almost all of the electron's energy for ionization is correct, however.

In this case, the ionization of the molecule is called **photoionization**. Ultraviolet radiation is capable of producing direct photoionization. The energy of a photon of visible light is insufficient to detach an electron from a molecule. Hence, visible radiation is not capable of producing direct photoionization. It may be the cause, however, of so-called **stepped photoionization**. This process is carried out in two steps. In the first one, a photon transfers the molecule to an excited state. In the second step, the excited molecule is ionized as a result of its colliding with another molecule.

Short-wave radiation may appear in a gas discharge that is capable of producing direct photoionization. A sufficiently fast electron may not only ionize a molecule when it collides with it, but also transfer the ion formed into an excited state. The transition of an ion to the ground state is attended by the emission of radiation having a higher frequency than that of a neutral molecule. The energy of a photon of such radiation is sufficient for direct photoionization.

Emission of Electrons by the Surface of Electrodes. Electrons may be supplied to a gas-discharge space as a result of their emission by the surface of the electrodes. Such kinds of emission as thermionic (thermoelectron), secondary electron, and autoelectronic emission play the main part in some kinds of discharge.

Thermionic emission is the name given to the emission of electrons by heated solid or liquid bodies. Owing to the free electrons in a metal having a variety of velocities in accordance with a distribution law, there is always a certain number of them whose energy is sufficient for them to overcome the potential barrier and leave the metal. The number of such electrons at room temperature is negligibly small. With elevation of the temperature, however, the number of electrons capable of leaving the metal grows very rapidly and becomes quite noticeable at a temperature of the order of 1000 K.

By **secondary electron emission** is meant the emission of electrons by the surface of a solid or a liquid body when it is bombarded with electrons or ions. The ratio of the number of emitted (secondary) electrons to the number of particles producing the emission is called the secondary electron emission coefficient. When electrons are used to bombard the surface of a metal, the values of this coefficient vary from 0.5 (for beryllium) to 1.8 (for platinum).

Autoelectronic (or **cold**) **emission** is the emission of electrons by the surface of a metal occurring when an electric field of a very high strength ($\sim 10^8$ V m^{-1}) is set up near the surface. This phenomenon is also sometimes called field induced electron emission.

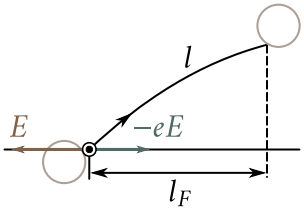


Fig. 12.7

12.5. Gas-Discharge Plasma

Some kinds of self-sustained discharge are characterized by a very high degree of ionization. A highly ionized gas, provided that the total charge of the electrons and ions in each elementary volume equals (or almost equals) zero, is called a **plasma**. A plasma is a special state of a substance. The matter in the interior of the Sun and other stars having a temperature of scores of millions of kelvins is in this state. A plasma produced owing to the high temperature of a substance is called **high-temperature** (or **isothermal**). A **gas-discharge** plasma, as its name implies, is one produced in a gas discharge.

For a plasma to be in a stationary state, processes are needed that replenish the stock of ions diminishing as a result of recombination. In high-temperature plasma, this is achieved as a result of thermal ionization, in gas-discharge plasma, as a result of collision ionization by electrons accelerated by an electric field. The ionosphere (one of the layers of the atmosphere) is a special variety of plasma. The high degree of ionization of the molecules ($\sim 1\%$) is maintained in the ionosphere by photoionization due to the Sun's shortwave radiation.

The electrons in a gas-discharge plasma participate in two motions—chaotic with a certain average velocity $\langle v \rangle$ and ordered motion in a direction opposite to $\boldsymbol{E}$ with the average velocity $\langle u \rangle$ much smaller than $\langle v \rangle$.

We shall prove that an electric field not only leads to ordered motion of the electrons of a plasma, but also increases the velocity $\langle v \rangle$ of their chaotic motion. Assume that at the moment when the field is switched on the gas contains a certain number of electrons whose average velocity corresponds to the gas temperature $T_g (m \langle v \rangle^2 / 2 = 3kT_g/2)$. In the interval between two successive collisions with molecules, an electron covers on an average the path l (Fig. 12.7; the trajectory of the electron is curved slightly under the action of the force $-eE$). The work done by the field on the electron is

$$A = eEl_F, \tag{12.17}$$

where l_F is the projection of the electron's path onto the direction of the force

exerted on it. Owing to collisions with molecules, the direction of motion of the electron constantly changes chaotically. The magnitude and sign of l, change accordingly. This is why the work given by Eq. (12.17) for separate portions of the path varies in magnitude and changes in its sign. On some sections, the field increases the energy of the electron, on others diminishes it. If ordered motion of the electrons were absent, the average value of l, and, consequently, the work given by Eq. (12.17) would be zero. The presence of ordered motion, however, leads to the average value of the work A differing from zero; it is positive and equals

$$\langle A \rangle = eE \langle u \rangle \tau = eE \langle u \rangle \frac{1}{\langle v \rangle}, \tag{12.18}$$

where τ is the average time needed by the electrons to cover their free path ($\langle u \rangle \ll \langle u \rangle$).

Thus, a field on an average increases the energy of the electrons. True, an electron upon colliding with a molecule gives up part of its energy to it. But, as we have seen in the preceding section, the fraction δ of the energy transferred in an elastic collision is very small—it averages[3] $\delta = 2(m/M)$ (here m is the mass of an electron, and M that of a molecule).

In a rarefied gas (in which l is greater) and with a sufficiently great field strength E, the work $\langle A \rangle$ [Eq. (12.18)] may exceed the energy $m \left\langle v^2 \right\rangle \langle \delta \rangle / 2$ transferred on an average to a molecule in each collision. The result will be a growth in the energy of chaotic motion of the electrons. It ultimately reaches values sufficient to excite or ionize a molecule. Beginning from this moment, part of the collisions stop being elastic and are attended by a large loss of energy. Therefore, the average fraction $\langle \delta \rangle$ of energy transferred increases.

Thus, the electrons acquire the energy needed for ionization not during one interval between collisions, but gradually in the course of a number of them. Ionization leads to the appearance of a large number of electrons and positive ions—a plasma is produced.

The energy of the electrons of a plasma is determined by the condition that the average value of the work done by the field on an electron during one interval between collisions equals the average value of the energy given up by the electron upon colliding with a molecule:

$$eE \langle u \rangle \frac{1}{\langle v \rangle} = \frac{m \left\langle v^2 \right\rangle}{2} \langle \delta \rangle .$$

Here, δ is an intricate function of $\langle v \rangle$.

Experiments show that the Maxwell distribution by velocities holds for the

[3] According to Eq. (12.12), in a central collision $\delta = 4(m/M)$. When the electron and the molecule only slightly touch each other, we have $\delta \approx 0$.

electrons in a gas-discharge plasma. Owing to the weak interaction of the electrons with the molecules (in an elastic collision δ is very small, while the relative number of inelastic collisions is negligible), the average velocity of chaotic motion of the electrons is many times greater than the velocity corresponding to the temperature T_g of the gas. If we introduce the temperature of the electrons T_g determining it from the equation $m\left\langle v^2\right\rangle = 3kT_e/2$, then we get a value of the order of several tens of thousands of kelvins for T_e. The failure of the temperatures T_g and T_e to coincide indicates that there is no thermodynamic equilibrium between the electrons and molecules in a gas-discharge plasma[4]. The concentration of the current carriers in a plasma is very high. Therefore, a plasma is an excellent conductor. The mobility of the electrons is about three orders of magnitude greater than that of the ions. Hence, the current in a plasma is mainly set up by its electrons.

12.6. Glow Discharge

A glow discharge appears at low pressures. It can be observed in a glass tube about 0.5 m long with flat metal electrodes soldered into its ends (Fig. 12.8). A voltage of $\sim$ 1000 V is supplied to the electrodes. There is virtually no current in the tube at atmospheric pressure. If the pressure is lowered, then approximately at 50 mmHg a discharge appears in the form of a glowing sinuous thin cord connecting the anode and the cathode. Lowering of the pressure is attended by thickening of the cord, and at about 5 mmHg the cord fills the entire cross section of the tube—a glow discharge sets in. Its principal parts are shown in Fig. 12.8. Near the cathode is a thin luminous layer called the **cathode luminous film**. Between the cathode and the luminous film is the **Aston dark space**. At the other side of the luminous film is a weakly luminous layer which by contrast appears to be dark and is accordingly known as the **cathode** (or **Crookes**) **dark space**. This layer bounds on a luminous region called the **negative glow**. All the above layers form the cathode part of the glow discharge.

The negative glow is followed by the **Faraday dark space**. The boundary between them is blurred. The remaining part of the tube is filled with a luminous gas; it is called the **positive column**. At a lower pressure, the cathode part of the discharge and the Faraday dark space become wider, while the positive column becomes shorter. At a pressure of the order of 1 mmHg, the positive column breaks up into a number of alternating dark and light bent layers-**strata**.

Measurements made with the aid of probes (thin wires soldered in at different

[4]The average energy of the molecules, electrons, and ions in a high-temperature plasma is the same. This explains its other name—isothermal plasma.

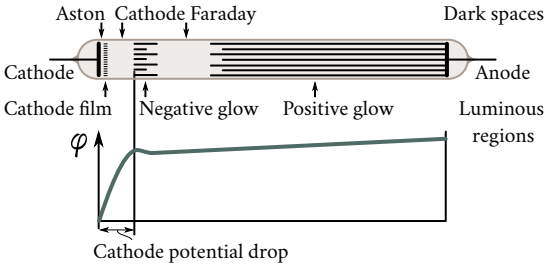


Fig. 12.8

points along the tube) and by other means have shown that the potential changes non-uniformly along a tube (see the graph in Fig. 12.8). Virtually the entire potential drop falls to the share of the first three parts of the discharge up to the cathode dark space inclusively. This portion of the voltage applied to a tube is called the **cathode potential drop**. The potential remains unchanged in the region of the negative glow—here the field strength is zero. Finally, the potential gradually grows in the Faraday dark space and in the positive column. Such a distribution of the potential is due to the formation in the cathode dark space of a positive space charge because of the increased concentration of the positive ions.

The main processes needed to maintain a glow discharge occur in its cathode part. The other parts of the discharge are not significant, they may even be absent (with a small spacing of the electrodes or at a low pressure). There are two main processes—secondary electron emission from the cathode produced by its bombardment with positive ions, and collision ionization of the gas molecules by electrons.

The positive ions accelerated by the cathode potential drop bombard the cathode and knock electrons out of it. These electrons are accelerated by the electric field in the Aston dark space. Acquiring sufficient energy, they begin to excite the gas molecules, owing to which the cathode luminous film appears. The electrons that fly without any collisions into the region of the cathode dark space have a high energy, and as a result they ionize the molecules more frequently than they excite them (see the graphs in Fig. 12.6). Thus, the intensity of glowing of the gas diminishes, but in return many electrons and positive ions appear. The ions produced first have a very low velocity. As a result, a positive space charge is formed in the cathode dark space. This leads to redistribution of the potential along the tube and to the appearance of the cathode potential drop.

The electrons appearing in the cathode dark space penetrate into the negative glow region that is characterized by a high concentration of electrons and positive ions and by a total space charge close to zero (a plasma). Therefore, the field strength

here is very low. Owing to the high concentration of electrons and ions, an intensive recombination process goes on in the negative glow region. It is attended by the emission of the energy liberated during this process. Thus, the negative glow is mainly a glow of recombination.

The electrons and ions penetrate from the negative glow region into the Faraday dark space because of diffusion (there is no field on the boundary between these regions, but in return there is a high gradient of electron and ion concentration). The lower concentration of the charged particles greatly diminishes the probability of recombination in the Faraday dark space. This is why the latter space seems to be dark.

A field is already present in the Faraday dark space. The electrons carried away by this field gradually accumulate energy so that the conditions needed for the existence of a plasma finally appear. The positive column is a gas-discharge plasma. It plays the part of a conductor joining the anode to the cathode parts of the discharge. The glow of the positive column is mainly due to transitions of excited molecules to their ground state. Molecules of different gases emit radiation of different wavelengths in such transitions. Therefore, the glow of the positive column has a characteristic colour for each gas. This circumstance is taken advantage of in glow tubes for manufacturing luminous inscriptions and advertisements. These inscriptions are the positive column of a glow discharge. Neon gas-discharge tubes produce a red glow, argon ones a bluish-green glow, etc.

If the electrode spacing is gradually diminished, the cathode part of the discharge remains unchanged whereas the length of the positive column diminishes until this column disappears completely. Next, the Faraday dark space disappears, and the length of the negative glow begins to decrease, the position of the boundary of this glow with the cathode dark space remaining unchanged. When the distance from the anode to this boundary becomes very small, the discharge stops.

If the pressure is gradually lowered, the cathode part of the discharge extends over a greater and greater part of the interelectrode space, and finally the cathode dark space extends over almost the entire tube. The glow of the gas in this case stops being noticeable but in return the tube walls begin to glow with a greenish colour. The majority of the electrons knocked out of the cathode and accelerated by the cathode potential drop reach the tube walls without colliding with molecules of the gas and cause the walls to glow upon striking them. For historical reasons, the stream of electrons emitted by the cathode of a gas-discharge tube at very low pressures was called **cathode rays**. The glow produced by bombardment with fast electrons is called **cathodoluminescence**.

If a narrow canal is made in the cathode of a gas-discharge tube, part of the positive ions penetrate into the space beyond the cathode and form a sharply

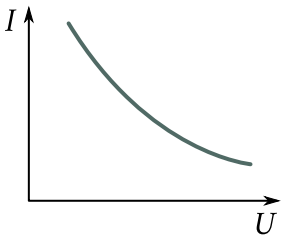


Fig. 12.9

bounded beam of ions called canal (or positive) rays. Beams of positive ions were first obtained in exactly this way.

12.7. Arc Discharge

In 1802, the Russian physicist Vasili Petrov (1761-1834) discovered that when contacting carbon electrodes connected to a large galvanic battery are moved apart, a concentrated light flares up between the electrodes. When the electrodes are horizontal, the heated luminescent gas bends in the shape of an arc. This is why the phenomenon discovered by Petrov was called an **electric arc**. The current in the arc may reach enormous values (from 10^3 A to 10^4 A) at a voltage of several scores of volts.

An arc discharge can proceed at both a low (of the order of several millimetres of mercury) and a high (up to 1000 atmospheres) pressure. The main processes maintaining the discharge are thermionic emission from the heated cathode surface and thermal ionization of the molecules due to the high temperature of the gas in the space between the electrodes. Almost the entire interelectrode space is filled with a high-temperature plasma. It is the conductor through which the electrons emitted by the cathode reach the anode. The temperature of the plasma is about 6000 K. In a superhigh-pressure arc, the temperature of the plasma may reach 10000 K (we remind our reader that the temperature of the Sun's surface is 5800 K). Owing to bombardment by positive ions, the cathode is heated to about 3500 K. The anode, bombarded by a powerful stream of electrons, is heated still more. As a result, the anode intensively evaporates, and a depression—a crater—is formed on its surface. The crater is the brightest place in an arc.

An arc discharge has a dropping volt-ampere characteristic (Fig. 12.9). The explanation is that a current increase is attended by a growth in the thermionic emission from the cathode and in the degree of ionization of the gas-discharge space. As a result, the resistance of this space diminishes at a greater rate than that of the current increase.

Apart from the thermionic arc described above (*i.e.*, a discharge due to thermionic emission from the heated surface of the cathode) an **arc with a cold cathode** is also encountered. Usually liquid mercury poured into a cylinder from which the air has been evacuated is the cathode of such an arc. The discharge occurs in the mercury vapour. The electrons fly out of the cathode as a result of autoelectronic emission. The strong field at the cathode surface needed for this to occur is set up by the positive space charge formed by the ions. The electrons are emitted not by the entire surface of the cathode, but by a small luminous and continuously moving cathode spot. The temperature of the gas in this case is not high. The molecules in the plasma are ionized, as in a glow discharge, as a result of collisions with the electrons.

12.8. Spark and Corona Discharges

A spark discharge is produced when the electric field strength reaches the breakdown value E_{br} for the given gas. The value of E_{br} depends on the gas pressure; it is about 3 MV m^{-1} (30 kV cm^{-1}) for air. The value of E_{br} varies with the pressure. According to the experimentally established **Paschen law**, the ratio of the breakdown field strength to the pressure is approximately constant:

$$\frac{E_{\text{br}}}{p} \approx \text{constant}.$$

A spark discharge is attended by the formation of a brightly luminous tortuous branched canal along which a short-time strong current pulse flows. An example is lightning; its length may be up to 10 km, the diameter of the canal up to 40 cm, the current may reach 100000 and more amperes, and the duration of the pulse is about 10^{-4} s. Every stroke of lightning consists of several (up to 50) pulses flowing along the same canal; their total duration (together with the intervals between the pulses) may reach several seconds. The temperature of the gas in the spark canal is up to 10000 K. The rapid strong heating of the gas leads to a sharp growth in the pressure and the production of shock and sound waves. This is why a spark discharge is attended by sound phenomena—from a weak crackling for a low-power spark to peals of thunder accompanying a stroke of lightning.

The appearance of a spark is preceded by the formation in the gas of a greatly ionized canal known as a streamer. The latter is obtained by overlapping of the separate electron avalanches appearing along the path of the spark. The forefather of each avalanche is an electron released by photoionization. How a streamer develops is shown in Fig. 12.10. Assume that the field strength has a value such that an electron flying out of the cathode as a result of some process or other acquires an

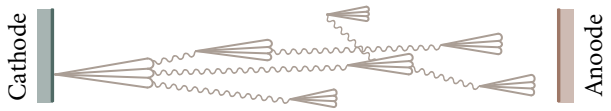


Fig. 12.10

energy sufficient for ionization along its free path. This causes multiplication of the electrons to occur—an avalanche is formed (the positive ions appearing during this process do not play a noticeable part owing to their much smaller mobility; they only set up the space charge resulting in redistribution of the potential). The short-wave radiation emitted by an atom that lost one of its inner electrons when ionized (this radiation is shown by wavy lines in the figure) produces photoionization of the molecules, the detached electrons giving birth to more and more new avalanches. After overlapping of the avalanches, a well-conducting canal—a streamer—is formed along which a powerful stream of electrons flows from the cathode to the anode—breakdown occurs.

If the electrodes have a shape at which the field in the space between them is approximately homogeneous (for example, they are spheres of a sufficiently great diameter), then breakdown occurs at a quite definite voltage U_{br} whose value depends on the distance between the spheres l ($U_{\text{br}} = E_{\text{br}}l$). This underlies the design of a spark voltmeter used to measure high voltages (from 10^3 V to 10^5 V). During such measurements, the maximum distance $l_{\max}$ is determined at which a spark appears. Next multiplying E_{br} by $l_{\max}$, we get the value of the voltage being measured.

If one of the electrodes (or both) has a very great curvature (for example, the electrode is a thin wire or a sharp point), then when the voltage is not too high, a so-called **corona discharge** is produced. When the voltage grows, this discharge transforms into a spark or an arc discharge.

In a corona discharge, the ionization and excitation of the molecules occur not in the entire interelectrode space, but only near an electrode having a small radius of curvature, where the field strength reaches values equal to or greater than E_{br}. The gas glows in this part of the discharge. The glow has the form of a corona surrounding the electrode, and this explains the name given to this kind of discharge. A corona discharge from a point has the form of a luminous brush, and for this reason it is sometimes known as a **brush discharge**. Positive and negative coronas are distinguished depending on the sign of the corona electrode. The **external corona** region is between the corona layer and the non-corona electrode. Breakdown conditions ($E \gg E_{\text{br}}$) exist only within the limits of the corona layer. We can, therefore, say that a corona discharge is incomplete breakdown of the gas

space.

With a negative corona, the phenomena at the cathode are similar to those at the cathode of a glow discharge. The positive ions accelerated by the field knock electrons out of the cathode. These electrons produce ionization and excitation of the molecules in the corona layer. In the external region of the corona, the field is not sufficient to impart the energy needed for ionization or excitation of the molecules to the electrons. For this reason, the electrons that penetrate into this region drift toward the anode under the action of the field. Part of the electrons are captured by the molecules, the result being the formation of negative ions. Thus, the current in the external region is due only to negative carriers-electrons and negative ions. The discharge in this region is of a semi-self-sustained nature.

In a positive corona, the electron avalanches are conceived at the outer boundary of the corona and fly toward the corona electrodethe anode. The appearance of electrons giving birth to avalanches is due to photoionization produced by the radiation of the corona layer. The current carriers in the external region of the corona are the positive ions that drift to the cathode under the action of the field.

If both electrodes have a great curvature (two corona electrodes), processes occur near each of them that are characteristic of a corona electrode of the given sign. Both corona layers are separated by an external region in which opposite streams of positive and negative current carriers travel. Such a corona is called a bipolar one.

The self-sustained gas discharge mentioned in Sec. 12.5 when treating counters is a corona discharge.

The thickness of the corona layer and the discharge current grow with an increasing voltage. At a low voltage, the size of the corona is small, and its glow is hard to notice. Such a microscopic corona is produced near a sharp point off which an electric wind flows (see Sec. 3.1).

The bluish electrical glow caused by corona discharge on masts and other high parts of a ship at sea before and after electrical storms was called St. Elmo's fire in olden days.

In high-voltage facilities, for example, in high-tension transmission lines, a corona discharge leads to the harmful leakage of current. Measures therefore have to be taken to prevent it. For this purpose, for instance, the wires of high-tension lines are taken of a sufficiently large diameter, which is the greater, the higher is the voltage of the line.

The corona discharge has found a useful application in engineering in electrical filters. The gas being purified flows through a tube along whose axis a negative corona electrode is arranged. The negative ions present in a great number in the external region of the corona settle on the particles or droplets polluting the gas and

are carried along with them to the external non-corona electrode. Upon reaching the latter, the particles become neutralized and settle on it. Later, blows are struck at the tube and the sediment formed by the precipitated particles drops into a collector.

Chapter 13
ELECTRICAL OSCILLATIONS

13.1. Quasistationary Currents

When considering electrical oscillations, we have to do with time-varying currents. Ohm's law and Kirchhoff's rules following from it were established for a steady current. They also hold, however, for the instantaneous values of a varying current and voltage if the changes are not too fast. Electromagnetic disturbances propagate along a circuit with a tremendous speed equal to the speed of light c. Assume that the length of a circuit is l. If during the time $\tau = l/c$ needed for the transmission of a disturbance to the farthest point of a circuit, the current changes insignificantly, then the instantaneous values of the current in all the cross sections of the circuit will be virtually identical. Currents obeying this condition are called **quasistationary**. For periodically varying currents, the condition for a quasistationary state is

$$\tau = \frac{l}{c} \ll T,$$

where T is the period of the changes.

The delay for a circuit 3 m long is $\tau = 10^{-8}$ s. Thus, up to values of T of the order of 10^{-6} s (which corresponds to a frequency of 10^6 Hz), the currents in such a circuit may be considered quasistationary. A current of industrial frequency ($\nu = 50$ or 60 Hz) is quasistationary for circuits up to about 100 km long.

The instantaneous values of quasistationary currents obey Ohm's law. Hence, Kirchhoff's rules also hold for them.

In the following when studying electrical oscillations, we shall always assume that the currents we are dealing with are quasistationary.

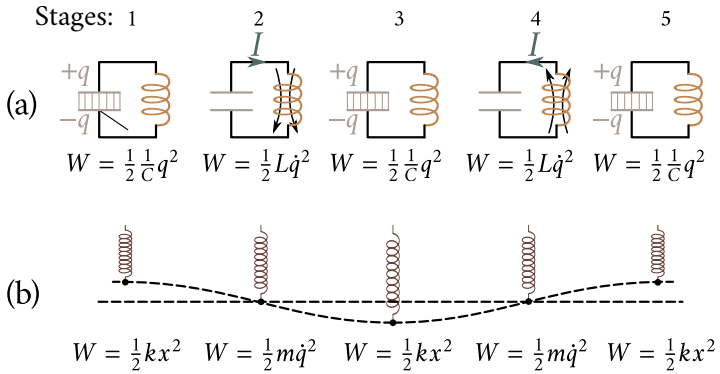


Fig. 13.1

13.2. Free Oscillations in a Circuit Without a Resistance

Electrical oscillations may appear in a circuit containing an inductance and a capacitance. Such a circuit is therefore called an **oscillatory circuit**. Figure 13.1a shows the consecutive stages of an oscillatory process in an idealized circuit containing no resistance.

Oscillations can be set up in the circuit either by supplying a certain initial charge to the capacitor plates or by producing a current in the inductance (for example, by switching off the external magnetic field passing through the coil turns). Let us use the first method. We shall connect the capacitor to a source of voltage after disconnecting it from the inductance. The result will be the appearance of unlike charges $+q$ and $-q$ on the plates (stage 1). An electric field will be set up between the plates, and its energy will be $(q^2/C)/2$ [see Eq. (4.5)]. If we next switch off the voltage source and connect the capacitor to the inductance, it will begin to discharge, and a current will flow through the circuit. The energy of the electric field will diminish as a result, but in return a constantly growing energy of the magnetic field set up by the current flowing through the inductance will appear. This energy is $LI^2/2$ [see Eq. (8.37)].

Since the resistance of the circuit is zero, the total energy consisting of the energies of the electric and magnetic fields is not used for heating the wires and will remain constant[1]. Therefore, at the moment when the voltage across the capacitor and, consequently, the energy of the electric field, vanish, the energy of the magnetic field and, consequently, the current reach their maximum value (stage 2; beginning

[1]Strictly speaking, in such an idealized circuit, energy would be lost on the radiation of electromagnetic waves. This loss grows with an increasing frequency of oscillations and when the circuit is more "open".

from this moment, the current flows at the expense of the self induced e.m.f.). After this, the current diminishes, and, when the charges on the plates reach their initial value q, the current will vanish (stage 3). Next, the same processes occur in the opposite direction (stages 4 and 5). After them, the system returns to its initial state (stage 5), and the entire cycle repeats again and again. The charge on the plates, the voltage across the capacitor, and the current flowing in the inductance periodically change (*i.e.*, oscillate) during the process. The oscillations are attended by mutual transformations of the electric and magnetic field energies.

Figure 13.1b compares the oscillations of a spring pendulum with those in the circuit. The supply of charges to the capacitor plates corresponds to bringing the pendulum out of its equilibrium position by exerting an external force on it and imparting the initial deviation x to it. The potential energy of elastic deformation of the spring equal to $kx^2/2$ is produced. Stage 2 corresponds to passing of the pendulum through its equilibrium position. At this moment, the quasi-elastic force vanishes, and the pendulum continues its motion by inertia. By this time, the energy of the pendulum completely transforms into kinetic energy and is determined by the expression $mx^2/2$. We shall let our reader compare the further stages.

It can be seen from a comparison of electrical and mechanical oscillations that the energy of an electric field $(q^2/C)/E$ is similar to the potential energy of elastic deformation, and the energy of a magnetic field $LI^2/2$ is similar to the kinetic energy. The inductance L plays the part of the mass m, and the reciprocal of the capacitance $(1/C)$ the part of the spring constant k. Finally, the displacement x of the pendulum from its equilibrium position corresponds to the charge q, and the speed $\dot{x}$ to the current $I = \dot{q}$. We shall see below that the analogy between electrical and mechanical oscillations also extends to the mathematical equations describing them.

Let us find an equation for the oscillations in a circuit without a resistance (an L-C circuit). We shall consider the current charging the capacitor to be positive[2] (Fig. 13.2). Hence, by Eq. (5.1),

$$I = \frac{\mathrm{d}q}{\mathrm{d}t} = \dot{q}.$$

Equation (5.27) of Ohm's law for circuit 1-3-2 is

$$IR = \varphi_1 - \varphi_2 + \mathscr{E}_{12}.$$

In our case, $R = 0$, $\varphi_1 - \varphi_2 = -q/C$, and $\mathscr{E}_{12} = \mathscr{E}_s = -L\,(\mathrm{d}I/\mathrm{d}t)$. Introducing these

[2] With such a choice of the direction of the current, the analogy between electrical and mechanical oscillations is more complete: $\dot{q}$ corresponds to the speed $\dot{X}$ (with a different choice, $-\dot{q}$ corresponds to the speed $\dot{x}$).

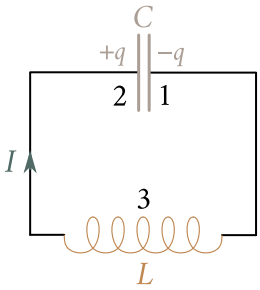


Fig. 13.2

values into Eq. (5.27), we get

$$0 = -\frac{q}{C} - L\frac{\mathrm{d}I}{\mathrm{d}t}. \tag{13.1}$$

Finally, replacing $\mathrm{d}I/\mathrm{d}t$ with $\ddot{q}$ [see Eq. (5.1)], we get

$$\ddot{q} + \frac{1}{LC}q = 0. \tag{13.2}$$

If we introduce the symbol

$$\omega_0 = \frac{1}{\sqrt{LC}}, \tag{13.3}$$

Eq. (13.2) becomes

$$\ddot{q} + \omega_0^2 q = 0, \tag{13.4}$$

which is our good acquaintance from the science of mechanical oscillations [see Eq. (7.7) of Vol. I]. The following function is a solution of this equation:

$$q = q_\mathrm{m}\cos(\omega_0 t + \alpha) \tag{13.5}$$

(the subscript "m" stands for maximum).

Thus, the charge on the capacitor plates changes according to a harmonic law with a frequency determined by Eq. (13.3). This frequency is called the **natural frequency of the circuit** (it corresponds to the natural frequency of a harmonic oscillator). We get the so-called **Thomson formula** for the period of the oscillations:

$$T = \frac{2\pi}{\sqrt{LC}}. \tag{13.6}$$

The voltage across the capacitor differs from the charge by the factor $1/C$:

$$U = \frac{q_\mathrm{m}}{C}\cos(\omega_0 t + \alpha) = U_\mathrm{m}\cos(\omega_0 t + \alpha). \tag{13.7}$$

Time differentiation of Eq. (13.5) yields an expression for the current:

$$I = -\omega_0 q_\mathrm{m}\sin(\omega_0 t + \alpha) = I_\mathrm{m}\cos(\omega_0 t + \alpha + \frac{\pi}{2}). \tag{13.8}$$

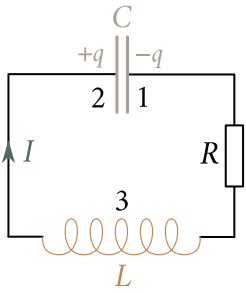


Fig. 13.3

Thus, the current leads the voltage across the capacitor in phase by $\pi/2$.

A comparison of Eqs. (13.5) and (13.7) with Eq. (13.8) shows that at the moment when the current reaches its maximum value, the charge and the voltage vanish, and vice versa. We have already established this relation between the charge and the current on the basis of energy considerations.

Examination of Eqs. (13.7) and (13.8) shows that

$$U_{\mathrm{m}} = \frac{q_{\mathrm{m}}}{C}, \quad I_{\mathrm{m}} = \omega_0 q_{\mathrm{m}}.$$

Taking the ratio of these amplitudes and substituting for ω_0 its value from Eq. (13.3), we get

$$U_{\mathrm{m}} = \left(\frac{L}{C}\right)^{1/2} I_{\mathrm{m}}. \tag{13.9}$$

We can also obtain this equation if we proceed from the fact that the maximum value of the energy of the electric field $CU_{\mathrm{m}}^2/2$ must equal the maximum value of the energy of the magnetic field $LI_{\mathrm{m}}^2/2$.

13.3. Free Damped Oscillations

Any real circuit has a resistance. The energy stored in the circuit is gradually spent in this resistance for heating, owing to which the free oscillations become damped. Equation (5.27) written for circuit 1-3-2 shown in Fig. 13.3 has the form

$$IR = -\frac{q}{C} - L\frac{\mathrm{d}I}{\mathrm{d}t} \tag{13.10}$$

[compare with Eq. (13.1)]. Dividing this equation by L and substituting $\dot{q}$ for I and $\ddot{q}$ for $\mathrm{d}I/\mathrm{d}t$, we obtain

$$\ddot{q} + \frac{R}{L}\dot{q} + \frac{1}{LC}q = 0. \tag{13.11}$$

Taking into account that the reciprocal of LC equals the square of the natural

frequency of the circuit ω_0 [see Eq. (13.3)], and introducing the symbol

$$\beta = \frac{R}{LC}, \tag{13.12}$$

Eq. (13.11) can be written in the form

$$\ddot{q} + 2\beta\dot{q} + \omega_0^2 q = 0. \tag{13.13}$$

This equation coincides with the differential equation of damped mechanical oscillations [see Eq. (7.11) of Vol. I].

When $\beta^2 < \omega_0^2$, *i.e.*, $R^2/(4L^2) < 1/(LC)$, the solution of Eq. (13.3) has the form

$$q = q_{\mathrm{m},0}\, e^{-\beta t} \cos(\omega t + \alpha), \tag{13.14}$$

where $\omega = \sqrt{\omega_0^2 - \beta^2}$. Substituting for ω_0 its value from Eq. (13.3) and for β its value from Eq. (13.12), we find that

$$\omega = \left(\frac{1}{LC} - \frac{R^2}{4L^2}\right). \tag{13.15}$$

Thus, the frequency of damped oscillations ω is smaller than the natural frequency ω_0. When $R = 0$, Eq. (13.13) transforms into Eq. (13.3).

Dividing Eq. (13.14) by the capacitance C, we get the voltage across the capacitor:

$$U = \frac{1}{C} q_{\mathrm{m},0}\, e^{-\beta t} \cos(\omega t + \alpha) = U_{\mathrm{m},0}\, e^{-\beta t} \cos(\omega t + \alpha). \tag{13.16}$$

To find the current, we shall differentiate Eq. (13.14) with respect to time

$$I = \dot{q} = q_{\mathrm{m},0}\, e^{-\beta t}[-\beta \cos(\omega t + \alpha) - \omega \sin(\omega t + \alpha)].$$

Multiplying the right-hand side of this equation by the expression

$$\frac{\omega_0}{\sqrt{\omega^2 - \beta^2}}$$

equal to unity, we get

$$I = \omega_0 q_{\mathrm{m},0}\, e^{-\beta t}\left[-\frac{\beta}{\sqrt{\omega^2 - \beta^2}} \cos(\omega t + \alpha) - \frac{\omega}{\sqrt{\omega^2 - \beta^2}} \sin(\omega t + \alpha)\right].$$

Introducing the angle ψ determined by the conditions

$$\cos\psi = -\frac{\beta}{\sqrt{\omega^2 - \beta^2}} = -\frac{\beta}{\omega_0}, \quad \sin\psi = \frac{\omega}{\sqrt{\omega^2 - \beta^2}} = \frac{\omega}{\omega_0},$$

we can write

$$I = \omega_0 q_{\mathrm{m},0}\, e^{-\beta t} \cos(\omega t + \alpha + \psi). \tag{13.17}$$

Since $\cos\psi < 0$ and $\sin\psi > 0$, the value of ψ is within the limits from $\pi/2$ to π (*i.e.*, $\pi/2 < \psi < \pi$). Thus, when a circuit contains a resistance, the current leads the voltage across the capacitor in phase by more than $\pi/2$ (when $R = 0$, the advance

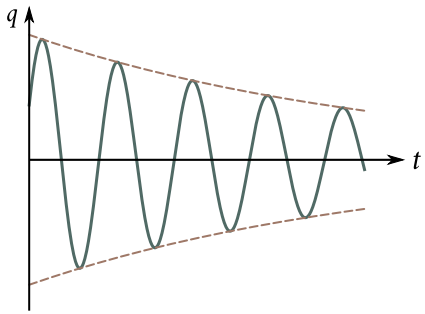


Fig. 13.4

in phase is $\pi/2$).

A plot of function (13.14) is depicted in Fig. 13.4. Plots of the voltage and current are similar to it.

It is customary practice to characterize the damping of oscillations by the **logarithmic decrement**

$$\lambda = \ln\left[\frac{A(t)}{A(t+T)}\right] = \beta T \tag{13.18}$$

[see Eq. (7.104) of Vol. I]. Here $A(t)$ is the amplitude of the relevant quantity (q, U, or I). We remind our reader that the logarithmic decrement is the reciprocal of the number of oscillations N_e performed during the time needed for the amplitude to decrease to $1/e$ of its initial value:

$$\lambda = \frac{1}{N_e}.$$

Using in Eq. (13.18) the value of β from Eq. (13.12) and substituting $2\pi/\omega$ for T, we get the following expression for A:

$$\lambda = \frac{R}{2L}\frac{2\pi}{\omega} = \frac{\pi R}{L\omega}. \tag{13.19}$$

The frequency ω, and, therefore, also A are determined by the parameters of a circuit L, C, and R. Thus, the logarithmic decrement is a characteristic of a circuit.

If the damping is not great ($\beta^2 \ll \omega_0^2$), we can assume in Eq. (13.19) that $\omega \approx \omega_0 = 1/\sqrt{LC}$. Hence,

$$\lambda \approx \frac{\pi R\sqrt{LC}}{L} = \pi R\left(\frac{C}{L}\right)^{1/2}. \tag{13.20}$$

An oscillatory circuit is often characterized by its quality, or simply Q, determined as a quantity that is inversely proportional to the logarithmic decrement:

$$Q = \frac{\pi}{\lambda} = \pi N_e. \tag{13.21}$$

It follows from Eq. (13.21) that the quality of a circuit is the higher, the greater is the number of oscillations completed before the amplitude diminishes to $1/e$ of its initial value.

For weak damping, we have

$$Q = \frac{1}{R}\left(\frac{L}{C}\right)^{1/2} \tag{13.22}$$

[see Eq. (13.20)].

In Sec. 7.10 of Vol. I, we showed that when the damping is weak, the quality of a mechanical oscillatory system equals the ratio of the energy stored in the system at a given moment to the decrement of this energy during one period of oscillations with an accuracy to the factor 2π. We shall show that this also holds for electrical oscillations. The amplitude of the current in a circuit diminishes according to the law $e^{-\beta t}$. The energy W stored in the circuit is proportional to the square of the current amplitude (or to the square of the amplitude of the voltage across the capacitor). Hence, W diminishes according to the law $e^{-2\beta t}$. The relative reduction in the energy during a period is

$$\frac{\Delta W}{W} = \frac{W(t) - W(t+T)}{W(t)} = \frac{1 - e^{-2\beta t}}{1} = 1 - e^{-2\lambda}.$$

With insignificant damping (*i.e.*, when $A \ll 1$), we may assume that $e^{-2\lambda}$ is approximately equal to $1 - 2\lambda$:

$$\frac{\Delta W}{W} = 1 - (1 - 2\lambda) = 2\lambda.$$

Finally, substituting the quality Q of the circuit for λ in this expression in accordance with Eq. (13.21) and solving the equation obtained relative to Q, we get

$$Q = 2\pi \frac{\Delta W}{W}. \tag{13.23}$$

We shall note in conclusion that when $R^2/(4L^2) \geqslant 1/(LC)$, *i.e.*, when $\beta^2 \geqslant \omega_0^2$, an aperiodic discharge of the capacitor occurs instead of oscillations. The resistance of a circuit at which an oscillatory process transforms into an aperiodic one is called **critical**. The value of the critical resistance R_{cr} is determined by the condition $R_{cr}^2/(4L^2) = 1/(LC)$, whence

$$R_{cr} = 2\left(\frac{L}{C}\right)^{1/2}. \tag{13.24}$$

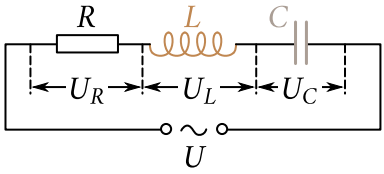


Fig. 13.5

13.4. Forced Electrical Oscillations

To produce forced oscillations of a system, an external periodically changing action must be exerted on it. This can be achieved for electrical oscillations if we connect a varying e.m.f. in series with the circuit elements or, if after breaking the circuit, we feed an alternating voltage to the contacts formed, *i.e.*, the voltage

$$U = U_{\mathrm{m}} \cos(\omega t) \tag{13.25}$$

(Fig. 13.5). This voltage must be added to the self-induced e.m.f.. As a result, Eq. (13.10) acquires the form

$$IR = -\frac{q}{C} - L\frac{\mathrm{d}I}{\mathrm{d}t} + U_{\mathrm{m}} \cos(\omega t). \tag{13.26}$$

After transformations, we get the equation

$$\ddot{q} + 2\beta\dot{q} + \omega_0^2 q = \frac{U_{\mathrm{m}}}{L} \cos(\omega t). \tag{13.27}$$

Here, ω_0^2 and β are determined by Eqs. (13.3) and (13.12).

Equation (13.27) coincides with the differential equation of forced mechanical oscillations [see Eq. (7.111) of Vol. I]. A partial solution of this equation has the form

$$q = q_{\mathrm{m}} \cos(\omega t - \psi), \tag{13.28}$$

where

$$q_{\mathrm{m}} = \frac{U_{\mathrm{m}}/L}{\sqrt{\left(\omega_0^2 - \omega^2\right)^2 + 4\beta^2\omega^2}}, \qquad \tan\psi = \frac{2\beta\omega}{\omega_0^2 - \omega^2}$$

[see Eq. (7.119) of Vol. I]. Substitution of their values for ω_0 and β gives

$$q_{\mathrm{m}} = \frac{U_{\mathrm{m}}}{\omega\sqrt{R^2 + [\omega L - 1/(\omega C)]^2}}, \tag{13.29}$$

$$\tan\psi = \frac{R}{[1/(\omega C) - \omega L]}. \tag{13.30}$$

A general solution is obtained if we add the general solution of the relevant homogeneous equation to partial solution (13.28). This solution was obtained in the preceding section [see Eq. (13.14)]. It contains the exponential factor $e^{-\beta t}$, therefore, after sufficient time elapses, becomes very small and it may be disregarded.

Consequently, stationary forced oscillations are described by the function (13.28).

Time differentiation of Eq. (13.28) gives the current in a circuit with stationary oscillations:

$$I = -\omega q_{\mathrm{m}} \sin(\omega t - \psi) = I_{\mathrm{m}} \cos\left(\omega t - \psi + \frac{\pi}{2}\right)$$

($I_{\mathrm{m}} = \omega q_{\mathrm{m}}$). Let us write this expression in the form[3]

$$I = I_{\mathrm{m}} \cos(\omega t - \varphi), \tag{13.31}$$

where $\varphi = \psi - \pi/2$ is the shift in phase between the current and the applied voltage [see Eq. (13.25)]. In accordance with Eq. (13.30):

$$\tan\varphi = \tan\left(\psi - \frac{\pi}{2}\right) = -\frac{q}{\tan\psi} = \frac{\omega L - 1/(LC)}{R}. \tag{13.32}$$

Inspection of this equation shows that the current lags in phase behind the voltage ($\varphi > 0$) when $\omega L > 1/(\omega C)$, and leads the voltage ($\varphi < 0$) when $\omega L < 1/(\omega C)$. According to Eq. (13.29):

$$I_{\mathrm{m}} = \omega q_{\mathrm{m}} = \frac{U_{\mathrm{m}}}{\sqrt{R^2 + [\omega L - 1/(\omega C)]^2}}. \tag{13.33}$$

Let us write Eq. (13.26) in the form

$$IR + \frac{q}{C} + L\frac{\mathrm{d}I}{\mathrm{d}t} = U_{\mathrm{m}} \cos(\omega t). \tag{13.34}$$

The product IR equals the voltage U_R across the resistance, q/C is the voltage across the capacitor U_C, and the expression $L\,(\mathrm{d}I/\mathrm{d}t)$ determines the voltage across the inductance U_L. Taking this into account, we can write

$$U_R + U_C + U_L = U_{\mathrm{m}} \cos(\omega t). \tag{13.35}$$

Thus, the sum of the voltages across the separate elements of a circuit at each moment of time equals the voltage applied from an external source (see Fig. 13.5).

According to Eq. (13.31)

$$U_R = RI_{\mathrm{m}} \cos(\omega t - \varphi). \tag{13.36}$$

Dividing Eq. (13.28) by the capacitance, we get the voltage across the capacitor

$$U_C = \frac{q_{\mathrm{m}}}{C} \cos(\omega t - \psi) = U_{C,\mathrm{m}} \cos\left(\omega t - \varphi - \frac{\pi}{2}\right). \tag{13.37}$$

Here,

$$U_{C,\mathrm{m}} = \frac{q_{\mathrm{m}}}{C} = \frac{U_{\mathrm{m}}}{\omega C\sqrt{R^2 + [\omega L - 1/(\omega C)]^2}} = \frac{I_{\mathrm{m}}}{\omega C} \tag{13.38}$$

[see Eq. (13.31)]. Multiplying the derivative of function (13.31) by L, we get the voltage

[3] We shall not encounter the concept of potential any more up to the end of this chapter. Therefore, no misunderstandings will appear if we use the symbol φ for the phase angle.

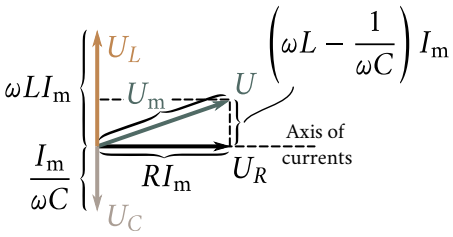


Fig. 13.6

across the inductance:

$$U_L = L\frac{\mathrm{d}I}{\mathrm{d}t} = -\omega L I_\mathrm{m}\sin(\omega t - \varphi) = U_{L,\mathrm{m}}\cos\left(\omega t - \varphi + \frac{\pi}{2}\right). \tag{13.39}$$

Here,

$$U_{L,\mathrm{m}} = \omega L I_\mathrm{m}. \tag{13.40}$$

A comparison of Eqs. (13.31), (13.36), (13.37), and (13.39) shows that the voltage across the capacitor lags in phase behind the current by $\pi/2$, while the voltage across the inductance leads the current by $\pi/2$. The voltage across the resistance changes in phase with the current. The phase relations can be shown very clearly with the aid of a vector diagram (see Sec. 7.7 of Vol. I). We remind our reader that a harmonic oscillation (or a harmonic function) can be shown with the aid of a vector whose length equals the amplitude of the oscillation, while the direction of the vector makes an angle equal to the initial phase of the oscillation with a certain axis. Let us take the axis of currents as the straight line from which the initial phase is counted. This gives us the diagram shown in Fig. 13.6.

According to Eq. (13.35), the sum of the three functions U_R, U_C, and U_L must equal the applied voltage U. The voltage U is accordingly shown in the diagram by a vector equal to the sum of the vectors U_R, U_C, and U_L. We must note that Eq. (13.33) is easily obtained from the right triangle formed in the vector diagram by the vectors U, U_R, and the difference $U_L - U_C$.

The resonance frequency for the charge q and the voltage U_C across the capacitor is

$$\omega_{q,\mathrm{res}} = \omega_{U,\mathrm{res}} = \left(\omega_0^2 - 2\beta^2\right)^{1/2} = \left(\frac{1}{LC} - \frac{R^2}{2L^2}\right)^{1/2} \leqslant \omega_0 \tag{13.41}$$

[see Eq. (7.127) of Vol. I].

Resonance curves for U_C are shown in Fig. 13.7 (resonance curves for q have the same form). They are similar to the resonance curves obtained for mechanical oscillations (see Fig. 7.24 of Vol. I). When $\omega \to 0$, the resonance curves converge at one point having the ordinate $U_{C,\mathrm{m}} = U_\mathrm{m}$, *i.e.*, the voltage appearing across the

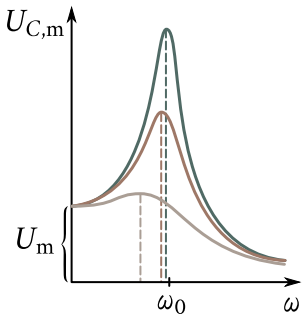


Fig. 13.7

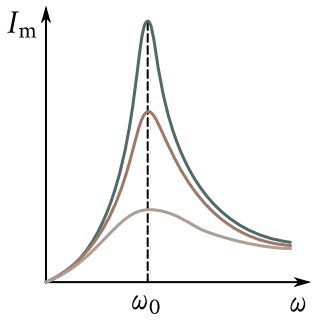


Fig. 13.8

capacitor when it is connected to a source of steady voltage U_m. The maximum in resonance will be the higher and the sharper, the smaller is $\beta = R/(2L)$, *i.e.*, the smaller is the resistance and the greater the inductance of the circuit.

Resonance curves for the current are shown in Fig. 13.8. They correspond to the resonance curves for the velocity in mechanical oscillations. The amplitude of the current has a maximum value at $\omega L - 1/(\omega C)$ [see Eq. (13.33)]. Consequently, the resonance frequency for the current coincides with the natural frequency of the circuit ω_0:

$$\omega_{I,\mathrm{res}} = \omega_0 = \frac{1}{\sqrt{LC}}. \tag{13.42}$$

The intercept formed by the resonance curves on the I_m-axis is zero—at a constant voltage, a steady current cannot flow in a circuit containing a capacitor.

At small damping (when $\beta^2 \ll \omega_0^2$), the resonance frequency for the voltage can be taken equal to ω_0 [see Eq. (13.41)]. Accordingly, we may consider that $\omega_\mathrm{res} L - 1/(\omega_\mathrm{res} C)$. By Eq. (13.38), the ratio of the amplitude of the voltage across the capacitor in resonance $U_{C,\mathrm{m,res}}$ to the amplitude of the external voltage U_m will in this case be

$$\frac{U_{C,\mathrm{m,res}}}{U_\mathrm{m}} = \frac{1}{\omega_0 CR} = \frac{\sqrt{LC}}{CR} = \frac{1}{R}\left(\frac{L}{C}\right)^{1/2} = Q \tag{13.43}$$

[we have assumed in Eq. (13.38) that $\omega = \omega_{U,\mathrm{res}} = \omega_0$. Here, Q is the quality of the circuit [see Eq. (13.22)]. Thus, the quality of a circuit shows how many times the voltage across a capacitor can exceed the applied voltage.

The quality of a circuit also determines the sharpness of the resonance curves. Figure 13.9 shows a resonance curve for the current in a circuit. Instead of laying off the values of I_m corresponding to a given frequency along the axis of ordinates, we have laid off the ratio of I_m to $I_{\mathrm{m,res}}$ (*i.e.*, to I_m in resonance). Let us consider the width of the curve $\Delta\omega$ taken at the height 0.7 (a power ratio of $0.7^2 \approx 0.5$

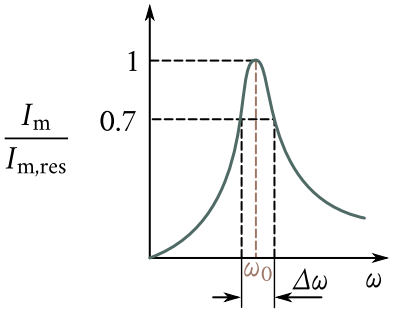


Fig. 13.9

corresponds to a ratio of the current amplitudes equal to 0.7). We can show that the ratio of this width to the resonance frequency equals a quantity that is the reciprocal of the quality of a circuit:

$$\frac{\Delta\omega}{\omega_0} = \frac{1}{Q}. \tag{13.44}$$

We remind our reader that Eqs. (13.43) and (13.44) hold only for large values of Q, *i.e.*, when the damping of the free oscillations in the circuit is small.

The phenomenon of resonance is used to separate the required component from a complex voltage. Assume that the voltage applied to a circuit is

$$U = U_{m,1}\cos(\omega_1 t + \alpha_1) + U_{m,2}\cos(\omega_2 t + \alpha_2) + \ldots.$$

By tuning the circuit to one of the frequencies ω_1, ω_2, etc. (*i.e.*, by correspondingly choosing its parameters C and L), we can obtain a voltage across the capacitor that exceeds the value of the given component Q times, whereas the voltage produced across the capacitor by the other components will be weak. Such a process is carried out, for example, when tuning a radio receiver to the required wavelength.

13.5. Alternating Current

The stationary forced oscillations described in the preceding section can be considered as the flow of an alternating current produced by the alternating voltage

$$U = U_m\cos(\omega t) \tag{13.45}$$

in a circuit including a capacitance, an inductance, and a resistance. According to Eqs. (13.31), (13.32), and (13.33), this current varies according to the law

$$I = I_m\cos(\omega t - \varphi). \tag{13.46}$$

The amplitude of the current is determined by the amplitude of the voltage U_m the

circuit parameters C, L, R, and the frequency ω:

$$I_{\mathrm{m}} = \frac{U_{\mathrm{m}}}{\sqrt{R^2 + [\omega L - 1/(\omega C)]^2}}. \tag{13.47}$$

The current lags in phase behind the voltage by the angle φ that depends on the parameters of the circuit and on the frequency:

$$\tan\varphi = \frac{\omega L - 1/(\omega C)}{R}. \tag{13.48}$$

When $\varphi < 0$, the current actually leads the voltage.

The expression

$$Z = \left(R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2\right)^{1/2} \tag{13.49}$$

in the denominator of Eq. (13.47) is called the **impedance**.

If a circuit consists only of a resistance R, the equation of Ohm's law has the form

$$IR = U_{\mathrm{m}} \cos(\omega t).$$

Hence, it follows that the current in this case varies in phase with the voltage, while the amplitude of the current is

$$I_{\mathrm{m}} = \frac{U_{\mathrm{m}}}{R}.$$

A comparison of this expression with Eq. (13.47) shows that the replacement of a capacitor with a shorted circuit section signifies a transition to $C \to \infty$ instead of to $C = 0$.

Any real circuit has finite values of R, L, and C. It may happen that some of these parameters are such that their influence on the current may be disregarded. Suppose that R of a circuit may be assumed equal to zero, and C equal to infinity. Now, we can see from Eqs. (13.47) and (13.48) that

$$I_{\mathrm{m}} = \frac{U_{\mathrm{m}}}{\omega L} \tag{13.50}$$

and that $\tan\varphi = \infty$ (accordingly, $\varphi = \pi/2$). The quantity

$$X_L = \omega L \tag{13.51}$$

is called the **inductive reactance**. If L is expressed in henries, and ω in rad s^{-1}, then X_L will be expressed in ohms. Examination of Eq. (13.51) shows that the inductive reactance grows with the frequency ω. An inductance does not react to a steady current ($\omega = 0$), *i.e.*, $X_L = 0$.

The current in an inductance lags behind the voltage by $\pi/2$. Accordingly, the voltage across the inductance leads the current by $\pi/2$ (see Fig. 13.6).

Now, let us assume that R and L both equal zero. Hence, according to Eqs. (13.47) and (13.48), we have

$$I_{\mathrm{m}} = \frac{U_{\mathrm{m}}}{1/(\omega C)} \tag{13.52}$$

$\tan\varphi = -\infty$ (*i.e.*, $\varphi = -\pi/2$). The quantity

$$X_C = \frac{1}{\omega C} \tag{13.53}$$

is called the **capacitive reactance**. If C is expressed in farads, and ω in rad s^{-1} then X_C will be expressed in ohms. It follows from Eq. (13.53) that the capacitive reactance diminishes with increasing frequency. For a steady current, $X_C = \infty$—a steady current cannot flow through a capacitor. Since $\varphi = -\pi/2$, the current flowing through a capacitor leads the voltage by $\pi/2$. Accordingly, the voltage across a capacitor lags behind the current by $\pi/2$ (see Fig. 13.6).

Finally, suppose that we may assume R to equal zero. In this case, Eq. (13.47) becomes

$$I_{\mathrm{m}} = \frac{U_{\mathrm{m}}}{|\omega L - 1/(\omega C)|}. \tag{13.54}$$

The quantity

$$X = \omega L - \frac{1}{\omega C} = X_L - X_C \tag{13.55}$$

is called the **reactance**.

Equations (13.48) and (13.49) can be written in the form

$$\tan\varphi = \frac{X}{R}, \quad Z = \sqrt{R^2 + X^2}.$$

Thus, if the values of the resistance R and the reactance X are laid off along the legs of a right triangle, then the length of the hypotenuse will numerically equal Z (see Fig. 13.6).

Let us find the power liberated in an alternating current circuit. The instantaneous value of the power equals the product of the instantaneous values of the voltage and current:

$$P(t) = U(t)I(t) = U_{\mathrm{m}}\cos(\omega t) \times I_{\mathrm{m}}\cos(\omega t - \varphi). \tag{13.56}$$

Taking advantage of the formula

$$\cos\alpha\cos\beta = \frac{1}{2}\cos(\alpha - \beta) + \frac{1}{2}\cos(\alpha + \beta),$$

we can write Eq. (13.56) in the form

$$P(t) = \frac{1}{2}U_{\mathrm{m}}I_{\mathrm{m}}\cos\varphi + \frac{1}{2}U_{\mathrm{m}}I_{\mathrm{m}}\cos(2\omega t - \varphi). \tag{13.57}$$

Of practical interest is the time-average value $P(t)$, which we shall denote

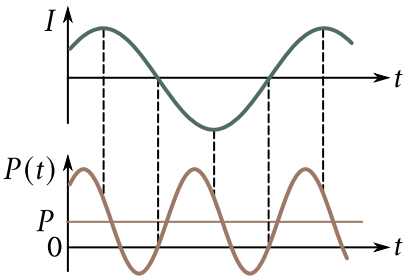


Fig. 13.10

simply by P. Since the average value of $\cos(2\omega t - \varphi)$ is zero, we have

$$P = \frac{U_m I_m}{2} \cos \varphi. \tag{13.58}$$

Inspection of Eq. (13.57) shows that the instantaneous power fluctuates about the average value with a frequency double that of the current (Fig. 13.10).

In accordance with Eq. (13.48),

$$\cos \varphi = \frac{R}{\sqrt{R^2 + [\omega L - 1/(\omega C)]^2}} = \frac{R}{Z}. \tag{13.59}$$

Using this value of $\cos \varphi$ in Eq. (13.48) and taking into account that $U_m/Z = I_m$, we get

$$P = \frac{R I_m^2}{2}. \tag{13.60}$$

The same power is developed by a direct current whose strength is

$$I = \frac{I_m}{\sqrt{2}}. \tag{13.61}$$

Quantity (13.61) is known as the **effective value of the current**. Similarly, the quantity

$$U = \frac{U_m}{\sqrt{2}}, \tag{13.62}$$

is called the **effective voltage**.

Expressing the average power through the effective current and voltage, we get

$$P = UI \cos \varphi. \tag{13.63}$$

The factor $\cos \varphi$ in this expression is called the **power factor**. Engineers try to make $\cos \varphi$ as high as possible. At a low value of $\cos \varphi$, a large current must be passed through a circuit to obtain the required power, and this results in greater losses in the feeder lines.

PART II

WAVES

Chapter 14
ELASTIC WAVES

14.1. Propagation of Waves in an Elastic Medium

If at any place of an elastic (solid or fluid) medium its particles are made to oscillate, then owing to interaction between the particles, this oscillation will propagate in the medium from particle to particle with a certain velocity v. The process of the propagation of oscillations in space is called a **wave**.

The particles of a medium in which a wave is propagating are not made to perform translational motion by the wave, they only oscillate about their equilibrium positions. Depending on the direction of oscillations of particles relative to the direction of propagation of the wave, **longitudinal** and **transverse** waves are distinguished. In the former, the particles of the medium oscillate along the direction of propagation of the wave. In transverse waves, the particles of the medium oscillate in directions at right angles to the direction of wave propagation. Elastic transverse waves can appear only in a medium having a resistance to shear. Therefore, only longitudinal waves can appear in fluids. Both longitudinal and transverse waves can appear in a solid.

Figure 14.1 shows the motion of the particles when a transverse wave propagates in a medium. The numbers 1, 2, etc. designate particles spaced at a distance of $vT/4$, *i.e.*, at the distance travelled by the wave during one-fourth of the period of the oscillations performed by the particles. At the moment of time taken as zero, the wave propagating along the axis from left to right reached particle 1. As a result, the particle began to move upward from its equilibrium position, carrying the following particles along. After one-fourth of a period, particle 1 reaches its extreme top position; simultaneously, particle 2 begins to move from its equilibrium position. After another fourth of a period elapses, the first particle will pass its equilibrium position moving downward, the second particle will reach its extreme top position,

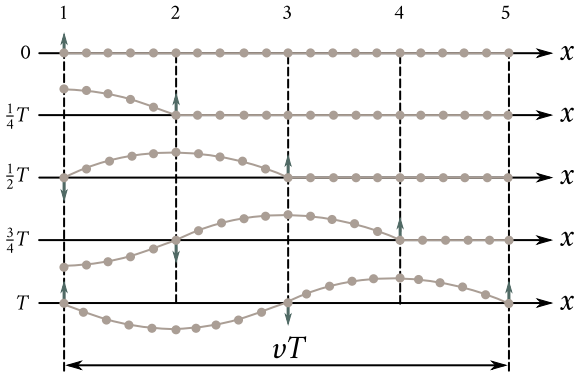


Fig. 14.1

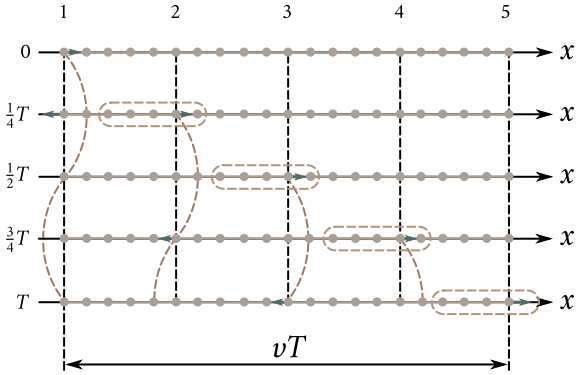


Fig. 14.2

and the third particle will begin to move upward from its equilibrium position. At the moment T, the first particle will complete a cycle of oscillation and will be in the same state of motion as at the initial moment. The wave by the moment T, having covered the path vT, will reach particle 5.

Figure 14.2 shows how the particles move when a longitudinal wave propagates in a medium. All the reasoning relating to the behaviour of particles in a transverse wave can also be related to the given case with displacements to the right and left substituted for the upward and downward ones. A glance at the figure shows that the propagation of a longitudinal wave in a medium is attended by alternating compensations and dilatations of the particles (the places of compensation of the particles are surrounded by a dash line in the figure). They move in the direction of wave propagation with the velocity v.

Figures 14.1 and 14.2 show oscillations of particles whose equilibrium positions are on the x-axis. Actually, not only the particles along the x-axis, but the entire

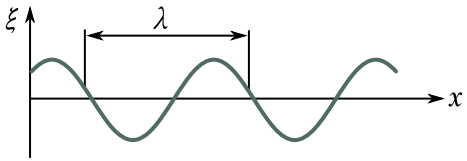


Fig. 14.3

collection of particles contained in a certain volume oscillate. Spreading from the source of oscillations, the wave process involves new and new parts of space. The locus of the points reached by the oscillations at the moment of time t is called the **wavefront**. The latter is the surface separating the part of space already involved in the wave process from the region in which oscillations have not yet appeared.

The locus of the points oscillating in the same phase is known as a **wave surface**. A wave surface can be drawn through any point of the space involved in a wave process. Hence, there is an infinitely great number of wave surfaces, whereas there is only one wavefront at each moment of time. Wave surfaces remain stationary (they pass through the equilibrium positions of particles oscillating in the same phase). A wavefront is in constant motion.

Wave surfaces can have any shape. In the simplest cases, they are planes or spheres. The wave in these cases is called plane or spherical, accordingly. In a plane wave, the wave surfaces are a multitude of parallel planes, in a spherical wave they are a multitude of concentric spheres.

Assume that a plane wave is propagating along the x-axis. Hence, all the points of the medium whose equilibrium positions have an identical coordinate x (but different values of y and z) oscillate in the same phase. Figure 14.3 shows a curve that produces the displacement ξ of points having different x's at a certain moment of time from their equilibrium position. This figure must not be understood as a visible image of a wave. It shows a graph of the function $\xi(x, t)$ for a certain fixed moment of time t. Such a graph can be constructed for both a longitudinal and a transverse wave.

The distance λ covered by a wave during the time equal to the period of oscillations of the particles of a medium is called the **wavelength**. It is obvious that

$$\lambda = vT, \tag{14.1}$$

where v is the velocity of the wave and T is the period of oscillations.

The wavelength can also be defined as the distance between the closest points of a medium that oscillate with a phase difference of 2π (see Fig. 14.3).

Substituting $1/\nu$ (ν is the frequency of oscillations) for T in Eq. (14.1), we get

$$\lambda\nu = v. \tag{14.2}$$

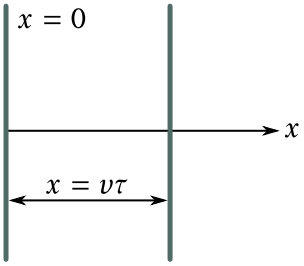


Fig. 14.4

We can also arrive at this equation from the following considerations. In one second, a wave source completes ν oscillations, producing during each oscillation one "crest" and one "trough" in the medium. By the moment when the source will complete its ν-th oscillation, the first crest will cover the path v. Consequently, the path v must contain ν crests and troughs of the wave.

14.2. Equations of a Plane and a Spherical Wave

A wave equation is an expression that gives the displacement of an oscillating particle as a function of its coordinates x, y, z, and the time t:

$$\xi = \xi(x, y, z, t) \tag{14.3}$$

(we have in mind the coordinates of the equilibrium position of the particle). This function must be periodical both relative to the time t and to the coordinates x, y, z. Its periodicity in time follows from the fact that ξ describes the oscillations of a particle having the coordinates x, y, z. Its periodicity with respect to the coordinates follows from the fact that points at a distance λ from one another oscillate in the same way.

Let us find the form of the function ξ for a plane wave assuming that the oscillations are harmonic. For simplicity, we shall direct the coordinate axes so that the x-axis coincides with the direction of propagation of the wave. The wave surfaces will therefore be perpendicular to the x-axis and, since all the points of the wave surface oscillate identically, the displacement ξ will depend only on x and on t, *i.e.*, $\xi = \xi(x, t)$. Let the oscillations of the points in the plane $x = 0$ (Fig. 14.4) have the form

$$\xi(0, t) = A\cos(\omega t + \alpha).$$

Let us find the form of the oscillations of the points in the plane corresponding to an arbitrary value of x. To travel the path from the plane $x = 0$ to this plane, the wave needs the time $T = x/v$ (here, v is the velocity of wave propagation).

Consequently, the oscillations of the particles in the plane x will lag in time by T behind the oscillations of the particles in the plane $x = 0$, *i.e.*, they will have the form

$$\xi(x,t) = A\cos[\omega(t-\tau)+\alpha] = A\cos\left[\omega\left(t-\frac{x}{v}\right)+\alpha\right].$$

Thus, the equation of a plane wave (both a longitudinal and a transverse one) propagating in the direction of the x-axis has the following form:

$$\xi = A\cos\left[\omega\left(t-\frac{x}{v}\right)+\alpha\right]. \tag{14.4}$$

The quantity A is the amplitude of a wave. The initial phase of the wave α is determined by our choice of the beginning of counting x and t. When considering one wave, the initial time and the coordinates are usually selected so that α is zero. This cannot be done, as a rule, when considering several waves jointly.

Let us fix a value of the phase in Eq. (14.4) by assuming that

$$\omega\left(t-\frac{x}{v}\right)+\alpha = \text{constant}. \tag{14.5}$$

This expression determines the relation between the time t and the place x where the phase has a fixed value. The value of $\mathrm{d}x/\mathrm{d}t$ ensuing from it gives the velocity with which the given value of the phase propagates. Differentiation of Eq. (14.5) yields

$$\mathrm{d}t - \frac{1}{v}\,\mathrm{d}x = 0,$$

whence

$$\frac{\mathrm{d}x}{\mathrm{d}t} = v. \tag{14.6}$$

Thus, the velocity of wave propagation v in Eq. (14.4) is the velocity of phase propagation, and in this connection it is called the **phase velocity**.

According to Eq. (14.6), we have $\mathrm{d}x/\mathrm{d}t > 0$. Hence, Eq. (14.4) describes a wave propagating in the direction of growing x. A wave propagating in the opposite direction is described by the equation

$$\xi = A\cos\left[\omega\left(t+\frac{x}{v}\right)+\alpha\right]. \tag{14.7}$$

Indeed, equating the phase of wave (14.7) to a constant and differentiating the equation obtained, we arrive at the expression

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -v,$$

from which it follows that the wave given by Eq. (14.7) propagates in the direction of diminishing x.

The equation of a plane wave can be given a symmetrical form relative to x and

t. For this purpose, let us introduce the quantity

$$k = \frac{2\pi}{\lambda}, \tag{14.8}$$

known as the **wave number**. Multiplying the numerator and the denominator of Eq. (14.8) by the frequency ν, we can represent the wave number in the form

$$k = \frac{\omega}{v} \tag{14.9}$$

[see Eq. (14.2)]. Opening the parentheses in Eq. (14.4) and taking Eq. (14.9) into account, we arrive at the following equation for a plane wave propagating along the x-axis:

$$\xi = A\cos(\omega t - kx + \alpha). \tag{14.10}$$

The equation of a wave propagating in the direction of diminishing x differs from Eq. (14.10) only in the sign of the term kx.

In deriving Eq. (14.10), we assumed that the amplitude of the oscillations does not depend on x. This is observed for a plane wave when the energy of the wave is not absorbed by the medium. When a wave propagates in a medium absorbing energy, the intensity of the wave gradually diminishes with an increasing distance from the source of oscillations—damping of the wave is observed. Experiments show that in a homogeneous medium such damping occurs according to an exponential law: $A = A_0 e^{-\gamma x}$ [compare with the diminishing of the amplitude of damped oscillations with time; see Eq. (7.102) of Vol. I]. Accordingly, the equation of a plane wave has the following form:

$$\xi = A_0\, e^{-\gamma x}\cos(\omega t - kx + \alpha) \tag{14.11}$$

(A_0 is the amplitude at points in the plane $x = 0$, and γ is the attenuation coefficient).

Now, let us find the equation of a spherical wave. Any real source of waves has a certain extent. But if we limit ourselves to considering a wave at distances from its source appreciably exceeding the dimensions of the source, then the latter may be treated as a point one. A wave emitted by a point source in an isotropic and homogeneous medium will be spherical. Assume that the phase of oscillations of the source is $(\omega t + \alpha)$. Hence, points on a wave surface of radius r will oscillate with the phase $\omega(t - r/v) + \alpha = \omega_0 t - kr + \alpha$ (the wave needs the time $\tau = r/v$ to travel the path r). The amplitude of the oscillations in this case, even if the energy of the wave is not absorbed by the medium, does not remain constant—it diminishes with the distance from the source as $1/r$ (see Sec. 14.6). Consequently, the equation of a spherical wave has the form

$$\xi = \frac{A}{r}\cos(\omega t - kr + \alpha), \tag{14.12}$$

where A is a constant quantity numerically equal to the amplitude at a distance of unity from the source. The dimension of A equals that of the oscillating quan-

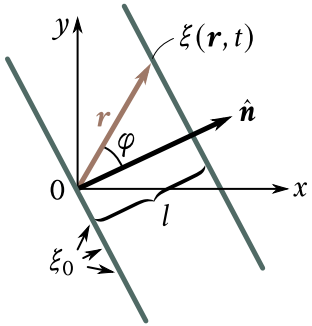


Fig. 14.5

tity multiplied by the dimension of length. The factor $e^{-\gamma r}$ must be multiplied to Eq. (14.12) for an absorbing medium.

We remind our reader that owing to the assumptions we have made, Eq. (14.12) holds only when r appreciably exceeds the dimensions of the source. When r tends to zero, the expression for the amplitude tends to infinity. The explanation of this absurd result is that the equation cannot be used for small r's.

14.3. Equation of a Plane Wave Propagating in an Arbitrary Direction

Let us find the equation of a plane wave propagating in a direction making the angles α, β, γ (not to be confused with the attenuation coefficient) with the coordinate axes x, y, z. We shall assume that the oscillations in a plane passing through the origin of coordinates (Fig. 14.5) have the form

$$\xi_0 = A\cos(\omega t + \alpha). \tag{14.13}$$

Let us take a wave surface (plane) at the distance l from the origin of coordinates. The oscillations in this plane will lag behind those expressed by Eq. (14.13) by the time $\tau = l/v$:

$$\xi = A\cos\left[\omega\left(t - \frac{l}{v}\right) + \alpha\right] = A\cos(\omega t - kl + \alpha) \tag{14.14}$$

[$k = \omega/v$; see Eq. (14.9)].

Let us express l through the position vector of points on the surface being considered. For this purpose, we shall introduce the unit vector $\hat{\boldsymbol{n}}$ of a normal to the wave surface. A glance at Fig. 14.5 shows that the scalar product of $\hat{\boldsymbol{n}}$ and the position vector $\boldsymbol{r}$ of any point on the surface is l:

$$\hat{\boldsymbol{n}} \cdot \boldsymbol{r} = r\cos\varphi = l.$$

Substitution of $\hat{\boldsymbol{n}} \cdot \boldsymbol{r}$ for l in Eq. (14.14) yields

$$\xi = A\cos[\omega t - k(\hat{\boldsymbol{n}} \cdot \boldsymbol{r}) + \alpha]. \tag{14.15}$$

The vector

$$\boldsymbol{k} = k\hat{\boldsymbol{n}}, \tag{14.16}$$

equal in magnitude to the wave number $k = 2\pi/\lambda$ and directed along a normal to the wave surface is called the **wave vector**. Thus, Eq. (14.15) can be written in the form

$$\xi(\boldsymbol{r}, t) = A\cos(\omega t - \boldsymbol{k} \cdot \boldsymbol{r} + \alpha). \tag{14.17}$$

We have obtained the equation for a plane undamped wave propagating in the direction determined by the wave vector $\boldsymbol{k}$. For a damped wave, the factor $e^{-\gamma l} = e^{-\gamma(\hat{\boldsymbol{n}} \cdot \boldsymbol{r})}$ must be added to the equation.

Function (14.17) gives the deviation of a point having the position vector $\boldsymbol{r}$ from its equilibrium position at the moment of time t (we remind our reader that $\boldsymbol{r}$ determines the equilibrium position of the point). To pass over from the position vector of a point to its coordinates x, y, z, let us express the scalar product $\boldsymbol{k} \cdot \boldsymbol{r}$ through the components of the vectors along the coordinate axes:

$$\boldsymbol{k} \cdot \boldsymbol{r} = k_x x + k_y y + k_z z.$$

The equation of a plane wave, therefore, becomes

$$\xi(x, y, z; t) = A\cos\left(\omega t - k_x x - k_y y - k_z z + \alpha\right). \tag{14.18}$$

Here,

$$k_x = \frac{2\pi}{\lambda}\cos\alpha, \quad k_y = \frac{2\pi}{\lambda}\cos\beta, \quad k_z = \frac{2\pi}{\lambda}\cos\gamma. \tag{14.19}$$

Function (14.18) gives the deviation of a point having the coordinates x, y, z at the moment of time t. When $\hat{\boldsymbol{n}}$ coincides with $\hat{\boldsymbol{e}}_x$, we have $k_x = x$, $k_y = k_z = 0$, and Eq. (14.18) transforms into Eq. (14.10). It is very convenient to write the equation of a plane wave in the form

$$\xi = \mathfrak{R}\left[Ae^{i(\omega t - \boldsymbol{k} \cdot \boldsymbol{r} + \alpha)}\right]. \tag{14.20}$$

The symbol $\mathfrak{R}$ is usually omitted, having in mind that only the real part of the relevant expression is taken. In addition, the complex number

$$\hat{A} = Ae^{i\alpha}, \tag{14.21}$$

called the **complex amplitude** is introduced. The magnitude of this number gives the amplitude, and the argument, the initial phase of the wave.

Thus, the equation of a plane undamped wave can be written in the form

$$\xi = \hat{A}e^{i(\omega t - \boldsymbol{k} \cdot \boldsymbol{r})}. \tag{14.22}$$

The advantages of writing the equation in this form will come to light later.

14.4. The Wave Equation

The equation of any wave is the solution of a differential equation called the **wave equation**. To establish the form of the wave equation, let us compare the second partial derivatives with respect to the coordinates and time of function (14.18) describing a plane wave. Differentiating this function twice with respect to each of the variables, we get

$$\frac{\partial^2 \xi}{\partial t^2} = -\omega^2 A \cos(\omega t - \boldsymbol{k}\cdot\boldsymbol{r} + \alpha) = -\omega^2 \xi,$$
$$\frac{\partial^2 \xi}{\partial x^2} = -k_x^2 A \cos(\omega t - \boldsymbol{k}\cdot\boldsymbol{r} + \alpha) = -k_x^2 \xi,$$
$$\frac{\partial^2 \xi}{\partial y^2} = -k_y^2 A \cos(\omega t - \boldsymbol{k}\cdot\boldsymbol{r} + \alpha) = -k_y^2 \xi,$$
$$\frac{\partial^2 \xi}{\partial t^2} = -k_z^2 A \cos(\omega t - \boldsymbol{k}\cdot\boldsymbol{r} + \alpha) = -k_z^2 \xi.$$

Summation of the derivatives with respect to the coordinates yields

$$\frac{\partial^2 \xi}{\partial x^2} + \frac{\partial^2 \xi}{\partial y^2} + \frac{\partial^2 \xi}{\partial z^2} = -(k_x^2 + k_y^2 + k_z^2)\xi = -k^2 \xi. \tag{14.23}$$

Comparing this sum with the time derivative and substituting $1/v^2$ for k^2/ω^2 [see Eq. (14.9)], we get the equation

$$\frac{\partial^2 \xi}{\partial x^2} + \frac{\partial^2 \xi}{\partial y^2} + \frac{\partial^2 \xi}{\partial z^2} = \frac{1}{v^2}\frac{\partial^2 \xi}{\partial t^2}. \tag{14.24}$$

This is exactly the wave equation. It can be written in the form

$$\Delta \xi = \frac{1}{v^2}\frac{\partial^2 \xi}{\partial t^2}, \tag{14.25}$$

where Δ is the Laplacian operator [see Eq. (1.104)].

It is easy to convince ourselves that the wave equation is satisfied not only by function (14.18), but also by any function of the form

$$f(x, y, z; t) = f(\omega t - k_x x - k_y y - k_z z + \alpha). \tag{14.26}$$

Indeed, denoting the expression in parentheses in the right-hand side of Eq. (14.26) by ζ, we have

$$\frac{\partial \zeta}{\partial t} = \frac{\partial f}{\partial \zeta}\frac{\partial \zeta}{\partial t} = f'\omega, \quad \frac{\partial^2 f}{\partial t^2} = \omega \frac{\partial f'}{\partial \zeta}\frac{\partial \zeta}{\partial t} = \omega^2 f''. \tag{14.27}$$

Similarly,

$$\frac{\partial^2 f}{\partial x^2} = k_x^2 f'', \quad \frac{\partial^2 f}{\partial y^2} = k_y^2 f'', \quad \frac{\partial^2 f}{\partial z^2} = k_z^2 f''. \tag{14.28}$$

Introducing Eqs. (14.27) and (14.28) into Eq. (14.24), we arrive at the conclusion that

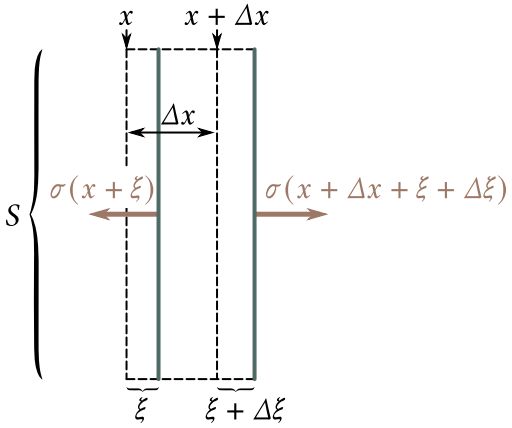


Fig. 14.6

function (14.26) satisfies the wave equation if we assume that $v = \omega/k$.

Any function satisfying an equation of the form of Eq. (14.24) describes a wave; the square root of the quantity that is the reciprocal of the coefficient of $\partial^2\xi/\partial t^2$ gives the phase velocity of this wave.

We must note that for a plane wave propagating along the x-axis, the wave equation has the form

$$\frac{\partial^2\xi}{\partial x^2} = \frac{1}{v^2}\frac{\partial^2\xi}{\partial t^2}. \tag{14.29}$$

14.5. Velocity of Elastic Waves in a Solid Medium

Assume that a longitudinal plane wave propagates in the direction of the x-axis. Let us separate in the medium a cylindrical volume with a base area of S and a height of Δx (Fig. 14.6). The displacements s of particles with different x's are different at each moment of time (see Fig. 14.3 showing ξ against x). If the base of the cylinder with the coordinate x has at a certain moment of time the displacement ξ, then the displacement of a base with the coordinate $x + \Delta x$ will be $\xi + \Delta\xi$. Therefore, the volume being considered will be deformed—it receives the elongation $\Delta\xi$ ($\Delta\xi$ is an algebraic quantity, $\Delta\xi < 0$ corresponds to compression of the cylinder) or the relative elongation $\Delta\xi/\Delta x$. The quantity $\Delta\xi/\Delta x$ gives the average deformation of the cylinder. Since ξ varies with x according to a non-linear law, the true deformation in different cross sections of the cylinder will differ. To obtain the deformation (strain) in the cross section x, we must make Δx tend to zero. Thus,

$$\varepsilon = \frac{\partial\xi}{\partial x} \tag{14.30}$$

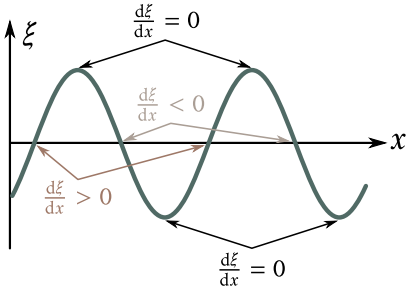


Fig. 14.7

(we have used the symbol of the partial derivative because ξ depends not only on x, but also on t).

The presence of tensile strain points to the existence of the normal stress σ which at small strains is proportional to the strain. According to Eq. (2.30) of Vol. I,

$$\sigma = E\varepsilon = E = \frac{\partial \xi}{\partial x} \tag{14.31}$$

(E is Young's modulus of the medium). We must note that the unit strain $\partial\xi/\partial x$ and, consequently, the stress σ at a fixed moment of time depend on x (Fig. 14.7). Where the deviations of the particles from their equilibrium position are maximum, the strain and the stress are zero. Where the particles are passing through their equilibrium position, the strain and stress reach their maximum values, the positive and negative strains (*i.e.*, tensions and compressions) alternating. Accordingly, as we have already noted in Sec. 14.1, a longitudinal wave consists of alternating compressions and dilatations of the medium.

Let us revert to the cylindrical volume depicted in Fig. 14.6 and write an equation of motion for it. Assuming that Δx is very small, we can consider that the projection of the acceleration onto the x-axis is the same for all points of the cylinder and is $\partial\xi/\partial x$. The mass of the cylinder is $\rho S \Delta x$, where ρ is the density of the undeformed medium. The projection onto the x-axis of the force acting on the cylinder equals the product of the area S of the cylinder base and the difference between the normal stresses in the cross sections $(x + \Delta x + \xi + \Delta\xi)$ and $(x + \xi)$:

$$F_x = SE\left[\left(\frac{\partial \xi}{\partial x}\right)_{x+\Delta x+\xi+\Delta\xi} - \left(\frac{\partial \xi}{\partial x}\right)_{x+\xi}\right]. \tag{14.32}$$

The value of the derivative $\partial\xi/\partial x$ in the section $X + \delta$ can be written with great accuracy for small values of δ in the form

$$\left(\frac{\partial \xi}{\partial x}\right)_{x+\delta} = \left(\frac{\partial \xi}{\partial x}\right)_x + \left[\frac{\partial}{\partial x}\left(\frac{\partial \xi}{\partial x}\right)\right]_x \delta = \left(\frac{\partial \xi}{\partial x}\right)_x + \frac{\partial^2 \xi}{\partial x^2}\,\delta, \tag{14.33}$$

where by $\partial^2\xi/\partial x^2$ is meant the value of the second partial derivative of ξ with respect

to x in the cross section x.

Owing to the smallness of the quantities Δx, S, and $\Delta\xi$, we can perform transformation (14.33) in Eq. (14.32):

$$F_x = SE\left\{\left[\left(\frac{\partial\xi}{\partial x}\right)_x + \frac{\partial^2\xi}{\partial x^2}(\Delta x + \xi + \Delta\xi)\right] - \left[\left(\frac{\partial\xi}{\partial x}\right)_x + \frac{\partial^2\xi}{\partial x^2}\xi\right]\right\}$$
$$= SE\frac{\partial^2\xi}{\partial x^2}(\Delta x + \Delta\xi) \approx SE\frac{\partial^2\xi}{\partial x^2}\Delta x \tag{14.34}$$

[the relative elongation $\partial\xi/\partial x$ in elastic deformations is much smaller than unity. Consequently, $\Delta\xi \ll \Delta x$ so that the addend $\Delta\xi$ in the sum $(\Delta x + \Delta\xi)$ may be disregarded].

Introducing the found values of the mass, acceleration, and force into the equation of Newton's second law, we get

$$\rho S\Delta x\frac{\partial^2\xi}{\partial t^2} = SE\frac{\partial^2\xi}{\partial x^2}\Delta x.$$

Finally, cancelling $S\Delta x$, we arrive at the equation

$$\frac{\partial^2\xi}{\partial x^2} = \frac{\rho}{E}\frac{\partial^2\xi}{\partial t^2}, \tag{14.35}$$

which is the wave equation for the case when ξ is independent of y and z. A comparison of Eqs. (14.29) and (14.35) shows that

$$v = \left(\frac{E}{\rho}\right)^{1/2}, \tag{14.36}$$

Thus, the phase velocity of longitudinal elastic waves equals the square root of Young's modulus divided by the density of the medium.

Similar calculations for transverse waves lead to the expression

$$v = \left(\frac{G}{\rho}\right)^{1/2}, \tag{14.37}$$

where G is the shear modulus.

14.6. Energy of an Elastic Wave

Assume that the plane longitudinal wave,

$$\xi = A\cos(\omega t - kx + \alpha)$$

[see Eq. (14.10)], is propagating in the direction of the x-axis in a certain medium. Let us separate in this medium an elementary volume ΔV so small that the velocity and the strain at all the points of this volume may be considered the same and equal, respectively, to $\partial\xi/\partial t$ and $\partial\xi/\partial x$.

The volume we have separated has the kinetic energy

$$\Delta W_{\mathrm{k}} = \frac{\rho}{2}\left(\frac{\partial \xi}{\partial t}\right)^2 \Delta V \tag{14.38}$$

($\rho \Delta V$ is the mass of the volume, and $\partial \xi / \partial t$ is its velocity).

According to Eq. (3.81) of Vol. I, the volume being considered also has the potential energy of elastic deformation

$$\Delta W_{\mathrm{p}} = \frac{E\varepsilon^2}{2}\Delta V = \frac{E}{2}\left(\frac{\partial \xi}{\partial x}\right)^2 \Delta V$$

($\varepsilon = \partial \xi / \partial x$ is the relative elongation of the cylinder, E is Young's modulus of the medium). Let us use Eq. (14.36) to substitute ρv^2 for Young's modulus (ρ is the density of the medium, and v is the phase velocity of the wave). Hence, the expression for the potential energy of the volume ΔV acquires the form

$$\Delta W_{\mathrm{p}} = \frac{E\varepsilon^2}{2}\left(\frac{\partial \xi}{\partial x}\right)^2 \Delta V. \tag{14.39}$$

The sum of Eqs. (14.38) and (14.39) gives the total energy

$$\Delta W = \Delta W_{\mathrm{k}} + \Delta W_{\mathrm{p}} = \frac{1}{2}\rho\left[\left(\frac{\partial \xi}{\partial t}\right)^2 + v^2\left(\frac{\partial \xi}{\partial x}\right)^2\right]\Delta V. \tag{14.40}$$

Dividing this energy by the volume ΔV in which it is contained, we get the energy density

$$w = \frac{1}{2}\rho\left[\left(\frac{\partial \xi}{\partial t}\right)^2 + v^2\left(\frac{\partial \xi}{\partial x}\right)^2\right]. \tag{14.41}$$

Differentiation of Eq. (14.10) once with respect to t and another time with respect to x yields:

$$\frac{\partial \xi}{\partial t} = -A\omega \sin(\omega t - kx + \alpha),$$

$$\frac{\partial \xi}{\partial x} = kA \sin(\omega t - kx + \alpha).$$

Introducing these equations into Eq. (14.41) and taking into account that $k^2 v^2 = \omega^2$, we get

$$w = \rho A^2 \omega^2 \sin^2(\omega t - kx + \alpha). \tag{14.42}$$

A similar expression for the energy density is obtained for a transverse wave.

It can be seen from Eq. (14.42) that the energy density at each moment of time is different at different points of space. At the same point, the energy density varies with time as the square of the sine. The average value of sine square is one-half. Accordingly, the time-averaged value of the energy density at each point of a medium

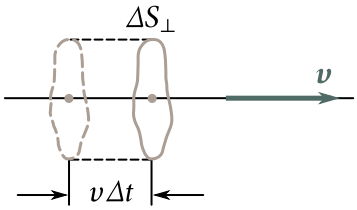


Fig. 14.8

is

$$\langle w \rangle = \frac{1}{2}\rho A^2 \omega^2. \tag{14.43}$$

The energy density given by Eq. (14.42) and its average value [Eq. (14.43)] are proportional to the density of the medium ρ, the square of the frequency ω, and the square of the wave amplitude A. Such a relation holds not only for an undamped plane wave, but also for other kinds of waves (a plane damped wave, a spherical wave, etc.).

Thus, a medium in which a wave is propagating has an additional store of energy. The latter is supplied to the different points of the medium from the source of oscillations by the wave itself; consequently, a wave carries energy with it. The amount of energy carried by a wave through a surface in unit time is called the **energy ftux** through this surface. If the energy $\mathrm{d}W$ is carried through a given surface during the time $\mathrm{d}t$, then the energy flux Φ is

$$\Phi = \frac{\mathrm{d}W}{\mathrm{d}t}. \tag{14.44}$$

The energy flux is a scalar quantity whose dimension equals that of energy divided by the dimension of time, *i.e.*, coincides with the dimension of power. Accordingly, Φ is measured in watts, erg s^{-1}, etc.

The energy flux at different points of a medium can have a different intensity. To characterize the flow of energy at different points of space, a vector quantity called the **density of the energy flux** is introduced. It numerically equals the energy flux through a unit area placed at the given point perpendicular to the direction in which the energy is being transferred. The direction of the vector of the energy flux density coincides with that of energy transfer.

Assume that the energy $\mathrm{d}W$ is transferred during the time $\mathrm{d}t$ through the area $\Delta S_\perp$ perpendicular to the direction of propagation of a wave. The energy flux density will therefore be

$$j = \frac{\Delta\Phi}{\Delta S_\perp} = \frac{\Delta W}{\Delta S_\perp \Delta t} \tag{14.45}$$

[see Eq. (14.44)]. The energy ΔW confined in a cylinder with the base $\Delta S_\perp$ and the

altitude $v\Delta t$ (v is the phase velocity of the wave) will be transferred through the area $\Delta S_\perp$ (Fig. 14.8) during the time Δt. If the dimensions of the cylinder are sufficiently small (as a result of the smallness of $\Delta S_\perp$ and Δt) to consider that the energy density at all points of the cylinder is the same, then ΔW can be found as the product of the energy density w and the volume of the cylinder equal to $\Delta S_\perp v\Delta t$:

$$\Delta W = w\Delta S_\perp v\Delta t.$$

Using this expression in Eq. (14.45), we get the following equation for the density of the energy:

$$j = wv. \tag{14.46}$$

Finally, introducing the vector $\boldsymbol{v}$ whose magnitude equals the phase velocity of the wave and whose direction coincides with that of wave propagation (and energy transfer), we can write

$$\boldsymbol{j} = w\boldsymbol{v}. \tag{14.47}$$

We have obtained an expression for the vector of the energy flux density. This vector was first introduced by the outstanding Russian physicist Nikolai Umov (1846-1915) and is called **Umov's vector**.

The vector given by Eq. (14.47), like the energy density w, is different at different points of space. At a given point, it varies in time according to a sine square law. Its average value is

$$\langle \boldsymbol{j} \rangle = \langle w \rangle \boldsymbol{v} = \frac{1}{2}\rho A^2\omega^2\boldsymbol{v} \tag{14.48}$$

[see Eq. (14.43)]. Equation (14.48), like Eq. (14.43), holds for a wave of any kind (spherical, damped, etc.). We shall note that when we speak of the intensity of a wave at a given point, we have in mind the time-averaged value of the density of the energy flux transferred by the wave.

Knowing $\boldsymbol{j}$ for all the points of an arbitrary surface S, we can calculate the energy flux through this surface. For this purpose, let us divide the surface into elementary areas $\mathrm{d}S$. During the time $\mathrm{d}t$, the energy $\mathrm{d}W$ confined in the oblique cylinder shown in Fig. 14.9 will pass through area $\mathrm{d}S$. The volume of this cylinder is $\mathrm{d}V = v\,\mathrm{d}t\,\mathrm{d}S\,\cos\varphi$. It contains the energy $\mathrm{d}W = w\,\mathrm{d}V = wv\,\mathrm{d}t\,\mathrm{d}S\,\cos\varphi$ (here, w is the instantaneous value of the energy density where area $\mathrm{d}S$ is). Taking into account that

$$wv\,\mathrm{d}S\,\cos\varphi = j\,\mathrm{d}S\,\cos\varphi = \boldsymbol{j}\cdot\mathrm{d}\boldsymbol{S}$$

($\mathrm{d}\boldsymbol{S} = \hat{\boldsymbol{n}}\,\mathrm{d}S$; see Fig. 14.9), we can write: $\mathrm{d}W = \boldsymbol{j}\cdot\mathrm{d}\boldsymbol{S}\,\mathrm{d}t$. Hence, we obtain the following equation for the energy flux $\mathrm{d}\Phi$ through area $\mathrm{d}S$:

$$\mathrm{d}\Phi = \frac{\mathrm{d}W}{\mathrm{d}t} = \boldsymbol{j}\cdot\mathrm{d}\boldsymbol{S} \tag{14.49}$$

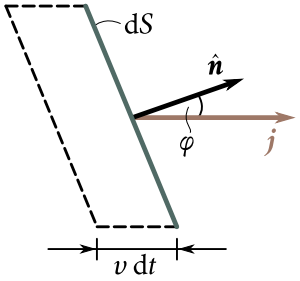


Fig. 14.9

[compare with Eq. (1.72)]. The total energy flux through a surface equals the sum of the elementary fluxes given by Eq. (14.49):

$$\Phi = \int_S \boldsymbol{j} \cdot \mathrm{d}\boldsymbol{S}. \tag{14.50}$$

We can say in accordance with Eq. (1.74) that the energy flux equals the flux of the vector $\boldsymbol{j}$ through surface S.

Substituting for the vector $\boldsymbol{j}$ in Eq. (14.50) its time-averaged value, we get the average value of Φ:

$$\langle \Phi \rangle = \int_S \langle \boldsymbol{j} \rangle \cdot \mathrm{d}\boldsymbol{S}. \tag{14.51}$$

Let us calculate the mean value of the energy flux through an arbitrary wave surface of an undamped spherical wave. At each point of this surface, the vectors $\boldsymbol{j}$ and $\mathrm{d}\boldsymbol{S}$ coincide in direction. In addition, the magnitude of the vector $\boldsymbol{j}$ for all points of the surface is identical. Hence,

$$\langle \Phi \rangle = \int_S \langle j \rangle \, \mathrm{d}S = \langle j \rangle \, S = \langle j \rangle \, 4\pi r^2$$

(r is the radius of the wave surface). According to Eq. (14.48), we have $\langle j \rangle = \rho A^2 \omega^2 v/2$. Thus,

$$\langle \Phi \rangle = 2\pi \rho \omega^2 A_r^2 r^2$$

(A_r is the amplitude of the wave at a distance r from its source). Since the energy of the wave is not absorbed by the medium, the average energy flux through a sphere of any radius must have the same value, *i.e.*, the condition

$$A_r^2 r^2 = \text{constant}$$

must be observed. It follows that the amplitude A_r of an undamped spherical wave is inversely proportional to the distance r from the wave source [see Eq. (14.12)]. Accordingly, the mean density of the energy flux $\langle j \rangle$ is inversely proportional to the square of the distance from the source.

For a plane damped wave, the amplitude diminishes with the distance according

to the law $A = A_0\, e^{-\gamma x}$ [see Eq. (14.11)]. The average density of the energy flux (*i.e.*, the wave intensity) correspondingly diminishes according to the law

$$j = j_0\, e^{-\varkappa x}. \tag{14.52}$$

Here, $\varkappa = 2\gamma$ is a quantity called the **wave absorption coefficient**. Its dimension is the reciprocal of that of length. It is easy to see that the reciprocal of $\varkappa$ equals the distance over which the intensity of a wave diminishes to $1/e$ of its initial value.

14.7. Standing Waves

If several waves propagate in a medium simultaneously, then the oscillations of the particles of the medium will be the geometrical sum of the oscillations which the particles would perform if each of the waves propagated separately. Hence, the waves are simply superposed onto one another without disturbing one another. This statement following from experiments is called the **principle of superposition of waves**.

When the oscillations due to separate waves at each point of a medium have a constant phase difference, the waves are called **coherent**. (A stricter definition of coherence will be given in Sec. 17.2). The summation of coherent waves gives rise to the phenomenon of **interference**, consisting in that the oscillations at some points amplify, and at other points weaken one another.

A very important case of interference is observed in the superposition of two plane waves having the same amplitude and approaching each other from opposite directions. The resulting oscillatory process is called a **standing wave**. Standing waves are produced when waves are reflected from obstacles. The wave striking an obstacle and the reflected wave travelling toward it in the opposite direction as a result of superposition produce a standing wave.

Let us write the equations of two plane waves propagating along the x-axis in opposite directions:

$$\xi_1 = A\cos(\omega t - kx + \alpha_1),$$
$$\xi_2 = A\cos(\omega t + kx + \alpha_2).$$

Adding these two equations and transforming the result according to the formula for the sum of cosines, we get

$$\xi = \xi_1 + \xi_2 = 2A\cos\left[kx + \left(\frac{\alpha_2 - \alpha_1}{2}\right)\right]\cos\left[\omega t + \left(\frac{\alpha_1 + \alpha_2}{2}\right)\right]. \tag{14.53}$$

Equation (14.53) is the equation of a standing wave. To simplify it, let us choose the beginning of reading x so that the difference $\alpha_2 - \alpha_1$ vanishes, and the beginning of reading t so that the sum $\alpha_1 + \alpha_2$ vanishes. We shall also substitute for the wave

number k its value $2\pi/\lambda$. Equation (14.53) now becomes

$$\xi = \left[2A\cos\left(\frac{2\pi x}{\lambda}\right)\right]\cos(\omega t). \tag{14.54}$$

A glance at Eq. (14.54) shows that at every point of a standing wave the oscillations have the same frequency as those of the opposite waves, the amplitude depending on x:

$$\text{amplitude} = \left|2A\cos\left(\frac{2\pi x}{\lambda}\right)\right|.$$

At the points whose coordinates comply with the condition

$$\frac{2\pi x}{\lambda} = \pm n\pi \quad (n = 0, 1, 2, \ldots), \tag{14.55}$$

the amplitude of the oscillations reaches its maximum value. These points are known as **antinodes** of the standing wave. We obtain the values of the antinode coordinates from Eq. (14.55):

$$x_{\text{anti}} = \pm n\frac{\lambda}{2} \quad (n = 0, 1, 2, \ldots). \tag{14.56}$$

It must be borne in mind that an antinode is not a single point, but a plane whose points have the value of the coordinate x determined by Eq. (14.56).

At the points whose coordinates comply with the condition

$$\frac{2\pi x}{\lambda} = \pm\left(n + \frac{1}{2}\right)\pi \quad (n = 0, 1, 2, \ldots),$$

the amplitude of the oscillations vanishes. These points are called the **nodes** of the standing wave. The points of the medium at the nodes do not oscillate. The coordinates of the nodes have the values

$$x_{\text{node}} = \pm\left(n + \frac{1}{2}\right)\frac{\lambda}{2} \quad (n = 0, 1, 2, \ldots). \tag{14.57}$$

A node, like an antinode, is not a single point, but a plane whose points have values of the coordinate x determined by Eq. (14.57).

Examination of Eqs. (14.56) and (14.57) shows that the distance between adjacent antinodes, like that between adjacent nodes, is $\lambda/2$. The antinodes and nodes are displaced relative to one another by a quarter of a wavelength.

Let us revert to Eq. (14.54). The factor $2A\cos(2\pi x/\lambda)$ changes its sign after passing through its zero value. Accordingly, the phase of the oscillations at different sides of a node differs by n. This signifies that points at different sides of a node oscillate in counterphase. All the points between two adjacent nodes oscillate in phase. Figure 14.10 contains a number of "instantaneous photographs" of the deviations of the points from their equilibrium position. The first of them corresponds to the moment when the deviations reach their greatest absolute value. The following

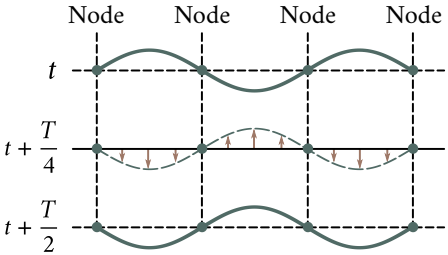


Fig. 14.10

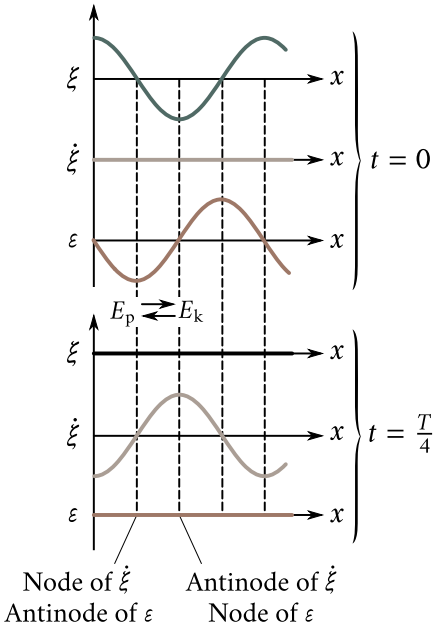


Fig. 14.11

"photographs" have been made at intervals of one-fourth of a period. The arrows show the velocities of the particles.

Differentiating Eq. (14.54) once with respect to t and once with respect to x, we find expressions for the velocity of the particles $\dot{\xi}$ and the deformation of the medium ε:

$$\dot{\xi} = \frac{\mathrm{d}\xi}{\mathrm{d}t} = -2\omega A \cos\left(\frac{2\pi x}{\lambda}\right) \sin(\omega t), \tag{14.58}$$

$$\varepsilon = \frac{\mathrm{d}\xi}{\mathrm{d}x} = -2\frac{2\pi}{\lambda} A \sin\left(\frac{2\pi x}{\lambda}\right) \cos(\omega t). \tag{14.59}$$

Equation (14.58) describes a standing wave of velocity, and Eq. (14.59) one of deformation.

Figure 14.11 compares "instantaneous photographs" of the displacement, velocity, and deformation for the time moments 0 and $T/4$. Inspection of the graphs shows that the nodes and antinodes of the velocity coincide with their displacement counterparts; the nodes and antinodes of the deformation, however, coincide with the antinodes and nodes of the displacement, respectively. When ξ and ε reach their maximum values, $\dot{\xi}$ becomes equal to zero, and vice versa. Accordingly, the energy of a standing wave transforms twice during a period, once completely into potential energy mainly concentrated near the nodes of the wave (where the deformation

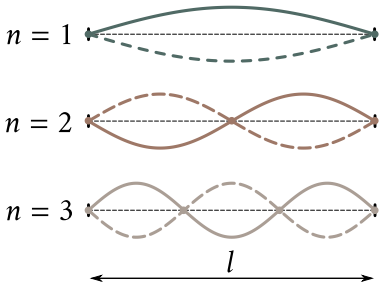


Fig. 14.12

antinodes are), and once completely into kinetic energy mainly concentrated near the antinodes of the wave (where the antinodes of the velocity are). The result is the transition of energy from each node to its adjacent antinodes and back. The time-averaged energy flux in any cross section of the wave is zero.

14.8. Oscillations of a String

When transverse oscillations are produced in a stretched string fastened at both ends, standing waves are set up in it, and there must be nodes at the places where the string is fastened. Hence, only such oscillations are produced with an appreciable intensity in a string when the length of the latter is an integer multiple of half their wavelength (Fig. 14.12). This gives the condition

$$l = n\frac{\lambda}{2} \quad \text{or} \quad \lambda_{\mathrm{n}} = \frac{2l}{n} \quad (n = 1, 2, 3, \ldots) \tag{14.60}$$

(l is the length of the string). The following frequencies correspond to the wavelengths given by Eq. (14.60):

$$\nu_{\mathrm{n}} = \frac{v}{\lambda_{\mathrm{n}}} = \frac{v}{2l}n \quad (n = 1, 2, 3, \ldots) \tag{14.61}$$

(v is the phase velocity of the wave determined by the string tension and the mass per unit length, *i.e.*, the linear density of the string).

The frequencies ν_{n} are called the **natural frequencies** of the string. The natural frequencies are integral multiples of the frequency

$$\nu_{\mathrm{i}} = \frac{v}{2l},$$

called the **fundamental frequency**.

Harmonic oscillations with frequencies according to Eq. (14.61) are called **natural** or **normal oscillations**. They are also known as **harmonics**. In the general case, the oscillation of a string is a superposition of various harmonics.

The oscillations of a string are remarkable in the respect that according to

classical notions, we get discrete values of one of the quantities characterizing the oscillations (their frequency). Such a discrete nature is an exception for classical physics. For quantum processes, it is the rule rather than an exception.

14.9. Sound

If elastic waves propagating in air have a frequency ranging from 16 Hz to 20000 Hz, then upon reaching the human ear, they cause a sound to be perceived. Accordingly, elastic waves in any medium having a frequency confined within the above limits are called **sound waves** or simply **sound**. Elastic waves with frequencies below 16 Hz are called **infrasound**, and those with frequencies above 20000 Hz are called **ultrasound**. The human ear does not hear infra- and ultrasounds.

People distinguish sounds they hear by **pitch**, **timbre** (**quality**), and **loudness**. A definite physical characteristic of a sound wave corresponds to each of these subjective appraisals.

Any real sound is not a simple harmonic oscillation, but is the superposition of harmonic oscillations with a definite set of frequencies. The collection of frequencies of the oscillations present in a given sound is called its **acoustic spectrum**. If a sound contains oscillations of all the frequencies within an interval from ν' to ν'', then, the spectrum is called **continuous**. If a sound consists of oscillations having the discrete frequencies ν_1, ν_2, ν_3, etc., then, the spectrum is known as **a line one**. Noises have a continuous acoustic spectrum. Oscillations with a line spectrum produce the sensation of a sound with a more or less definite pitch. Such a sound is called a **tone sound**, or simply a **tone**.

The pitch of a tone is determined by its fundamental (lowest) frequency. The relative intensity of the **overtones** (*i.e.*, of the oscillations of the frequencies ν_2, ν_3, etc.) determines the timbre, or quality, of the sound. The different spectral composition of sounds produced by various musical instruments makes it possible to distinguish by ear, for example, a flute from a violin or a piano.

By the intensity of a sound is meant the time-averaged value of the density of the energy flux carried by a sound wave. To be audible, a wave must have a certain minimum intensity known as the **threshold of hearing**. This threshold differs somewhat for different persons and depends quite greatly on the frequency of the sound. The human ear is most sensitive to frequencies from 1000 Hz to 4000 Hz. In this region of frequencies, the threshold of hearing averages about 10^{-12} W m^{-2}. At other frequencies, it is higher (see the bottom curve in Fig. 14.13).

At intensities of the order of 1 W m^{-2} to 10 W m^{-2}, a wave stops being perceived as a sound and produces only a feeling of pain and pressure in the ear. The value of the intensity at which this occurs is known as the **threshold of pain** (or the

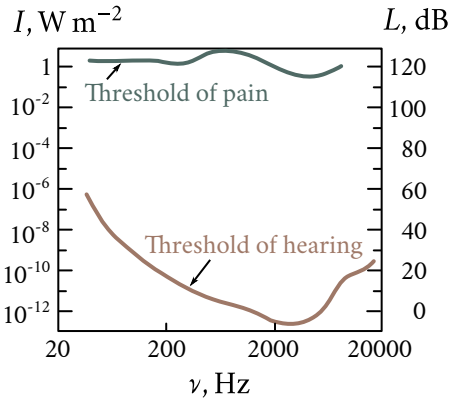


Fig. 14.13

threshold of feeling). The pain threshold, like the hearing one, depends on the frequency (see the top curve in Fig. 14.13; the data given in this figure relate to the average normal hearing).

The subjectively estimated loudness of a sound grows much more slowly than the intensity of the sound waves. When the intensity grows in a geometric progression, the loudness grows approximately in an arithmetical progression, *i.e.*, linearly. On these grounds, the **loudness level** L is determined as the logarithm of the ratio between the intensity of the given sound I and the intensity I_0 taken as the initial one:

$$L = \log\left(\frac{I}{I_0}\right). \tag{14.62}$$

The initial intensity I_0 is taken equal to 10^{-12} W m^{-2} so that the hearing threshold at a frequency of the order of 1000 Hz is at the zero level ($L = 0$).

The unit of loudness level L determined by Eq. (14.62) is called the **bell** (B). Generally the **decibel** (dB), which is one-tenth of a bell is preferred. The value of L in decibels is determined by the equation

$$L = 10\log\left(\frac{I}{I_0}\right). \tag{14.63}$$

The ratio of two intensities I_1 and I_2 can also be expressed in decibels:

$$L_{12} = 10\log\left(\frac{I_1}{I_2}\right). \tag{14.64}$$

This equation can be used to express the reduction in the intensity (the damping) of a wave over a certain path in decibels. Thus, for example, a damping of 20 dB signifies that the intensity has dropped to one-hundredth of its initial value.

The entire range of intensities at which a wave produces a feeling of sound

in the human ear (from 10^{-12} W m^{-2}), corresponds to values of the loudness level from 0 dB to 130 dB. Table 14.1 gives approximate values of the loudness level for selected sounds.

The energy which sound waves convey with them is extremely small. If we assume, for example, that a glass of water completely absorbs the entire energy of a sound wave with a loudness level of 70 dB falling on it (in this case the amount of energy absorbed per second will be about 2×10^{-7} W), then, to heat the water from room temperature to boiling about ten thousand years will be needed.

Ultrasonic waves can be produced in the form of directed beams like beams of light. Directed ultrasonic beams have found a widespread application for locating objects and determining the distance to them in water. The first to put forward the idea of ultrasonic location was the outstanding French physicist Paul Langevin. He implemented this idea during the first world war for detecting submarines.

At present, ultrasonic locators are used for detecting icebergs, fish shoals, and the like.

It is general knowledge that by shouting and determining the time that elapses until the echo arrives, *i.e.*, the sound reflected by an obstacle—a mountain, forest, the surface of the water in a well, etc.—we can find the distance to the obstacle by multiplying half of this time by the speed of sound. This principle underlies the locator (sonar) mentioned above, and also the ultrasonic echo sounder used to measure the depth and determine the relief of the sea bottom.

Ultrasonic location permits bats to orient themselves very well when flying in the dark. A bat periodically emits pulses of an ultrasonic frequency and according

Table 14.1

Sound	**Loudness level**, dB
Ticking of a clock	20
Whisper at a distance of 1 m	30
Quiet conversation	40
Speech of a moderate loudness	60
Loud speech	70
Shout	80
Noise of an aircraft engine:	
at a distance of 5 m	120
at a distance of 3 m	130

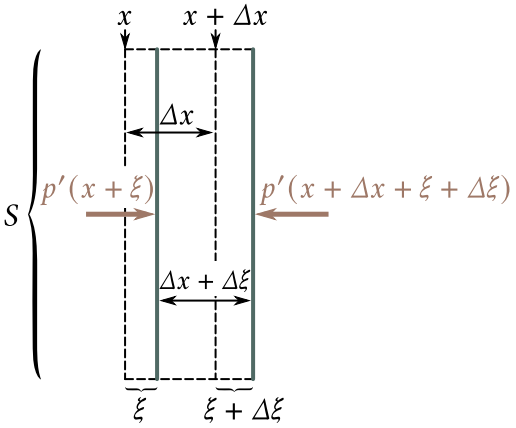


Fig. 14.14

to the reflected signals received by its ears assesses the distances to surrounding objects with a high accuracy.

14.10. The Velocity of Sound in Gases

A sound wave in a gas is a sequence of alternating regions of compression and rarefaction of the gas propagating in space. Hence, the pressure at every point of space experiences a periodically changing deflection Δp from its average value p coinciding with the pressure existing in the gas when waves are absent. Thus, the instantaneous value of the pressure at a point of space can be written in the form

$$p' = p + \Delta p.$$

Assume that a wave is propagating along the x-axis. Let us consider the volume of a gas in the form of a cylinder with a base area of S and an altitude of Δx (Fig. 14.14), as we did in Sec. 14.5 when finding the velocity of elastic waves in a solid medium. The mass of the gas confined in this volume is $\rho S \Delta x$, where ρ is the density of the gas undisturbed by the wave. Owing to the smallness of Δx, the projection of the acceleration onto the x-axis for all the points of the cylinder may be considered the same and equal to $\partial^2 \xi / \partial t^2$.

To find the projection onto the x-axis of the force exerted on the volume being considered, we must take the product of the cylinder base area S and the difference between the pressures in the cross sections $(x + \xi)$ and $(x + \Delta x + \xi + \Delta\xi)$. Repeating the reasoning that led us to Eq. (14.34), we get

$$F_x = -\frac{\partial p'}{\partial x} S \Delta x$$

[we remind our reader that when deriving Eq. (14.34) we took advantage of the assumption $\Delta\xi \ll \Delta x$].

Thus, we have found the mass of the separated volume of gas, its acceleration, and the force exerted on it. Now let us write the equation of Newton's second law for this volume of gas:

$$(\rho S\Delta x)\frac{\partial^2\xi}{\partial t^2} = -\frac{\partial p'}{\partial x}S\Delta x.$$

After cancelling $S\Delta x$, we get

$$\rho\frac{\partial^2\xi}{\partial t^2} = -\frac{\partial p'}{\partial x}. \tag{14.65}$$

The differential equation we have obtained contains two unknown functions, namely, ξ and p'. Let us express one of them through the other. To do this, we shall find the relation between the pressure of a gas and the relative change in its volume $\partial\xi/\partial x$. This relation depends on the nature of the process of compression (or rarefaction) of the gas. The compressions and rarefactions of a gas in a sound wave follow one another so frequently that adjacent portions of the medium do not manage to exchange heat, and the process can be considered as an adiabatic one. In an adiabatic process, the pressure and volume of a given mass of a gas are related by the equation

$$pV^\gamma = \text{constant}, \tag{14.66}$$

where γ is the ratio between the heat capacities of the gas at constant pressure and at constant volume [see Eq. (10.42) of Vol. I].

In accordance with Eq. (14.66):

$$p(S\Delta x)^\gamma = p'[S(\Delta x + \Delta\xi)]^\gamma = p'\left[S\left(\Delta x + \frac{\partial\xi}{\partial x}\Delta x\right)\right]^\gamma = p'(S\Delta x)^\gamma\left(1 + \frac{\partial\xi}{\partial x}\right)^\gamma.$$

Cancelling $(S\Delta x)^\gamma$ yeilds:

$$p = p'\left(1 + \frac{\partial\xi}{\partial x}\right)^\gamma.$$

Taking advantage of the assumption $\partial\xi/\partial x \ll 1$, let us expand the expression $(1 + \partial\xi/\partial x)^\gamma$ into a series by powers of $\partial\xi/\partial x$ and disregard the terms of the higher orders of smallness. The result is

$$p = p'\left(1 + \gamma\frac{\partial\xi}{\partial x}\right).$$

Let us solve this equation with respect to p':

$$p' = \frac{p}{\left(1 + \gamma\dfrac{\partial\xi}{\partial x}\right)} \approx p\left(1 - \gamma\frac{\partial\xi}{\partial x}\right) \tag{14.67}$$

[we have used the formula $1/(1+x) \approx 1 - x$ holding for $x \ll 1$]. It is a simple matter to obtain an expression for Δp from the relation we have found:

$$\Delta p = p' - p = -\gamma p \frac{\partial \xi}{\partial x}. \tag{14.68}$$

Since the order of magnitude of γ is near unity, it follows from Eq. (14.68) that $|\partial\xi/\partial x| \approx |\Delta p/p|$. Thus, the condition, $\partial\xi/\partial x \ll 1$, signifies that the deviation of the pressure from its average value is much smaller than the pressure itself. This is indeed true: for the loudest sounds, the amplitude of oscillations of the air pressure does not exceed 1 mmHg, whereas the atmospheric pressure p has a value of the order of 10^3 mmHg.

Differentiating Eq. (14.67) with respect to x, we find that

$$\frac{\partial p'}{\partial x} = -\gamma p \frac{\partial^2 \xi}{\partial x^2}.$$

Finally, using this value of $\partial p'/\partial x$ in Eq. (14.65), we get the differential equation

$$\frac{\partial^2 \xi}{\partial x^2} = \frac{\rho}{\gamma p} \frac{\partial^2 \xi}{\partial t^2}.$$

Comparing it with wave equation (14.29), we get the following expression for the velocity of sound waves in a gas:

$$v = \left(\gamma \frac{p}{\rho}\right)^{1/2} \tag{14.69}$$

(we remind our reader that p and ρ are the pressure and the density of the gas undisturbed by a wave).

At atmospheric pressure and conventional temperatures, most gases are close in their properties to an ideal gas. Therefore, we can assume that the ratio p/ρ for them equals RT/M, where R is the molar gas constant, T is the absolute temperature, and M is the mass of a mole of a gas [see Eq. (10.22) of Vol. I]. Introducing this value into Eq. (14.69), we get the following equation for the velocity of sound in a gas:

$$v = \left(\frac{\gamma RT}{M}\right)^{1/2}. \tag{14.70}$$

Examination of this equation shows that the velocity of sound is proportional to the square root of the temperature and does not depend on the pressure.

The average velocity of thermal motion of gas molecules is determined by the formula

$$\langle v_{\text{mol}} \rangle = \left(\frac{8RT}{\pi M}\right)^{1/2}$$

[see Eq. (11.70) of Vol. I]. A comparison of this equation with Eq. (14.70) shows that the velocity of sound in a gas is related to the average velocity of thermal motion of

its molecules by the expression

$$v = \langle v_{\text{mol}} \rangle \left(\frac{\gamma\pi}{8} \right)^{1/2}. \tag{14.71}$$

Substitution for γ of its value for air equal to 1.4 yields the expression $v \approx 3 \langle v_{\text{mol}} \rangle / 4$. The maximum possible value of γ is 5/3. In this case, $v \approx 4 \langle v_{\text{mol}} \rangle / 5$. Thus, the velocity of sound in a gas is of the same order of magnitude as the average velocity of thermal motion of the molecules, but is always somewhat lower than $\langle v_{\text{mol}} \rangle$.

Let us calculate the value of the velocity of sound in air at a temperature of 290 K (room temperature). For air, we have $\gamma = 1.40$, and $M = 29 \times 10^{-3}$ kg mol^{-1}. The molar gas constant is $R = 8.31$ J mol^{-1} K^{-1}. Introducing these values into Eq. (14.70), we get

$$v = \left(\frac{\gamma RT}{M} \right)^{1/2} = \left(\frac{1.4 \times 8.31 \times 290}{29 \times 10^{-3}} \right)^{1/2} = 340 \text{ m s}^{-1}.$$

The value of the sound velocity in air which we have found agrees quite well with the value found experimentally.

Let us find the relation between the intensity of a sound wave I and the amplitude of the pressure oscillations $(\Delta p)_{\text{m}}$. We mentioned in Sec. 14.9 that by the intensity of sound is meant the average value of the density of the energy flux. Hence,

$$I = \frac{1}{2} \rho A^2 \omega^2 v \tag{14.72}$$

[see Eq. (14.48)]. Here, p is the density of the undisturbed gas, A is the amplitude of oscillations of the particles of the medium, *i.e.*, the amplitude of the oscillations of the displacement ξ, ω is the frequency, and v the phase velocity of the wave. We must note that in the given case the particles of the medium are understood to be macroscopic (*i.e.*, including a great number of molecules) volumes, and not molecules; the linear dimensions of these volumes are much smaller than the wavelength.

Assume that ξ changes according to the law $\xi = A \cos(\omega t - kx + \alpha)$. Hence,

$$\frac{\partial \xi}{\partial x} = Ak \sin(\omega t - kx + \alpha) = A \frac{\omega}{v} \sin(\omega t - kx + \alpha).$$

Introducing this value into Eq. (14.68), we obtain

$$\Delta p = -\gamma p A \frac{\omega}{v} \sin(\omega t - kx + \alpha) = -(\Delta p)_{\text{m}} \sin(\omega t - kx + \alpha),$$

whence

$$A = \frac{(\Delta p)_{\text{m}} v^2}{\gamma p \omega}. \tag{14.73}$$

Using this expression in Eq. (14.72), we get

$$I = \frac{1}{2}\rho\frac{(\Delta p)_{\mathrm{m}}^2 v^2}{\gamma^2 p^2 \omega^2}\omega^2 v = \frac{(\Delta p)_{\mathrm{m}}^2}{2\gamma^2 \rho v}\left(\frac{\rho}{p}\right)^2 v^4.$$

Taking into account that $v^4 = (\gamma RT/M)^2$, and $(p/\rho)^2 = (RT/M)^2$ [see Eq. (14.70) and the text preceding it], we can write that

$$I = \frac{(\Delta p)_{\mathrm{m}}^2}{2\rho v}. \tag{14.74}$$

We can use this equation to calculate the approximate values of the amplitude of air pressure oscillations corresponding to the range of loudness levels from 0 dB to 130 dB. These values range from 3×10^{-5} Pa (2×10^{-7} mmHg) to 100 Pa (about 1 mmHg).

Let us assess the amplitude of oscillations of the particles A and that of the velocity of the particles $(\dot{\xi})_{\mathrm{m}}$. We shall begin with an assessment of the quantity A determined by Eq. (14.73). Taking into account that $v/\omega = \lambda/(2\pi)$, we get the expression

$$\frac{A}{\lambda} = \frac{1}{2\pi\gamma}\frac{(\Delta p)_{\mathrm{m}}^2}{p} \approx 0.1\frac{(\Delta p)_{\mathrm{m}}^2}{p} \tag{14.75}$$

($\gamma \approx 1.5$, consequently, $2\pi\gamma \approx 10$). At a loudness of 130 dB, the ratio $(\Delta p)_{\mathrm{m}}^2/p$ has a value of the order of 10^{-3}, while at a loudness of 60 dB this ratio is about 2×10^{-7}. The lengths of sound waves in air range from 21 m (at $\nu = 16$ Hz) to 17 mm (at $\nu = 20000$ Hz). Inserting these values into Eq. (14.75), we find that at a loudness of 60 dB the amplitude of oscillations of the particles is about 4×10^{-4} mm for the longest waves and about 3×10^{-7} mm for the shortest ones. At a loudness of 130 dB, the amplitude of oscillations for the longest waves is about 2 mm.

For harmonic oscillations, the amplitude of the velocity $(\dot{\xi})_{\mathrm{m}}$ equals that of the displacement A multiplied by the cyclic frequency ω: $(\dot{\xi})_{\mathrm{m}} = A\omega$. Multiplying Eq. (14.75) by ω, we get

$$\frac{(\dot{\xi})_{\mathrm{m}}}{v} = \frac{1}{\gamma}\frac{(\Delta p)_{\mathrm{m}}}{p} \approx \frac{(\Delta p)_{\mathrm{m}}}{p}. \tag{14.76}$$

Consequently, at a loudness of 130 dB, the amplitude of the velocity is about $340\ \mathrm{m\,s^{-1}} \times 10^{-3} = 0.34\ \mathrm{m\,s^{-1}}$. At a loudness of 60 dB, the amplitude of the velocity will be of the order of $0.1\ \mathrm{mm\,s^{-1}}$. We must note that unlike the displacement amplitude, the velocity amplitude does not depend on the wavelength.

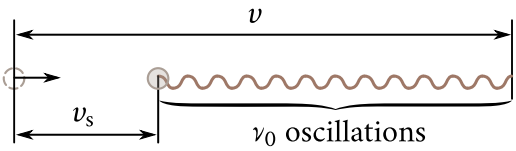


Fig. 14.15

14.11. The Doppler Effect for Sound Waves

Assume that a device sensing the oscillations of the medium, which we shall call a receiver, is placed in a fluid at a certain distance from the wave source. If the source and the receiver of the waves are stationary relative to the medium in which the wave is propagating, then the frequency of the oscillations picked up by the receiver will equal the frequency ν_0 of the oscillations of the source. If the source or the receiver or both are moving relative to the medium, then the frequency ν picked up by the receiver may differ from ν_0. This phenomenon is called the **Doppler effect**. [It is named after the Austrian scientist Christian Doppler (1803-1853) who described the effect for light waves.]

Let us assume that the source and the receiver move along the straight line joining them. We shall assume the velocity of the source v_s to be positive if it moves toward the receiver and negative if it moves away from the receiver. Similarly, we shall assume the velocity of the receiver v_r to be positive if the latter moves toward the source and negative if it moves away from the source.

If the source is stationary and oscillates with the frequency ν_0, then by the moment when the source will complete its ν_0-th oscillation, the "crest" of the wave produced by the first oscillation will travel the path v in the medium (v is the velocity of propagation of the wave relative to the medium). Hence, the ν_0 "crests" and "troughs" of the wave produced by the source in one second will cover the length v. If the source is moving relative to the medium with the velocity v_s, then at the moment when the source completes its ν_0-th oscillation, the crest produced by the first oscillation will be at a distance of $v - v_s$ from the source (Fig. 14.15). Hence, the length $v - v_s$, will contain ν_0 crests and troughs of a wave, so that the wavelength will be

$$\lambda = \frac{v - v_s}{\nu_0}. \tag{14.77}$$

The stationary receiver will be passed in one second by the crests and troughs accommodated on the length v. If the receiver is moving with the velocity v_r, then at the end of a time interval of one second it will pick up the trough which at the beginning of this interval was at a distance numerically equal to v from its present position. Thus, in one second, the receiver will pick up the oscillations

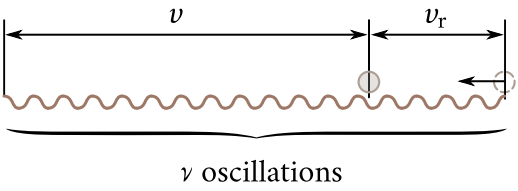


Fig. 14.16

corresponding to the crests and troughs accommodated on a length numerically equal to $v + v_r$ (Fig. 14.16) and will oscillate with the frequency

$$\nu = \frac{v + v_r}{\lambda}.$$

Substituting for λ its value from Eq. (14.77), we get

$$\nu = \nu_0 \left(\frac{v + v_r}{v - v_s} \right). \tag{14.78}$$

It follows from Eq. (14.78) that upon such motion of the source and the receiver when the distance between them diminishes, the frequency ν picked up by the receiver will be greater than that of the source ν_0. If the distance between the source and the receiver increases, ν will be less than ν_0.

If the directions of the velocities $\boldsymbol{v}_s$ and $\boldsymbol{v}_r$ do not coincide with the straight line passing through the source and the receiver, then, the projections of the vectors $\boldsymbol{v}_s$ and $\boldsymbol{v}_r$ onto the direction of this straight line must be substituted for v_s and v_s in Eq. (14.78).

Inspection of Eq. (14.78) shows that the Doppler effect for sound waves is determined by the velocities of the source and the receiver relative to the medium in which the sound propagates. The Doppler effect is also observed for light waves, but the equation for the change in the frequency differs from Eq. (14.78). This is due to the fact that no material medium exists for light waves whose oscillations would be "light". Therefore, the velocities of the source and the receiver of light relative to the "medium" are deprived of a meaning. For light, we can speak only of the relative velocity of the receiver and the source. The Doppler effect for light waves depends on the magnitude and direction of this velocity. This effect will be considered for light waves in Sec. 21.4.

Chapter 15
ELECTROMAGNETIC WAVES

15.1. The Wave Equation for an Electromagnetic Field

We established in Chapter 9 that a varying electric field sets up a magnetic one which, generally speaking, is also varying. This varying magnetic field sets up an electric field, and so on. Thus, if we use oscillating charges to produce a varying (alternating) electromagnetic field, then, in the space surrounding the charges a sequence of mutual transformations of an electric and a magnetic field propagating from point to point will appear. This process will be periodic in both time and space and, consequently, will be a wave.

We shall show that the existence of electromagnetic waves follows from Maxwell's equations. For a homogeneous, neutral ($\rho = 0$), non-conducting ($\boldsymbol{j} = 0$) medium with a constant permittivity ε and a constant permeability μ, we have

$$\frac{\partial \boldsymbol{B}}{\partial t} = \mu\mu_0 \frac{\partial \boldsymbol{H}}{\partial t}, \quad \frac{\partial \boldsymbol{D}}{\partial t} = \varepsilon\varepsilon_0 \frac{\partial \boldsymbol{E}}{\partial t},$$

$$\nabla \cdot \boldsymbol{B} = \mu\mu_0(\nabla \cdot \boldsymbol{H}), \quad \nabla \cdot \boldsymbol{D} = \varepsilon\varepsilon_0(\nabla \cdot \boldsymbol{E}).$$

Consequently, Eqs. (9.5), (7.3), (9.13), and (2.23) can be written as follows:

$$\nabla \times \boldsymbol{E} = -\mu\mu_0 \frac{\partial \boldsymbol{H}}{\partial t}, \tag{15.1}$$

$$\nabla \cdot \boldsymbol{H} = 0, \tag{15.2}$$

$$\nabla \times \boldsymbol{H} = \varepsilon\varepsilon_0 \frac{\partial \boldsymbol{E}}{\partial t}, \tag{15.3}$$

$$\nabla \cdot \boldsymbol{E} = 0. \tag{15.4}$$

Let us take a curl of both sides of Eq. (15.1):

$$\nabla \times (\nabla \times \boldsymbol{E}) = -\mu\mu_0 \nabla \times \left(\frac{\partial \boldsymbol{H}}{\partial t} \right). \tag{15.5}$$

The symbol ∇ denotes differentiation by coordinates. A change in the sequence of differentiation with respect to the coordinates and time leads to the equation

$$\nabla \times \left(\frac{\partial \boldsymbol{H}}{\partial t} \right) = \frac{\partial}{\partial t} (\nabla \times \boldsymbol{H}).$$

Making such a substitution in Eq. (15.5) and introducing the value given by Eq. (15.3) for the curl of $\boldsymbol{H}$ into the equation obtained, we have

$$\nabla \times (\nabla \times \boldsymbol{E}) = -\varepsilon \varepsilon_0 \mu \mu_0 \frac{\partial^2 \boldsymbol{E}}{\partial t^2}. \tag{15.6}$$

According to Eq. (1.107), $\nabla \times (\nabla \times \boldsymbol{E}) = \nabla(\nabla \cdot \boldsymbol{E}) - \Delta \boldsymbol{E}$. Because of Eq. (15.4), the first term of this expression is zero. Consequently, the left-hand side of Eq. (15.6) is $-\Delta \boldsymbol{E}$. Thus, omitting the minus signs at both sides of the equation, we obtain

$$\Delta \boldsymbol{E} = \varepsilon \varepsilon_0 \mu \mu_0 \frac{\partial^2 \boldsymbol{E}}{\partial t^2}.$$

According to Eq. (6.15), we have $\varepsilon_0 \mu_0 = 1/c$. The equation can, therefore, be written in the form

$$\Delta \boldsymbol{E} = \frac{\varepsilon \mu}{c^2} \frac{\partial^2 \boldsymbol{E}}{\partial t^2}. \tag{15.7}$$

Expanding the Laplacian operator, we get

$$\frac{\partial \boldsymbol{E}}{\partial x} 2 + \frac{\partial \boldsymbol{E}}{\partial y} 2 + \frac{\partial \boldsymbol{E}}{\partial z} 2 = \frac{\varepsilon \mu}{c^2} \frac{\partial^2 \boldsymbol{E}}{\partial t^2}. \tag{15.8}$$

Taking a curl of both sides of Eq. (15.3) and performing similar transformations, we arrive at the equation

$$\frac{\partial \boldsymbol{H}}{\partial x} 2 + \frac{\partial \boldsymbol{H}}{\partial y} 2 + \frac{\partial \boldsymbol{H}}{\partial z} 2 = \frac{\varepsilon \mu}{c^2} \frac{\partial^2 \boldsymbol{H}}{\partial t^2}. \tag{15.9}$$

Equations (15.8) and (15.9) are inseparably related to each other because they have been obtained from Eqs. (15.1) and (15.3) each of which contains both $\boldsymbol{E}$ and $\boldsymbol{H}$.

Equations (15.8) and (15.9) are typical wave equations [see Eq. (14.24)]. Any function satisfying such an equation describes a wave. The square root of the quantity that is the reciprocal of the coefficient of the time derivative gives the phase velocity of this wave. Hence, Eqs. (15.8) and (15.9) point to the fact that electromagnetic fields can exist in the form of electromagnetic waves whose phase velocity is

$$v = \frac{c}{\sqrt{\varepsilon \mu}}. \tag{15.10}$$

In a vacuum (*i.e.*, when $\varepsilon = \mu = 1$), the velocity of electromagnetic waves coincides with that of light in free space c.

15.2. Plane Electromagnetic Wave

Let us investigate a plane electromagnetic wave propagating in a neutral non-conducting medium with a constant permittivity ε and permeability μ ($\rho = 0$, $\boldsymbol{j} = 0$, $\varepsilon =$ constant, $\mu =$ constant). We shall direct the x-axis at right angles to the wave surfaces. Hence, $\boldsymbol{E}$ and $\boldsymbol{H}$, and, consequently, their components along the coordinate axes will not depend on the coordinates y and z. For this reason, Eqs. (9.15)-(9.18) can be simplified as follows:

$$0 = \mu\mu_0 \frac{\partial H_x}{\partial t}, \quad \frac{\partial E_z}{\partial x} = \mu\mu_0 \frac{\partial H_y}{\partial t}, \quad \frac{\partial E_y}{\partial x} = -\mu\mu_0 \frac{\partial H_z}{\partial t} \tag{15.11}$$

$$\frac{\partial B_x}{\partial x} = \mu\mu_0 \frac{\partial H_x}{\partial x} = 0, \tag{15.12}$$

$$0 = \varepsilon\varepsilon_0 \frac{\partial E_x}{\partial t}, \quad \frac{\partial H_z}{\partial x} = -\varepsilon\varepsilon_0 \frac{\partial E_y}{\partial t}, \quad \frac{\partial H_y}{\partial x} = \varepsilon\varepsilon_0 \frac{\partial E_z}{\partial t} \tag{15.13}$$

$$\frac{\partial D_x}{\partial x} = \varepsilon\varepsilon_0 \frac{\partial E_x}{\partial x} = 0. \tag{15.14}$$

Equation (15.14) and the first of Eqs. (15.13) show that E_x can depend neither on x nor on t. Equation (15.12) and the first of Eqs. (15.11) give the same result for H_x. Hence, E_x and H_x differing from zero can be due only to constant homogeneous fields superposed onto the electromagnetic field of a wave. The wave field itself cannot have components along the x-axis. It thus follows that the vectors $\boldsymbol{E}$ and $\boldsymbol{H}$ are perpendicular to the direction of propagation of the wave, *i.e.*, that electromagnetic waves are transverse. We shall assume in the following that the constant fields are absent and that $E_x = H_x = 0$.

The last two equations (15.11) and the last two equations (15.13) can be combined into two independent groups

$$\frac{\partial E_y}{\partial x} = -\mu\mu_0 \frac{\partial H_z}{\partial t}, \quad \frac{\partial H_z}{\partial x} = -\varepsilon\varepsilon_0 \frac{\partial E_y}{\partial t}, \tag{15.15}$$

$$\frac{\partial E_z}{\partial x} = \mu\mu_0 \frac{\partial H_y}{\partial t}, \quad \frac{\partial H_y}{\partial x} = \varepsilon\varepsilon_0 \frac{\partial E_z}{\partial t}. \tag{15.16}$$

The first group of equations relates the components E_y and H_z, and the second group, the components E_z and H_y. Assume that there was initially set up a varying electric field E_y directed along the y-axis. According to the second of Eqs. (15.15), this field produces the magnetic field H_z directed along the z-axis. In accordance with the first of Eqs. (15.15), the field H_z produces the electric field E_y, and so on. Neither the field E_z nor the field H_y is produced. Similarly, if the field E_z was produced initially, then according to Eqs. (15.16) the field H_y will appear that will set up the field E_z, etc. In this case, the fields E_y and H_z are not produced. Thus, to

describe a plane electromagnetic wave, it is sufficient to take one of the systems of equations (15.15) or (15.16) and to assume that the components in the other system equal zero.

Let us take Eqs. (15.15) to describe a wave, assuming that $E_z = H_y = 0$. We shall differentiate the first equation with respect to x and make the substitution $(\partial/\partial x)(\partial H_z/\partial t) = (\partial/\partial t)(\partial H_z/\partial x)$. Next introducing $\partial H_z/\partial x$ from the second equation, we get a wave equation for E_y:

$$\frac{\partial^2 E_y}{\partial x^2} = \frac{\varepsilon\mu}{c^2}\frac{\partial^2 E_y}{\partial t^2} \tag{15.17}$$

(we have substituted $1/c^2$ for $\varepsilon_0\mu_0$). Differentiating the second of Eqs. (15.15) with respect to x, we find a wave equation for H_z after similar transformations:

$$\frac{\partial^2 H_z}{\partial x^2} = \frac{\varepsilon\mu}{c^2}\frac{\partial^2 H_z}{\partial t^2}. \tag{15.18}$$

The equations obtained are a particular case of Eqs. (15.8) and (15.9).

We remind our reader that $E_x = E_z = 0$ and $H_x = H_y = 0$, so that $E_y = E$ and $H_z = H$. We have retained the subscripts y and z of E and H to stress the circumstance that the vectors $\boldsymbol{E}$ and $\boldsymbol{H}$ are directed along mutually perpendicular axes y and z.

The simplest solution of Eq. (15.17) is the function

$$E_y = E_\mathrm{m}\cos(\omega t - kx + \alpha_1). \tag{15.19}$$

The solution of Eq. (15.18) is similar:

$$H_z = H_\mathrm{m}\cos(\omega t - kx + \alpha_2). \tag{15.20}$$

In these equations, ω is the frequency of the wave, k is the wave number equal to ω/v, and α_1 and α_2 are the initial phases of the oscillations at points with the coordinate $x = 0$.

Introducing functions (15.19) and (15.20) into Eqs. (15.15), we get

$$kE_\mathrm{m}\sin(\omega t - kx + \alpha_1) = \mu\mu_0\omega H_\mathrm{m}\sin(\omega t - kx + \alpha_2),$$

$$kH_\mathrm{m}\sin(\omega t - kx + \alpha_2) = \varepsilon\varepsilon_0\omega E_\mathrm{m}\sin(\omega t - kx + \alpha_1).$$

For these equations to be satisfied, equality of the initial phases α_1 and α_2 is needed. In addition, the following relations must be observed

$$kE_\mathrm{m} = \mu\mu_0\omega H_\mathrm{m},$$

$$kH_\mathrm{m} = \varepsilon\varepsilon_0\omega E_\mathrm{m}.$$

Multiplying these two equations, we find that

$$\varepsilon\varepsilon_0 E_\mathrm{m}^2 = \mu\mu_0 H_\mathrm{m}^2. \tag{15.21}$$

Thus, the oscillations of the electric and magnetic vectors in an electromagnetic

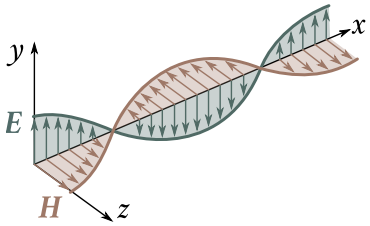


Fig. 15.1

wave occur with the same phase ($\alpha_1 = \alpha_2$), while the amplitudes of these vectors are related by the expression

$$E_\mathrm{m}\sqrt{\varepsilon\varepsilon_0} = H_\mathrm{m}\sqrt{\mu\mu_0}. \tag{15.22}$$

For a wave propagating in a vacuum, we have

$$\frac{E_\mathrm{m}}{H_\mathrm{m}} = \left(\frac{\mu_0}{\varepsilon_0}\right)^{1/2} = \sqrt{4\pi \times 10^{-7} \times 4\pi \times 9 \times 10^9} = 120\pi \approx 377. \tag{15.23}$$

In the Gaussian system of units, Eq. (15.22) becomes

$$E_\mathrm{m}\sqrt{\varepsilon} = H_\mathrm{m}\sqrt{\mu}. \tag{15.24}$$

Consequently, for a vacuum, we have $E_\mathrm{m} = H_\mathrm{m}$ (E_m is measured in cgse units, and H_m in cgsm ones).

Multiplying Eq. (15.19) by the unit vector $\hat{\boldsymbol{e}}_y$ of the y-axis ($E_y\hat{\boldsymbol{e}}_y = \boldsymbol{E}$), and Eq. (15.20) by the unit vector $\hat{\boldsymbol{e}}_z$ of the z-axis ($H_z\hat{\boldsymbol{e}}_z = \boldsymbol{H}$), we get equations for a plane electromagnetic wave in the vector form

$$\begin{aligned} \boldsymbol{E} &= \boldsymbol{E}_\mathrm{m}\cos(\omega t - kx) \\ \boldsymbol{H} &= \boldsymbol{H}_\mathrm{m}\cos(\omega t - kx) \end{aligned} \tag{15.25}$$

(we have assumed that $\alpha_1 = \alpha_2 = 0$).

Figure 15.1 shows an "instantaneous photograph" of a plane electromagnetic wave. A glance at the figure shows that the vectors $\boldsymbol{E}$ and $\boldsymbol{H}$ form a right-handed system with the direction of propagation of the wave. At a fixed point of space, the vectors $\boldsymbol{E}$ and $\boldsymbol{H}$ vary with time according to a harmonic law. They simultaneously grow from zero, and next reach their maximum value in one-fourth of a period; if $\boldsymbol{E}$ is directed upward, then, $\boldsymbol{H}$ is directed to the right (we look along the direction of propagation of the wave). In another one-fourth of a period, both vectors simultaneously vanish. Next, they again reach their maximum value, but this time, $\boldsymbol{E}$ is directed downward, and $\boldsymbol{H}$ to the left. And, finally, upon completion of a period of oscillation, the vectors again vanish. Such changes in the vectors $\boldsymbol{E}$ and $\boldsymbol{H}$ occur at all points of space, but with a shift in phase determined by the distance between the points measured along the x-axis.

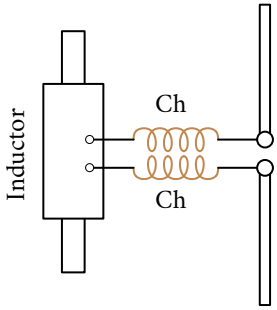


Fig. 15.2

15.3. Experimental Investigation of Electromagnetic Waves

The first experiments with non-optical electromagnetic waves were conducted in 1888 by the German physicist Heinrich Hertz (1857-1894). Hertz produced waves with the aid of a vibrator which he had invented. The vibrator consisted of two rods separated by a spark gap. When a high voltage was fed to the vibrator from an induction coil, a spark jumped through the gap. It shorted the latter, and damped electrical oscillations were set up in the vibrator (Fig. 15.2; the chokes shown in the figure were intended to prevent the high-frequency current from branching off into the inductor winding). During the time the spark burned, a great number of oscillations were completed. They produced a train of electromagnetic waves whose length was approximately twice that of the vibrator. By placing vibrators of various length at the focus of a concave parabolic mirror, Hertz obtained directed plane waves whose length ranged from 0.6 m to 10 m.

Hertz also studied the emitted wave with the aid of a half-wave vibrator having a small spark gap at its middle. When such a vibrator was placed parallel to the electric field strength vector of the wave, oscillations of the current and voltage were produced in it. Since the length of the vibrator was equal to $\lambda/2$, the oscillations in it owing to resonance reached such an intensity that they caused small sparks to jump across the spark gap.

Hertz reflected and refracted electromagnetic waves with the aid of large metal mirrors and an asphalt prism (over 1 m in size and with a mass of 1200 kg). He discovered that both these phenomena obey the laws established in optics for light waves. By reflecting a running plane wave with the aid of a metal mirror to the opposite direction, Hertz obtained a standing wave. The distance between the nodes and antinodes of the wave made it possible to find its length λ. By multiplying λ by the frequency of oscillations ν of the vibrator, the velocity of the electromagnetic waves was determined, and it was found to be close to c. By placing a grate of

parallel copper wires in the path of waves, Hertz discovered that the intensity of the waves passing through the grate changes very greatly when the grate is rotated about the beam. When the wires forming the grate were perpendicular to the vector $\boldsymbol{E}$, the wave passed through the grate without any hindrance. When the wires were arranged parallel to $\boldsymbol{E}$, the wave did not pass through the grate. Thus, the transverse nature of electromagnetic waves was proved.

Hertz's experiments were continued by the Russian physicist Pyotr Lebedev (1866-1912), who in 1894 obtained electromagnetic waves 6 mm long and studied how they travel in crystals. He detected double refraction of the waves (see Sec. 19.3).

In 1896, the Russian inventor Aleksandr Popov (1859-1905) for the first time in history transmitted a message over a distance of about 250 m with the aid of electromagnetic waves (the words "Heinrich Hertz" were transmitted). This laid the foundation of radio engineering.

15.4. Energy of Electromagnetic Waves

Electromagnetic waves transfer energy. According to Eq. (14.46), the density of the energy flux can be obtained by multiplying the energy density by the wave velocity.

The density of the energy of an electromagnetic field w consists of the density of the energy of the electric field [determined by Eq. (4.10)] and that of the energy of the magnetic field [determined by Eq. (8.40)]:

$$w = w_E + w_H = \frac{\varepsilon\varepsilon_0 E^2}{2} + \frac{\mu\mu_0 H^2}{2}. \tag{15.26}$$

The vectors $\boldsymbol{E}$ and $\boldsymbol{H}$ at a given point of space vary in the same phase[1]. Therefore, Eq. (15.22) giving the relation between the amplitude values of E and H also holds for their instantaneous values. It thus follows that the densities of the energy of the electric and magnetic fields of a wave are identical at each moment of time: $w_E = w_H$. We can, therefore, write that

$$w = 2w_E = \varepsilon\varepsilon_0 E^2. \tag{15.27}$$

Taking advantage of the fact that $E\sqrt{\varepsilon\varepsilon_0} = H\sqrt{\mu\mu_0}$, we can write Eq. (15.27) in the form

$$w = \sqrt{\varepsilon\varepsilon_0\mu\mu_0}EH = \frac{1}{v}EH,$$

where v is the velocity of an electromagnetic wave [see Eq. (15.10)].

Multiplying the expression found for w by the wave velocity v, we get the

[1]This holds only for a non-conducting medium. The phases of $\boldsymbol{E}$ and $\boldsymbol{B}$ do not coincide in a conducting medium.

magnitude of the energy flux density vector

$$S = wv = EH. \tag{15.28}$$

The vectors $\boldsymbol{E}$ and $\boldsymbol{H}$ are mutually perpendicular and form a right-handed system with the direction of propagation of the wave. For this reason, the direction of the vector $\boldsymbol{E} \times \boldsymbol{H}$ coincides with that of energy transfer, and the magnitude of this vector is EH. Hence, the vector of the density of the electromagnetic energy flux can be written as the vector product of $\boldsymbol{E}$ and $\boldsymbol{H}$:

$$\boldsymbol{S} = \boldsymbol{E} \times \boldsymbol{H}. \tag{15.29}$$

The vector $\boldsymbol{S}$ is known as the **Poynting vector**.

By analogy with Eq. (14.50), the flux Φ of electromagnetic energy through surface A_s can be found by integration:

$$\Phi = \oint_{A_s} \boldsymbol{S} \cdot \mathrm{d}\boldsymbol{A}_s \tag{15.30}$$

[in Eq. (14.50) the surface area was designated by the symbol S; since this symbol is used to designate the Poynting vector, we were forced to introduce the symbol A_s for the surface area].

Let us consider a portion of a homogeneous cylindrical conductor through which a steady current is flowing (Fig. 15.3) as an example of applying Eqs. (15.29) and (15.30). We shall first consider that extraneous forces are absent on this portion of the conductor. Hence, according to Eq. (5.22), the following relation is observed at each point of the conductor:

$$\boldsymbol{j} = \sigma \boldsymbol{E} = \frac{1}{\rho}\boldsymbol{E}.$$

The steady current is distributed over the cross section of the conductor with an identical density $\boldsymbol{j}$. Hence, the electric field within the limits of the portion of the conductor shown in Fig. 15.3 will be homogeneous. Let us mentally separate a cylindrical volume of radius r and length l inside the conductor. At each point on the side surface of this cylinder, the vector $\boldsymbol{H}$ is perpendicular to the vector $\boldsymbol{E}$ and is directed tangentially to the surface. The magnitude of $\boldsymbol{H}$ is $jr/2$ [according to Eq. (7.10), we have $2\pi rH = j\pi r^2$]. Thus, the Poynting vector given by Eq. (15.29) is directed toward the axis of the conductor at each point on the surface and has the magnitude $S = EH = Ejr^2/2$. Multiplying S by the side surface area of the cylinder A_s equal to $2\pi rl$, we find that the following flux of electromagnetic energy enters the volume we are considering:

$$\Phi = SA_s = \frac{1}{2}Ejr \times 2\pi rl = Ej \times \pi r^2 l = Ej \times V, \tag{15.31}$$

where V is the volume of the cylinder.

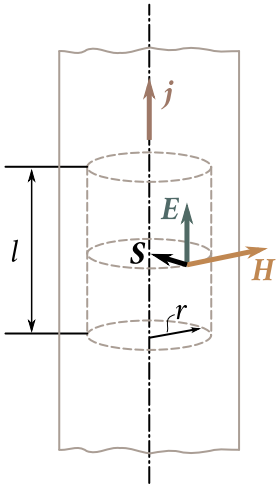


Fig. 15.3

According to Eq. (6.4), $Ej = pj^2$ is the amount of heat liberated in unit time per unit volume of the conductor. Consequently, Eq. (15.31) indicates that the energy liberated in the form of Lenz-Joule heat is supplied to the conductor through its side surface in the form of energy of an electromagnetic field. The energy flux gradually weakens with deeper penetration into the conductor (both the Poynting vector and the surface through which the flux passes diminish) as a result of absorption of energy and its conversion into heat.

Now, let us assume that extraneous forces whose field is homogeneous are exerted within the limits of the portion of the conductor we are considering ($\boldsymbol{E}^* =$ constant). In this case according to Eq. (5.25), at each point of the conductor we have

$$\boldsymbol{j} = \sigma\,(\boldsymbol{E} + \boldsymbol{E}^*) = \frac{1}{\rho}\,(\boldsymbol{E} + \boldsymbol{E}^*)\,,$$

whence

$$\boldsymbol{E} = \rho\boldsymbol{j} - \boldsymbol{E}^*. \tag{15.32}$$

We shall consider that the extraneous forces on the portion of the circuit being considered do not hamper the flow of the current, but facilitate it. This signifies that the direction of $\boldsymbol{E}^*$ coincides with that of $\boldsymbol{j}$. Let us assume that the relation $\rho j = E^*$ is observed. Hence, according to Eq. (15.32), the electrostatic field strength $\boldsymbol{E}$ at each point vanishes, and there is no flux of electromagnetic energy through the side surface. In this case, heat is liberated at the expense of the work of the extraneous forces.

If the relation $E^* > \rho j$ holds, then, as can be seen from Eq. (15.32), the vector

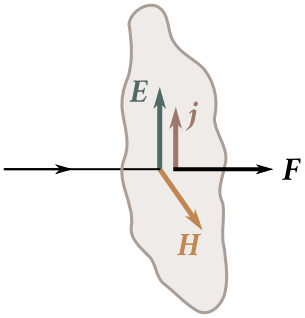


Fig. 15.4

$\boldsymbol{E}$ will be directed oppositely to the vector $\boldsymbol{j}$. In this case, the vectors $\boldsymbol{E}$ and $\boldsymbol{S}$ will have directions opposite to those shown in Fig. 15.3. Hence, instead of flowing in, electromagnetic energy flows out through the side surface of the conductor into the space surrounding it.

In summarizing, we can say that in the closed circuit of a steady current, the energy from the sections where extraneous forces act is transmitted to other sections of the circuit not along the conductors, but through the space surrounding the conductors in the form of a flux of electromagnetic energy characterized by the vector $\boldsymbol{S}$.

15.5. Momentum of Electromagnetic Field

An electromagnetic wave absorbed in a body imparts a momentum to the body, *i.e.*, exerts a pressure on it. This can be shown by the following example. Assume that a plane wave impinges normally onto a flat surface of a weakly conducting body with ε and μ equal to unity (Fig. 15.4). The electric field of the wave produces a current of density $\boldsymbol{j} = \sigma \boldsymbol{E}$ in the body. The magnetic field of the wave will act on the current with a force whose value per unit volume of the body can be found by Eq. (6.43):

$$\boldsymbol{F}_{\text{u.v}} = \boldsymbol{j} \times \boldsymbol{B} = \mu_0 (\boldsymbol{k} \times \boldsymbol{H}).$$

The direction of this force, as can be seen from Fig. 15.4, coincides with the direction of propagation of the wave.

The momentum

$$\mathrm{d}K = F_{\text{u.v}}\, \mathrm{d}l = \mu_0 j H\, \mathrm{d}l \tag{15.33}$$

is imparted to a surface layer having a unit area and a thickness of $\mathrm{d}l$ in unit time (the vectors $\boldsymbol{j}$ and $\boldsymbol{H}$ are mutually perpendicular). The energy absorbed in this layer in unit time is

$$\mathrm{d}W = jE\, \mathrm{d}l. \tag{15.34}$$

It is liberated in the form of heat.

The momentum given by Eq. (15.33) and the energy [Eq. (15.34)] are imparted to the layer by the wave. Let us take their ratio, omitting the differential symbols as superfluous:

$$\frac{K}{W} = \mu_0 \frac{H}{E}.$$

Taking into account that $\mu_0 H^2 = \varepsilon_0 E^2$, we get

$$\frac{K}{W} = \sqrt{\varepsilon_0 \mu_0} = \frac{1}{c}.$$

It thus follows that an electromagnetic wave carrying the energy W has the momentum

$$K = \frac{1}{c} W. \tag{15.35}$$

The same relation between the energy and the momentum holds for particles with a zero rest mass [see Eq. (8.57) of Vol. I]. This is not surprising because according to quantum notions, an electromagnetic wave is equivalent to a flux of photons, *i.e.*, particles whose mass (we have in mind their rest mass) is zero.

Examination of Eq. (15.35) shows that the density of the momentum (*i.e.*, the momentum of unit volume) of an electromagnetic field is

$$K_{\text{u.v}} = \frac{1}{c} w. \tag{15.36}$$

The energy density is related to the magnitude of the Poynting vector by the expression $S = wc$. Substituting S/c for w in Eq. (15.36) and taking into account that the directions of the vectors $\boldsymbol{K}$ and $\boldsymbol{S}$ coincide, we can write that

$$\boldsymbol{K}_{\text{u.v}} = \frac{1}{c^2} \boldsymbol{S} = \frac{1}{c^2} (\boldsymbol{E} \times \boldsymbol{H}). \tag{15.37}$$

We shall note that when energy of any kind is transferred, the density of the energy flux equals the density of the momentum multiplied by c^2. Let us consider, for example, a collection of particles distributed in space with the density n and flying with a velocity v identical in magnitude and direction. In this case, the density of the momentum

$$\boldsymbol{K}_{\text{u.v}} = n \frac{m\boldsymbol{v}}{\sqrt{1 - (v^2/c^2)}}. \tag{15.38}$$

The particles carry along energy whose density flux $\boldsymbol{j}_W$ equals the density of the particle flux multiplied by the energy of one particle:

$$\boldsymbol{j}_W = n\boldsymbol{v} \frac{mc^2}{\sqrt{1 - (v^2/c^2)}}. \tag{15.39}$$

It follows from Eqs. (15.38) and (15.39) that

$$\boldsymbol{K}_{\text{u.v}} = \frac{1}{c^2}\boldsymbol{j}_W. \tag{15.40}$$

Assume that an electromagnetic wave falling normally on a body is completely absorbed by the body. Hence, a unit of surface area of the body receives in unit time the momentum of the wave enclosed in a cylinder with a base area of unity and an altitude of c. According to Eq. (15.36), this momentum is $(w/c)c = w$. At the same time, the momentum imparted to a unit surface area in unit time equals the pressure p on the surface. Hence, for an absorbing surface, we have $p = w$. This quantity pulsates with a very high frequency. We can, therefore, measure its time-averaged value in practice. Thus,

$$p = \langle w \rangle . \tag{15.41}$$

For an ideally reflecting surface, the pressure will be double this value.

The value of the pressure calculated by Eq. (15.41) is very small. For example, at a distance of 1 m from a source of light having an intensity of a million candelas, the pressure is only about 10^{-7} Pa (about 10^{-9} gf cm^{-2}). Pyotr Lebedev succeeded in measuring the pressure of light. By carrying out experiments requiring great inventiveness and skill, Lebedev measured the pressure of light on solids in 1900, and on gases in 1910. The results of the measurements completely agreed with Maxwell's theory.

15.6. Dipole Emission

An oscillating electric dipole is the simplest system emitting electromagnetic waves. An example of such a dipole is the system formed by a fixed point charge $+q$ and a point charge $-q$ oscillating near it (Fig. 15.5). The dipole electric moment of this system varies in time according to the law

$$\boldsymbol{p} = -q\boldsymbol{r} = -ql\hat{\boldsymbol{e}}\cos(\omega t) = \boldsymbol{p}_{\text{m}}\cos(\omega t), \tag{15.42}$$

where $\boldsymbol{r}$ is the position vector of the charge $-q$, l the amplitude of oscillations, $\hat{\boldsymbol{e}}$ is the unit vector directed along the dipole axis, and $\boldsymbol{p}_{\text{m}} = -gl\hat{\boldsymbol{e}}$.

Acquaintance with such an emitting system is especially important in connection with the fact that many questions of the interaction of radiation with a substance can be explained classically proceeding from the notion of atoms as of systems of charges containing electrons that are capable of performing harmonic oscillations about their equilibrium position.

Let us consider the radiation of a dipole whose dimensions are small in comparison with the wavelength ($l \ll \lambda$). Such a dipole is called **elementary**. The pattern of the electromagnetic field in direct proximity to the dipole is very complicated.

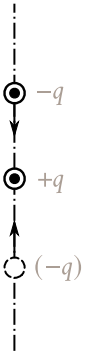


Fig. 15.5

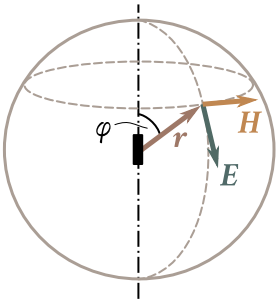


Fig. 15.6

It becomes simplified quite greatly in the so-called **wave zone** of the dipole that begins at distances r considerably exceeding the wavelength ($r \gg \lambda$). If a wave is propagating in a homogeneous isotropic medium, then its wavefront in the wave zone will be spherical (Fig. 15.6). The vectors $\boldsymbol{E}$ and $\boldsymbol{H}$ at each point are mutually perpendicular and are perpendicular to the ray, *i.e.*, to the position vector drawn to the given point from the centre of the dipole.

Let us call sections of the wavefront by planes passing through the dipole axis **meridians**, and by planes perpendicular to the dipole axis **parallels**. We can now say that the vector $\boldsymbol{E}$ at each point of a wave zone is directed along a tangent to the meridian, and the vector $\boldsymbol{H}$ along a tangent to the parallel. If we look along the ray r, then the instantaneous pattern of the wave will be the same as shown in Fig. 15.5, the only difference being that the amplitude in motion along the ray gradually diminishes.

At each point, the vectors $\boldsymbol{E}$ and $\boldsymbol{H}$ oscillate according to the law $\cos(\omega t - kr)$. The amplitudes $\boldsymbol{E}_\mathrm{m}$ and $\boldsymbol{H}_\mathrm{m}$ depend on the distance r to the emitter and on the angle a between the direction of the position vector $\boldsymbol{r}$ and the dipole axis (see Fig. 15.6). This dependence has the following form for a vacuum:

$$E_\mathrm{m} \propto H_\mathrm{m} \propto \frac{1}{r} \sin\theta.$$

The average value of the density of the energy flux $\langle S \rangle$ is proportional to the product $E_\mathrm{m} H_\mathrm{m}$, consequently,

$$\langle S \rangle \propto \frac{1}{r^2} \sin^2\theta. \tag{15.43}$$

A glance at this expression shows that the wave intensity changes along the ray (at θ = constant) in inverse proportion to the square of the distance from the emitter. In addition, it depends on the angle θ. The emission of a dipole is the greatest in directions at right angles to its axis ($\theta = \pi/2$). There is no emission in the directions

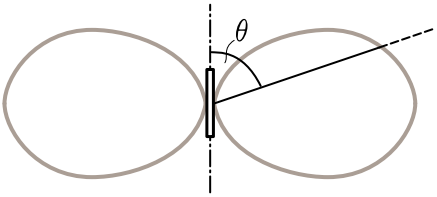


Fig. 15.7

coinciding with the axis ($\theta = 0$ and π). How the intensity depends on the angle θ is shown very illustratively with the aid of a **dipole directional diagram** (Fig. 15.7). This diagram is constructed so that the length of the segment it intercepts on a ray conducted from the centre of the dipole gives the intensity of emission at the angle θ.

The corresponding calculations show that the **radiant power** P of a dipole (*i.e.*, the energy emitted in all directions in unit time) is proportional to the square of the second time derivative of the dipole moment:

$$P \propto \ddot{\boldsymbol{p}}^2. \tag{15.44}$$

According to Eq. (15.42), $\ddot{\boldsymbol{p}}^2 = p_{\mathrm{m}}^2 \omega^4 \cos^2(\omega t)$. Introduction of this value into expression (15.44) yields

$$P \propto p_{\mathrm{m}}^2 \omega^4 \cos(\omega t). \tag{15.45}$$

Time averaging of this expression gives

$$\langle P \rangle \propto p_{\mathrm{m}}^2 \omega^4. \tag{15.46}$$

Thus, the average radiant power of a dipole is proportional to the square of the amplitude of the electric dipole moment and to the fourth power of the frequency. Therefore, at a low frequency, the emission of electrical systems (for instance, industrial frequency alternating current transmission lines) is insignificant.

According to Eq. (15.42), we have $\ddot{\boldsymbol{p}} = -q\ddot{\boldsymbol{r}} = -q\boldsymbol{a}$, where $\boldsymbol{a}$ is the acceleration of an oscillating charge. Substitution of this expression for $\ddot{\boldsymbol{p}}$ in Eq. (15.44) yields[2]:

$$P \propto q^2 \boldsymbol{a}^2. \tag{15.47}$$

Expression (15.47) determines the radiant power not only for oscillations, but also for arbitrary motion of a charge. A charge travelling with acceleration produces electromagnetic waves, and the radiated power is proportional to the square of the charge and the square of the acceleration. For example, the electrons accelerated in a betatron (see Sec. 10.5) lose energy as a result of radiation mainly due to centripetal acceleration $a_{\mathrm{n}} = v^2/r$. According to expression (15.47), the amount of energy

[2]The constant of proportionality when SI units are used is $\sqrt{\mu_0/\varepsilon_0}/(6\pi c^2)$, and when units of the Gaussian system are used is $2/(3c^3)$.

lost grows greatly with an increasing velocity of the electrons in the betatron (in proportion to v^4). Hence, the possible acceleration of electrons in a betatron is limited to about 500 MeV (at a velocity corresponding to this value, the losses due to radiation become equal to the energy imparted to the electrons by the vortex electric field).

A charge performing harmonic oscillations emits a monochromatic wave with a frequency equal to that of the charge oscillations. If the acceleration $\boldsymbol{a}$ of the charge does not change according to a harmonic law, then the radiation consists of a set of waves of different frequencies.

According to expression (15.47), the intensity vanishes when $\boldsymbol{a} = 0$. Consequently, *an electron travelling with a constant velocity does not emit electromagnetic waves*. This holds, however, only for the case when the velocity of an electron v_{el} does not exceed the speed of light $v_{\mathrm{l}} = c/\sqrt{\varepsilon\mu}$ in the medium in which the electron is travelling. When $v_{\mathrm{el}} > v_{\mathrm{l}}$, radiation is observed. It was discovered in 1934 by the Soviet physicists Sergei Vavilov (1891-1951) and Pavel Cerenkov (born 1904).

PART III

OPTICS

Chapter 16
OPTICS

16.1. The Light Wave

Light is a complicated phenomenon: in some cases it behaves like an electromagnetic wave, in others like a stream of special particles (photons). In the present volume, we shall treat wave optics, *i.e.*, the range of phenomena based on the wave nature of light. The collection of phenomena due to the corpuscular (particulate) nature of light will be dealt with in Volume III.

What oscillates in an electromagnetic wave are the vectors $\boldsymbol{E}$ and $\boldsymbol{H}$. Experiments show that the physiological, photochemical, photoelectrical and other actions of light are due to the oscillations of the electric vector. Accordingly, we shall speak in the following of the **light vector**, having in mind the electric field strength vector. We shall meanwhile make no mention of the magnetic vector of a light wave.

We shall designate the magnitude of the light vector amplitude, as a rule, by the letter A (sometimes by the symbol E_{m}). Hence, the change in space and time of the projection of the light vector onto the direction along which it oscillates will be described by the equation

$$E = A\cos(\omega t - kr + \alpha), \tag{16.1}$$

where k is the wave number and r is the distance measured along the direction of propagation of the light wave.

For a plane wave propagating in a non-absorbing medium, A = constant, for a spherical wave, A diminishes in proportion to $1/r$, and so on.

The ratio of the speed of a light wave in a vacuum to the phase velocity v in a medium is known as the **absolute refractive index** of the medium and is designated by the letter n. Thus,

$$n = \frac{c}{v}. \tag{16.2}$$

A comparison with Eq. (15.10) shows that $n = \sqrt{\varepsilon\mu}$. For the overwhelming majority of transparent substances, μ does not virtually differ from unity. We can therefore consider that

$$n = \sqrt{\varepsilon}. \tag{16.3}$$

Equation (16.3) relates the optical and the electrical properties of a substance. It may seem on the face of it that this equation is wrong. For example, for water $\varepsilon = 81$, whereas $n = 1.33$. It must be borne in mind, however, that the value $\varepsilon = 81$ has been obtained from electrostatic measurements. A different value of ε is obtained for fast-varying electric fields, and it depends on the frequency of oscillations of the field. This explains the **dispersion** of light, *i.e.*, the dependence of the refractive index (or speed of light) on the frequency (or wavelength). Using the value of ε obtained for the relevant frequency in Eq. (16.3) leads to the correct value of n.

The values of the refractive index characterize the **optical density** of the medium. A medium with a greater n is called optically denser than one with a smaller n. Conversely, a medium with a lower n is called optically less dense than one with a greater n.

The wavelengths of visible light are within the following limits:

$$\lambda_0 = 0.40\ \mu\text{m to } 0.76\ \mu\text{m} (4000\ \text{Å to } 7600\ \text{Å}). \tag{16.4}$$

These values relate to light waves in a vacuum. The lengths of light waves in substances will have other values. For oscillations of frequency ν, the wavelength in a vacuum is $\lambda_0 = c/\nu$. In a medium in which the phase velocity of a light wave is $v = c/n$, the wavelength has the value $\lambda = v/\nu = c/(\nu n) = \lambda_0/n$. Thus, the length of a light wave in a medium with the refractive index n is related to the wavelength in a vacuum by the expression

$$\lambda = \frac{\lambda_0}{n}. \tag{16.5}$$

The frequencies of visible light waves are within the limits

$$\nu = 0.39 \times 10^{15}\ \text{Hz to } 0.75 \times 10^{15}\ \text{Hz}. \tag{16.6}$$

The frequency of the changes in the vector of the energy flux density carried by a wave will be still greater (it equals 2ν). Neither our eye nor any other receiver of luminous energy can track such frequent changes in the energy flux, hence, they register the time-averaged flux. The magnitude of the time-averaged energy flux density carried by a light wave is called the **light intensity** I at the given point of space. The density of the flux of electromagnetic energy is determined by the Poynting vector $\boldsymbol{S}$. Hence,

$$I = |\langle \boldsymbol{S} \rangle| = |\langle \boldsymbol{E} \times \boldsymbol{H} \rangle|. \tag{16.7}$$

Averaging is performed over the time of "operation" of the instrument, which, as

we have already noted, is much greater than the period of oscillations of the wave. The intensity is measured either in energy units (for example, in W m^{-2}), or in light units named "lumen per square metre" (see Sec. 16.5).

According to Eq. (15.22), the magnitudes of the amplitudes of the vectors $\boldsymbol{E}$ and $\boldsymbol{H}$ in an electromagnetic wave are related by the expression

$$E_\mathrm{m}\sqrt{\varepsilon\varepsilon_0} = H_\mathrm{m}\sqrt{\mu\mu_0} = H_\mathrm{m}\sqrt{\mu_0}$$

(we have assumed that $\mu = 1$). It thus follows that

$$H_\mathrm{m} = \sqrt{\varepsilon}\left(\frac{\varepsilon_0}{\mu_0}\right)^{1/2} E_\mathrm{m} = nE_\mathrm{m}\left(\frac{\varepsilon_0}{\mu_0}\right)^{1/2},$$

where n is the refractive index of the medium in which the wave propagates. Thus, H_m is proportional to E_m and n:

$$H_\mathrm{m} \propto nE_\mathrm{m}. \tag{16.8}$$

The magnitude of the average value of the Poynting vector is proportional to $H_\mathrm{m}E_\mathrm{m}$. We can therefore write that

$$I \propto nE_\mathrm{m}^2 = nA^2 \tag{16.9}$$

(the constant of proportionality is $\sqrt{\varepsilon_0/\mu_0}$). Hence, the light intensity is proportional to the refractive index of the medium and the square of the light wave amplitude.

We must note that when considering the propagation of light in a homogeneous medium, we may assume that the intensity is proportional to the square of the light wave amplitude

$$I \propto A^2. \tag{16.10}$$

For light passing through the interface between media, however, the expression for the intensity, which does not take the factor n into account, leads to non-conservation of the light flux.

The lines along which light energy propagates are called **rays**. The averaged Poynting vector $\langle \boldsymbol{S} \rangle$ is directed at each point along a tangent to a ray. The direction of $\langle \boldsymbol{S} \rangle$ in isotropic media coincides with a normal to the wave surface, *i.e.*, with the direction of the wave vector $\boldsymbol{k}$. Hence, the rays are perpendicular to the wave surfaces. In anisotropic media, a normal to the wave surface generally does not coincide with the direction of the Poynting vector so that the rays are not orthogonal to the wave surfaces.

Although light waves are transverse, they usually do not display asymmetry relative to a ray. The explanation is that in **natural** light (*i.e.*, in light emitted by conventional sources) there are oscillations that occur in the most diverse directions perpendicular to a ray (Fig. 16.1). The radiation of a luminous body consists of the waves emitted by its atoms. The process of radiation in an individual atom continues

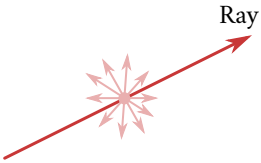


Fig. 16.1

about 10^{-8} s. During this time, a sequence of crests and troughs (or, as is said, a **wave train**) of about three metres in length is formed. The atom "dies out", and then "flares up" again after a certain time elapses. Many atoms "flare up" at the same time. The wave trains they emit are superposed on one another and form the light wave emitted by the relevant body. The plane of oscillations is oriented randomly for each wave train. Therefore, the resultant wave contains oscillations of different directions with an equal probability.

In natural light, the oscillations in different directions follow one another rapidly and without any order. Light in which the direction of the oscillations has been brought into order in some way or other is called **polarized**. If the oscillations of the light vector occur only in a single plane passing through a ray, the light is called **plane** (or **linearly**) **polarized**. The order may consist in that the vector $\boldsymbol{E}$ rotates about a ray while simultaneously pulsating in magnitude. The result is that the tip of the vector $\boldsymbol{E}$ describes an ellipse. Such light is called **elliptically polarized**. If the tip of the vector $\boldsymbol{E}$ describes a circle, the light is called **circularly polarized**.

We shall deal with natural light in Chapters 17 and 18. For this reason, we shall display no interest in the direction of the light vector oscillations. The ways of obtaining polarized light and its properties are considered in Chapter 19.

16.2. Representation of Harmonic Functions Using Exponents

Let us form the sum of two complex numbers $z_1 = x_1 + iy_1$ and $z_2 = x_2 + iy_2$:

$$z = z_1 + z_2 = (x_1 + iy_1) + (x_2 + iy_2) = (x_1 + x_2) + i(y_1 + y_2). \tag{16.11}$$

It can be seen from Eq. (16.11) that the real part of the sum of complex numbers equals the sum of the real parts of the addends:

$$\Re\{(z_1 + z_2)\} = \Re\{z_1\} + \Re\{z_2\}. \tag{16.12}$$

Let us assume that a complex number is a function of a certain parameter, for example, of the time t:

$$z(t) = x(t) + iy(t).$$

Differentiating this function with respect to t, we get

$$\frac{\mathrm{d}z}{\mathrm{d}t} = \frac{\mathrm{d}x}{\mathrm{d}t} + i\frac{\mathrm{d}y}{\mathrm{d}t}.$$

It thus follows that the real part of the derivative of z with respect to t equals the derivative of the real part of z with respect to t:

$$\Re\left\{\frac{\mathrm{d}z}{\mathrm{d}t}\right\} = \frac{\mathrm{d}}{\mathrm{d}t}\Re\{z\}. \tag{16.13}$$

A similar relation holds upon integration of a complex function. Indeed,

$$\int z(t)\,\mathrm{d}t = \int x(t)\,\mathrm{d}t + i\int y(t)\,\mathrm{d}t,$$

whence it can be seen that the real part of the integral of $z(t)$ equals the integral of the real part of $z(t)$:

$$\Re\left\{\int z(t)\,\mathrm{d}t\right\} = \int \Re\left[z(t)\,\mathrm{d}t\right]. \tag{16.14}$$

It is evident that relations similar to Eqs. (16.12), (16.13), and (16.14) also hold for the imaginary parts of complex functions.

It follows from the above that when the operations of addition, differentiation, and integration are performed with complex functions, and also linear combinations of these operations, the real (imaginary) part of the result coincides with the result that would be obtained when similar operations are performed with the real (imaginary) parts of the same functions[1]. Using the symbol $\widetilde{L}$ to denote a linear combination of the operations listed above, we can write:

$$\Re\left\{\widetilde{L}(z_1, z_2, \ldots)\right\} = \widetilde{L}(\Re\{z_1\}, \Re\{z_2\}, \ldots). \tag{16.15}$$

The property of linear operations we have established makes it possible to use the following procedure in calculations: when performing linear operations with harmonic functions of the form $A\cos(\omega t - k_x x - k_y y - k_z z + \alpha)$, we can replace these functions with the exponents

$$A\exp[i(\omega t - k_x x - k_y y - k_z z + \alpha)] = \hat{A}\exp[i(\omega t - k_x x - k_y y - k_z z)], \tag{16.16}$$

where $\hat{A} = A\,e^{i\alpha}$ is a complex number called the **complex amplitude**. With such representation, we can add functions, differentiate them with respect to the variables t, x, y, z, and also integrate over these variables. In performing the calculations, we must take the real part of the result obtained. The expediency of this procedure is explained by the fact that calculations with exponents are considerably simpler than calculations performed with trigonometric functions.

[1]We must note that this rule cannot be applied to non-linear operations, for example, to the multiplication of functions and squaring them.

Passing over to representation (16.16), we in essence add to all functions of the kind $A\cos(\omega t - k_x x - k_y y - k_z z + \alpha)$ the addends $iA\sin(\omega t - k_x x - k_y y - k_z z + \alpha)$. We remind our reader that we have used a similar procedure when studying forced oscillations (see Sec. 7.12 of Vol. I).

16.3. Reflection and Refraction of a Plane Wave at the Interface Between Two Dielectrics

Assume that a plane electromagnetic wave falls on the plane interface between two homogeneous and isotropic dielectrics. The dielectric in which the incident wave is propagating is characterized by the permittivity ε_1, and the second dielectric by the permittivity ε_2. We assume that the permeabilities are unity. Experiments show that in this case, apart from the plane refracted wave propagating in the second dielectric, a plane reflected wave propagating in the first dielectric is produced.

Let us determine the direction of propagation of the incident wave with the aid of the wave vector $\boldsymbol{k}$, of the reflected wave with the aid of the vector $\boldsymbol{k}'$ and, finally, of the refracted wave with the aid of the vector $\boldsymbol{k}''$. We shall find how the directions of $\boldsymbol{k}'$ and $\boldsymbol{k}''$ are related to the direction of $\boldsymbol{k}$. We can do this by taking advantage of the fact that the following condition must be observed at the interface between the two dielectrics:

$$E_{1,\tau} = E_{2,\tau}. \tag{16.17}$$

Here $E_{1,\tau}$ and $E_{2,\tau}$ are the tangential components of the electric field strength in the first and second medium, respectively.

In Sec. 2.7, we proved Eq. (16.17) for electrostatic fields [see Eq. (2.44)]. It can easily be extended, however, to time-varying fields. According to Eq. (9.5), the circulation of $\boldsymbol{E}$ determined by Eq. (2.42) for varying fields must be not zero, but equal to the integral $\int(-\dot{\boldsymbol{B}})\cdot \mathrm{d}\boldsymbol{S}$ taken over the area of the loop shown in Fig. 2.9:

$$\oint E_l \,\mathrm{d}l = E_{1,x}a - E_{2,x}a + \langle E_b\rangle 2b = -\int_{S=ab} \dot{\boldsymbol{B}}\cdot \mathrm{d}\boldsymbol{S}.$$

Since $\dot{\boldsymbol{B}}$ is finite, in the limit transition $b \to 0$ the integral in the right-hand side vanishes, and we arrive at condition (2.43), from which follows Eq. (2.44).

Assume that the vector $\boldsymbol{k}$ determining the direction of propagation of the incident wave is in the plane of the drawing (Fig. 16.2). The direction of a normal to the interface will be characterized by the vector $\hat{\boldsymbol{n}}$. The plane in which the vectors $\boldsymbol{k}$ and $\hat{\boldsymbol{n}}$ are is called the **plane of incidence** of the wave. Let us take the line of intersection of the plane of incidence with the interface between the dielectrics as the x-axis. We shall direct the y-axis at right angles to the plane of the dielectric interface. The z-axis will, therefore, be perpendicular to the plane of incidence,

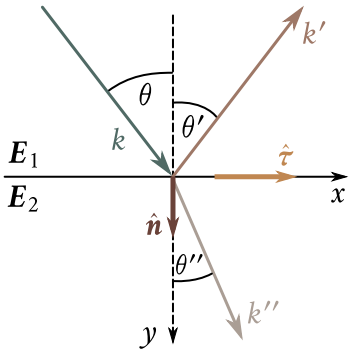


Fig. 16.2

while the vector $\hat{\boldsymbol{\tau}}$ will be directed along the x-axis (see Fig. 16.2).

It is obvious from considerations of symmetry that the vectors $\boldsymbol{k}'$ and $\boldsymbol{k}''$ can only be in the plane of incidence (the media are homogeneous and isotropic). Indeed, assume that the vector $\boldsymbol{k}'$ has deflected from this plane "toward us". There are no grounds, however, to give such a deflection priority over an equal deflection "away from us". Consequently, the only possible direction of $\boldsymbol{k}'$ is that in the plane of incidence. Similar reasoning also holds for the vector $\boldsymbol{k}''$.

Let us separate from a naturally falling ray a plane-polarized component in which the direction of oscillations of the vector $\boldsymbol{E}$ makes an arbitrary angle with the plane of incidence. The oscillations of the vector $\boldsymbol{E}$ in the plane electromagnetic wave propagating in the direction of the vector $\boldsymbol{k}$ are described by the function[2]

$$\boldsymbol{E} = \boldsymbol{E}_{\mathrm{m}} \exp[i(\omega t - \boldsymbol{k} \cdot \boldsymbol{r})] = \boldsymbol{E}_{\mathrm{m}} \exp[i(\omega t - k_x x - k_y y)]$$

(with our choice of the coordinate axes, the projection of the vector $\boldsymbol{k}$ onto the z-axis is zero, therefore, the addend $-k_z z$ is absent in the exponent). By correspondingly choosing the beginning of reading t, we have made the initial phase of the wave equal zero.

The field strengths in the reflected and refracted waves are determined by similar expressions

$$\boldsymbol{E}' = \boldsymbol{E}'_{\mathrm{m}} \exp[i(\omega t - k'_x x - k'_y y + \alpha')]$$

$$\boldsymbol{E}'' = \boldsymbol{E}''_{\mathrm{m}} \exp[i(\omega t - k''_x x - k''_y y + \alpha'')],$$

where α' and α'' are the initial phases of the relevant waves.

[2] More exactly, the real part of this function, but we shall say simply function for brevity's sake.

The resultant field in the first medium is

$$\boldsymbol{E}_1 = \boldsymbol{E} + \boldsymbol{E}' = \boldsymbol{E}_\mathrm{m} \exp[i(\omega t - k_x x - k_y y + \alpha)] + \boldsymbol{E}'_\mathrm{m} \exp[i(\omega t - k'_x x - k'_y y + \alpha')]. \tag{16.18}$$

In the second medium

$$\boldsymbol{E}_2 = \boldsymbol{E}'' = \boldsymbol{E}''_\mathrm{m} \exp[i(\omega t - k''_x x - k''_y y + \alpha'')]. \tag{16.19}$$

According to Eq. (16.17), the tangential components of Eqs. (16.18) and (16.19) must be the same at the interface, *i.e.*, when $y = 0$. We thus arrive at the expression

$$E_{\mathrm{m},\tau} \exp[i(\omega t - k_x x)] + E'_{\mathrm{m},\tau} \exp[i(\omega' t - k_x x' + \alpha')] = E''_{\mathrm{m},\tau} \exp[i(\omega'' t - k_x x + \alpha'')]. \tag{16.20}$$

For condition (16.20) to be observed at any t, all the frequencies must be the same:

$$\omega = \omega' = \omega''. \tag{16.21}$$

To convince ourselves that this is true, let us write Eq. (16.20) in the form

$$a \exp(i\omega t) + b \exp(i\omega' t) = c \exp(i\omega'' t),$$

where the coefficients a, b, and c are independent of t. The equation which we have written is equivalent to the following two:

$$a \cos(\omega t) + b \cos(\omega' t) = c \cos(\omega'' t)$$
$$a \sin(\omega t) + b \sin(\omega' t) = c \sin(\omega'' t).$$

The sum of two harmonic functions will also be a harmonic function only if the functions being added have the same frequencies. The harmonic function obtained as a result of addition will have the same frequency as the summated functions. Hence, follows Eq. (16.21). We have, thus, arrived at the conclusion that the frequencies of the reflected and refracted waves coincide with that of the incident wave.

For condition (16.20) to be observed at any x, the projections of the wave vectors onto the x-axis must be equal:

$$k_x = k'_x = k''_x. \tag{16.22}$$

The angles θ, θ', and θ'' shown in Fig. 16.2 are called the **angle of incidence**, the **angle of reflection**, and the **angle of refraction**. A glance at the figure shows that $k_x = k \sin\theta$, $k'_x = k' \sin\theta'$, $k''_x == k'' \sin\theta''$. Equation (16.22) can therefore be written in the form

$$k \sin\theta = k' \sin\theta' = k'' \sin\theta''.$$

The vectors $\boldsymbol{k}$ and $\boldsymbol{k}'$ have the same magnitude equal to ω/v_1; the magnitude of the

vector $\boldsymbol{k}''$ equals w/v_2. Hence,

$$\frac{\omega}{v_1}\sin\theta = \frac{\omega}{v_1}\sin\theta' = \frac{\omega}{v_2}\sin\theta''.$$

It thus follows that

$$\theta' = \theta, \tag{16.23}$$

$$\frac{\sin\theta}{\sin\theta''} = \frac{v_1}{v_2} = n_{12}. \tag{16.24}$$

The relations we have obtained are obeyed for any plane-polarized component of a natural ray. Hence, they also hold for a natural ray as a whole.

Equation (16.23) expresses the **law of reflection of light**, according to which *the reflected ray lies in one plane with the incident ray and the normal to the point of incidence; the angle of reflection equals the angle of incidence*.

Equation (16.24) expresses the **law of refraction of light**, according to which *the refracted ray lies in one plane with the incident ray and the normal to the point of incidence; the ratio of the sine of the angle of incidence to the sine of the angle of refraction is constant for given substances*.

The quantity n_{12} in Eq. (16.24) is known as the **relative refractive index** of the second substance with respect to the first one. Let us write this quantity in the form

$$n_{12} = \frac{v_1}{v_2} = \frac{c}{v_2}\frac{v_1}{c} = \frac{c/v_2}{c/v_1} = \frac{n_2}{n_1}. \tag{16.25}$$

Thus, the relative refractive index of two substances equals the ratio of their absolute refractive indices.

Substituting the ratio n_2/n_1 for n_{12} in Eq. (16.24), we can write the law of refraction in the form

$$n_1\sin\theta = n_2\sin\theta''. \tag{16.26}$$

Inspection of this equation shows that when light passes from an optically denser medium to an optically less dense one, the rays move away from a normal to the interface of the media. An increase in the angle of incidence θ is attended by a more rapid growth in the angle of refraction θ'', and when the angle θ reaches the value

$$\theta_{cr} = \arcsin n_{12}, \tag{16.27}$$

the angle θ'' becomes equal to $\pi/2$. The angle determined by Eq. (16.27) is called the **critical angle**.

The energy carried by an incident ray is distributed between the reflected and the refracted rays. As the angle of incidence grows, the intensity of the reflected ray increases, while that of the refracted ray diminishes and vanishes at the critical angle. At angles of incidence within the limits from θ_{cr} to $\pi/2$, the light wave penetrates

into the second medium to a distance of the order of a wavelength λ and then returns to the first medium. This phenomenon is called **total internal reflection**.

Let us find the relations between the amplitudes and phases of the incident, reflected, and refracted waves. For simplicity, we shall limit ourselves to the normal incidence of a wave onto the interface between dielectrics (we remind our reader that the dielectrics are assumed to be homogeneous and isotropic). Assume that the oscillations of the vector $\boldsymbol{E}$ in the falling wave occur along the direction which we shall take as the x-axis. It follows from considerations of symmetry that the oscillations of the vectors $\boldsymbol{E}'$ and $\boldsymbol{E}''$ also occur along the x-axis. In the given case, the unit vector $\hat{\boldsymbol{\tau}}$ coincides with the unit vector $\hat{\boldsymbol{e}}_x$. Therefore, the condition of continuity of the tangential component of the electric field strength has the form

$$E_x + E'_x = E''_x. \tag{16.28}$$

Expression (16.8) obtained for the amplitude values of E and H also holds for their instantaneous values: $H \propto nE$. It thus follows that the instantaneous value of the energy flux density is proportional to nE^2. Thus, the law of energy conservation leads to the equation

$$n_1 E_x^2 = n_1 E_x'^2 + n_2 E_x''^2. \tag{16.29}$$

We must note that the quantities E_x, E'_x and E''_x in Eqs. (16.28) and (16.29) are the instantaneous values of the projections.

Introducing $E''_x - E_x$ into Eq. (16.29) instead of E'_x [see Eq. (16.28)], it is easy to see that

$$E''_x = \left(\frac{2n_1}{(n_1 + n_2)}\right) E_x. \tag{16.30}$$

Using this value of E''_x in Eq. (16.28), we find that

$$E'_x = \left(\frac{n_1 - n_2}{n_1 + n_2}\right) E_x. \tag{16.31}$$

Examination of Eq. (16.30) shows that the projections of the vectors $\boldsymbol{E}$ and $\boldsymbol{E}''$ have identical signs at each moment of time. Hence, we conclude that the oscillations in the incident wave and in the one passing into the second medium occur at the interface in the same phase—when a wave passes through the interface there is no jump in the phase.

It can be seen from Eq. (16.31) that when $n_2 < n_1$, the sign of E'_x coincides with that of E_x. This signifies that the oscillations in the incident and reflected waves occur at the interface in the same phase—the phase of a wave does not change upon reflection. If $n_2 > n_1$, then the sign of E'_x is opposite to that of E_x, the oscillations in the incident and reflected waves occur at the interface in counterphase—the phase of the wave changes in a jump by π upon reflection. The result obtained also holds

upon the inclined falling of a wave at the interface between two transparent media.

Thus, when a light wave is reflected from an interface between an optically less dense medium and an optically denser one (when $n_1 < n_2$), the phase of oscillations of the light vector changes by π. Such a phase change does not occur upon reflection from an interface between an optically denser medium and an optically less dense one (when $n_1 > n_2$).

Equations (16.30) and (16.31) have been obtained for the instantaneous values of the projections of the light vectors. Similar relations also hold for the amplitudes of the light vectors:

$$E''_{\mathrm{m}} = \left(\frac{2n_1}{(n_1 + n_2)}\right) E_{\mathrm{m}}, \quad E'_{\mathrm{m}} = \left|\frac{n_1 - n_2}{n_1 + n_2}\right| E_{\mathrm{m}}. \tag{16.32}$$

These relations make it possible to find the reflection coefficient ρ and the transmission coefficient τ of a light wave (for normal incidence at the interface between two transparent media). Indeed, by definition

$$\rho = \frac{I'}{I} = \frac{n_1 E'^2_{\mathrm{m}}}{n_1 E^2_{\mathrm{m}}},$$

where I' is the intensity of the reflected wave, and I is the intensity of the incident one. Using in this equation the ratio $E'_{\mathrm{m}}/E_{\mathrm{m}}$ obtained from Eq. (16.32), we arrive at the formula

$$\rho = \left(\frac{n_{12} - 1}{n_{12} + 1}\right)^2. \tag{16.33}$$

Here, $n_{12} = n_2/n_1$ is the refractive index of the second medium relative to the first one.

We get the following expression for the transmission coefficient:

$$\tau = \frac{I''}{I} = \frac{n_2 E''^2_{\mathrm{m}}}{n_1 E^2_{\mathrm{m}}} = n_{12}\left(\frac{2}{n_{12} + 1}\right)^2. \tag{16.34}$$

We must note that the substitution for n_{12} in Eq. (16.33) of its reciprocal $n_{21} = 1/n_{12}$ does not change the value of ρ. Hence, the coefficient of reflection of the interface between two given media has the same value for both directions of propagation of light.

The index of refraction for glass is close to 1.5. Introducing $n_{12} = 1.5$ into Eq. (16.33), we get $\rho = 0.04$. Thus, each surface of a glass plate reflects (with incidence close to normal) about four per cent of the luminous energy falling on it.

16.4. Luminous Flux

A real light wave is a superposition of waves with lengths confined within the interval $\Delta\lambda$. The latter is finite even for monochromatic (single-coloured) light. In white light, $\Delta\lambda$ covers the entire range of electromagnetic waves perceived by the eye, *i.e.*, it ranges from 0.40 μm to 0.76 μm.

The distribution of the energy flux by wavelengths can be characterized with the aid of the distribution function

$$\varphi(\lambda) = \frac{\mathrm{d}\Phi_{\mathrm{en}}}{\mathrm{d}\lambda}, \tag{16.35}$$

where $\mathrm{d}\Phi_{\mathrm{en}}$ is the energy flux falling to the wavelengths from λ to $\lambda + \Delta\lambda$. Knowing the form of function (16.35), we can calculate the energy flux transferred by waves whose lengths are within the finite interval from λ_1 to λ_2:

$$\Phi_{\mathrm{en}} = \int_{\lambda_1}^{\lambda_2} \varphi(\lambda)\, \mathrm{d}\lambda. \tag{16.36}$$

The action of light on the eye (the perception of light) depends quite greatly on the wavelength. This is easy to understand if we take into account that electromagnetic waves with λ below 0.40 μm and above 0.76 μm are not perceived at all by the human eye. The sensitivity of an average normal human eye to radiation of various wavelengths can be depicted graphically by a **curve of relative spectral sensitivity** (Fig. 16.3). The wavelength λ is laid off along the horizontal axis, and the relative spectral sensitivity $V(\lambda)$ along the vertical one. The eye is most sensitive to radiation of the wavelength 0.555 μm[3] (the green part of the spectrum). The function $V(\lambda)$ for this wavelength is taken equal to unity. The luminous intensity estimated visually for other wavelengths is lower, although the energy flux is the same. Accordingly, $V(\lambda)$ for these wavelengths is also less than unity. The values of the function $V(\lambda)$ are inversely proportional to the values of the energy fluxes producing a visual sensation identical in intensity:

$$\frac{V(\lambda_1)}{V(\lambda_2)} = \frac{(\mathrm{d}\Phi_{\mathrm{en}})_2}{(\mathrm{d}\Phi_{\mathrm{en}})_1}.$$

For example, $V(\lambda) = 0.5$ signifies that for obtaining a visual sensation of the same intensity, light of the given wavelength must have a density of the energy flux twice that of light for which $V(\lambda) = 1$. Outside of the interval of visible wavelengths, the function $V(\lambda)$ is zero.

The quantity Φ called the **luminous flux** is introduced to characterize the luminous intensity with account of its ability to produce a visual sensation. For the

[3]It is interesting to note that this wavelength is represented with the greatest intensity in solar radiation.

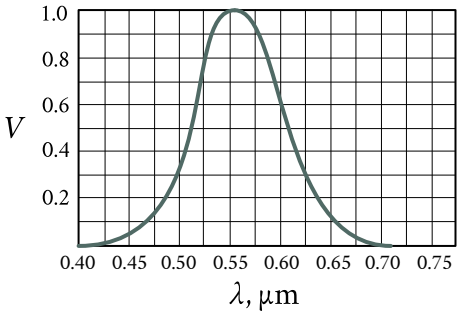


Fig. 16.3

interval $d\lambda$, the luminous flux is determined as the product of the energy flux and the corresponding value of the function $V(\lambda)$:

$$d\Phi = V(\lambda)\, d\Phi_{en}. \tag{16.37}$$

Expressing the energy flux through the function of energy distribution by wavelengths [see Eq. (16.35)], we get

$$d\Phi = V(\lambda)\varphi(\lambda)\, d\lambda. \tag{16.38}$$

The total luminous flux is

$$\Phi = \int_0^\infty V(\lambda)\varphi(\lambda)\, d\lambda. \tag{16.39}$$

The function $V(\lambda)$ is a dimensionless quantity. Consequently, the dimension of luminous flux coincides with that of energy flux. This makes it possible to define the luminous flux as the flux of luminous energy assessed according to its visual sensation.

16.5. Photometric Quantities and Units

Photometry is the branch of optics occupied in measuring luminous fluxes and quantities related to them.

Luminous Intensity. A source of light whose dimensions may be disregarded in comparison with the distance from the place of observation to the source is called a **point source**. In a homogeneous and isotropic medium, the wave emitted by a point source will be spherical. Point sources of light are characterized by the luminous intensity I determined as the luminous flux emitted by a source per unit solid angle:

$$I = \frac{d\Phi}{d\Omega} \tag{16.40}$$

$d\Phi$ is the luminous flux emitted by a source within the limits of the solid angle $d\Omega$).

In the general case, the luminous intensity depends on the direction: $I = I(\theta, \varphi)$ (here θ and φ are the polar and the azimuth angles in a spherical system of coordinates). If I does not depend on the direction, the light source is called **isotropic**. For an isotropic source

$$I = \frac{\Phi}{4\pi}, \tag{16.41}$$

where Φ is the total luminous flux emitted by the source in all directions.

When dealing with an extended source, we can speak of the luminous intensity of an element of its surface dS. Now by dΦ in Eq. (16.40) we must understand the luminous flux emitted by the surface element dS within the limits of the solid angle dΩ.

The unit of luminous intensity—the candela (cd) is one of the basic SI units. It is defined as the luminous intensity, in the perpendicular direction, of a surface of 1/600000 square metre of a complete radiator at the temperature of freezing platinum under a pressure of 101325 pascals. By a complete radiator is meant a device having the properties of a blackbody (see Vol. III).

Luminous Flux. The unit of luminous flux is the lumen (lm). It equals the luminous flux emitted by an isotropic source with a luminous intensity of 1 candela within a solid angle of one steradian:

$$1\ \text{lm} = 1\ \text{cd} \cdot 1\ \text{sr}. \tag{16.42}$$

It has been established experimentally that an energy flux of 0.0016 W corresponds to a luminous flux of 1 lm formed by radiation having a wavelength of $\lambda = 0.555\ \mu\text{m}$. The energy flux

$$\Phi_{\text{en}} = \frac{0.0016}{V(\lambda)}\text{W}, \tag{16.43}$$

corresponds to a luminous flux of 1 lm formed by radiation of a different wavelength.

Illuminance. The degree of illumination of a surface by the light falling on it is characterized by the quantity

$$E = \frac{\text{d}\Phi_{\text{inc}}}{\text{d}S}, \tag{16.44}$$

known as the **illuminance** or **illumination** (dΦ_{inc} is the luminous flux incident on the surface element dS).

The unit of illuminance is the lux (lx) equal to the illuminance produced by a flux of 1 lm uniformly distributed over a surface having an area of 1 m^2:

$$1\ \text{lx} = 1\ \text{lm} : 1\ \text{m}^2. \tag{16.45}$$

The illuminance E produced by a point source can be expressed through the luminous intensity I, the distance r from the surface to the source, and the angle α between a normal to the surface $\hat{\boldsymbol{n}}$ and the direction to the source. The flux incident

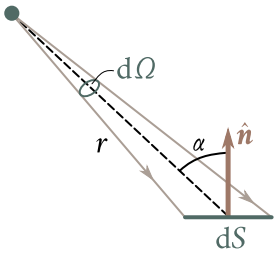


Fig. 16.4

on the area dS (Fig. 16.4) is $d\Phi_{inc} = I\, d\Omega$ and it is confined within the solid angle $d\Omega$ subtended by dS. The angle $d\Omega$ is $dS\cos\alpha/r^2$. Hence, $d\Phi_{inc} = I\, dS\cos\alpha/r^2$. Dividing this flux by dS, we get

$$E = \frac{I\cos\alpha}{r^2}. \tag{16.46}$$

Luminous Emittance. An extended source of light can be characterized by the luminous emittance M of its various sections, by which is meant the luminous flux emitted outward by unit area in all directions (within the limits of values of θ from 0 to $\pi/2$, where θ is the angle made by the given direction with an external normal to the surface):

$$M = \frac{d\Phi_{em}}{dS} \tag{16.47}$$

($d\Phi_{em}$ is the flux emitted outward in all directions by the surface elements dS of the source).

Luminous emittance may appear as a result of a surface reflecting the light falling on it. Here, by $d\Phi_{em}$ in Eq. (16.47), we must understand the flux reflected by the surface element dS in all directions.

The unit of luminous emittance is the lumen per square metre ($\mathrm{lm\,m^{-2}}$).

Luminance. Luminous emittance characterizes radiation (or reflection) of light by a given place of a surface in all directions. The radiation (reflection) of light in a given direction is characterized by the luminance L. The direction can be given by the polar angle θ (measured from the outward normal $\hat{\boldsymbol{n}}$ to the emitting surface area ΔS) and the azimuth angle φ. Luminance is defined as the ratio of the luminous intensity of an elementary surface area ΔS in a given direction to the projection of the area ΔA onto a plane perpendicular to the chosen direction.

Let us consider the elementary solid angle $d\Omega$ subtended by the luminous area dS and oriented in the direction (θ, φ) (Fig. 16.5). The luminous intensity of area ΔS in the given direction, according to Eq. (16.40), is $I = d\Phi/d\Omega$, where $d\Phi$ is the luminous flux propagating within the limits of the angle $d\Omega$. The projection of ΔS onto a plane normal to the direction (θ, φ) (in Fig. 16.5 the trace of this plane is

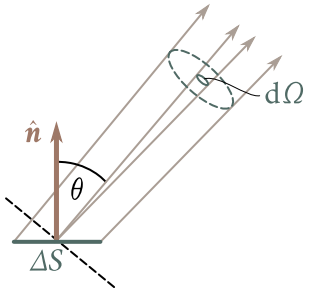


Fig. 16.5

depicted by a dash line) is $\Delta S \cos\theta$. Hence, the luminance is

$$L = \frac{\mathrm{d}\Phi}{\mathrm{d}\Omega\, \Delta S \cos\theta}. \tag{16.48}$$

In the general case, the luminance differs for different directions: $L = L(\theta, \varphi)$. Like the luminous emittance, the luminance can be used to characterize a surface that reflects the light falling on it.

In accordance with Eq. (16.48), the flux emitted by the area ΔS within the limits of the solid angle $\mathrm{d}\Omega$ in the direction determined by θ and φ is

$$\mathrm{d}\Phi = L(\theta, \varphi)\, \mathrm{d}\Omega\, \Delta S \cos\theta. \tag{16.49}$$

A source whose luminance is identical in all directions (L = constant) is called a **Lambertian source** (obeying Lambert's law) or a cosine source (the flux emitted by a surface element of such a source is proportional to $\cos\theta$). Only a blackbody strictly observes Lambert's law.

The luminous emittance M and luminance L of a Lambertian source are related by a simple expression. To find it, let us introduce $\mathrm{d}\Omega = \sin\theta\, \mathrm{d}\theta\, \mathrm{d}\varphi$ into Eq. (16.49) and integrate the expression obtained with respect to φ within the limits from 0 to 2π and with respect to θ from 0 to $\pi/2$, taking into account that L = constant. As a result, we shall find the total light flux emitted by surface element ΔS of a Lambertian source outward in all directions:

$$\Delta\Phi_{\mathrm{em}} = L\Delta S \int_0^{2\pi} \mathrm{d}\varphi \int_0^{\pi/2} \sin\theta \cos\theta\, \mathrm{d}\theta = \pi L \Delta S.$$

We get the luminous emittance by dividing this flux by ΔS. Thus, for a Lambertian source, we have

$$M = \pi L. \tag{16.50}$$

The unit of luminance is the candela per square metre (cd m^{-2}). A uniformly luminous plane surface has a luminance of 1 cd m^{-2} in a direction normal to it if in this direction the luminous intensity of one square metre of surface is one candela.

16.6. Geometrical Optics

The lengths of light waves perceived by the human eye are very small (of the order of 10^{-7} m). For this reason, the propagation of visible light in a first approximation can be considered without giving attention to its wave nature and assuming that light propagates along lines called **rays**. In the limiting case corresponding to $\lambda \to 0$, the laws of optics can be formulated using the language of geometry.

Accordingly, the branch of optics in which the finiteness of the wavelengths is disregarded is known as **geometrical optics**. Another name for it is **ray optics**.

Geometrical optics is based on four laws: (1) the law of propagation of light along a straight line; (2) the law of independence of light rays; (3) the law of light reflection; and (4) the law of refraction.

The **law of straight-line propagation** states that *in a homogeneous medium light propagates in a straight line*. This law is approximate—when light passes through very small openings, deviations from a straight line are observed that increase with a diminishing size of the opening.

The **law of independence of light rays** states that *rays do not disturb one another when they intersect*. The intersection of rays does not hinder each of them from propagating independently of the others. This law holds only at not too great luminous intensities. At intensities reached with the aid of lasers, the independence of light rays stops being observed.

The laws of reflection and refraction of light were formulated in Sec. 16.3 [see Eqs. (16.23) and (16.24) and the text following them].

Geometrical optics can be based on the principle established by the French mathematician Pierre de Fermat (1601-1665). It underlies the laws of straight-line propagation, reflection, and refraction of light. As formulated by Fermat himself, this principle states that *any light ray will travel between two end points along a line requiring the minimum transit time*.

Light needs the time $dt = ds/v$, where v is the speed of light at the given point of the medium, to cover the distance ds (Fig. 16.6). Replacing v with c/n [see Eq. (16.2)], we find that $dt = (1/c)n\,ds$. Consequently, the time τ spent by light in covering the distance from point 1 to point 2 is

$$\tau = \frac{1}{c}\int_1^2 n\,ds. \tag{16.51}$$

The quantity

$$L = \int_1^2 n\,ds \tag{16.52}$$

having the dimension of length is called the **optical path**. In a homogeneous

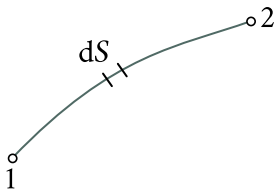


Fig. 16.6

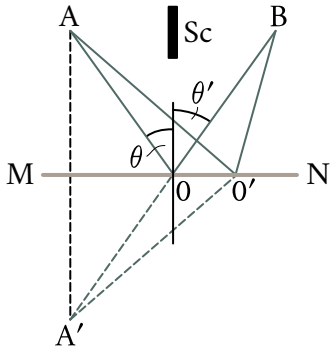


Fig. 16.7

medium, the optical path equals the product of the geometrical path s and the index of refraction n of the medium:

$$L = ns. \tag{16.53}$$

According to Eqs. (16.51) and (16.52), we have

$$\tau = \frac{L}{c}. \tag{16.54}$$

The proportionality of the time τ of covering a path to the optical path L makes it possible to word Fermat's principle as follows: *light travels along a path whose optical length is minimum*. More exactly, the optical path must be extremal, *i.e.*, either minimum or maximum, or stationary—identical for all possible paths. In the last case, all the paths of light between two points are **tautochronous** (requiring the same time for covering them).

The reversibility of light rays ensues from Fermat's principle. Indeed, the optical path that is minimum when light travels from point 1 to point 2 is also minimum when light travels in the opposite direction. Consequently, a ray emitted toward another one that has travelled from point 1 to point 2 will cover the same path, but in the opposite direction.

Let us use Fermat's principle to obtain the laws of reflection and refraction of light. Assume that a light ray reaches point B from point A after being reflected from surface MN (Fig. 16.7, the straight path from A to B is blocked by opaque screen Sc). The medium in which the ray travels is homogeneous. Therefore, the minimality of the optical length consists in the minimality of its geometrical length. The geometrical length of an arbitrarily taken path is A0′B = A′0′B (auxiliary point A′ is a mirror image of point A). A glance at the figure shows that the path of the ray reflected at point 0 will be the shortest. At this point the angle of reflection equals the angle of incidence. We must note that when point 0′ moves away from point 0,

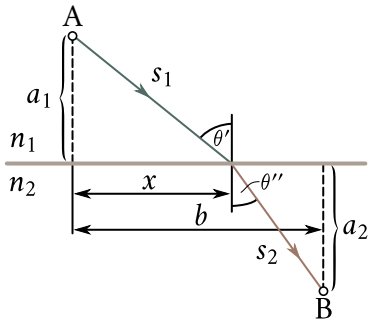


Fig. 16.8

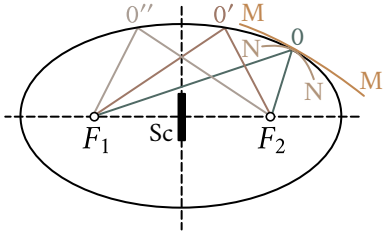


Fig. 16.9

the geometrical path grows unlimitedly so that in the given case we have only one extreme—a minimum.

Now let us find the point at which a ray travelling from A to B must be refracted for the optical path to be extremal (Fig. 16.8). The optical path for an arbitrary ray is

$$L = n_1 s_1 + n_2 s_2 + n_1\sqrt{a_1^2 + x^2} + n_2\sqrt{a_2^2 + (b-x)^2}.$$

To find the extreme value, let us differentiate L with respect to x and equate the derivative to zero:

$$\frac{\mathrm{d}L}{\mathrm{d}x} = \frac{n_1 x}{\sqrt{a_1^2 + x^2}} - \frac{n_2(b-x)}{\sqrt{a_2^2 + (b-x)^2}} = n_1\frac{x}{s_1} - n_2\frac{(b-x)}{s_2} = 0.$$

The factors of n_1 and n_2 equal $\sin\theta$ and $\sin\theta''$, respectively. We, thus, get the relation

$$n_1 \sin\theta = n_2 \sin\theta'',$$

expressing the law of refraction [see Eq. (16.26)].

Let us consider reflection from the inner surface of an ellipsoid of revolution (Fig. 16.9; F_1 and F_2 are the foci of the ellipsoid). According to the definition of an ellipse, the paths $F_1 0 F_2$, $F_1 0' F_2$, $F_1 0'' F_2$, etc. are identical in length. Hence, all the rays leaving focus F_1 and arriving after reflection at focus F_2 are tautochronous. In this case, the optical path is stationary. If we replace the surface of the ellipsoid with surface MM having a smaller curvature and oriented so that a ray leaving point F_1 arrives at point F_2 after being reflected from MM, then path $F_1 0 F_1$ will be minimum. For surface NN whose curvature is greater than that of the ellipsoid, path $F_1 0 F_2$ will be maximum.

The optical paths are also stationary when the rays pass through a lens (Fig. 16.10). Ray P0P′ has the shortest path in air (where the index of refraction n is virtually equal to unity) and the longest path in glass ($n \sim 1.5$). Ray PQQ′P′ has the longest path in air, but a shorter one in glass. As a result, the optical paths will be the

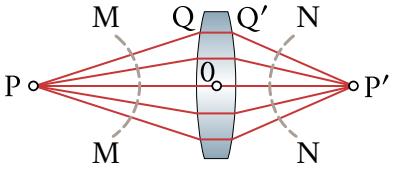


Fig. 16.10

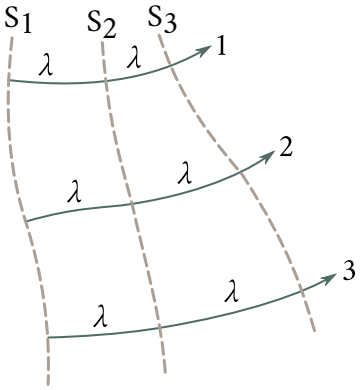


Fig. 16.11

same for all the rays. Hence, the latter are tautochronous, and the optical path is stationary.

Let us consider a wave propagating in a non-homogeneous isotropic medium along rays 1, 2, 3, etc. (Fig. 16.11). We shall consider that the non-homogeneity is sufficiently small for us to assume the index of refraction to be constant on sections of the rays of length λ. We shall construct wave surfaces S_1, S_2, S_3, etc., so that the oscillations at the points of each following surface, lag in phase by 2π behind the oscillations at the points on the preceding surface. The oscillations at points on the same ray are described by the equation $\xi = A\cos(\omega t - kr + \alpha)$ (here, r is the distance measured along the ray). The lag in phase is determined by the expression $k\Delta r$, where Δr is the distance between adjacent surfaces. From the condition $k\Delta r = 2\pi$, we find that $\Delta r = 2\pi/k = \lambda$. The optical length of each of the paths of geometrical length λ is $n\lambda = \lambda_0$ [see Eq. (16.5)]. According to Eq. (16.54), the time τ during which light covers a path is proportional to the optical length of the path. Consequently, the equality of the optical paths signifies equality of the times needed for light to travel the relevant paths. We, thus, arrive at the conclusion that sections of rays confined between two wave surfaces have identical optical paths and are tautochronous. In particular, the sections of the rays between wave surfaces MM and NN depicted by dash lines in Fig. 16.10 are tautochronous.

It can be seen from our treatment that the lag in phase δ appearing on the optical path L is determined by the expression

$$\delta = \frac{L}{\lambda_0} 2\pi \tag{16.55}$$

(λ_0 is the length of a wave in a vacuum).

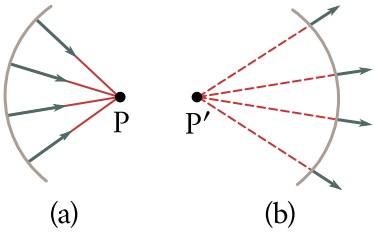


Fig. 16.12

16.7. Centered Optical System

A collection of rays forms a beam. If rays when continued intersect at one point, the beam is called homocentric. A spherical wave surface corresponds to a homocentric beam of rays. Figure 16.12a shows a converging, and Fig. 16.12b a diverging homocentric beam. A particular case of a homocentric beam is a beam of parallel rays; a plane light wave corresponds to it.

Any optical system transforms light beams. If the system does not violate the homocentricity of the beams, then the rays emerging from point P intersect at one point P′. This point is the **optical image** of point P. If a point of an object is depicted in the form of a point, the image is called a **point** or a **stigmatic** one.

An image is called **real** if the light rays actually intersect at point P′ (see Fig. 16.12a), and virtual if the continuations of the rays in a direction opposite to the direction of propagation of the light intersect at P′ (see Fig. 16.12b).

Owing to the reversibility of light rays, light source P and image P′ may exchange roles—a point source placed at P′ will have its image at P. For this reason, P and P′ are called **conjugate points**.

An optical system that produces a stigmatic image which is geometrically similar to the object it depicts is called **ideal**. With the aid of such a system, a space continuity of points P is depicted in the form of a space continuity of points P′. The first continuity of points is known as the **object space**, and the second one as the **image space**. In both spaces, points, straight lines, and planes uniquely correspond to one another. Such a relation of two spaces is called **collinear correspondence** in geometry.

An optical system is a collection of reflecting and refracting surfaces separating optically homogeneous media from one another. These surfaces are usually spherical or plane (a plane can be considered as a sphere of infinite radius). More complicated surfaces such as an ellipsoid, hyperboloid or paraboloid of revolution are used much less frequently.

An optical system formed by spherical (in particular, by plane) surfaces is called

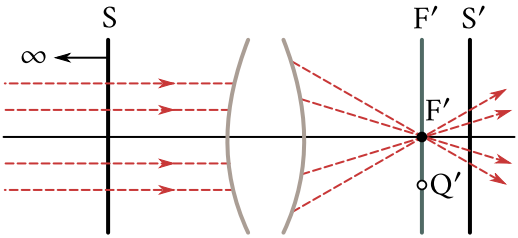


Fig. 16.13

centered if the centres of all the surfaces are on a single straight line. This line is called the **optical axis** of the system. To each point P or plane S in object space there corresponds its conjugate point P′ or plane S′ in image space. The infinite multitude of conjugate points and conjugate planes includes points and planes having special properties. Such points and planes are called **cardinal** ones. Among them are the **focal**, **principal**, and **nodal** points and planes. Setting of the cardinal points or planes completely determines the properties of an ideal centered optical system.

Focal Planes and Focal Points of an Optical System. Figure 16.13 shows the external refracting surfaces and the optical axis of an ideal centered optical system. Let us take plane S perpendicular to the optical axis in the object space of this system. It follows from considerations of symmetry that plane S′ conjugate to S is also perpendicular to the optical axis. Displacement of plane S relative to the system will produce a corresponding displacement of plane S′. When plane S is very far, a further increase in its distance from the system will produce virtually no change in the position of plane S′. This signifies that as a result of removing plane S to infinity, plane S′ will be in a definite extreme position F′. Pland F′ coinciding with the extreme position of plane S′ is called the **second** (or **hack**) **focal plane** of the optical system. We can say briefly that the second focal plane F′ is defined as a plane conjugate to plane S_∞ perpendicular to the axis of the system and at infinity in the object space.

The point of intersection of the second focal plane with the optical axis is known as the **second** (or **hack**) **focal point** (**focus**) of the system. It is also designated by the letter F′. This point is conjugate to point P_∞ on the axis of the system at infinity. Rays emerging from P_∞ form a beam parallel to the axis (see Fig. 16.13). When they leave the system, these rays form a beam converging at focal point F′. A parallel beam impinging on the system may leave it not in the form of a converging beam (as in Fig. 16.13), but in the form of a diverging one. Hence, what intersects at point F′ will be not the actual rays that emerge, but their extensions in the reverse direction. Accordingly, the second focal plane will he in front (in the direction of the rays) of

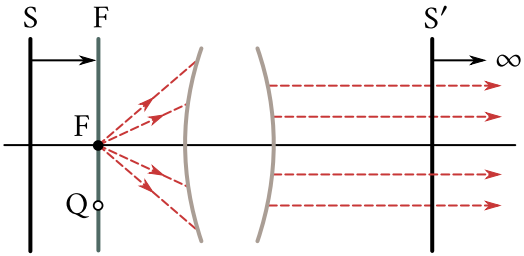


Fig. 16.14

the system or inside it.

The rays emanating from an infinitely remote point Q_∞, not lying on the axis of the system form a parallel beam directed at an angle to the axis of the system. Upon emerging from the system, these rays form a beam converging at point Q′ belonging to the second focal plane, but not coinciding with focal point F′ (see point Q′ in Fig. 16.13). It follows from the above that the image of an infinitely remote object will be in the focal plane.

If we remove plane S′ perpendicular to the axis to infinity (Fig. 16.14), its conjugate plane S will advance to its extreme position F called the **first** (or **front**) **focal plane** of the system. We can say for short that the first focal plane F is a plane conjugate to planes S'_∞ perpendicular to the axis of the system and at infinity in the image space.

The point of intersection of first focal plane F with the optical axis is called the **first** (or **front**) **focal point** (**focus**) of the system. This point is also designated by the symbol F. The rays emerging from focal point F form a beam of rays parallel to the axis after leaving the system. The rays emerging from point Q belonging to focal plane F (see Fig. 16.14) form a parallel beam directed at an angle to the axis of the system after passing through the latter. It may happen that a beam which is parallel upon leaving a system is obtained when a converging beam of light falls on the system instead of a diverging one (as in Fig. 16.14). In this case, the first focal point is either beyond the system or inside it.

Principal Planes and Points. Let us consider two conjugate planes at right angles to the optical axis of the system. Arrow y (Fig. 16.15) in one of these planes will have as its image arrow y' in the other plane. It follows from axial symmetry of the system that arrows y and y' must be in the same plane passing through the optical axis (in the plane of the drawing). The image y' may be in the same direction as object y (see Fig. 16.15a), or in the opposite direction (see Fig. 16.15b). In the first case, the image is called **erect**, in the second—**inverted**. Segments laid off upward from an optical axis are considered to be positive, and those laid off

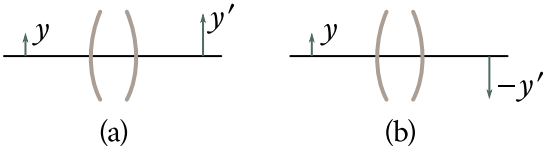


Fig. 16.15

downward—negative. The actual lengths of the segments are shown in drawings, *i.e.*, the positive quantities $(-y)$ and $(-y')$ for negative segments.

The ratio of the linear dimensions of an image and an object is called the **linear** (**longitudinal**) or the **lateral magnification**. Designating it by the symbol M, we can write

$$M = \frac{y'}{y}. \tag{16.56}$$

The linear magnification is an algebraic quantity. It is positive if the image is erect (the signs of y and y' are the same) and negative if the image is inverted (the signs of y and y' are opposite).

We can prove that there are two conjugate planes which reflect each other with a linear magnification of $M = +1$. These planes are known as the **principal ones**. The plane belonging to the object space is called the **first** (or **front**) principal plane of a system. It is designated by the symbol H. The plane belonging to the image space is called the **second** (or **back**) **principal plane**. Its symbol is H′. The points of intersection of the principal planes with the optical axis are called the **principal points** of the system (first and second, respectively). They are designated by the same symbols H and H′. Depending on the design of a system, its principal planes and points may be either outside or inside the system. One of the planes may be outside and the other inside a system. Finally, both planes may be outside a system at the same side of it.

It can be seen from the definition of the principal planes that ray 1 intersecting (actually—Fig. 16.16a, or when virtually continued inside the system—Fig. 16.16b) the first principal plane H at point Q has as its conjugate ray 1′ that intersects (directly or upon virtual continuation) principal plane H′ at point Q′. The latter is in the same direction and at the same distance from the axis as point Q. This is easy to understand if we remember that Q and Q′ are conjugate points, and take into account that any ray passing through point Q must have as its conjugate a ray passing through point Q′.

Nodal Planes and Nodal Points. Conjugate points N and N' lying on the optical axis and having the property that the conjugate rays passing through them (actually or when imaginarily continued inside the system) are parallel to each

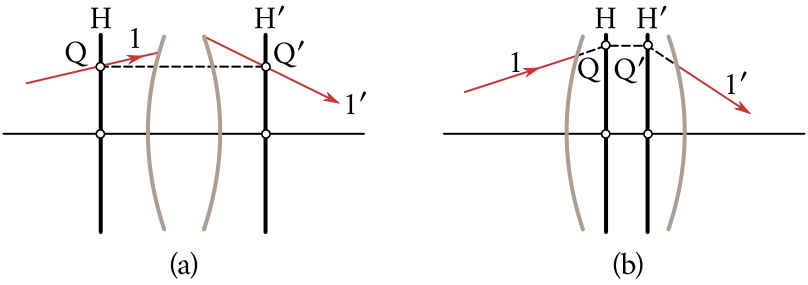


Fig. 16.16

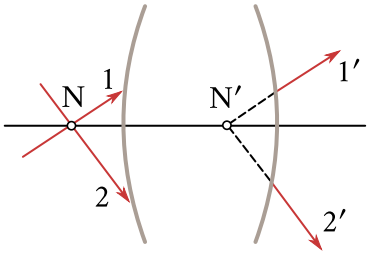


Fig. 16.17

other are called **nodal points** or **nodes** (see rays 1-1′ and 2-2′ in Fig. 16.17). Planes perpendicular to the axis and passing through the nodal points are called **nodal planes** (first and second).

The distance between the nodal points always equals that between the principal points. When the optical properties of the media at both sides of the system are the same (*i.e.*, $n = n'$), the nodal and principal points coincide.

Focal Lengths and Optical Power of a System. The distance from first principal point H to first focal point F is called the **first focal length** f of the system. The distance from H' to F' is known as the **second focal length** f'. The focal lengths f and f', are algebraic quantities. They are positive if a given focal point is at the right of the relevant principal point, and negative in the opposite case. For example, for the system shown in Fig. 16.18 (see below), the second focal length f' is positive, and the first focal length f is negative. The figure depicts the true length of HF, *i.e.*, the positive quantity $(-f)$ equal to the absolute value of f.

We can show that the following relation holds between the focal lengths f and f' of a centered optical system formed by spherical refracting surfaces:

$$\frac{f}{f'} = -\frac{n}{n'}, \tag{16.57}$$

where n is the refractive index of the medium in front of the optical system, and n' is the refractive index of the medium behind the system. Examination of Eq. (16.57)

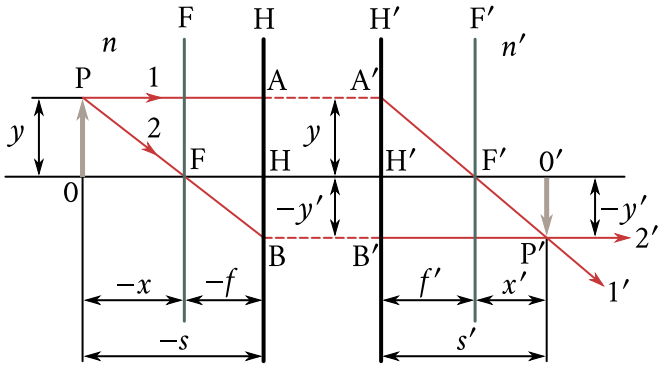


Fig. 16.18

shows that when the refractive indices of the media at both sides of an optical system are the same, the focal lengths differ only in their sign:

$$f' = -f. \tag{16.58}$$

The quantity

$$P = \frac{n'}{f'} = -\frac{n}{f}, \tag{16.59}$$

is known as the **optical power** of a system. When P grows, the focal length f' diminishes, and, consequently, the rays are refracted by the optical system to a greater extent. The optical power is measured in dioptres (D). To obtain P in dioptres, the focal length in Eq. (16.59) must be taken in metres. When P is positive, the second focal length f' is also positive; hence, the system produces a real image of an infinitely remote point—a parallel beam of rays is transformed into a converging one. In this case, the system is called **converging**. When P is negative, the image of an infinitely remote point will be virtual—a parallel beam of rays is transformed by the system into a diverging one. Such a system is called **diverging**.

Formula of a System. We completely determine the properties of an optical system by setting its cardinal planes or points. In particular, knowing the position of the cardinal planes, we can construct the optical image produced by a system. Let us take segment OP perpendicular to the optical axis in the object space (Fig. 16.18, the nodal points are not shown in the figure). The position of this segment can be set either by the distance x measured from point F to point 0, or by the distances from H to 0. The quantities x and s, like the focal lengths f and f', are algebraic ones (their magnitudes are shown in figures).

Let us draw ray 1 parallel to the optical axis from point P. It will intersect plane H at point A. In accordance with the properties of principal planes, ray 1′ conjugate to ray 1 must pass through point A′ of plane H′ conjugate to point A. Since ray 1

is parallel to the optical axis, then ray 1′ conjugate to it will pass through second focal point F′. Now let us draw ray 2 passing through the first focal point F from point P. It will intersect plane H at point B. Ray 2′ conjugate to it will pass through point B′ of plane H′ conjugate to B and will be parallel to the optical axis. Point P′ of intersection of rays 1′ and 2′ is the image of point P. Image 0′P′, like object OP, is perpendicular to the optical axis.

The position of image 0′P′ can be characterized either by the distance x' from point F′ to point 0′ or by the distance s' from H′ to 0′. The quantities x' and s' are algebraic ones. For the case shown in Fig. 16.18, they are positive.

The quantity x' determining the position of the image is related to the quantity x determining the position of the object and to the focal lengths f and f'. For the right triangles with a common apex at point F (see Fig. 16.18), we can write the relation

$$\frac{\mathrm{OP}}{\mathrm{HB}} = \frac{y}{-y'} = \frac{-x}{-f}. \tag{16.60}$$

Similarly, for the triangles with their common apex at point F′, we have

$$\frac{\mathrm{H'A'}}{\mathrm{O'P'}} = \frac{y}{-y'} = \frac{f'}{x'}. \tag{16.61}$$

Combining both relations, we find that $(-x)/(-f) = f'/x'$, whence

$$xx' = ff'. \tag{16.62}$$

This equation is known as **Newton's formula**. For the condition that $n = n'$, Newton's formula has the form

$$xx' = -f^2 \tag{16.63}$$

[see Eq. (16.57)].

It is easy to pass over from the formula relating the distances x and x' to the object and to the image from the focal points of a system to a formula establishing the relation between the distances s and s' from the principal points. A glance at Fig. 16.18 shows that $(-x) == (-s) - (-f)$ (*i.e.*, $x = s - f$), and $x' = s' - f'$. Introducing these expressions for x and x' into Eq. (16.62) and making the relevant transformations, we get

$$\frac{f}{s} + \frac{f'}{s'} = 1. \tag{16.64}$$

When the condition is observed that $f' = -f$ [see Eq. (16.58)], Eq. (16.64) is simplified as follows:

$$\frac{1}{s} - \frac{1}{s'} = \frac{1}{f}. \tag{16.65}$$

Equations (16.62)-(16.65) are equations of a centered optical system.

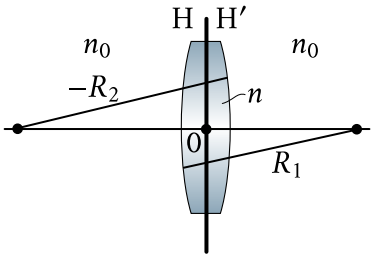


Fig. 16.19

16.8. Thin Lenses

A **lens** is a very simple centered optical system. It is a transparent (usually glass) body bounded by two spherical surfaces[4] (in a particular case one of the surfaces can be plane). The points of intersection of the surfaces with the optical axis of a lens are called the **apices** of the refracting surfaces. The distance between the apices is named the **thickness** of the lens. If the lens thickness may be ignored in comparison with the smaller of the radii of curvature of the surfaces bounding a lens, the latter is called **thin**.

Calculations which we do not give here show that for a thin lens the principal planes H and H′ may be considered to coincide and pass through the centre 0 of the lens (Fig. 16.19). The following expression is obtained for the focal lengths of a thin lens:

$$f' = -f = \left(\frac{n_0}{n - n_0}\right)\left(\frac{R_1 R_2}{R_2 - R_1}\right), \tag{16.66}$$

where n is the refractive index of the lens, n_0 is the refractive index of the medium surrounding the lens, R_1 and R_2 are the radii of curvature of the lens surfaces.

The radii of curvature must be treated as algebraic quantities: for a convex surface (*i.e.*, when the centre of curvature is to the right of the apex), the radius of curvature must be considered positive, and for a concave surface (*i.e.*, when the centre of curvature is to the left of the apex) the radius must be considered negative. The magnitude of the radius of curvature is shown in drawings, *i.e.*, $-R$ if $R < 0$.

If the refractive indices of the media at both sides of a thin lens are the same, then the nodal points N and N′ coincide with the principal points, *i.e.*, are at the centre 0 of the lens. Hence, in this case, any ray passing through the centre of the lens does not change its direction. If the refractive indices of the media before and after a lens are different, then the nodal points do not coincide with the principal points and a ray passing through the centre of the lens changes its direction.

[4]There are also lenses with surfaces having a more intricate shape.

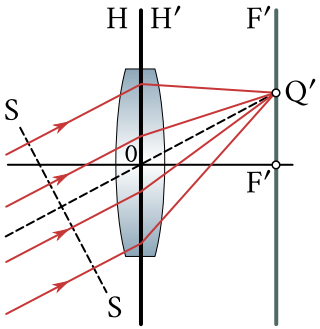


Fig. 16.20

A parallel beam of rays after passing through a lens converges at a point on the focal plane (see point Q′ in Fig. 16.20). To determine the position of this point, we must continue the ray passing through the centre of the lens up to its intersection with the focal plane (see ray 0Q′ shown by a dash line). The other rays will gather at the point of intersection too. Such a method is suitable when the optical properties of the medium at each side of a lens are identical ($n = n'$). Otherwise, a ray passing through the centre will change its direction. To find point Q′ in this case, we must know the position of the nodal points of the lens.

We must note that the optical paths laid off along the rays, beginning at wave surface SS (see Fig. 16.20) and terminating at point Q′ are identical and are tautochronous (see the end of Sec. 16.6).

In concluding, we must say that a lens is far from ideal optical system. The images of objects it produces have a number of errors. But a consideration of them is beyond the scope of the present book.

16.9. Huygens' Principle

In the following two chapters, we shall have to do with processes taking place behind an opaque barrier with apertures when a light wave impinges on the barrier. In the approximation of geometrical optics, no light ought to penetrate beyond the barrier into the region of the geometrical shadow. Actually, however, a light wave in principle propagates throughout the entire space behind the barrier and penetrates into the region of the geometrical shadow, this penetration being the more noticeable, the smaller are the dimensions of the apertures. With a diameter of the apertures or a width of slits comparable with the length of a light wave, the approximation of geometrical optics is absolutely illegitimate.

The behaviour of light behind a barrier with an aperture can be explained

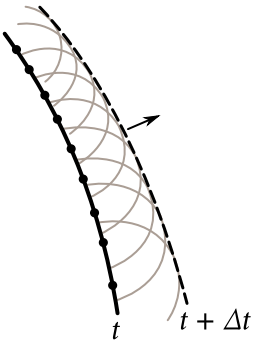


Fig. 16.21

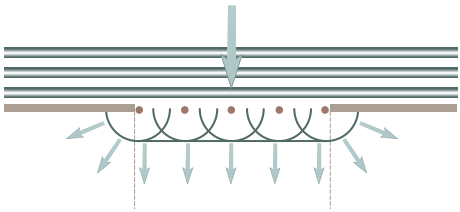


Fig. 16.22

qualitatively with the aid of **Huygens' principle**, named in honour of the Dutch physicist Christian Huygens (1629-1696) who discovered it. This principle establishes the way of constructing a wavefront at the moment of time $t + \Delta t$ according to the known position of the wavefront at the moment t. According to Huygens' principle, every point on an advancing wavefront can be considered as a source of secondary wavelets, and the envelope of these wavelets defines a new wavefront (Fig. 16.21; the medium is assumed to be non-homogeneous—the velocity of the wave in the lower part of the figure is greater than in the upper one).

Assume that a plane barrier with an aperture is struck by a wavefront parallel to it (Fig. 16.22). According to Huygens, every point on the portion of the wavefront bordering on the aperture is a centre of secondary wavelets which will be spherical in a homogeneous and isotropic medium. Constructing the envelope of these wavelets, we shall see that the wave penetrates beyond the aperture into the region of the geometrical shadow (these regions are shown by dash lines in the figure), bending around the edges of the barrier.

Huygens' principle gives no information on the intensity of waves propagating in various directions. This shortcoming was eliminated by the French physicist Augustin Fresnel (1788-1827). The improved Huygens-Fresnel principle is treated in Sec. 18.1, where a physical substantiation of the principle is also given.

Chapter 17

INTERFERENCE OF LIGHT

17.1. Interference of Light Waves

Let us assume that two waves of the same frequency, being superposed on each other, produce oscillations of the same direction, namely,

$$A_1 \cos(\omega t + \alpha_1), \quad A_2 \cos(\omega t + \alpha_2),$$

at a certain point in space. The amplitude of the resultant oscillation at the given point is determined by the expression

$$A^2 = A_1^2 + A_2^2 + 2A_1 A_2 \cos\delta,$$

where $\delta = \alpha_2 - \alpha_1$ [see Eq. (7.84) of Vol. I].

If the phase difference δ of the oscillations set up by the waves remains constant in time, then the waves are called **coherent**[1].

The phase difference δ for incoherent waves varies continuously and takes on any values with an equal probability. Hence, the time-averaged value of $\cos\delta$ equals zero. Therefore,

$$\langle A^2 \rangle = \langle A_1^2 \rangle + \langle A_2^2 \rangle .$$

Taking into account Eq. (16.10), we thus conclude that the intensity observed upon the superposition of incoherent waves equals the sum of the intensities produced by each of the waves individually:

$$I = I_1 + I_2. \tag{17.1}$$

For coherent waves, $\cos\delta$ has a time-constant value (but a different one for each point of space), so that,

$$I = I_1 + I_2 + 2\sqrt{I_1 I_2}\cos\delta. \tag{17.2}$$

At the points of space for which $\cos\delta > 0$, the intensity I will exceed $I_1 + I_2$; at the

[1]We shall discuss the concept of coherence in greater detail in the following.

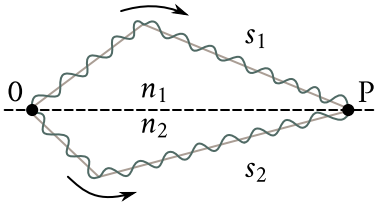


Fig. 17.1

points for which $\cos\delta < 0$, it will be smaller than $I_1 + I_2$. Thus, the superposition of coherent light waves is attended by redistribution of the light flux in space. As a result, maxima of the intensity will appear at some spots and minima at others. This phenomenon is called the interference of waves. Interference manifests itself especially clearly when the intensity of both interfering waves is the same: $I_1 = I_2$. Hence, according to Eq. (17.2), at the maxima $I = 4I_1$, while at the minima $I = 0$. For incoherent waves in the same condition, we get the same intensity $I = 2I_1$ everywhere [see Eq. (17.1)].

It follows from what has been said above that when a surface is illuminated by several sources of light (for example, by two lamps), an interference pattern ought to be observed with a characteristic alternation of maxima and minima of intensity. We know from our everyday experience, however, that in this case the illumination of the surface diminishes monotonously with an increasing distance from the light sources, and no interference pattern is observed. The explanation is that natural light sources are not coherent.

The incoherence of natural light sources is due to the fact that the radiation of a luminous body consists of the waves emitted by many atoms. The individual atoms emit wave trains with a duration of about 10^{-8} s and a length of about 3 m (see Sec. 16.1). The phase of a new train is not related in any way to that of the preceding one. In the light wave emitted by a body, the radiation of one group of atoms after about 10^{-8} s is replaced by the radiation of another group, and the phase of the resultant wave undergoes random changes.

Coherent light waves can be obtained by splitting (by means of reflections or refractions) the wave emitted by a single source into two parts. If these waves are made to cover different optical paths and are then superposed onto each other, interference is observed. The difference between the optical paths covered by the interfering waves must not be very great because the oscillations being added must belong to the same resultant wave train. If this difference will be of the order of one metre, oscillations corresponding to different trains will be superposed, and the phase difference between them will continuously change in a chaotic way.

Assume that the splitting into two coherent waves occurs at point 0 (Fig. 17.1).

Up to point P, the first wave travels the path s_1 in a medium of refractive index n_1, and the second wave travels the path s_2, in a medium of refractive index n_2. If the phase of oscillations at point 0 is ωt, then the first wave will produce the oscillation $A_1 \cos \omega(t - s_1/v_1)$ at point P, and the second wave, the oscillation $A_2 \cos \omega(t - s_2/v_2)$ at this point; $v_1 = c/n_1$ and $v_2 = c/n_2$ are the phase velocities of the waves. Hence, the difference between the phases of the oscillations produced by the waves at point P will be

$$\delta = \omega \left(\frac{s_2}{v_2} - \frac{s_1}{v_1} \right) = \frac{\omega}{c}(n_2 s_2 - n_1 s_1).$$

Replacing ω/c with $2\pi\nu/c = 2\pi/\lambda_0$ (where λ_0 is the wavelength in a vacuum), the expression for the phase difference can be written in the form

$$\delta = \frac{2\pi}{\lambda_0} \Delta, \tag{17.3}$$

where

$$\Delta = n_2 s_2 - n_1 s_1 = L_1 - L_2, \tag{17.4}$$

is a quantity equal to the difference between the optical paths travelled by the waves and is called the **difference in optical path** [compare with Eq. (16.55)].

A glance at Eq. (17.3) shows that if the difference in the optical path equals an integral number of wavelengths in a vacuum:

$$\Delta = \pm m\lambda_0 \quad (m = 0, 1, 2, \ldots), \tag{17.5}$$

then the phase difference δ is a multiple of 2π, and the oscillations produced at point P by both waves will occur with the same phase. Thus, Eq. (17.5) is the condition for an interference maximum, *i.e.*, for **constructive interference**.

If Δ equals a half-integral number of wavelengths in a vacuum:

$$\Delta = \pm \left(m + \frac{1}{2} \right) \lambda_0 \quad (m = 0, 1, 2, \ldots), \tag{17.6}$$

then, $\delta = \pm(2m + 1)\pi$, so that the oscillations at point P are in counterphase. Thus, Eq. (17.6) is the condition for an interference minimum, *i.e.*, for **destructive interference**.

Let us consider two cylindrical coherent light waves emerging from sources S_1 and S_2 having the form of parallel thin luminous filaments or narrow slits (Fig. 17.2). The region in which these waves overlap is called the **interference field**. Within this entire region, there are observed alternating places with maximum and minimum intensity of light. If we introduce a screen into the interference field, we shall see on it an interference pattern having the form of alternating light and dark fringes. Let us calculate the width of these fringes, assuming that the screen is parallel to a plane passing through sources S_1 and S_2. We shall characterize the

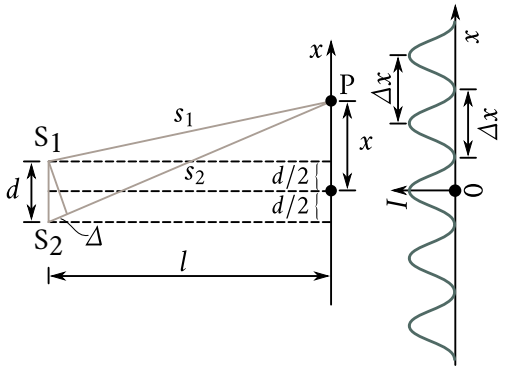


Fig. 17.2

position of a point on the screen by the coordinate x measured in a direction at right angles to lines S_1 and S_2. We shall choose the beginning of our readings at point 0 relative to which S_1 and S_2 are arranged symmetrically. We shall consider that the sources oscillate in the same phase. Examination of Fig. 17.2 shows that

$$s_1^2 = l^2 + \left(x - \frac{d}{2}\right)^2, \quad s_2^2 = l^2 + \left(x + \frac{d}{2}\right)^2.$$

Hence,

$$s_2^2 - s_1^2 = (s_2 + s_1)(s_2 - s_1) = 2xd.$$

It will be established somewhat later that to obtain a distinguishable interference pattern, the distance between the sources d must be considerably smaller than the distance to the screen l. The distance x within whose limits interference fringes are formed is also considerably smaller than l. In these conditions, we can assume that $s_2 + s_1 \approx 2l$. Thus, $s_2 - s_1 = xd/l$. Multiplying $s_2 - s_1$ by the refractive index of the medium n, we get the difference in the optical path

$$\Delta = n\frac{xd}{l}. \tag{17.7}$$

The introduction of this value of Δ into condition (17.5) shows that intensity maxima will be observed at values of x equal to

$$x_{\max} = \pm m\frac{l}{d}\lambda \quad (m = 0, 1, 2, \ldots). \tag{17.8}$$

Here $\lambda = \lambda_0/n$ is the wavelength in the medium filling the space between the sources and the screen.

Using the value of Δ given by Eq. (17.7) in condition (17.6), we get the coordinates of the intensity minima:

$$x_{\min} = \pm\left(m + \frac{1}{2}\right)\frac{l}{d}\lambda \quad (m = 0, 1, 2, \ldots). \tag{17.9}$$

Let us call the distance between two adjacent intensity maxima the **distance between interference fringes**, and the distance between adjacent intensity minima the **width of an interference fringe**. It can be seen from Eqs. (17.8) and (17.9) that the distance between fringes and the width of a fringe have the same value equal to

$$\Delta x = \frac{l}{d}\lambda. \tag{17.10}$$

According to Eq. (17.10), the distance between the fringes grows with a decreasing distance d between the sources. If d were comparable with l, the distance between the fringes would be of the same order as λ, *i.e.*, would be several scores of micrometres. In this case, the separate fringes would be absolutely indistinguishable. For an interference pattern to become distinct, the above-mentioned condition $d \ll l$ must be observed.

If the intensity of the interfering waves is the same ($I_1 = I_2 = I_0$), then according to Eq. (17.2) the resultant intensity at the points for which the phase difference is δ is determined by the expression

$$I = 2I_0(1 + \cos\delta) = 4I_0 \cos^2\left(\frac{\delta}{2}\right).$$

Since δ is proportional to Δ [see Eq. (17.3)], then, in accordance with Eq. (17.7), δ grows proportionally to x. Hence, the intensity varies along the screen in accordance with the law of cosine square. The right-hand part of Fig. 17.2 shows the dependence of I on x obtained in monochromatic light.

The width of the interference fringes and their spacing depend on the wavelength λ. The maxima of all wavelengths will coincide only at the centre of a pattern when $x = 0$. With an increasing distance from the centre of the pattern, the maxima of different colours become displaced from one another more and more. The result is blurring of the interference pattern when it is observed in white light. The number of distinguishable interference fringes appreciably grows in monochromatic light.

Having measured the distance between the fringes Δx and knowing l and d, we can use Eq. (17.10) to find λ. It is exactly from experiments involving the interference of light that the wavelengths for light rays of various colours were determined for the first time.

We have considered the interference of two cylindrical waves. Let us see what happens when two plane waves are superposed. Assume that the amplitudes of these waves are the same, and the directions of their propagation make the angle 2φ (Fig. 17.3). We shall consider that the directions of oscillations of the light vector are perpendicular to the plane of the drawing. The wave vectors $\boldsymbol{k}_1$ and $\boldsymbol{k}_2$ are in

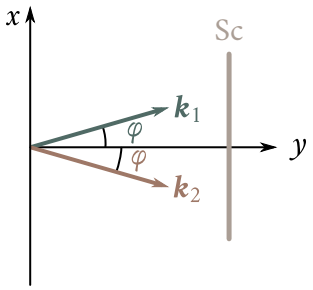


Fig. 17.3

the plane of the drawing and have the same magnitude equal to $k = 2\pi/\lambda$. Let us write the equations of these waves:

$$A\cos(\omega t - \boldsymbol{k}_1 \cdot \boldsymbol{r}) = A\cos(\omega t - k\sin\varphi \times x - k\cos\varphi \times y),$$

$$A\cos(\omega t - \boldsymbol{k}_2 \cdot \boldsymbol{r}) = A\cos(\omega t + k\sin\varphi \times x - k\cos\varphi \times y).$$

The resultant oscillation at points with the coordinates x and y has the form

$$A\cos(\omega t - k\sin\varphi \times x - k\cos\varphi \times y) + A\cos(\omega t + k\sin\varphi \times x - k\cos\varphi \times y)$$
$$= 2A\cos(k\sin\varphi \times x)\cos(\omega t - k\cos\varphi \times y). \qquad (17.11)$$

It follows from this equation that at points where $k\sin\varphi \times x = \pm m\pi (m = 0, 1, 2, \ldots)$, the amplitude of the oscillations is $2A$; where $k\sin\varphi \times x = \pm(m + 1/2)\pi$, the amplitude of the oscillations is zero. No matter where we place screen Sc, which is perpendicular to the y-axis, we shall observe on it a system of alternating light and dark fringes parallel to the z-axis (this axis is perpendicular to the plane of the drawing). The coordinates of the intensity maxima will be

$$x_{\max} = \pm\frac{m\pi}{k\sin\varphi} = \pm\frac{m\lambda}{2\sin\varphi}. \qquad (17.12)$$

Only the phase of the oscillations depends on the position of the screen (on the coordinate y) [see Eq. (17.11)].

We have assumed for simplicity that the initial phases of interfering waves are zero. If the difference between these phases is other than zero, a constant addend will appear in Eq. (17.12)—the fringe pattern will move along the screen.

17.2. Coherence

By **coherence** is meant the coordinated proceeding of several oscillatory or wave processes. The degree of coordination may vary. We can accordingly introduce the concept of the **degree of coherence** of two waves.

Temporal and **spatial coherence** are distinguished. We shall begin with a

discussion of temporal coherence.

Temporal Coherence. The process of interference described in the preceding section is idealized. This process is actually much more complicated. The reason is that a monochromatic wave described by the expression

$$A\cos(\omega t - kr + \alpha),$$

where A, ω, and α are constants, is an abstraction. A real light wave is formed by the superposition of oscillations of all possible frequencies (or wavelengths) confined within a more or less narrow but finite range of frequencies $\Delta\omega$ (or the corresponding range of wavelengths $\Delta\lambda$). Even for light considered to be monochromatic (single-coloured), the frequency interval $\Delta\omega$ is finite[2]. In addition, the amplitude of the wave A and the phase α undergo continuous random (chaotic) changes with time. Hence, the oscillations produced at a certain point of space by two superposed light waves have the form

$$A_1(t)\cos[\omega_1(t)t + \alpha_1(t)], \quad A_2(t)\cos[\omega_2(t)t + \alpha_2(t)], \tag{17.13}$$

the chaotic changes in the functions $A_1(t)$, $\omega_1(t)$, $\alpha_1(t)$, $A_2(t)$, $\omega_2(t)$, and $\alpha_2(t)$ being absolutely independent.

We shall assume for simplicity's sake that the amplitudes A_1 and A_2 are constant. Changes in the frequency and phase can be re duced either to a change only in the phase, or to a change only in the frequency. Let us write the function

$$f(t) = A\cos[\omega(t)t + \alpha(t)], \tag{17.14}$$

in the form

$$f(t) = A\cos\{\omega_0 t + [\omega(t) - \omega_0]t + \alpha(t)\},$$

where ω_0 is a certain average value of the frequency, and introduce the notation $[\omega(t) - \omega_0]t + \alpha(t) = \alpha'(t)$. Equation (17.14) will, thus, become

$$f(t) = A\cos[\omega_0 t + \alpha'(t)]. \tag{17.15}$$

We have obtained a function in which only the phase of the oscillation changes chaotically.

On the other hand, it is proved in mathematics that an inharmonic function, for example, function (17.14), can be represented in the form of the sum of harmonic functions with frequencies confined within a certain interval $\Delta(\omega)$ [see Eq. (17.16)].

Thus, when considering the matter of coherence, two approaches are possible: a "phase" one and a "frequency" one. Let us begin with the phase approach. Assume that the frequencies ω_1 and ω_2 in Eqs. (17.13) satisfy the condition $\omega_1 = \omega_2 = \text{constant}$. Now let us find the influence of a change in the phases α_1 and α_2. According to

[2]The spectral lines emitted by atoms have a "natural" width of the order of 10^{-8} rad s^{-1} ($\Delta\lambda \sim 10^{-4}$ Å).

Eq. (17.12), with our assumptions, the intensity of light at a given point is determined by the expression

$$I = I_1 + I_2 + 2\sqrt{I_1 I_2}\cos[\delta(t)],$$

where $\delta(t) = \alpha_2(t) - \alpha_1(t)$. The last addend in this equation is called the **interference term**.

An instrument that can be used to observe an interference pattern (the eye[3], a photographic plate, etc.) has a certain inertia. In this connection, it registers a pattern averaged over the time interval t_{instr} needed for "operation" of the instrument. If during the time t_{instr} the factor $\cos[\delta(t)]$ takes on all the values from -1 to $+1$, the average value of the interference term will be zero. Therefore, the intensity registered by the instrument will equal the sum of the intensities produced at a given point by each of the waves separately—interference is absent, and we are forced to acknowledge that the waves are incoherent.

If during the time t_{instr}, however, the value of $\cos[\delta(t)]$ remains virtually constant[4], the instrument will detect interference, and the waves must be acknowledged as coherent.

It follows from the above that the concept of coherence is relative: two waves can behave like coherent ones when observed using one instrument (having a low inertia), and like incoherent ones when observed using another instrument (having a high inertia). The coherent properties of waves are characterized by introducing the **coherence time** t_{coh}. It is defined as the time during which a chance change in the wave phase $\alpha(t)$ reaches a value of the order of π. During the time t_{coh}, an oscillation, as it were, forgets its initial phase and becomes incoherent with respect to itself.

Using the concept of the coherence time, we can say that when the instrument time is much greater than the coherence time of the superposed waves ($t_{\text{instr}} \gg t_{\text{coh}}$), the instrument does not register interference. When $t_{\text{instr}} \ll t_{\text{coh}}$, the instrument will detect a sharp interference pattern. At intermediate values of t_{instr}, the sharpness of the pattern will diminish as t_{instr} grows from values smaller than t_{coh} to values greater than it.

The distance $l_{\text{coh}} = ct_{\text{coh}}$ over which a wave travels during the time l_{coh} is called the **coherence length** (or the **train length**). The coherence length is the distance over which a chance change in the phase reaches a value of about π. To obtain an interference pattern by splitting a natural wave into two parts, it is essential that the

[3]We remind our reader that the showing of motion picture films is based on the inertia of visual perception, which is about 0.1 s.

[4]The phase difference $\delta(t)$ varies for different points of space. The influence of the interference term manifests itself at the points where it differs from zero.

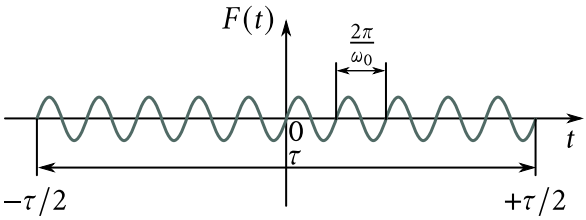


Fig. 17.4

optical path difference Δ be smaller than the coherence length. This requirement limits the number of visible interference fringes observed when using the layout shown in Fig. 17.2. An increase in the fringe number m is attended by a growth in the path difference. As a result, the sharpness of the fringes becomes poorer and poorer.

Let us pass over to a consideration of the part of the non-monochromatic nature of light waves. Assume that light consists of a sequence of identical trains of frequency ω_0 and duration T. When one train is replaced with another one, the phase experiences disordered changes. As a result, the trains are mutually incoherent. With these assumptions, the duration of a train τ virtually coincides with the coherence time t_{coh}.

In mathematics, the Fourier theorem is proved, according to which any finite and integrable function $F(t)$ can be represented in the form of the sum of an infinite number of harmonic components with a continuously changing frequency:

$$F(t) = \int_{-\infty}^{+\infty} A(\omega)\, e^{i\omega t}\, \mathrm{d}\omega. \tag{17.16}$$

Expression (17.16) is known as the **Fourier integral**. The function $A(\omega)$ inside the integral is the amplitude of the relevant monochromatic component. According to the theory of Fourier integrals, the analytical form of the function $A(\omega)$ is determined by the expression

$$A(\omega) = 2\pi \int_{-\infty}^{+\infty} F(\xi)\, e^{-i\omega\xi}\, \mathrm{d}\xi, \tag{17.17}$$

where ξ is an auxiliary integration variable.

Assume that the function $F(t)$ describes a light disturbance at a certain point at the moment of time t due to a single wave train. Hence, it is determined by the conditions

$$F(t) = A_0 \exp(i\omega_0 t) \quad \text{at } |t| \ll \frac{\tau}{2}$$

$$F(t) = 0 \quad \text{at } |t| > \frac{\tau}{2}.$$

A graph of the real part of this function is given in Fig. 17.4.

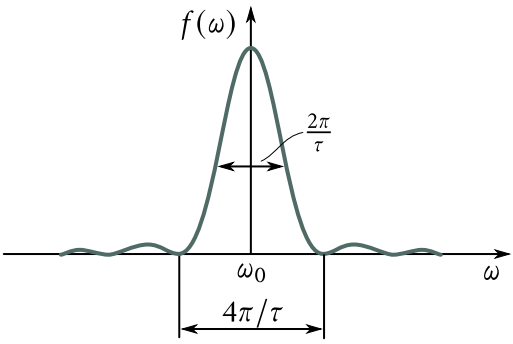


Fig. 17.5

Outside the interval from $-\tau/2$ to $+\tau/2$, the function $F(t)$ is zero. Therefore, expression (17.17) determining the amplitude of the harmonic components has the form

$$A(\omega) = 2\pi \int_{-\tau/2}^{+\tau/2} [A_0 \exp(i\omega_0\xi)] \exp(-i\omega\xi)\, \mathrm{d}\xi$$
$$= 2\pi A_0 \int_{-\tau/2}^{+\tau/2} \exp[i(\omega_0 - \omega)\xi]\, \mathrm{d}\xi = 2\pi A_0 \left.\frac{\exp[i(\omega_0 - \omega)\xi]}{i(\omega_0 - \omega)}\right|_{-\tau/2}^{+\tau/2}.$$

After introducing the integration limits and simple transformations, we arrive at the equation

$$A(\omega) = \pi A_0 \tau \frac{\sin[(\omega - \omega_0)\tau/2]}{(\omega - \omega_0)\tau/2}.$$

The intensity $I(\omega)$ of a harmonic wave component is proportional to the square of the amplitude, *i.e.*, to the expression

$$f(\omega) = \frac{\sin^2[(\omega - \omega_0)\tau/2]}{[(\omega - \omega_0)\tau/2]^2}. \tag{17.18}$$

A graph of function (17.18) is shown in Fig. 17.5. A glance at the figure shows that the intensity of the components whose frequencies are within the interval of width $\Delta\omega = 2\pi/\tau$ considerably exceeds the intensity of the remaining components. This circumstance allows us to relate the duration of a train τ to the effective frequency range $\Delta\omega$ of a Fourier spectrum:

$$\tau = \frac{2\pi}{\Delta\omega} = \frac{1}{\Delta\nu}.$$

Identifying τ with the coherence time, we arrive at the relation

$$t_{\mathrm{coh}} \sim \frac{1}{\Delta\nu} \tag{17.19}$$

(The sign $\sim$ stands for "equal to in the order of magnitude").

It can be seen from expression (17.19) that the broader the interval of frequencies present in a given light wave, the smaller is the coherence time of this wave.

The frequency is related to the wavelength in a vacuum by the expression $\nu = c/\lambda_0$. Differentiation of this expression yields $\Delta\nu = c\Delta\lambda_0/\lambda_0^2 \approx c\Delta\lambda/\lambda^2$ (we have omitted the minus sign obtained in differentiation and also assumed that $\lambda_0 \approx \lambda$). Substituting for $\Delta\nu$ in Eq. (17.19) its expression through λ and $\Delta\lambda$, we obtain the following expression for the coherence time:

$$t_{\text{coh}} \sim \frac{\lambda^2}{c\Delta\lambda}. \tag{17.20}$$

Hence, we get the following value for the coherence length:

$$l_{\text{coh}} = ct_{\text{coh}} \sim \frac{\lambda^2}{\Delta\lambda}. \tag{17.21}$$

Examination of Eq. (17.5) shows that the path difference at which a maximum of the m-th order is obtained is determined by the relation

$$\Delta_m = \pm m\lambda_0 \approx \pm m\lambda.$$

When this path difference reaches values of the order of the coherence length, the fringes become indistinguishable. Consequently, the extreme interference order observed is determined by the condition

$$m_{\text{extr}}\lambda \sim l_{\text{coh}} \sim \frac{\lambda^2}{\Delta\lambda},$$

whence

$$m_{\text{extr}} \sim \frac{\lambda}{\Delta\lambda}. \tag{17.22}$$

It follows from Eq. (17.22) that the number of interference fringes observed according to the layout shown in Fig. 17.2 grows when the wavelength interval in the light used diminishes.

Spatial Coherence. According to the equation $k = \omega/v = n\omega/c$, scattering of the frequencies $\Delta\omega$ results in scattering of the values of k. We have established that the temporal coherence is determined by the value of $\Delta\omega$. Consequently, the temporal coherence is associated with scattering of the values of the magnitude of the wave vector $\boldsymbol{k}$. Spatial coherence is associated with scattering of the directions of the vector $\boldsymbol{k}$ that is characterized by the quantity $\Delta\hat{\boldsymbol{e}}_k$.

The setting up at a certain point of space of oscillations produced by waves with different values of $\hat{\boldsymbol{e}}_k$ is possible if these waves are emitted by different sections of an extended (not a point) light source. Let us assume for simplicity's sake that the source has the form of a disk visible from a given point at the angle φ. It can be seen from Fig. 17.6 that the angle φ characterizes the interval confining the unit vectors $\hat{\boldsymbol{e}}_k$. We shall consider that this angle is small.

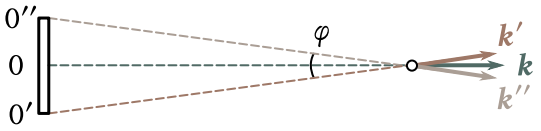


Fig. 17.6

Assume that the light from the source falls on two narrow slits behind which there is a screen (Fig. 17.7). We shall consider that the interval of frequencies emitted by the source is very small. This is needed for the degree of temporal coherence to be sufficient for obtaining a sharp interference pattern. The wave arriving from the section of the surface designated in Fig. 17.7 by 0 produces a zero-order maximum M at the middle of the screen. The zero-order maximum M′ produced by the wave arriving from section 0′ will be displaced from the middle of the screen by the distance x'. Owing to the smallness of the angle φ and of the ratio d/l, we can consider that $x' = l\varphi/2$. The zero-order maximum M″ produced by the wave arriving from section 0″ is displaced in the opposite direction from the middle of the screen over the distance x'' equal to x'. The zero-order maxima from the other sections of the source will be between the maxima M′ and M″.

The separate sections of the light source produce waves whose phases are in no way related to one another. For this reason, the interference pattern appearing on the screen will be a superposition of the patterns produced by each section separately. If the displacement x' is much smaller than the width of an interference fringe $\Delta x = l\lambda/d$ [see Eq. (17.10)], then, the maxima from different sections of the source will practically be superposed on one another, and the pattern will be like the one produced by a point source. When $x' \approx \Delta x$, the maxima from some sections will coincide with the minima from others, and no interference pattern will be observed. Thus, an interference pattern will be distinguishable provided that $x' < \Delta x$, *i.e.*,

$$\frac{l\varphi}{2} < \frac{l\lambda}{d}, \tag{17.23}$$

or

$$\varphi < \frac{\lambda}{d}. \tag{17.24}$$

We have omitted the factor 2 when passing over from expression (17.23) to (17.24).

Formula (17.24) determines the angular dimensions of a source at which interference is observed. We can also use this formula to find the greatest distance between the slits at which interference from a source with the angular dimension φ can still be observed. Multiplying inequality (17.24) by d/φ, we arrive at the condition

$$d < \frac{\lambda}{\varphi}. \tag{17.25}$$

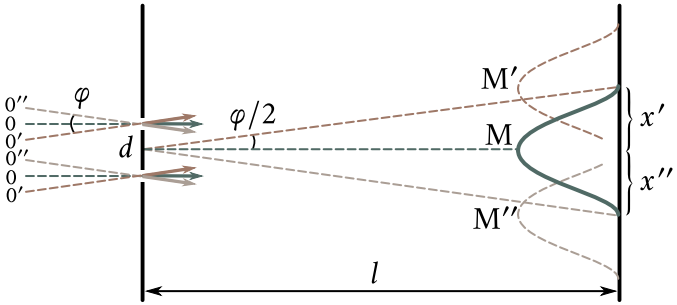


Fig. 17.7

A collection of waves with different values of $\hat{\boldsymbol{e}}_k$ can be replaced with the resultant wave falling on a screen with slits. The absence of an interference pattern signifies that the oscillations produced by this wave at the places where the first and second slits are situated are incoherent. Consequently, the oscillations in the wave itself at points at a distance d apart are incoherent too. If the source were ideally monochromatic (this means that $\Delta\nu = 0$ and $t_{\text{coh}} = \infty$), the surface passing through the slits would be a wave one, and the oscillations at all the points of this surface would occur in the same phase. We have established that when $\Delta\nu \neq 0$ and the source has finite dimensions ($\varphi \neq 0$), the oscillations at points of a surface at a distance of $d > \lambda/\varphi$ are incoherent.

We shall call a surface, which would be a wave one if the source were monochromatic, a pseudowave surface[5] for brevity. We could satisfy condition (17.24) by reducing the distance d between the slits, *i.e.*, by taking closer points of the pseudowave surface. Consequently, oscillations produced by a wave at adequately close points of a pseudowave surface are coherent. Such coherence is called **spatial**. Thus, the phase of an oscillation changes chaotically when passing from one point of a pseudowave surface to another. Let us introduce the distance ρ_{coh}, upon displacement by which along a pseudowave surface a random change in the phase reaches a value of about π. Oscillations at two points of a pseudowave surface spaced apart at a distance less than ρ_{coh} will be approximately coherent. The distance ρ_{coh} is called the **spatial coherence length** or the **coherence radius**. It can be seen from expression (17.25) that

$$\rho_{\text{coh}} \sim \frac{\lambda}{\varphi}. \tag{17.26}$$

The angular dimension of the Sun is about 0.01 radians, and the length of its light waves is about 0.5 μm. Hence, the coherence radius of the light waves arriving

[5]It must be borne in mind that this term is not used in scientific publications. The author has coined it for conditional use only to make the treatment more illustrative.

from the Sun has a value of the order of

$$\rho_{\text{coh}} = \frac{0.5}{0.01} = 50\ \mu\text{m} = 0.05\ \text{mm}. \tag{17.27}$$

The entire space occupied by a wave can be divided into parts in each of which the wave approximately retains coherence. The volume of such a part of space, called the **coherence volume**, in its order of magnitude equals the product of the temporal coherence length and the area of a circle of radius ρ_{coh}.

The spatial coherence of a light wave near the surface of the heated body emitting it is restricted by a value of ρ_{coh} of only a few wavelengths. With an increasing distance from the source, the degree of spatial coherence grows. The radiation of a laser[6] has an enormous temporal and spatial coherence. At the outlet opening of a laser, spatial coherence is observed throughout the entire cross section of the light beam.

It would seem possible to observe interference by passing light propagating from an arbitrary source through two slits in an opaque screen. With a small spatial coherence of the wave falling on the slits, however, the beams of light passing through them will be incoherent, and an interference pattern will be absent. The English scientist Thomas Young (1773-1829) in 1802 obtained interference from two slits by increasing the spatial coherence of the light falling on the slits. Young achieved such an increase by first passing the light through a small aperture in an opaque screen. This light was used to illuminate the slits in a second opaque screen. Thus, for the first time in history, Young observed the interference of light waves and determined the lengths of these waves.

17.3. Ways of Observing the Interference of Light

Let us consider two concrete interference layouts of which one uses reflection for splitting a light wave into two parts, and the other refraction of light.

Fresnel's Double Mirror. Two plane contacting mirrors 0M and 0N are arranged so that their reflecting surfaces form an obtuse angle close to π (Fig. 17.8). Hence, the angle φ in the figure is very small. A straight light source S (for example, a narrow luminous slit) is placed parallel to the line of intersection of the mirrors 0 (perpendicular to the plane of the drawing) at a distance r from it. The mirrors reflect two cylindrical coherent waves onto screen Sc. They propagate as if they were emitted by virtual sources S_1 and S_2. Opaque screen Sc_1 prevents the direct propagation of the light from source S to screen Sc.

Ray 0Q is the reflection of ray S0 from mirror 0M, and ray 0P is the reflection

[6] Lasers will be treated in Vol. III of our course.

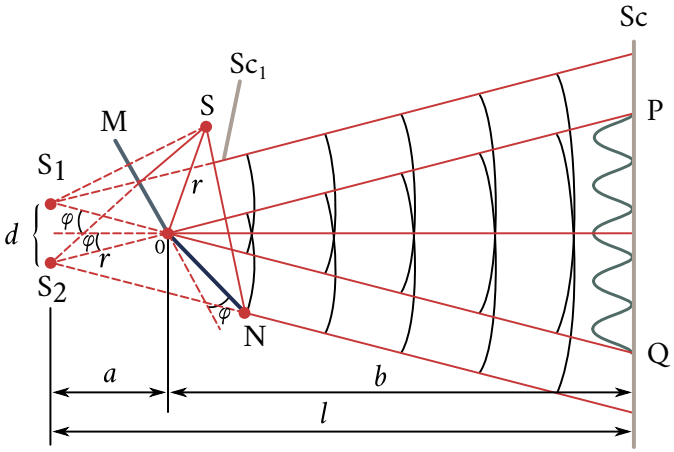


Fig. 17.8

of ray S0 from mirror 0N. It is easy to see that the angle between rays 0P and 0Q is 2φ. Since S and S_1 are symmetrical relative to 0M, the length of segment $0S_1$ equals 0S, *i.e.*, r. Similar reasoning leads to the same result for segment $0S_2$. Thus, the distance between sources S_1 and S_2 is

$$d = 2r \sin \varphi \approx 2r\varphi.$$

Inspection of Fig. 17.8 shows that $a = r \cos \varphi \approx r$. Hence,

$$l = r + b,$$

where b is the distance from the line of intersection of the mirrors 0 to screen Sc.

Using the values of d and l we have found in Eq. (17.10), we obtain the width of an interference fringe:

$$\Delta x = \frac{r + b}{2r\varphi}\lambda. \tag{17.28}$$

The region of wave overlapping PQ has a length of $2b \tan \varphi \approx 2b\varphi$. Dividing this length by the width of a fringe Δx, we find the maximum number of interference fringes that can be observed with the aid of Fresnel's double mirror at the given parameters of a layout:

$$N = \frac{4br\varphi^2}{\lambda(r + b)}. \tag{17.29}$$

For all these fringes to be visible indeed, it is essential that $N/2$ be not greater than m_{extr} determined by expression (17.22).

Fresnel's Biprism. Two prisms with a small refracting angle θ made from a single piece of glass have one common face (Fig. 17.9). A straight light source S is arranged parallel to this face at a distance a from it.

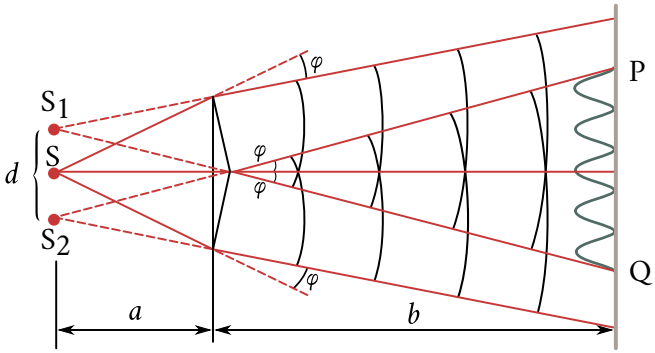


Fig. 17.9

It can be shown that when the refracting angle a of the prism is very small and the angles of incidence of the rays on the face of the prism are not very great, all the rays are deflected by the prism through a practically identical angle equal to

$$\varphi = (n-1)\theta$$

(n is the refractive index of the prism). The angle of incidence of the rays on the biprism is not great. Therefore, all the rays are deflected by each half of the biprism through the same angle. As a result, two coherent cylindrical waves are formed emerging from virtual sources S_1 and S_2 in the same plane as S. The distance between the sources is

$$d = 2a \sin\varphi \approx 2a\varphi = 2a(n-1)\theta.$$

The distance from the sources to the screen is

$$l = a + b.$$

We find the width of an interference fringe by Eq. (17.10):

$$\Delta x = \frac{(a+b)}{2a(n-1)\theta}\lambda. \tag{17.30}$$

The region of overlapping of the waves PQ has the length

$$2b \tan\varphi \approx 2b\varphi = 2b(n-1)\theta.$$

The maximum number of fringes observed is

$$N = \frac{4ab(n-1)^2\theta^2}{\lambda(a+b)}. \tag{17.31}$$

17.4. Interference of Light Reflected from Thin Plates

When a light wave falls on a thin transparent plate (or film), reflection occurs from both surfaces of the plate. The result is the production of two light waves that in

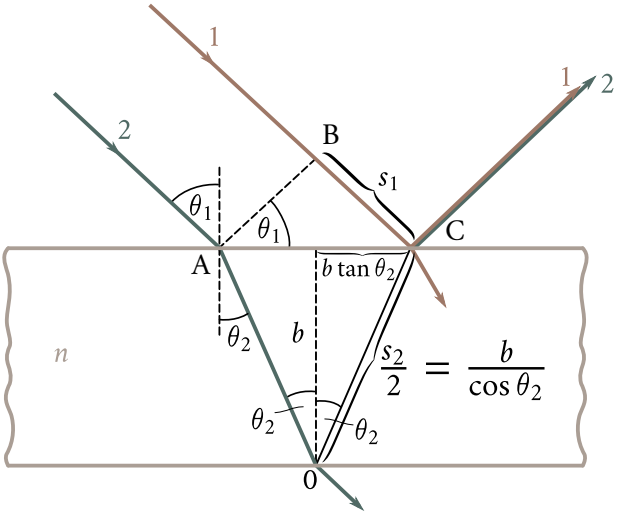


Fig. 17.10

known conditions can interfere.

Assume that a plane light wave which can be considered as a parallel beam of rays falls on a transparent plane-parallel plate (Fig. 17.10). The plate reflects upward two parallel beams of light. One of them was formed as a result of reflection from the top surface of the plate and the second as a result of reflection from its bottom surface (in Fig. 17.10 each of these beams is represented by only one ray). The second beam is refracted when it enters the plate and leaves it. In addition to these two beams, the plate throws upward beams produced as a result of three-, five-fold, etc. reflection from the plate surfaces. Owing to their small intensity, however, we shall take no account of these beams[7]. We shall also display no interest in the beams passing through the plate.

The path difference acquired by rays 1 and 2 before they meet at point C is

$$\Delta = ns_2 - s_1, \tag{17.32}$$

where s_1 is the length of segment BC, s_2 is the total length of segments AO and OC and n the refractive index of the plate.

We assume that the refractive index of the medium surrounding the plate is unity. A glance at Fig. 17.10 shows that $s_1 = 2b \tan \theta_2 \times \sin \theta_1$, and $s_2 = 2b/\cos \theta_2$ (here, b is the thickness of the plate).

[7] At $n = 1.5$, about 5% of the incident luminous flux is reflected from the surface of the plate (see the last paragraph of Sec. 16.3). After two reflections, the intensity will be 0.05×0.05 or 0.25% of the intensity of the initial beam. After three reflections, the relevant figure is $0.05 \times 0.05 \times 0.05$, or 0.0125%, which is 1/400 of the intensity of the singly reflected beam.

Using these values in Eq. (17.32), we get

$$\Delta = \frac{2bn}{\cos\theta_2} - 2b\tan\theta_2 \sin\theta_1 = 2b\frac{n^2 - n\sin\theta_2 \sin\theta_1}{n\cos\theta_2}.$$

Substituting $\sin\theta_1$ for $n\sin\theta_1$ and taking into account that

$$n\cos\theta_2 = \sqrt{n^2 - n^2\sin^2\theta_2} = \sqrt{n^2 - \sin^2\theta_1},$$

it is easy to give the equation for Δ the form

$$\Delta = 2b\sqrt{n^2 - \sin^2\theta_1}. \tag{17.33}$$

When calculating the phase difference δ between the oscillations in rays 1 and 2, it is necessary, in addition to the optical path difference Δ, to take into account the possibility of a change in the phase of the wave upon reflection (see Sec. 16.3). At point A (see Fig. 17.10), reflection occurs from the interface between the optically less dense medium and the optically denser one. Consequently, the wave phase experiences a change by π. At point 0, reflection occurs from the interface between the optically denser medium and the optically less dense one, so that there is no jump in the phase. Hence, an additional phase difference equal to π is produced between rays 1 and 2. It can be taken into account by adding to Δ (or subtracting from it) half a wavelength in a vacuum. The result is

$$\Delta = 2b\sqrt{n^2 - \sin^2\theta_1} - \frac{\lambda_0}{2}. \tag{17.34}$$

Thus, when a plane wave falls on the plate, two reflected waves are formed, and their path difference is determined by Eq. (17.34). Let us determine the conditions in which these waves will be coherent and can interfere. We shall consider two cases.

1. A Plane-Parallel Plate. Both plane reflected waves propagate in one direction making an angle equal to the angle of incidence θ_1 with a normal to the plate. These waves can interfere if conditions of both temporal and spatial coherence are observed.

For temporal coherence to take place, the path difference given by Eq. (17.34) must not exceed the coherence length equal to $\lambda^2/\Delta\lambda \approx \lambda_0^2/\Delta\lambda_0$ [see expression (17.21)]. Consequently, the condition

$$2b\sqrt{n^2 - \sin^2\theta_1} - \frac{\lambda_0}{2} < \frac{\lambda_0^2}{\Delta\lambda_0},$$

or

$$b < \frac{\lambda_0(\lambda_0/\Delta\lambda_0 + 1/2)}{2\sqrt{n^2 - \sin^2\theta_1}},$$

must be observed. In the obtained relation, we may disregard $1/2$ in comparison

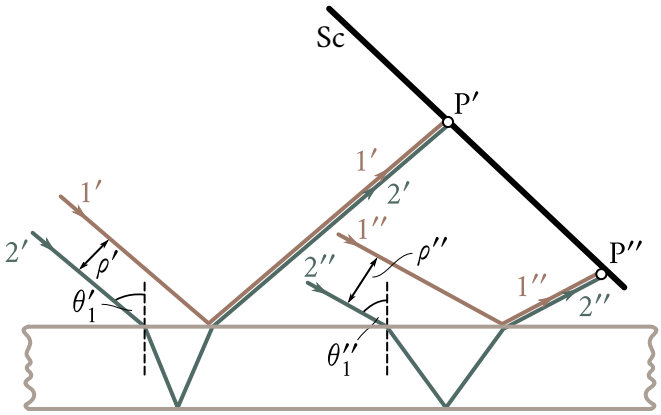


Fig. 17.11

with $\lambda_0/\Delta\lambda_0$. The expression $\sqrt{n^2 - \sin^2\theta_1}$ has a magnitude of the order of unity[8]. We can therefore write

$$b < \frac{\lambda_0^2}{2\Delta\lambda_0} \tag{17.35}$$

(the double plate thickness must be less than the coherence length).

Thus, the reflected waves will be coherent only if the plate thickness b does not exceed the value determined by expression (17.35). Assuming that $\lambda_0 = 5000$ Å and $\Delta\lambda_0 = 20$ Å, we get the extreme value of the thickness equal to

$$\frac{5000^2}{2 \times 20} \approx 6 \times 10^5 \text{ Å} = 0.06 \text{ mm}. \tag{17.36}$$

Now, let us consider the conditions for observance of spatial coherence. Let us place screen Sc in the path of the reflected beams (Fig. 17.11). Rays 1′ and 2′ arriving at point P′ will be at a distance ρ' apart in the incident beam. If this distance does not exceed the coherence radius ρ_{coh} of the incident wave, rays 1′ and 2′ will be coherent and will produce at point P′ an illumination determined by the value of the path difference Δ corresponding to the angle of incidence θ_1. The other pairs of rays travelling at the same angle θ_1' will produce the same illumination at the other points of the screen. The screen will thus be uniformly illuminated (in the particular case when $\Delta = (n + 1/2)\lambda_0$, the screen will be dark). When the inclination of the beam is changed (*i.e.*, when the angle θ_1 is changed), the illumination of the screen will change too.

A glance at Fig. 17.10 shows that the distance between the incident rays 1 and 2

[8]For $n = 1.5$, the magnitude of this expression varies within the limits from 1.12 (at $\theta_1 = \pi/2$) to 1.5 (at $\theta_1 = 0$).

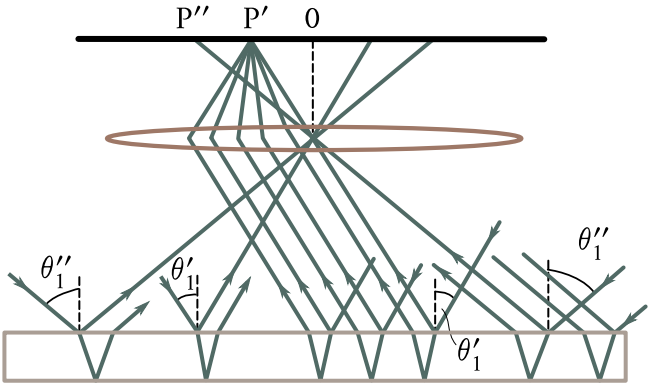


Fig. 17.12

is

$$\rho = 2b \tan\theta_2 \sin\theta_1 = \frac{b \sin(2\theta_1)}{\sqrt{n^2 - \sin^2\theta_1}}. \tag{17.37}$$

If we assume that $n = 1.5$, then, for $\theta_1 = 45°$ we get $\rho = 0.8b$, and for $\theta_1 = 10°$ we get $\rho = 0.1b$. For normal incidence ($\theta_1 = 0$), we have $\rho = 0$ at any n.

The coherence radius of sunlight has a value of the order of 0.05 mm [see Eq. (17.27)]. At an angle of incidence of 45°, we may assume that $\rho \approx b$. Hence, for interference to occur in these conditions, the relation

$$b < 0.05 \text{ mm} \tag{17.38}$$

must be observed [compare with Eq. (17.36)]. For an angle of incidence of about 10°, spatial coherence will be retained at a plate thickness not exceeding 0.5 mm. We thus arrive at the conclusion that owing to the restrictions imposed by temporal and spatial coherence, interference is observed when a plate is illuminated by sunlight only if the thickness of the plate does not exceed a few hundredths of a millimetre. Upon illumination with light having a greater degree of coherence, interference is also observed in reflection from thicker plates or films.

Interference from a plane-parallel plate is observed in practice by placing in the path of the reflected beams a lens that gathers the rays at one of the points of the screen in the focal plane of the lens (Fig. 17.12). The illumination at this point depends on the value of quantity (17.34). When $\Delta = m\lambda_0$, we get maxima, and when $\Delta = (m + 1/2)\lambda_0$—minima of the intensity (m is an integer). The condition for the maximum intensity has the form

$$2b\sqrt{n^2 - \sin^2\theta_1} = \left(m + \frac{1}{2}\right)\lambda_0. \tag{17.39}$$

Assume that a thin plane-parallel plate is illuminated by diffuse monochromatic

light (see Fig. 17.12). Let us arrange a lens parallel to the plate and put a screen in the focal plane of the lens. Diffuse light contains rays of the most diverse directions. The rays parallel to the plane of the drawing and falling on the plate at the angle θ_1' after reflection from both surfaces of the plate will be gathered by the lens at point P′ and will set up at this point an illumination determined by the value of the optical path difference. Rays propagating in other planes but falling on the plate at the same angle θ_1' will be gathered by the lens at other points at the same distance as point P′ from centre 0 of the screen. The illumination at all these points will be the same. Thus, the rays falling on the plate at the same angle θ_1' will produce on the screen a collection of identically illuminated points arranged along a circle with its centre at 0. Similarly, the rays falling at a different angle θ_1'' will produce on the screen a collection of identically (but different in value because Δ is different) illuminated points arranged along a circle of another radius. The result will be the appearance on the screen of a system of alternating bright and dark circular fringes with a common centre at point 0. Each fringe is formed by the rays falling on the plate at the same angle θ_1. This is why interference fringes produced in such conditions are known as fringes of equal inclination. When the lens is arranged differently relative to the plate (the screen must coincide with the focal plane of the lens in all cases), the fringes of equal inclination will have another shape.

Every point of an interference pattern is due to rays which formed a parallel beam before passing through the lens. Hence, in observing fringes of equal inclination, the screen must be placed in the focal plane of the lens, *i.e.*, in the same way in which it is arranged to produce an image of infinitely remote objects on it. Accordingly, fringes of equal inclination are said to be localized at infinity. The part of the lens can be played by the crystalline lens, and that of the screen by the retina of the eye. In this case for observing fringes of equal inclination, the eye must be accommodated as when looking at very remote objects.

According to Eq. (17.39), the position of the maxima depends on the wavelength λ_0. Therefore, in white light, we get a collection of fringes displaced relative to one another and formed by rays of different colours; the interference pattern acquires the colouring of a rainbow. The possibility of observing an interference pattern in white light is determined by the ability of the eye to distinguish light tints of close wavelengths. The average human eye perceives rays differing in wavelength by less than 20 Å as having the same colour. Therefore, to assess the conditions in which interference from plates can be observed in white light, we must assume that $\Delta\lambda_0$ equals 20 Å. We took exactly this value in assessing the thickness of a plate [see Eq. (17.36)].

2. Plate of Varying Thickness. Let us take a plate in the form of a wedge with an apex angle of φ (Fig. 17.13). Assume that a parallel beam of rays falls on it. Now

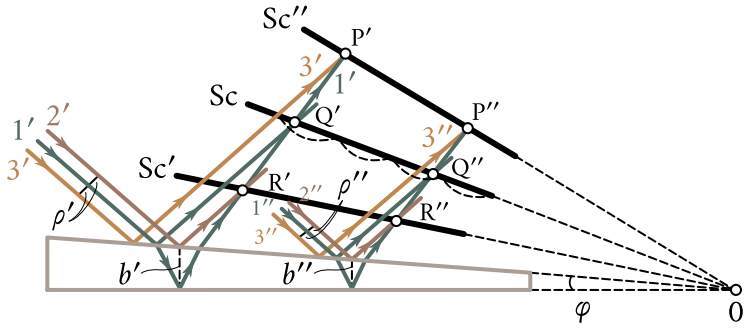


Fig. 17.13

the rays reflected from different surfaces of the plate will not be parallel. Two rays that practically merge before falling on the plate (in Fig. 17.13 they are depicted in the form of a single straight line designated by the figure 1′) intersect after reflection at point Q'. The two rays 1″ practically merging intersect at point Q″ after reflection. It can be shown that points Q′, Q″ and other points similar to them lie in one plane passing through apex 0 of the wedge. Ray 1′ reflected from the bottom surface of the wedge and ray 2′ reflected from its top surface will intersect at point R′ that is closer to the wedge than Q′. Similar rays 1′ and 3′ will intersect at point P′ that is farther from the wedge surface than Q′.

The directions of propagation of the waves reflected from the top and bottom surfaces of the wedge do not coincide. Temporal coherence will be observed only for the parts of the waves reflected from places of the wedge for which the thickness satisfies condition (17.35). Assume that this condition is observed for the entire wedge. In addition, assume that the coherence radius is much greater than the wedge length. Hence, the reflected waves will be coherent in the entire space over the wedge, and no matter at what distance from the wedge the screen is, an interference pattern will be observed on it in the form of fringes parallel to the wedge apex 0 (see the last three paragraphs of Sec. 17.1). This, particularly, is how matters are when a wedge is illuminated by light emitted by a laser.

With restricted spatial coherence, the region of localization of the interference pattern (*i.e.*, the region of space in which an interference pattern can be seen on a screen placed in it) will be restricted too. If we arrange a screen so that it pass.3s through points Q′, Q″, . . . (see screen Sc in Fig. 17.13), an interference pattern will appear on it even if the spatial coherence of the falling wave is extremely small (rays that coincided before falling on the wedge will intersect at points on the screen). At a small wedge angle φ, the path difference of the rays can be calculated with sufficient accuracy by Eq. (17.34) taking as b the thickness of the plate at the

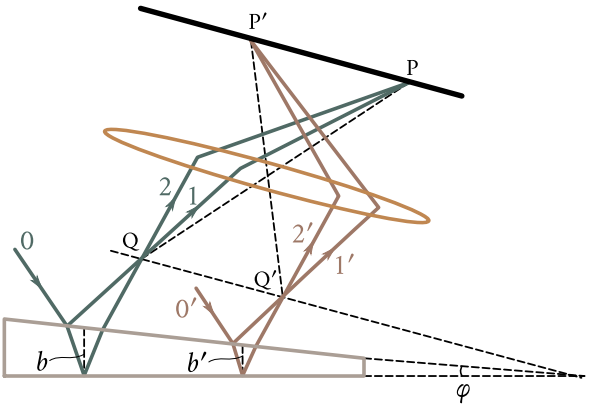


Fig. 17.14

place where the rays fall on it. Since the path difference for the rays reflected from different sections of the wedge is now different, the illumination of the screen will be non-uniform—bright and dark fringes will appear on it (see the dash curve showing the illumination of screen Sc in Fig. 17.13). Each of these fringes is produced as a result of reflection from sections of the wedge having the same thickness. This is why they are known as **fringes of equal thickness**.

Upon displacement of the screen from position Sc in a direction away from the wedge or toward it, the degree of spatial coherence of the incident wave begins to tell. If in the position of the screen denoted in Fig. 17.13 by Sc′, the distance ρ' between the incident rays 1′ and 2′ becomes of the order of the coherence radius, no interference pattern will be observed on screen Sc′. Similarly, the pattern vanishes when the screen is at position Sc″.

Thus, the interference pattern produced when a plane wave is reflected from a wedge is localized in a certain region near the surface of the wedge. This region becomes narrower when the degree of spatial coherence of the incident wave diminishes. Inspection of Fig. 17.13 shows that the conditions for both temporal and spatial coherence become more favourable nearer to the apex of the wedge. Therefore, the distinctness of the interference pattern diminishes when moving from the apex of the wedge to its base. A pattern may be observed only for the thinner part of the wedge. For its remaining part, the screen will be uniformly illuminated.

Practically, fringes of equal thickness are observed by placing a lens near a wedge, and a screen behind the lens (Fig. 17.14). The part of the lens can be played by the crystalline lens, and of the screen by the retina of the eye. If the screen behind the lens is in a plane conjugated with the plane designated by Sc in Fig. 17.13 (the

eye is accordingly accommodated to this plane), the pattern will be most distinct. When the screen onto which the image is projected is moved (or when the lens is moved), the pattern will become less distinct and will vanish completely if the plane conjugated with the screen passes beyond the limits of the region of localization of the interference pattern observed without a lens.

When observed in white light, the fringes will be coloured, so that the surface of a plate or film will have rainbow colouring. For example, thin films of oil on the surface of water and soap films have such colouring. The temper colours appearing on the surface of steel articles when they are hardened are also due to interference from a film of transparent oxides.

Let us compare the two cases of interference upon reflection from thin films which we have considered. Fringes of equal inclination are obtained when a plate of constant thickness (b = constant) is illuminated by diffuse light containing rays of various directions (θ_1 is varied within more or less broad limits). Fringes of equal inclination are localized at infinity. Fringes of equal thickness are observed when a plate of varying thickness (b varies) is illuminated by a parallel beam of light (θ_1 = constant). Fringes of equal thickness are localized near the plate. In real conditions, for example, when observing rainbow colours on a soap or oil film, both the angle of incidence of the rays and the thickness of the film are varied. In this case, fringes of a mixed type are observed.

We must note that interference from thin films can be observed not only in reflected, but also in transmitted light.

Newton's Rings. A classical example of fringes of equal thickness are **Newton's rings**. They are observed when light is reflected from a thick plane-parallel glass plate in contact with a plano-convex lens having a large radius of curvature (Fig. 17.15). The part of a thin film from whose surfaces coherent waves are reflected is played by the air gap between the plate and the lens (owing to the great thickness of the plate and the lens, no interference fringes appear as a result of reflections from other surfaces). With normal incidence of the light, fringes of equal thickness have the form of concentric rings, and with inclined incidence, of ellipses. Let us find the radii of Newton's rings produced when light falls along a normal to the plate. In this case, $\sin\theta_1 = 0$, and the optical path difference equals the double thickness of the gap [see Eq. (17.33), it is assumed that $n = 1$ in the gap]. It follows from Fig. 17.15 that

$$R^2 = (R - b)^2 + r^2 \approx R^2 - 2Rb + r^2, \tag{17.40}$$

where R is the radius of curvature of the lens, r is the radius of a circle with the identical gap b corresponding to all of its points.

Owing to the smallness of b, in expression (17.40) we have disregarded the

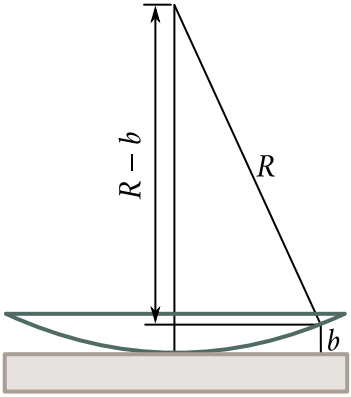


Fig. 17.15

quantity b^2 in comparison with $2Rb$. In accordance with expression (17.40), $b = r^2/(2R)$. To take account of the change in the phase by π occurring upon reflection from the plate, we must add $\lambda_0/2$ to $2b = r^2/R$. The result is

$$\Delta = \frac{r^2}{R} + \frac{\lambda_0}{2}. \tag{17.41}$$

At points for which $\Delta = m'\lambda_0 = 2m'(\lambda_0/2)$, maxima appear, and at points for which $\Delta = (m' + 1/2)\lambda_0 = (2m' + 1)(\lambda_0/2)$, minima of the intensity appear. Both conditions can be combined into the single one

$$\Delta = m\frac{\lambda_0}{2}$$

maxima corresponding to even values of m, and minima of the intensity, to odd values. Introducing into this expression Eq. (17.41) for Δ and solving the resulting equation relative to r, we find the radii of bright and dark Newton's rings:

$$r = \left(\frac{R\lambda_0(m-1)}{2}\right)^{1/2} \qquad (m = 1, 2, 3, \ldots). \tag{17.42}$$

Radii of bright rings correspond to even m's, and radii of dark rings to odd ones. The value $r = 0$ corresponds to $m = 1$, *i.e.*, to the point at the place of contact of the plate and the lens. A minimum of intensity is observed at this point. It is due to the change in the phase by π when a light wave is reflected from the plate.

Coating of Lenses. The coating of lenses is based on the interference of light when reflected from thin films. The transmission of light through each refracting surface of a lens is attended by the reflection of about four per cent of the incident light. In multicomponent lenses, such reflections occur many times, and the total loss of the light flux reaches an appreciable value. In addition, the reflections from the lens surfaces result in the appearance of highlights. The reflection of light is

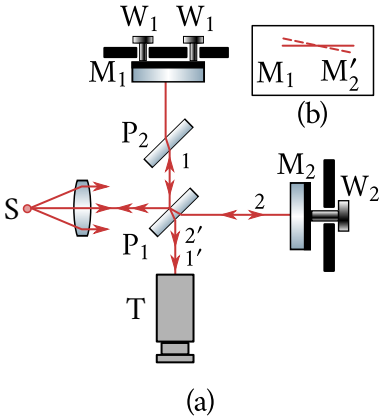


Fig. 17.16

eliminated by applying a thin film of a substance having a refractive index other than that of the lens to each free surface of the latter. The components obtained in this way are called **coated lenses**. The thickness of the coating is chosen so that the waves reflected from both its surfaces interfere destructively. An especially good result is obtained if the refractive index of the film equals the square root of the refractive index of the lens. When this condition is satisfied, the intensity of both waves reflected from the film surfaces is the same.

17.5. The Michelson Interferometer

Many varieties of interference instruments called **interferometers** are in use. Figure 17.16 is a schematic view of a **Michelson interferometer**[9]. A light beam from source S falls on semitransparent plate P_1 coated with a thin layer of silver (this layer is depicted by dots in the figure). Half of the incident light flux is reflected by plate P_1 in the direction of ray 1 and half passes through the plate and propagates in the direction of ray 2. Beam 1 is reflected from mirror M_1 and returns to P_1, where it is split into two beams of equal intensity. One of them passes through the plate and forms beam 1′, and the second one is reflected in the direction of S. The latter beam will no longer interest us. Beam 2 after being reflected by mirror M_2 also returns to plate P_1 where it is divided into two parts: beam 2′ reflected from the semitransparent layer, and the beam transmitted through the layer, which will also no longer interest us. Light beams 1′ and 2′ have the same intensity.

If conditions of temporal and spatial coherence are observed, beams 1′ and 2′

[9]Named after its inventor, the American physicist Albert Michelson (1852-1931).

will interfere. The result of this interference depends on the optical path difference from plate P_1 to mirrors M_1 and M_2, and back. Ray 2 passes through the plate three times, and ray 1 only once. To compensate the resulting change in the optical path difference (owing to dispersion) for waves of different lengths, plate P_1 is placed in the path of ray 1. Plates P_1 and P_2 are identical, except for the silver coating on the former. This arrangement makes the paths of rays 1 and 2 in glass equal. The interference pattern is observed with the aid of telescope T.

Let us mentally replace mirror M_2 with its virtual image M_2' in semitransparent plate P_1. Beams 1′ and 2′ can thus be considered as due to reflection from a transparent plate contained between planes M_1 and M_2. We can use adjusting screws W_1 to change the angle between these planes; in particular, they can be arranged strictly parallel to each other. By rotating micrometric screw W_2, we can smoothly move mirror M_1 without changing its inclination. We can thus change the thickness of the "plate"; in particular, we can make planes M_1 and M_2 intersect (Fig. 17.16b).

The nature of the interference pattern depends on the adjustment of the mirrors and on the divergence of the beam of light falling on the instrument. If the beam is parallel, and planes M_1 and M_2 make an angle other than zero, then straight fringes of equal thickness parallel to the lines of intersection of planes M_1 and M_2 will be observed in the field of vision of the telescope. In white light, all the fringes except the one coinciding with the line of intersection of the zero-order fringe will be coloured. The zero-order fringe will be black because beam 1 is reflected from plate P_1 from the outside, and beam 2 from the inside. As a result, a phase difference equal to π is produced between them. In white light, fringes are observed only with a small thickness of "plate" M_1M_2' [see Eq. (17.36)]. In monochromatic light corresponding to the red line of cadmium, Michelson observed a distinct interference pattern at a path difference of the order of 500000 wavelengths (the distance between M_1 and M_2' in this case is about 150 mm).

With a slightly diverging beam of light and a strictly parallel arrangement of planes M_1 and M_2', fringes of equal inclination are obtained that have the form of concentric rings. When micrometric screw W_2 is rotated, the diameter of the rings grows or diminishes. Either new rings appear at the centre of the pattern, or the diminishing rings shrink to a point and then vanish. Displacement of the pattern by one fringe corresponds to movement of mirror M_2 through half a wavelength.

Michelson used the instrument described above to carry out several experiments that entered the annals of physics. The most famous of them, performed together with the American chemist Edward Morley (1838-1923) in 1887, had the aim of detecting motion of the Earth relative to the hypothetic ether (we shall treat this experiment in Sec. 21.3). In 1890-1895, Michelson used the interferometer he had

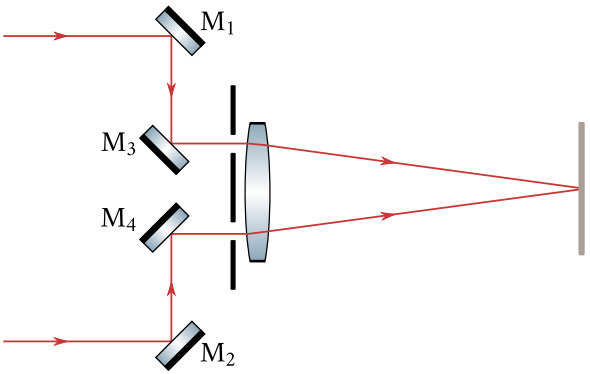


Fig. 17.17

invented to make the first comparison of the wavelength of the red line of cadmium with the length of the standard metre.

In 1920, Michelson constructed a **stellar interferometer** which he used to measure the angular dimensions of stars. This instrument was mounted on a telescope. A screen with two slits was installed in front of the objective of the telescope (Fig. 17.17). The light from a star was reflected from a symmetrical system of mirrors M_1, M_2, M_3 and M_4, installed on a rigid frame fastened on a carriage. The inner mirrors M_3 and M_4, were fixed, and the outer ones M_1 and M_2, could move symmetrically away from or toward mirrors M_3 and M_4. The path of the rays is clear from the figure. Interference fringes were produced in the focal plane of the telescope objective. Their visibility[10] depended on the distance between the outer mirrors. By moving these mirrors, Michelson determined the distance l between them at which the visibility of the fringes vanishes. This distance must be of the order of the coherence radius of a light wave arriving from a star. According to expression (17.26), the coherence radius is $l = \lambda/\varphi$. The condition $l = \lambda/\varphi$ gives the angular diameter of a star

$$\varphi = \frac{\lambda}{l}.$$

Accurate calculations give the formula

$$\varphi = A\frac{\lambda}{l},$$

where $A = 1.22$ for a source in the form of a uniformly illuminated disk. If the disk is darker at its edges than at the centre, the coefficient exceeds 1.22, its value

[10]The visibility of a fringe is defined as the quantity $V = (I_{max} - I_{min})/(I_{max} + I_{min})$, where I_{max} and I_{min} are the maximum and minimum intensities of the light in the vicinity of the given fringe, respectively.

depending on the rate of diminishing of the illumination in the direction from the centre toward the edge. In addition, accurate calculations show that after vanishing at a certain value of l, the visibility upon a further increase in l again becomes other than zero; however, the values it reaches are not great.

The maximum distance between the outer mirrors in the stellar interferometer constructed by Michelson was 6.1 m (the diameter of the telescope was 2.5 m). A minimum measurable angular diameter of about 0.02′ corresponded to this distance. The first star whose angular diameter was measured was Betelgeuse (alpha Orion). The value of φ obtained for it was 0.047′.

17.6. Multibeam Interference

Up to now, we have dealt with two-beam interference. Now let us investigate the interference of many light rays.

Assume that N rays of the same intensity arrive at a given point of a screen, the phase of each following ray being shifted relative to that of the preceding one by the same value δ. Let us represent the oscillations set up by the rays in the form of exponents:

$$E_1 = ae^{i\omega t},\ E_2 = ae^{i(\omega t+\delta)}, \ldots, E_m = ae^{i[\omega t+(m-1)\delta]}, \ldots, E_N = ae^{i[\omega t+(N-1)\delta]},$$

where a is the amplitude of an oscillation. The resultant oscillation is determined by the formula

$$E = \sum_{m=1}^{N} E_m = ae^{i\omega t} \sum_{m=1}^{N} e^{i(m-1)\delta}.$$

The expression obtained is the sum of N terms of a geometrical progression with its first term equal to unity and its common ratio equal to $e^{i\delta}$. Hence,

$$E = ae^{i\omega t}\left(\frac{1-e^{iN\delta}}{1-e^{i\delta}}\right) = \hat{A}\, e^{i\omega t},$$

where

$$\hat{A} = a\left(\frac{1-e^{iN\delta}}{1-e^{i\delta}}\right), \tag{17.43}$$

is the complex amplitude that can be represented in the form

$$\hat{A} = Ae^{i\alpha}, \tag{17.44}$$

(A is the usual amplitude of the resultant oscillation, and α is its initial phase).

The product of quantity (17.44) and its complex conjugate gives the square of the amplitude of the resultant oscillation:

$$\hat{A}\hat{A}^* = Ae^{i\alpha}Ae^{-i\alpha} = A^2. \tag{17.45}$$

Substituting for A in Eq. (17.45) its value from Eq. (17.43), we get the following expression for the square of the amplitude:

$$A^2 = \hat{A}\hat{A}^* = a^2 \frac{\left(1 - e^{iN\delta}\right)\left(1 - e^{-iN\delta}\right)}{\left(1 - e^{i\delta}\right)\left(1 - e^{-i\delta}\right)} = a^2 \frac{\left(2 - e^{iN\delta} - e^{-iN\delta}\right)}{\left(2 - e^{i\delta} - e^{-i\delta}\right)}$$

$$= a^2 \left[\frac{1 - \cos(N\delta)}{1 - \cos\delta}\right] = a^2 \frac{\sin^2(N\delta/2)}{\sin^2(\delta/2)}. \tag{17.46}$$

The intensity is proportional to the square of the amplitude. Hence, the intensity produced upon the interference of the N rays being considered is determined by the expression

$$I(\delta) = Ka^2 \frac{\sin^2(N\delta/2)}{\sin^2(\delta/2)} = I_0 \frac{\sin^2(N\delta/2)}{\sin^2(\delta/2)} \tag{17.47}$$

(K is a constant of proportionality, $I_0 = Ka^2$ is the intensity produced by each of the rays separately).

At the values

$$\delta = 2\pi m \quad (m = 0, \pm 1, \pm 2, \ldots), \tag{17.48}$$

Eq. (17.47) becomes indeterminate. For this reason, we apply L'Hospital's rule:

$$\lim_{\delta \to 2\pi m} \frac{\sin^2(N\delta/2)}{\sin^2(\delta/2)} = \lim_{\delta \to 2\pi m} \frac{2\sin(N\delta/2)\cos(N\delta/2)(N/2)}{2\sin(\delta/2)\cos(\delta/2)(1/2)} = \lim_{\delta \to 2\pi m} N \frac{\sin(N\delta)}{\sin\delta}.$$

The expression obtained is also indeterminate. For this reason, we apply L'Hospital's rule again:

$$\lim_{\delta \to 2\pi m} \frac{\sin^2(N\delta/2)}{\sin^2(\delta/2)} = \lim_{\delta \to 2\pi m} N \frac{\sin(N\delta)}{\sin\delta} = \lim_{\delta \to 2\pi m} N \frac{\cos(N\delta)}{\cos\delta} = N^2.$$

Thus, when $\delta = 2\pi m$ (or when the path differences $\Delta = m\lambda_0$), the resultant intensity is

$$I = I_0 N^2. \tag{17.49}$$

This result could have been predicted. Indeed, all the oscillations arrive at points for which $\delta = 2\pi m$ in the same phase. Hence, the resultant amplitude is N times the amplitude of a separate oscillation, and the intensity is N^2 times that of a separate oscillation.

Let us call the spots where the intensity determined by Eq. (17.49) is observed the **principal maxima**. Their position is determined by condition (17.48). The number m is called the **order** of the principal maximum. It can be seen from Eq. (17.47) that the space between two adjacent principal maxima accommodates $N - 1$ minima of the intensity. To verify this statement, let us consider, for example, the interval between the maxima of the zero ($m = 0$) and of the first ($m = 1$) order. In this interval, δ changes from zero to 2π, and $\delta/2$ from zero to π. The denominator

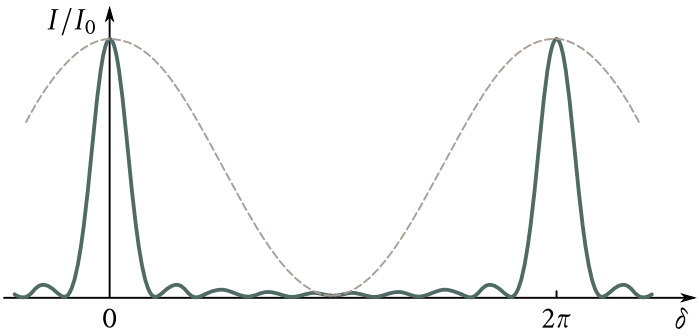


Fig. 17.18

of Eq. (17.47) is other than zero everywhere except for the ends of the interval. It reaches its maximum value equal to unity at the middle of the interval. The quantity $N\delta/2$ takes on all the values from zero to $N\pi$ within the interval being considered. At values of $\pi, 2\pi, \ldots, (N-1)\pi$, the numerator of Eq. (17.47) becomes equal to zero. Here, we have minima of the intensity. Their positions correspond to values of δ equal to

$$\delta = \frac{k'}{N} 2\pi \quad (k' = 1, 2, \ldots, N-1). \tag{17.50}$$

There are $N-2$ secondary maxima in the intervals between the $N-1$ minima. The secondary maxima closest to the principal maxima have the greatest intensity. The secondary maximum closest to the principal zero-order maximum is between the first ($k' = 1$) and second ($k' = 2$) minima. Values of δ equal to $2\pi/N$ and $4\pi/N$ correspond to these minima. Hence, $\delta = 3\pi/N$ corresponds to the secondary maximum being considered. Introduction of this value into Eq. (17.47) yields

$$I(3\pi/N) = Ka^2 \frac{\sin^2(3\pi/N)}{\sin^2(3\pi/2N)}.$$

The numerator equals unity. At a great value of N, we may assume that the sine in the denominator equals its argument [$\sin(3\pi/2N) \approx 3\pi/2N$]. Hence,

$$I(3\pi/N) = Ka^2 \frac{1}{(3\pi/2N)^2} = \frac{Ka^2N^2}{(3\pi/2)^2}.$$

The quantity in the numerator is the intensity of the principal maximum [see Eq. (17.49)]. Thus, at a great value of N, the secondary maximum closest to the principal maximum has an intensity that is $1/(3n/2)^2 \approx 1/22$ of the intensity of the principal maximum. The other secondary maxima are still weaker.

Figure 17.18 shows a plot of the function $I(\delta)$ for $N = 10$. For comparison, a plot of the intensity for $N = 2$ [two-beam interference; see the curve $I(x)$ in Fig. 17.2] is shown by a dash line. Inspection of the figure shows that the principal maxima

become narrower and narrower with an increase in the number of interfering rays. The secondary maxima are so weak that the interference pattern practically has the form of narrow bright lines on a dark background.

Now, let us consider the interference of a very great number of rays whose intensity diminishes in a geometrical progression. The oscillations being added have the form

$$E_1 = ae^{i\omega t},\ E_2 = a\rho e^{i(\omega t+\delta)}, \ldots, E_m = a\rho^{m-1}e^{i[\omega t+(m-1)\delta]}, \ldots, \tag{17.51}$$

(ρ is a constant quantity less than unity). The resultant oscillation is described by the equation

$$E = \sum_{m=1}^{N} E_m = ae^{i\omega t}\sum_{m-1}^{N} \rho^{m-1}e^{i(m-1)\delta}.$$

Using the expression for the sum of the terms of a geometrical progression, we get

$$E = ae^{i\omega t}\left(\frac{1-\rho e^{iN\delta}}{1-\rho e^{i\delta}}\right) = \hat{A}e^{i\omega t}.$$

Thus, the complex amplitude is

$$\hat{A} = a\left(\frac{1-\rho e^{iN\delta}}{1-\rho e^{i\delta}}\right). \tag{17.52}$$

If N is very great, the complex number $\rho N e^{iN\delta}$ may be disregarded in comparison with unity (we shall indicate as an example that $0.9^{100} \approx 4\times 10^{-4}$). Equation (17.52) is thus simplified as follows:

$$\hat{A} = a\left(\frac{1}{1-\rho e^{i\delta}}\right).$$

Multiplying this equation by its complex conjugate, we get the square of the ordinary amplitude of the resultant oscillation:

$$\begin{aligned} A^2 = \hat{A}\hat{A}^* &= \frac{a^2}{\left(1-\rho e^{i\delta}\right)\left(1-\rho e^{-i\delta}\right)} = \frac{a^2}{1+\rho^2-\rho\left(e^{i\delta}+e^{-i\delta}\right)} \\ &= \frac{a^2}{1+\rho^2-2\rho\cos\delta} = \frac{a^2}{(1-\rho)^2+2\rho(1-\cos\delta)} \\ &= \frac{a^2}{(1-\rho^2)+4\rho\sin^2(\delta/2)}. \end{aligned}$$

Hence,

$$I(\delta) = \frac{Ka^2}{(1-\rho^2)+4\rho\sin^2(\delta/2)} = \frac{I_1}{(1-\rho^2)+4\rho\sin^2(\delta/2)}, \tag{17.53}$$

where $I_1 = Ka^2$ is the intensity of the first (most intensive) ray.

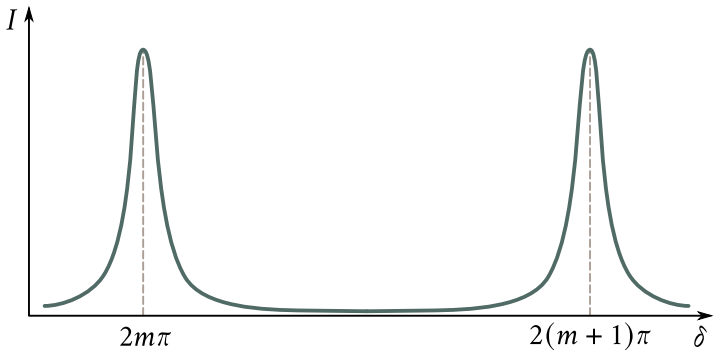


Fig. 17.19

At values of

$$\delta = 2\pi m \quad (m = 0, \pm 1, \pm 2, \ldots), \tag{17.54}$$

Eq. (17.53) has maxima equal to

$$I_{\max} = \frac{I_1}{(1-\rho)^2}. \tag{17.55}$$

In the intervals between maxima, the function changes monotonously, reaching a value equal to

$$I_{\min} = \frac{I_1}{(1-\rho)^2 + 4\rho} = \frac{I_1}{(1+\rho)^2} \tag{17.56}$$

at the middle of the interval. Thus, the ratio of the intensity at a maximum to that at a minimum

$$\frac{I_{\max}}{I_{\min}} = \left(\frac{1+\rho}{1-\rho}\right)^2 \tag{17.57}$$

is the greater, the closer ρ is to unity, *i.e.*, the slower is the rate of diminishing of the intensity of the interfering rays. Figure 17.19 shows a graph of function (17.53) for $\rho = 0.8$. It can be seen from the figure that the interference pattern has the form of narrow sharp lines on a virtually dark background. Unlike Fig. 17.18, secondary maxima are absent.

A practical case of a great number of rays with a diminishing intensity is encountered in the **Fabry-Perot interferometer**. This instrument consists of two glass or quartz plates separated by an air gap (Fig. 17.20). The internal surfaces of the plates are thoroughly polished so that the irregularities on them do not exceed several hundredths of the length of a light wave. Next partly transparent metal layers or dielectric films[11] are applied to these surfaces. The outer surfaces of the plates are

[11] Metal layers have the shortcoming that they absorb light rays to a great extent. This is why recent years have seen their replacement with multilayer dielectric coatings having a high reflectivity.

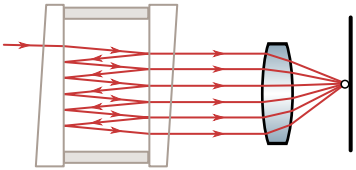


Fig. 17.20

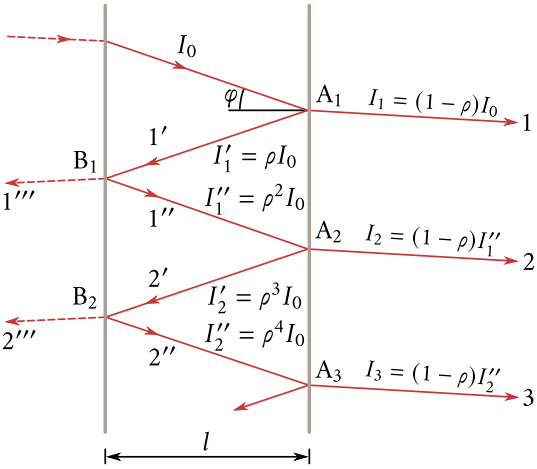


Fig. 17.21

at a slight angle relative to the inner ones to eliminate the highlights due to the reflection of light from these surfaces. In the original design of the interferometer, one of the plates could be moved relative to the other stationary one with the aid of a micrometric screw. The unreliability of this design, however, resulted in its coming out of use. In modern designs, the plates are secured rigidly. The parallelity of the internal working planes is achieved by installing an invar or quartz ring[12] between the plates. This ring has three projections with thoroughly polished edges at each side. The plates are pressed against the ring by springs. This design reliably ensures strict parallelity of the internal planes of the plates and constancy of the distance between them. Such an interferometer with a fixed distance between its plates is known as a **Fabry-Perot etalon**.

Let us see what happens to a ray entering the gap between the plates (Fig. 17.21). Assume that the intensity of the entering ray is I_0. At point A_1, this ray is divided into ray 1 emerging outward and reflected ray 1′. If the coefficient of reflection from the surface of the plate is ρ, then the intensity of ray 1 will be $I_1 = (1-p)I_0$, and the intensity of the reflected ray will be $I_1' = \rho I_0$[13]. At point B_1, ray 1′ is divided into two. Ray 1″ shown by a dash line will drop out of consideration, while reflected ray 1″ will have an intensity of $I_1'' = \rho I_1' = \rho^2 I_0$. At point A_2, ray 1″ will be divided into two rays—ray 2 emerging outward having an intensity of $I_2 = (1-\rho)I_1'' = (1-\rho)\rho^2/I_0$ and reflected ray 2′, and so on. Thus, the following

[12] Both these materials are distinguished by their extremely low temperature coefficient of expansion.

[13] We disregard the absorption of light in the reflecting layers and inside the plates.

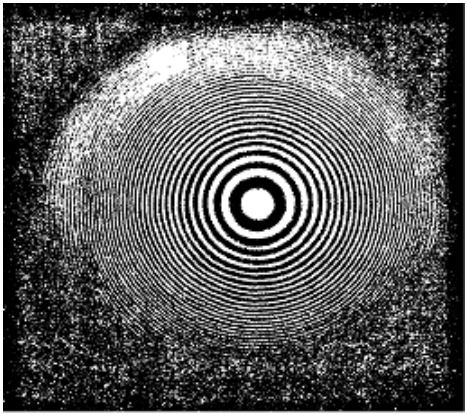


Fig. 17.22

relation holds for the intensities of rays 1, 2, 3, etc. emerging from the instrument:

$$I_1 : I_2 : I_3 : \ldots = 1 : \rho^2 : \rho^4 : \ldots .$$

Accordingly, for the amplitudes of the oscillations we have

$$A_1 : A_2 : A_3 : \ldots = 1 : \rho : \rho^2 : \ldots$$

[compare with Eq. (17.51)].

The oscillation in each of the rays 2, 3, 4, . . ., lags in phase behind the oscillation in the preceding ray by the same amount δ determined by the optical path difference Δ appearing on the path A_1-B_1-A_2 or A_2-B_2-A_3, etc. (see Fig. 17.21). A glance at the figure shows that $\Delta = 2l/\cos\varphi$, where φ is the angle of incidence of the rays on the reflecting layers.

If we gather rays 1, 2, 3, . . ., with the aid of a lens at point P of its focal plane (see Fig. 17.20), then the oscillations produced by these rays will have the form given by Eq. (17.51). Hence, the intensity at point P is determined by Eq. (17.53), in which ρ has the meaning of the coefficient of reflection, and

$$\delta = \frac{2\pi}{\lambda} = \frac{2l}{\cos\varphi}.$$

When a diverging beam of light is passed through the instrument, fringes of equal inclination having the form of sharp rings (Fig. 17.22) will be produced in the focal plane of the lens.

The Fabry-Perot interferometer is used in spectroscopy to study the fine structure of spectral lines. It has also come into great favour in metrology for comparing the length of the standard metre with the wavelengths of individual spectral lines.

Chapter 18
DIFFRACTION OF LIGHT

18.1. Introduction

By diffraction is meant the combination of phenomena observed when light propagates in a medium with sharp heterogeneities[1] and associated with deviations from the laws of geometrical optics. Diffraction, in particular, leads to light waves bending around obstacles and to the penetration of light into the region of a geometrical shadow. The bending of sound waves around obstacles (*i.e.*, the diffraction of sound waves) is constantly observed in our everyday life. To observe the diffraction of light waves, special conditions must be set up. This is due to the smallness of the lengths of light waves. We know that in the limit, when $\lambda \to 0$, the laws of wave optics transform into those of geometrical optics. Hence, other conditions being equal, the deviations from the laws of geometrical optics decrease with a diminishing wavelength.

There is no appreciable physical difference between interference and diffraction. Both phenomena consist in the redistribution of the light flux as a result of superposition of the waves. For historical reasons, the redistribution of the intensity produced as a result of the superposition of waves emitted by a finite number of discrete coherent sources has been called the interference of waves. The redistribution of the intensity produced as a result of the superposition of waves emitted by coherent sources arranged continuously has been called the diffraction of waves. We, therefore, speak about the interference pattern from two narrow slits and about the diffraction pattern from one slit.

Diffraction is usually observed by means of the following set-up. An opaque barrier closing part of the wave surface of the light wave is placed in the path of a light wave propagating from a certain source. A screen on which the diffraction

[1] For example, near the boundaries of opaque or transparent bodies, through small holes, etc.

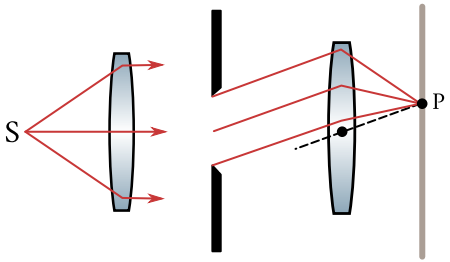


Fig. 18.1: Fraunhofer diffraction observed by placing a lens after light source S and another one in front of point of observation P. Points S and P lie in the focal plane of each lens.

pattern appears is placed after the barrier.

Two kinds of diffraction are distinguished. If the light source S and the point of observation P are so far from a barrier that the rays falling on the barrier and those travelling to point P form virtually parallel beams, we have to do with diffraction in parallel rays or with **Fraunhofer diffraction**. Otherwise, we have to do with Fresnel diffraction. Fraunhofer diffraction can be observed by placing a lens after light source S and another one in front of point of observation P, so that points S and P lie in the focal plane of the relevant lens (Fig. 18.1).

The criterion allowing us to determine the kind of diffraction we are dealing with—Fresnel or Fraunhofer—in each specific case will be given in Sec. 18.5.

18.2. Huygens-Fresnel Principle

The penetration of light waves into the region of a geometrical shadow can be explained with the aid of Huygens' principle (see Sec. 16.9). This principle, however, gives no information on the amplitude and, consequently, on the intensity of waves propagating in different directions. The French physicist Augustin Fresnel (1788-1827) supplemented Huygens' principle with the concept of the interference of secondary waves. Taking into account the amplitudes and phases of the secondary waves makes it possible to find the amplitude of the resultant wave for any point of space. Huygens' principle developed in this way was named the **Huygens-Fresnel principle**.

According to the Huygens-Fresnel principle, every element of wave surface S (Fig. 18.2) is the source of a secondary spherical wave whose amplitude is proportional to the size of element $\mathrm{d}S$. The amplitude of a spherical wave diminishes with the distance r from its source according to the law $1/r$ [see Eq. (14.12)]. Consequently, the oscillation

$$\mathrm{d}E = K\frac{a_0\,\mathrm{d}S}{r}\cos(\omega t - kr + \alpha_0), \tag{18.1}$$

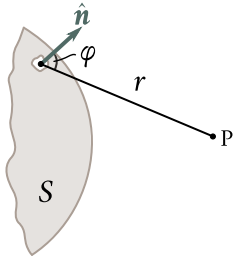


Fig. 18.2: Huygens-Fresnel principle: every element of wave surface S is the source of a secondary spherical wave whose amplitude is proportional to the size of element dS.

arrives from each section dS of a wave surface at point P in front of this surface. In Eq. (18.1), $(\omega t + \alpha_0)$ is the phase of the oscillation where wave surface S is, k is the wave number, r is the distance from surface element dS to point P. The factor α_0 is determined by the amplitude of the light oscillation at the location of dS. The coefficient K depends on the angle φ between a normal $\hat{\boldsymbol{n}}$ to area ds and the direction from dS to point P. When $\varphi = 0$, this coefficient is maximum; when $\varphi = \pi/2$, it vanishes.

he resultant oscillation at point P is the superposition of the oscillations given by Eq. (18.1) taken for the entire wave surface S:

$$E = \int_S K(\varphi) \frac{a_0}{r} \cos(\omega t - kr + \alpha_0)\, \mathrm{d}S. \tag{18.2}$$

This equation is an analytical expression of the Huygens-Fresnel principle.

The Huygens-Fresnel principle can be substantiated by the following reasoning. Assume that thin opaque screen Sc (Fig. 18.3) is placed in the path of a light wave (we shall consider it plane for simplicity's sake). The intensity of the light everywhere after the screen will be zero. The reason is that the light wave falling on the screen produces oscillations of the electrons in the material of the screen. The oscillating electrons emit electromagnetic waves. The field after the screen is a superposition of the primary wave (falling on the screen) and all the secondary waves. The amplitudes and phases of the secondary waves are such that upon superposition of these waves with the primary one, a zero amplitude is obtained at any point P after the screen. Consequently, if the primary wave produces the oscillation

$$A_{\text{prim}} \cos(\omega t + \alpha)$$

at point P, then the resultant oscillation produced by the secondary waves at the same point has the form

$$A_{\text{sec}} \cos(\omega t + \alpha - \pi).$$

Here, $A_{\text{sec}} = A_{\text{prim}}$.

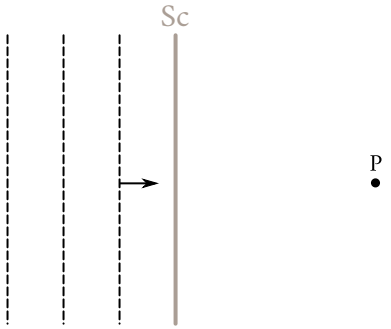


Fig. 18.3: A light wave (primary wave) falling on a thin opaque screen Sc produces oscillations of the electrons in the material of the screen. The oscillating electrons emit electromagnetic waves (secondary wave). The amplitudes and phases of the secondary waves are such that, upon superposition of these waves with the primary one, a zero amplitude is obtained at any point P after the screen.

What has been said above signifies that when calculating the amplitude of an oscillation set up at point P by a light wave propagating from a real source, we can replace this source with a collection of secondary sources arranged along the wave surface. This is exactly the essence of Huygens-Fresnel principle.

Let us divide the opaque barrier into two parts. One of them, which we shall call a stopper, has finite dimensions and an arbitrary shape (a circle, rectangle, etc.). The other part includes the entire remaining surface of the infinite barrier. As long as the stopper is in place, the resultant oscillation at point P after the barrier is zero. It can be represented as the sum of the oscillations set up by the primary wave, the wave produced by the stopper, and the wave produced by the remaining part of the barrier:

$$A_{\text{prim}} \cos(\omega t + \alpha) + A_{\text{stop}} \cos(\omega t + \alpha') + A_{\text{bar}} \cos(\omega t + \alpha'') = 0. \tag{18.3}$$

If the stopper is removed, *i.e.*, the wave is transmitted through the aperture in the opaque barrier, then the oscillation at point P will have the form

$$\begin{aligned} E_{\text{P}} &= A_{\text{prim}} \cos(\omega t + \alpha) + A_{\text{bar}} \cos(\omega t + \alpha'') \\ &= -A_{\text{stop}} \cos(\omega t + \alpha') = A_{\text{stop}} \cos(\omega t + \alpha' - \pi). \end{aligned}$$

We have used condition (18.3) and assumed that removal of the stopper does not change the nature of the oscillations of the electrons in the remaining part of the barrier.

We can, thus, consider that the oscillations at point P are produced by a collection of sources of secondary waves on the surface of the aperture formed after removal of the stopper.

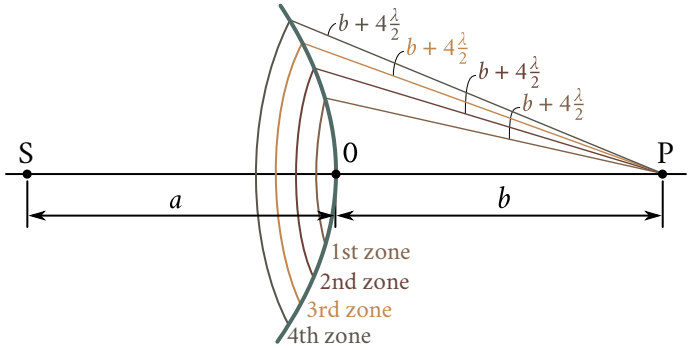


Fig. 18.4: Fresnel zones obtained by division of a wave surface, travelling from S to P, into annular zones constructed so that the distances from the edges of each zone to point P differ by $\lambda/2$, where λ is the length of the wave in the medium of propagation.

18.3. Fresnel Zones

The performance of calculations by Eq. (18.2) is a very difficult task in the general case. As Fresnel showed, however, the amplitude of the resultant oscillation can be found by simple algebraic or geometrical summation in cases distinguished by symmetry.

To understand the essence of the method developed by Fresnel, let us determine the amplitude of the light oscillation set up at point P by a spherical wave propagating in an isotropic homogeneous medium from point source S (Fig. 18.4). The wave surfaces of such waves are symmetrical relative to straight line SP. Taking advantage of this circumstance, let us divide the wave surface shown in the figure into annular zones constructed so that the distances from the edges of each zone to point P differ by $\lambda/2$ (λ is the length of the wave in the medium in which it is propagating). Zones having this property are known as **Fresnel zones**.

A glance at Fig. 18.4 shows that the distance b_m from the outer edge of the m-th zone to point P is

$$b_m = b + m\frac{\lambda}{2} \tag{18.4}$$

(b is the distance from the crest 0 of the wave surface to point P).

The oscillations arriving at point P from similar points of two adjacent zones (*i.e.*, from points at the middle of the zones, or at the outer edges of the zones, etc.) are in counterphase. Therefore, the resultant oscillations produced by each of the zones as a whole will differ in phase for adjacent zones by π too.

Let us calculate the areas of the zones. The outer boundary of the m-th zone separates a spherical segment of height h_m on the wave surface (Fig. 18.5). Let the

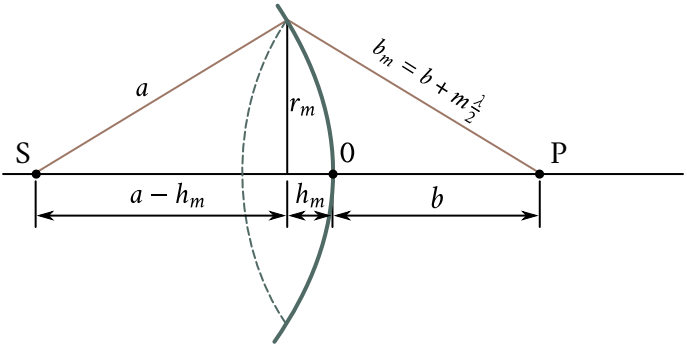


Fig. 18.5: Area of a Fresnel zone. The outer boundary of the m-th zone separates a spherical segment of height h_m on the wave surface. The area of this segment be S_m.

area of this segment be S_m. Hence, the area of the m-th zone can be written as

$$\Delta S_m = S_m - S_{m-1},$$

where S_{m-1} is the area of the spherical segment separated by the outer boundary of the $(m-1)$-th zone.

It can be seen from Fig. 18.5 that

$$r_m^2 = A^2 - (a - h_m)^2 = \left(b + m\frac{\lambda}{2}\right)^2 - (b + h_m)^2,$$

where a is the radius of the wave surface and r_m is the radius of the outer boundary of the m-th zone.

Squaring the terms in parentheses, we get

$$r_m^2 = 2ah_m - h_m^2 = bm\lambda + m^2\left(\frac{\lambda}{2}\right)^2 - 2bh_m - h_m^2, \tag{18.5}$$

whence,

$$h_m = \frac{bm\lambda + m^2(\lambda/2)^2}{2(a+b)}. \tag{18.6}$$

Restricting ourselves to a consideration of not too great m's, we may disregard the addend containing λ^2 owing to the smallness of λ. In this approximation

$$h_m = \frac{bm\lambda}{2(a+b)}. \tag{18.7}$$

The area of a spherical segment is $S = 2\pi Rh$ (here, R is the radius of the sphere and h is the height of the segment). Hence,

$$S_m = 2\pi ah_m = \left[\frac{\pi ab}{(a+b)}\right] m\lambda,$$

and the area of the m-th zone is

$$\Delta S_m = S_m - S_{m-1} = \frac{\pi ab\lambda}{(a+b)}.$$

The expression we have obtained does not depend on m. This signifies that when m is not too great, the areas of the Fresnel zones are approximately identical.

We can find the radii of the zones from Eq. (18.5). When m is not too great, the height of a segment $h_m \ll a$, and we can, therefore, consider that $r_m^2 = 2ah_m$. Substituting for h_m its value from Eq. (18.7), we get the following expression for the radius of the outer boundary of the m-th zone:

$$r_m = \left[\left(\frac{ab}{a+b}\right) m\lambda\right]^{1/2}. \tag{18.8}$$

If we assume that $a = b = 1$ m and $A = 0.5$ μm, then, we get a value of $r_1 = 0.5$ mm for the radius of the first (central) zone. The radii of the following zones grow as $\sqrt{m}$.

Thus, the areas of the Fresnel zones are approximately the same. The distance b_m from a zone to point P slowly increases with the zone number m. The angle φ between a normal to the zone elements and the direction toward point P also grows with m. All this leads to the fact that the amplitude A_m of the oscillation produced by the m-th zone at point P diminishes monotonously with increasing m. Even at very high values of m, when the area of a zone begins to grow appreciably with m [see Eq. (18.6)], the decrease in the factor $K(\varphi)$ overbalances the increase in ΔS_m, so that A_m continues to diminish. Thus, the amplitudes of the oscillations produced at point P by Fresnel zones form a monotonously diminishing sequence:

$$A_1 > A_2 > A_3 > \ldots > A_{m-1} > A_m > \ldots.$$

The phases of the oscillations produced by adjacent zones differ by π. Therefore, the amplitude A of the resultant oscillation at point P can be represented in the form

$$A = A_1 - A_2 + A_3 - A_4 + \ldots. \tag{18.9}$$

This expression includes all the amplitudes from odd zones with one sign and from even zones with the opposite one.

Let us write Eq. (18.9) in the form

$$A = \frac{A_1}{2} + \left(\frac{A_1}{2} - A_2 + \frac{A_3}{2}\right) + \left(\frac{A_3}{2} - A_4 + \frac{A_5}{2}\right) + \ldots. \tag{18.10}$$

Owing to the monotonous diminishing of A_m, we can approximately assume that

$$A_m = \frac{A_{m-1} + A_{m+1}}{2}.$$

The expressions in parentheses will therefore vanish, and Eq. (18.10) will be simplified

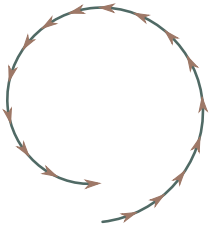


Fig. 18.6: Vector diagram obtained when the oscillations produced by the separate zones are added. The vectors form a broken spiral-shaped line instead of a closed figure.

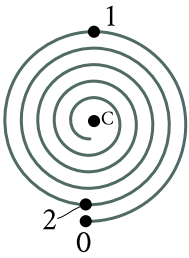


Fig. 18.7: The phases of the oscillations at points 0 and 1 differ by π (the infinitely small vectors forming the spiral have opposite directions at these points).

as follows:

$$A = \frac{A_1}{2}. \tag{18.11}$$

According to Eq. (18.11), the amplitude produced at a point P by an entire spherical wave surface equals half the amplitude produced by the central zone alone. If we put in the path of a wave an opaque screen having an aperture that leaves only the central Fresnel zone open, the amplitude at point P will equal A_1, *i.e.*, it will be double the amplitude given by Eq. (18.11). Accordingly, the intensity of the light at point P will in this case be four times greater than when there are no barriers between points S and P.

Now, let us solve the problem on the propagation of light from source S to point P by the method of graphical addition of amplitudes. We shall divide the wave surface into annular zones similar to Fresnel zones, but much smaller in width (the path difference from the edges of a zone to point P is a small fraction of λ the same for all zones). We shall depict the oscillation produced at point P by each of the zones in the form of a vector whose length equals the amplitude of the oscillation, while the angle made by the vector with the direction taken as the beginning of measurement gives the initial phase of the oscillation (see Sec. 7.7 of Vol. I). The amplitude of the oscillations produced by such zones at point P slowly diminishes from zone to zone. Each following oscillation lags behind the preceding one in phase by the same magnitude. Hence, the vector diagram obtained when the oscillations produced by the separate zones are added has the form shown in Fig. 18.6.

If the amplitudes produced by the individual zones were the same, the tail of the last of the vectors shown in Fig. 18.6 would coincide with the tip of the first vector. Actually, the value of the amplitude diminishes, although very slightly. Hence, the vectors form a broken spiral-shaped line instead of a closed figure.

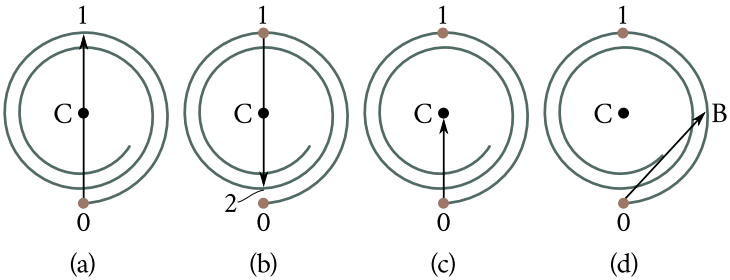


Fig. 18.8: (a, b) First and second Fresnel zones. (c) The oscillation produced at point P by the entire wave surface is depicted by vector OC. The amplitude here equals half the amplitude produced by the first zone. (d) The oscillation produced by the inner half of the first Fresnel zone is depicted by vector 0B, which is $\sqrt{2}$ times greater than vector 0C. Thus, the intensity of light in the inner half of the First zone is twice that of the entire wave surface.

In the limit when the widths of the annular zones tend to zero (their number will grow unlimitedly), the vector diagram has the form of a spiral winding toward point C (Fig. 18.7). The phases of the oscillations at points 0 and 1 differ by π (the infinitely small vectors forming the spiral have opposite directions at these points).

Consequently, part 0-1 of the spiral corresponds to the first Fresnel zone. The vector drawn from point 0 to point 1 (Fig. 18.8a) depicts the oscillation produced at point P by this zone. Similarly, the vector drawn from point 1 to point 2 (Fig. 18.8b) depicts the oscillation produced by the second Fresnel zone. The oscillations from the first and second zones are in counterphase; accordingly, vectors 01 and 12 have opposite directions.

The oscillation produced at point P by the entire wave surface is depicted by vector 0C (Fig. 18.8c). Inspection of the figure shows that the amplitude in this case equals half the amplitude produced by the first zone. We have obtained this result algebraically earlier [see Eq. (18.11)]. We shall note that the oscillation produced by the inner half of the first Fresnel zone is depicted by vector 0B (Fig. 18.8d). Thus, the action of the inner half of the first Fresnel zone is not equivalent to half the action of the first zone. Vector 0B is $\sqrt{2}$ times greater than vector 0C. Consequently, the intensity of the light produced by the inner half of the first Fresnel zone is double the intensity produced by the entire wave surface.

The oscillations from the even and odd Fresnel zones are in counterphase and, therefore, mutually weaken one another. If we would place in the path of the light wave a plate that would cover all the even or odd zones, the intensity of the light at point P would sharply grow. Such a plate, known as a zone one, functions like a converging lens. Figure 18.9 shows a plate covering the even zones. A still greater effect can be achieved by changing the phase of the even (or odd) zone

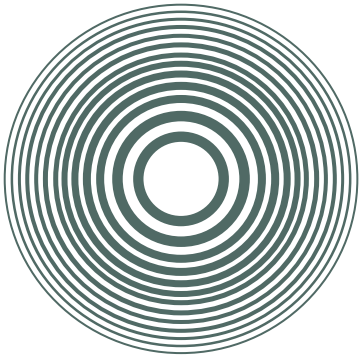


Fig. 18.9: Plate covering the even Fresnel zones. The oscillations from the even and odd Fresnel zones are in counterphase and, therefore, mutually weaken one another.

oscillations by π instead of covering these zones. This can be done with the aid of a transparent plate whose thickness at the places corresponding to the even or odd zones differs by a properly selected value. Such a plate is called a **phase zone plate**. In comparison with the **amplitude zone plate** covering zones, a phase plate produces an additional two-fold increase in the amplitude, and a four-fold increase in the light intensity.

18.4. Fresnel Diffraction from Simple Barriers

The methods of algebraic and graphical addition of amplitudes treated in the preceding section make it possible to solve a number of problems involving the diffraction of light.

Diffraction from a Round Aperture. Let us put an opaque screen with a round aperture of radius r_0 cut out in it in the path of a spherical light wave. We shall arrange the screen so that a perpendicular dropped from light source S passes through the centre of the aperture (Fig. 18.10). Let us take point P on the continuation of this perpendicular. At an aperture radius r_0 considerably smaller than the lengths a and b shown in the figure, the length a can be considered equal to the distance from source S to the barrier, and the length b, to the distance from the barrier to point P. If the distances a and b satisfy the relation

$$r_0 = \left[\left(\frac{ab}{a+b}\right) m\lambda\right]^{1/2}, \tag{18.12}$$

where m is an integer, then the aperture will leave open exactly m first Fresnel zones constructed for point P [see Eq. (18.8)]. Hence, the number of open Fresnel zones is

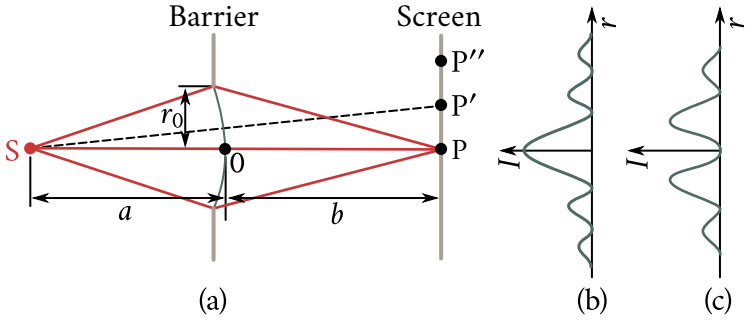


Fig. 18.10: Opaque screen with a round aperture of radius r_0 cut out in it in the path of a spherical light wave. The screen is arranged so that a perpendicular dropped from light source S passes through the centre of the aperture. The diffraction patterns produced by the round aperture are shown for when m is odd (b) and when m is even (c).

determined by the expression

$$m = \frac{r_0^2}{\lambda}\left(\frac{1}{a} + \frac{1}{b}\right). \tag{18.13}$$

According to Eq. (18.9), the amplitude at point P will be

$$A = A_1 - A_2 + A_3 - A_4 + \ldots \pm A_m. \tag{18.14}$$

The amplitude A_m is taken with a plus sign if m is odd and with a minus sign if m is even. Writing Eq. (18.14) in a form similar to Eq. (18.10) and assuming that the expressions in parentheses equal zero, we arrive at the equations

$$A = \frac{A_1}{2} + \frac{A_m}{2} \quad (m \text{ is odd})$$

$$A = \frac{A_1}{2} + \frac{A_{m-1}}{2} - A_m \quad (m \text{ is even}).$$

The amplitudes from two adjacent zones are virtually the same. We may therefore replace $(A_{m-1}/2) - A_m$ with $-A_m/2$. The result is

$$A = \frac{A_1}{2} \pm \frac{A_m}{2}, \tag{18.15}$$

where the plus sign is taken for odd and the minus sign for even m's.

The amplitude A_m differs only slightly from A_1 for small m's. Hence, with odd m's, the amplitude at point P will approximately equal A_1, and at even m's, zero. It is easy to obtain this result with the aid of the vector diagram shown in Fig. 18.7.

If we remove the barrier, the amplitude at point P will become equal to $A_1/2$ [see Eq. (18.11)]. Thus, a barrier with an aperture opening a small odd number of zones not only does not weaken the illumination at point P but, on the contrary, leads to an increase in the amplitude almost twice, and of the intensity, almost four

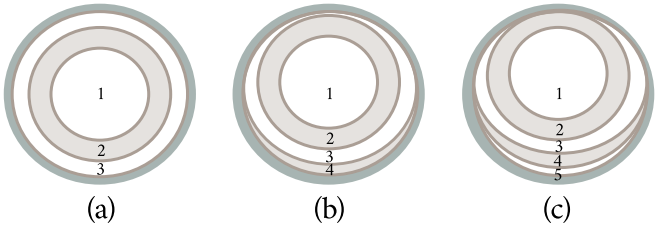


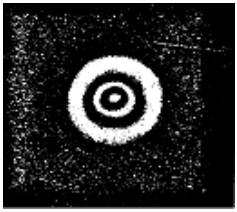
Fig. 18.11: (a, b, c) Diffraction patterns for different numbers of Fresnel zones obtained at points P, P′, and P″ of Fig. 18.10a, respectively. The patterns in (b) and (c) are limited by the edges of the aperture.

times.

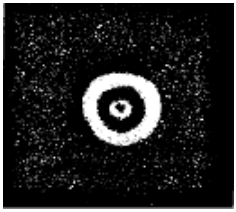
Let us determine the nature of the diffraction pattern that will be observed on a screen placed after the barrier (see Fig. 18.10). Owing to the symmetrical arrangement of the aperture relative to straight line SP, the illumination at various points of the screen will depend only on the distance r from point P. At this point itself, the intensity will reach a maximum or a minimum depending on whether the number of open (effective) Fresnel zones is even or odd. Assume, for example, that this number is three. In this case, we get a maximum of intensity at the centre of the diffraction pattern. A pattern of the Fresnel zones for point P is given in Fig. 18.11a.

Now let us move along the screen to point P′. The pattern of the Fresnel zones for point P′ limited by the edges of the aperture has the form shown in Fig. 18.11b. The edges of the aperture will obstruct a part of the third zone, and simultaneously the fourth zone will be partly opened. As a result, the intensity of the light diminishes, and reaches a minimum at a certain position of point P′. If we move along the screen to point P″, the edges of the aperture will partly obstruct not only the third, but also the second Fresnel zone, simultaneously partly opening the fifth zone (Fig. 18.11c). The result will be that the action of the open sections of the odd zones will overbalance the action of the open sections of the even zones and the intensity will reach a maximum. True, this maximum will be weaker than that observed at point P.

Thus, the diffraction pattern produced by a round aperture has the form of alternating bright and dark concentric rings. There will be either a bright (m is odd) or a dark (m is even) spot at the centre of the pattern (Fig. 18.12). The variation in the intensity I with the distance r from the centre of the pattern is shown in Fig. 18.10b (for an odd m) and in Fig. 18.10c (for an even m). When the screen is moved parallel to itself along straight line SP, the patterns shown in Fig. 18.12 will replace one another [according to Eq. (18.13), when b changes, the value of m becomes odd and even alternately].



Odd m



Even m

Fig. 18.12: Diffraction patterns produced by a round aperture alternating bright and dark concentric rings. At the centre of the pattern, a bright spot results for an odd m value, while a dark spot for an even m.

If the aperture opens only a part of the central Fresnel zone, a blurred bright spot is obtained on the screen; there is no alternation of bright and dark rings in this case. If the aperture opens a great number of zones, the alternation of the bright and dark rings is observed only in a very narrow region on the boundary of the geometrical shadow; inside this region the illumination is virtually constant.

Diffraction from a Disk. Let us place an opaque disk of radius r_0 between light source S and observation point P (Fig. 18.13). If the disk covers m first Fresnel zones, the amplitude at point P will be

$$A = A_{m+1} - A_{m+2} + A_{m+3} - \ldots = \frac{A_{m+1}}{2} + \left(\frac{A_{m+1}}{2} - A_{m+2} + \frac{A_{m+3}}{2} \right) + \ldots .$$

The expressions in parentheses can be assumed to equal zero, consequently

$$A = \frac{A_{m+1}}{2}. \tag{18.16}$$

Let us determine the nature of the pattern obtained on the screen (see Fig. 18.13). It is obvious that the illumination can depend only on the distance r from point P. With a small number of covered zones, the amplitude A_{m+1} differs slightly from A_1. The intensity at point P will therefore be almost the same as in the absence of a barrier between source S and point P [see Eq. (18.11)]. For point P′, displaced relative to point P in any radial direction, the disk will cover a part of the $(m+1)$-th Fresnel zone and part of the m-th zone will be opened simultaneously. This will cause the intensity to diminish. At a certain position of point P′, the intensity will reach its minimum. If the distance from the centre of the pattern is still greater, the disk will cover additionally a part of the $(m+2)$-th zone, and a part of the $(m-1)$-th zone will be opened simultaneously. As a result, the intensity grows and reaches a maximum at point P″.

Thus, the diffraction pattern for an opaque disk has the form of alternating bright and dark concentric rings. The centre of the pattern contains a bright spot (Fig. 18.14). The light intensity I varies with the distance r from point P as shown in

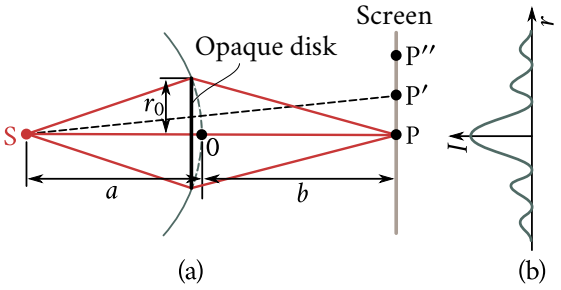


Fig. 18.13: (a) Opaque disk of radius r_0 between light source S and observation point P. (b) Variation of the light intensity with the distance r from point P.

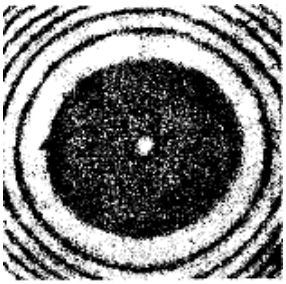


Fig. 18.14: Diffraction pattern for an opaque disk alternating bright and dark concentric rings.

Fig. 18.13b.

If the disk covers only a small part of Fig. 18.14 the central Fresnel zone, it does not form a shadow at all—the illumination of the screen everywhere is the same as in the absence of barriers. If the disk covers many Fresnel zones, alternation of the bright and dark rings is observed only in a narrow region on the boundary of the geometrical shadow. In this case, $A_{m+1} \ll A_1$, so that the bright spot at the centre is absent, and the illumination in the region of the geometrical shadow equals zero practically everywhere.

The bright spot at the centre of the shadow formed by a disk was the cause of an incident between Simeon Poisson and Augustin Fresnel. The Paris Academy of Sciences announced the diffraction of light as the topic for its 1818 prize. The organizers of the competition were advocates of the corpusculate theory of light and were sure that the papers submitted to the competition would bring a final victory to their theory. Fresnel submitted a paper, however, in which all the optical phenomena known at that time were explained from the wave viewpoint. In considering this paper, Poisson, who was a member of the competition committee, gave attention to the fact that an "absurd" conclusion follows from Frenel's theory: there must be a bright spot at the centre of the shadow formed by a small disk. D. Arago immediately conducted an experiment and found that such a spot does indeed exist. This brought victory and all-round recognition to the wave theory of light.

Diffraction from the Straight Edge of a Half-Plane. Let us put an opaque half-plane with a straight edge in the path of a light wave (which we shall consider to be plane for simplicity). We shall arrange this half-plane so that it coincides with one of the wave surfaces. We shall place a screen parallel to the half plane at a distance b behind it and take point P on the screen (Fig. 18.15). Let us divide the open part of the wave surface into zones having the form of very narrow straight

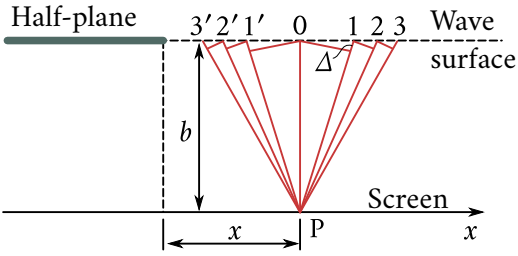


Fig. 18.15: Opaque half-plane with a straight edge in the path of a light wave, arranged to coincide with one of the wave surfaces. A screen parallel to the half-plane is at a distance b behind it.

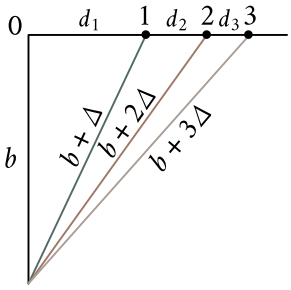


Fig. 18.16: Picture that helps to establish the dependence of the amplitude on the zone number m.

strips parallel to the edge of the half-plane. We shall choose the width of the zones so that the distances from point P to the edges of any zone measured in the plane of the drawing differ by the same amount Δ. When this condition is observed, the oscillations set up at point P by the adjacent zones will differ in phase by a constant value.

We shall assign the numbers 1, 2, 3, etc. to the zones at the right of point P, and the numbers $1'$, $2'$, $3'$, etc. to those at the left of this point. The zones numbered m and m' have an identical width and are symmetrical relative to point P. Therefore, the oscillations produced by them at P coincide in amplitude and in phase.

To establish the dependence of the amplitude on the zone number m, let us assess the areas of the zones. A glance at Fig. 18.16 shows that the total width of the first m zones is

$$d_1 + d_2 + \ldots + d_m = \sqrt{(b + m\Delta)^2 - b^2} = \sqrt{2bm\Delta + m^2\Delta^2}.$$

Since the zones are narrow, we have $\Delta \ll b$. Consequently, when m is not very great, we may ignore the quadratic term in the radicand. This yields

$$d_1 + d_2 + \ldots + d_m = \sqrt{2bm\Delta}.$$

Assuming in this equation that $m = 1$, we find that $d_1 = \sqrt{2b\Delta}$. Hence, we can write the expression for the total width of the first m zones as follows

$$d_1 + d_2 + \ldots + d_m = \sqrt{m},$$

whence

$$d_m = d_1 \left(\sqrt{m} - \sqrt{m-1}\right). \tag{18.17}$$

Calculations by Eq. (18.17) show that

$$d_1 : d_2 : d_3 : d_4 : \ldots = 1 : 0.41 : 0.32 : 0.27 : \ldots. \tag{18.18}$$

The areas of the zones are in the same proportion. Examination of Eq. (18.18) shows

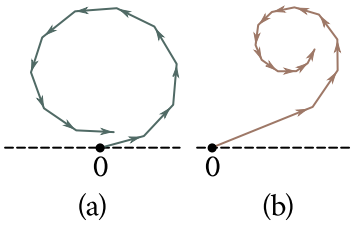


Fig. 18.17: Approximate vector diagrams showing the graphical addition of the oscillations produced by straight lines. The amplitudes are constant in (a) and variable in accordance to Eq. (18.18) in (b).

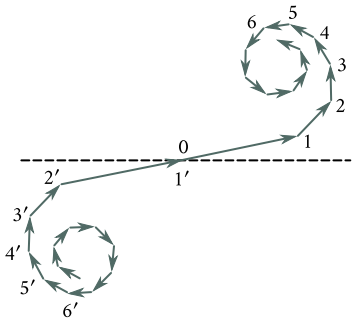


Fig. 18.18: Complete diagram vectors depicting the oscillations corresponding to these zones symmetrically relative to the origin of coordinates 0.

that the amplitude of the oscillations set up at point P by the individual zones initially (for the first zones) diminishes very rapidly, and then this diminishing becomes slower. For this reason, the broken line obtained in the graphical addition of the oscillations produced by the straight zones first has a gentler slope than that for annular zones (the areas of which in a similar construction are approximately equal). Both vector diagrams are compared in Fig. 18.17. In both cases, the lag in phase of each following oscillation has been taken the same. The value of the amplitude for the annular zones (Fig. 18.17a) has been taken constant, and for the straight zones (Fig. 18.17b), diminishing in accordance with proportion (Fig. 18.18). The graphs in Fig. 18.17 are approximate. In an exact construction of these graphs, account must be taken of the dependence of the amplitude on r and φ [see Eq. (18.1)]. This does not affect the general nature of the diagrams, however.

Figure Fig. 18.17b shows only the oscillations produced by the zones to the right of point P. The zones numbered m and m' are symmetrical relative to P. It is therefore natural, when constructing the diagram, to arrange the vectors depicting the oscillations corresponding to these zones symmetrically relative to the origin of coordinates 0 (Fig. 18.18). If the width of the zones is made to tend to zero, the broken line shown in Fig. 18.18 will transform into a smooth curve (Fig. 18.19) called a **Cornu spiral**.

The equation of a Cornu spiral in the parametric form is

$$\xi = \int_0^v \cos\left(\frac{\pi u^2}{2}\right) \mathrm{d}u, \quad \eta = \int_0^v \sin\left(\frac{\pi u^2}{2}\right) \mathrm{d}u. \tag{18.19}$$

These integrals are known as Fresnel integrals. They can be solved only numerically, and tables are available that can be used to find the values of the integrals for various

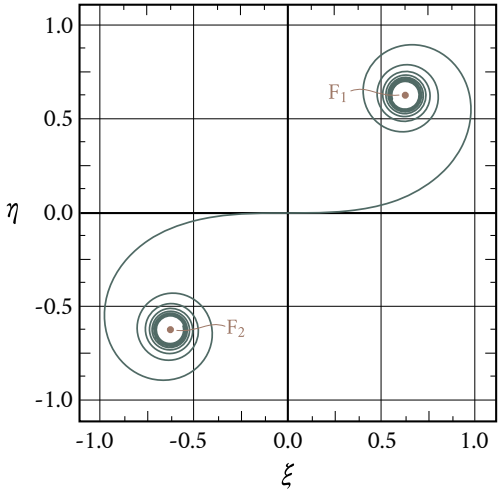


Fig. 18.19: Cornu spiral. When the width of the zones is made tend to zero, the broken line shown in Fig. 18.18 transforms into a smooth curve. This makes it possible to find the amplitude of a light oscillation at any point on a screen.

v's. The meaning of the parameter v is that $|v|$ gives the length of the arc of the Cornu spiral measured from the origin of coordinates.

The figures along the curve in Fig. 18.19 give the values of the parameter v. Points F_1 and F_2, which are asymptotically approached by the curve when v tends to $+\infty$ and $-\infty$, are called the **focal points** or **poles** of the Cornu spiral. Their coordinates are

$$\xi = +\frac{1}{2}, \quad \eta = +\frac{1}{2}, \quad \text{for point } F_1,$$
$$\xi = -\frac{1}{2}, \quad \eta = -\frac{1}{2}, \quad \text{for point } F_2.$$

The right-hand curl of the spiral (section $0F_1$) corresponds to zones to the right of point P, and the left-hand curl (section $0F_2$) to zones to the left of this point.

Let us find the derivative $d\eta/d\xi$ for the point of the curve corresponding to a given value of the parameter v. According to Eq. (18.19), the values

$$d\xi = \cos\left(\frac{\pi v^2}{2}\right) dv \qquad d\eta = \sin\left(\frac{\pi v^2}{2}\right) dv,$$

correspond to the increment dx of v. Consequently, $d\eta/d\xi = \tan(\pi v^2/2)$. At the same time, $d\eta/d\xi = \tan\theta$, where θ is the angle of inclination of a tangent to the curve at the given point. Thus,

$$\theta = \frac{\pi}{2} v^2. \tag{18.20}$$

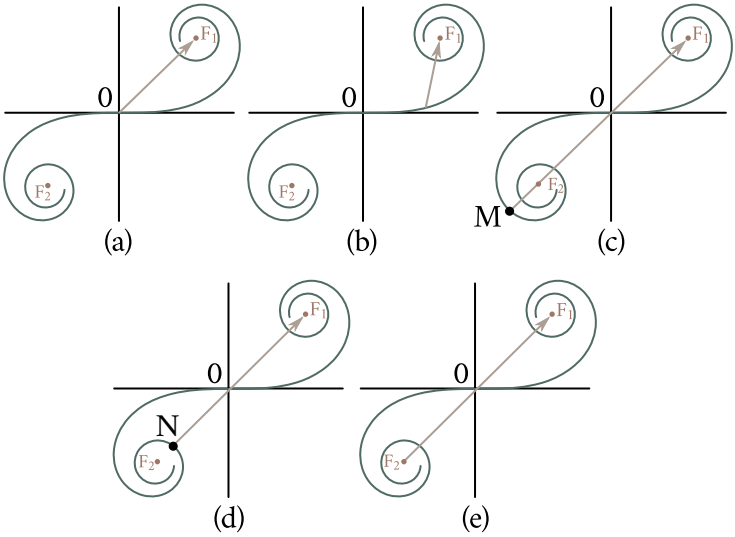


Fig. 18.20: (a) The right-hand curl of the spiral corresponds to oscillations from the unhatched zones depicted by a vector whose origin is at point o and whose end is at point F_1. (b) When point P is displaced to the region of the geometrical shadow, the half-plane covers a greater and greater number of unhatched zones. The beginning of the resultant vector moves along the right-hand curl in the direction of pole F_1. If point P is displaced from the boundary of the geometrical shadow to the right, the tip of the resultant vector slides along the left-hand curl of the spiral in a direction to pole F_2. The amplitude passes through a number of maxima [the first one equals the length of the segment MF_1 (c)] and minima [the first one equals the length of segment NF_1 (d)].

It thus follows that at the point corresponding to $v = 1$ a tangent to the Cornu spiral is perpendicular to the ξ-axis. When $v = 2$, the angle θ is 2π, so that a tangent is parallel to the ξ-axis. When $v = 3$, the angle θ is $9\pi/2$, so that a tangent is again perpendicular to the ξ-axis, and so on.

The Cornu spiral makes it possible to find the amplitude of a light oscillation at any point on a screen. We shall characterize the position of the point by the coordinate x measured from the boundary of the geometrical shadow (see Fig. 18.15). All the hatched zones will be covered for point P on the boundary of the geometrical shadow ($x = 0$). The right-hand curl of the spiral corresponds to oscillations from the unhatched zones. Hence, the resultant oscillation will be depicted by a vector whose origin is at point 0 and whose end is at point F_1 (Fig. 18.20a). When point P is displaced to the region of the geometrical shadow, the half-plane covers a greater and greater number of unhatched zones. Therefore, the beginning of the resultant vector moves along the right-hand curl in the direction of pole F_1 (Fig. 18.20b). As a result, the amplitude of the oscillation monotonously tends to zero.

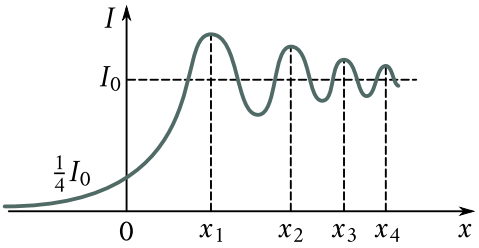


Fig. 18.21: Dependence of light intensity with the coordinate x. Upon a transition to the region of the geometrical shadow, the intensity gradually tends to zero instead of changing in a jump. A number of alternating maxima and minima of the intensity are to the right of the boundary of the geometrical shadow.

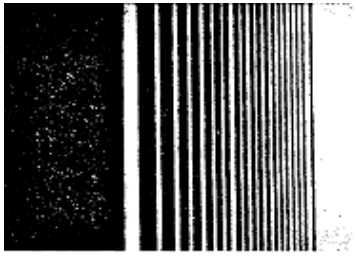


Fig. 18.22: Photograph of the diffraction pattern produced by the edge of a half-plane.

If point P is displaced from the boundary of the geometrical shadow to the right, in addition to the unhatched zones a constantly growing number of hatched ones will be uncovered. Therefore, the tip of the resultant vector slides along the left-hand curl of the spiral in a direction to pole F_2. The amplitude passes through a number of maxima (the first of them equals the length of segment MF_1 in Fig. 18.20c) and minima (the first of them equals the length of segment NF_1 in Fig. 18.20d). When the wave surface is completely uncovered, the amplitude equals the length of F_2F_1 (Fig. 18.20e), *i.e.*, is exactly double the amplitude on the boundary of the geometrical shadow (see Fig. 18.20a). Accordingly, the intensity on the boundary of the geometrical shadow is one-fourth of the intensity I_0 obtained on the screen in the absence of barriers.

The dependence of light intensity I on the coordinate z is shown in Fig. 18.21. Inspection of the figure shows that upon a transition to the region of the geometrical shadow, the intensity gradually tends to zero instead of changing in a jump. A number of alternating maxima and minima of the intensity are to the right of the boundary of the geometrical shadow. Calculations show that at $b = 1$ m and $\lambda = 0.5$ μm the coordinates of the maxima (see Fig. 18.21) have the following values: $x_1 = 0.61$ mm, $x_2 = 1.17$ mm, $x_3 = 1.54$ mm, $x_4 = 1.85$ mm, etc. With a change in the distance b from the half-plane to the screen, the values of the coordinates of the maxima and minima vary as $\sqrt{b}$. It can be seen from the above data that the maxima are quite dense. The Cornu curve can also be used to find the relative value of the intensity at the maxima and minima. We get the value of $1.37I_0$ for the first maximum and $0.78I_0$ for the first minimum.

Figure 18.22 contains a photograph of the diffraction pattern produced by the

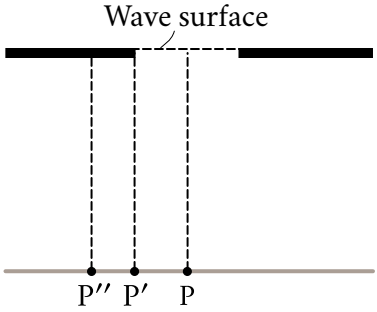


Fig. 18.23: Infinitely long slit formed by placing two half-planes facing opposite directions next to each other.

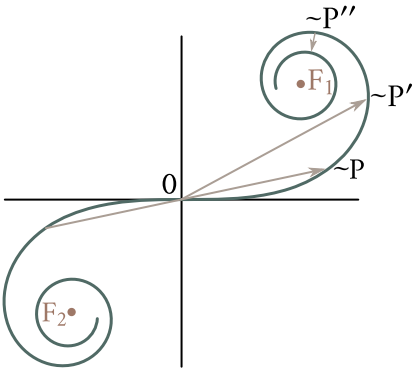


Fig. 18.24: Cornu spiral with the vectors corresponding to the infinitely long slit formed from two-half-planes as in Fig. 18.23.

edge of a half-plane.

Diffraction from a Slit. An infinitely long slit can be formed by placing two half-planes facing opposite directions next to each other. Therefore, the problem on the Fresnel diffraction from a slit can be solved with the aid of a Cornu spiral. We shall consider that the wave surface of the incident light, the plane of the slit, and the screen on which a diffraction pattern is observed are parallel to one another (Fig. 18.23).

For point P opposite the middle of the slit, the tip and the tail of the resultant vector are at points on the spiral that are symmetrical relative to the origin of coordinates (Fig. 18.24). If we move to point P' opposite an edge of the slit, the tip of the resultant vector will move to the middle of the spiral 0. The tail of the vector will move along the spiral in the direction of pole F_1. Upon motion into the region of the geometrical shadow, the tip and the tail of the resultant vector will slide along the spiral and in the long run will be at the smallest distance apart (see the vector in Fig. 18.24 corresponding to point P″). The intensity of the light reaches a minimum here. Upon further sliding along the spiral, the tip and tail of the vector will move apart again, and the intensity will grow. The same will occur when we move from point P in the opposite direction because the diffraction pattern is symmetrical relative to the middle of the slit.

If we change the width of the slit by moving the half-planes in opposite directions, the intensity at middle point P will pulsate, alternately passing through maxima (Fig. 18.25a) and minima that differ from zero (Fig. 18.25b).

Thus, a Fresnel diffraction pattern from a slit is either a bright (for the case shown in Fig. 18.25a) or a relatively dark (for the case shown in Fig. 18.25b) cen-

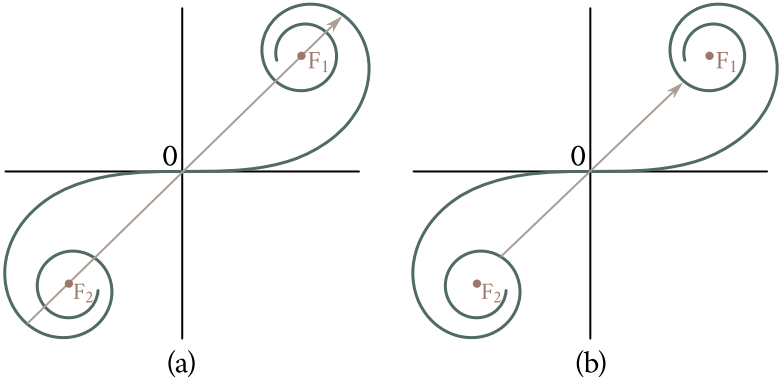


Fig. 18.25: Fresnel diffraction pattern from a slit is either a bright (a) or a relatively dark (b) central fringe at both sides of which there are alternating dark and bright fringes symmetrical relative to it.

tral fringe at both sides of which there are alternating dark and bright fringes symmetrical relative to it.

With a great width of the slit, the tip and tail of the resultant vector for point P are on the internal turns of the spiral near poles F_1 and F_2. Therefore, the intensity of the light at the points opposite the slit will be virtually constant. A system of closely spaced narrow bright and dark fringes is formed only on the boundaries of the geometrical shadow.

We must note that all the results obtained in the present section hold provided that the coherence radius of the light wave falling on the barrier greatly exceeds the characteristic dimension of the barrier (the diameter of the aperture or disk, the width of the slit, etc.).

18.5. Fraunhofer Diffraction from a Slit

Assume that a plane light wave falls on an infinitely long[2] slit (Fig. 18.26). Let us place a converging lens behind the slit and a screen in the focal plane of the lens. The wave surface of the incident wave, the plane of the slit, and the screen are parallel to one another. Since the slit is infinite, the pattern observed in any plane at right angles to the slit will be the same. It is therefore sufficient to investigate the nature of the pattern in one such plane, for example, in that of Fig. 18.26. All the quantities introduced in the following, in particular the angle φ made by a ray with the optical axis of the lens, relate to this plane.

Let us divide the open part of the wave surface into elementary zones of width

[2] In practice, it is sufficient that the length of the slit he many times its width.

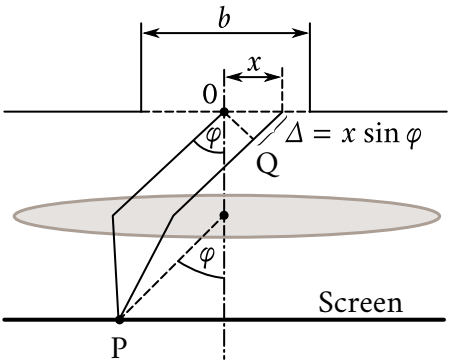


Fig. 18.26: A plane light wave hitting on an infinitely long slit. A converging lens is placed behind the slit and a screen in the focal plane of the lens. The wave surface of the incident wave, the plane of the slit, and the screen are parallel to one another.

dx parallel to the edges of the slit. The secondary waves emitted by the zones in the direction determined by the angle φ will gather at point P of the screen. Each elementary zone will produce the oscillation dE at point P. The lens will gather plane (and not spherical) waves in the focal plane. Therefore, the factor $1/r$ in Eq. (18.1) for dE will be absent for Fraunhofer diffraction. Limiting ourselves to a consideration of not too great angles φ, we can assume that the coefficient K in Eq. (18.1) is constant. Hence, the amplitude of the oscillation produced by a zone at any point of the screen will depend only on the area of the zone. The area is proportional to the width dx of a zone. Consequently, the amplitude dA of the oscillation dE produced by a zone of width dx at any point of the screen will have the form

$$\mathrm{d}A = C\,\mathrm{d}x,$$

where C is a constant.

Let A_0 stand for the algebraic sum of the amplitudes of the oscillations produced by all the zones at a point of the screen. We can find A_0 by integrating dA over the entire width of the slit b:

$$A_0 = \int \mathrm{d}A = \int_{-b/2}^{+b/2} C\,\mathrm{d}x = Cb.$$

Hence, $C = A_0/b$ and, therefore,

$$\mathrm{d}A = \frac{A_0}{b}\,\mathrm{d}x.$$

Now let us find the phase relations between the oscillations dE. We shall compare the phases of the oscillations produced at point P by the elementary zones having the coordinates 0 and x (Fig. 18.26). The optical paths 0P and QP

are tautochronous (see Fig. 16.20). Therefore, the phase difference between the oscillations being considered is formed on the path Δ equal to $x \sin\varphi$. If the initial phase of the oscillation produced at point P by the elementary zone at the middle of the slit ($\varphi = 0$) is assumed to equal zero, then the initial phase of the oscillation produced by the zone with the coordinate x will be

$$-2\pi\frac{\Delta}{\lambda} = -\frac{2\pi}{\lambda}x\sin\varphi,$$

where λ is the wavelength in the given medium.

Thus, the oscillation produced by the elementary zone with the coordinate x at point P (whose position is determined by the angle φ) can be written in the form

$$\mathrm{d}E_\varphi = \left(\frac{A_0}{b}\,\mathrm{d}x\right)\exp\left[i\left(\omega t - \frac{2\pi}{\lambda}x\sin\varphi\right)\right] \tag{18.21}$$

(we have in mind the real part of this expression).

Integrating Eq. (18.21) over the entire width of the slit, we shall find the resultant oscillation produced at point P by the part of the wave surface uncovered by the slit:

$$E_\varphi = \int_{-b/2}^{+b/2}\left(\frac{A_0}{b}\right)\exp\left[i\left(\omega t - \frac{2\pi}{\lambda}x\sin\varphi\right)\right]\mathrm{d}x.$$

Let us put the multipliers not depending on x outside the integral In addition, we shall introduce the symbol

$$\gamma = \frac{\pi}{\lambda}\sin\varphi. \tag{18.22}$$

As a result, we get

$$E_\varphi = \frac{A_0}{b}e^{i\omega t}\int_{-b/2}^{+b/2}e^{-2i\gamma x}\,\mathrm{d}x = \frac{A_0}{b}e^{i\omega t}\left(-\frac{1}{2i\gamma}\right)e^{-2i\gamma x}\Big|_{-b/2}^{+b/2}$$
$$= e^{i\omega t}\left[\frac{A_0}{\gamma b}\left(-\frac{1}{2i}\right)\left(e^{-i\gamma b} - e^{i\gamma b}\right)\right] = e^{i\omega t}\left[\frac{A_0}{\gamma b}\frac{1}{2i}\left(e^{i\gamma b} - e^{-i\gamma b}\right)\right].$$

The expression in brackets determines the complex amplitude $\hat{A}_\varphi$ of the resultant oscillation. Taking into account that the difference between the exponents divided by $2i$ is $\sin(\gamma b)$ (see Sec. 7.3 of Vol. I), we can write

$$\hat{A}_\varphi = A_0\frac{\sin(\gamma b)}{\gamma b} = A_0\frac{\sin[(\pi b\sin\varphi)/\lambda]}{[(\pi b\sin\varphi)/\lambda]} \tag{18.23}$$

[we have introduced the value of γ from Eq. (18.22)].

Equation (18.23) is a real one. Its magnitude is the usual amplitude of the resultant oscillation:

$$A_\varphi = \left|A_0\frac{\sin[(\pi b\sin\varphi)/\lambda]}{[(\pi b\sin\varphi)/\lambda]}\right|. \tag{18.24}$$

For a point opposite the centre of the lens, $\varphi = 0$. Introduction of this value into Eq. (18.24) gives the value A_0 for the amplitude[3]. This result can be obtained in a simpler way. When $\varphi = 0$, the oscillations from all the elementary zones arrive at point P in the same phase. Therefore, the amplitude of the resultant oscillation equals the algebraic sum of the amplitudes of the oscillations being added.

At values of φ satisfying the condition $(\pi b \sin\varphi)/\lambda = \pm k\pi$, *i.e.*, when

$$b \sin\varphi = \pm k\lambda \quad (k = 1, 2, 3, \ldots), \tag{18.25}$$

the amplitude A_φ vanishes. Thus, condition (18.25) determines the positions of the minima of intensity. We must note that $b \sin\varphi$ is the path difference Δ of the rays travelling to point P from the edges of the slit (see Fig. 18.26).

It is easy to obtain condition (18.25) from the following considerations. If the path difference Δ from the edges of the slit is $\pm k\lambda$, the uncovered part of the wave surface can be divided into $2k$ zones equal in width, and the path difference from the edges of each zone will be $\lambda/2$ (see Fig. 18.27 for $k = 2$). The oscillations from each pair of adjacent zones mutually destroy each other, so that the resultant amplitude vanishes. If the path difference Δ for point P is $+(k + 1/2)\lambda$, the number of zones will be odd, the action of one of them will not be compensated, and a maximum of intensity is observed.

The intensity of light is proportional to the square of the amplitude. Hence, in accordance with Eq. (18.24),

$$I_\varphi = I_0 \frac{\sin^2[(\pi b \sin\varphi)\lambda]}{[(\pi b \sin\varphi)\lambda]^2}, \tag{18.26}$$

where I_0 is the intensity at the middle of the diffraction pattern (opposite the centre of the lens), and I_φ is the intensity at the point whose position is determined by the given value of φ.

We find from Eq. (18.26) that $I_{-\varphi} = I_\varphi$. This signifies that the diffraction pattern is symmetrical relative to the centre of the lens. We must note that when the slit is displaced parallel to the screen (along the x-axis in Fig. 18.26), the diffraction pattern observed on the screen remains stationary (its middle is opposite the centre of the lens). Conversely, displacement of the lens with the slit stationary is attended by the same displacement of the pattern on the screen.

A graph of function (18.26) is depicted in Fig. 18.28. The values of $\sin\varphi$ are laid off along the axis of abscissas, and the intensity I_φ along the axis of ordinates. The number of intensity minima is determined by the ratio of the width of a slit b to the wavelength λ. It can be seen from condition (18.25) that $\sin\varphi = \pm k\lambda/b$. The

[3]We remind our reader that $\lim_{u\to 0} \sin u/u = 1$ (at small values of u we may assume that $\sin \sin u \approx u$).

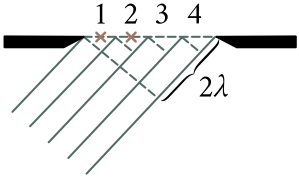


Fig. 18.27: Destruction and construction of oscillations for a path difference Δ from the edges of the slit equal to $\pm k\lambda$ and to $+(k + 1/2)\lambda$, respectively, with $k = 2$.

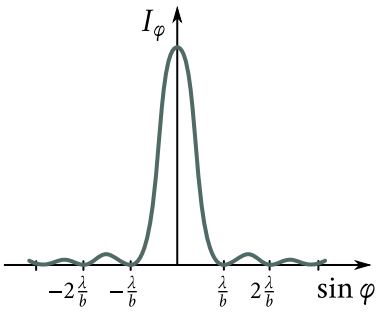


Fig. 18.28: Graph of function (18.26). The number of intensity minima is determined by the ratio of the width of a slit b to the wavelength λ.

magnitude of $\sin\varphi$ cannot exceed unity. Hence, $k\lambda/b \geqslant 1$, whence

$$k \geqslant \frac{b}{\lambda}. \tag{18.27}$$

At a slit width less than a wavelength, minima do not appear at all. In this case, the intensity of the light monotonously diminishes from the middle of the pattern toward its edges.

The values of the angle φ obtained from the condition $b\sin\varphi = \pm\lambda$ correspond to the edges of the central maximum. These values are $\pm\arcsin(\lambda/b)$. Consequently, the angular width of the central maximum is

$$\delta\varphi = 2\arcsin\left(\frac{\lambda}{b}\right). \tag{18.28}$$

When $b \gg \lambda$, the value of $\sin(\lambda/b)$ can be assumed equal to λ/b. The equation for the angular width of the central maximum is thus simplified as follows:

$$\delta\varphi = \frac{2\lambda}{\varphi}. \tag{18.29}$$

Let us solve the problem on the Fraunhofer diffraction from a slit by the method of graphical summation of the amplitudes. We divide the open part of the wave surface into very narrow zones of an identical width. The oscillation produced by each of these zones has the same amplitude ΔA and lags in phase behind the preceding oscillation by the same value δ that depends on the angle φ determining the direction to the point of observation P. When $\varphi = 0$, the phase difference δ vanishes, and the vector diagram has the form shown in Fig. 18.29a. The amplitude of the resultant oscillation A_0 equals the algebraic sum of the amplitudes of the oscillations being added. When $\Delta = b\sin\varphi = \lambda/2$, the oscillations from edges of the slit are in counterphase. Accordingly, the vectors $\Delta\boldsymbol{A}$ arrange themselves along a

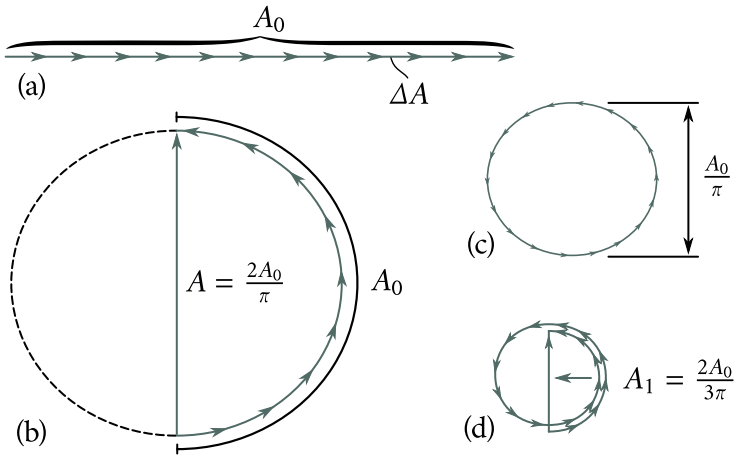


Fig. 18.29: Solution of the problem on the Fraunhofer diffraction from a slit by the method of graphical summation of the amplitudes. The open part of the wave surface is divided into very narrow zones of an identical width. The oscillation produced by each of these zones has the same amplitude ΔA and lags in phase behind the preceding oscillation by the same value δ that depends on the angle φ determining the direction to the point of observation P. (a) Vector diagram with $\varphi = 0, \delta = 0$. (b) When $\Delta = b \sin\varphi = \lambda/2$, the vectors $\Delta \boldsymbol{A}$ form a semicircle of length A_0. (c) When $\Delta = b \sin\varphi = \lambda$, the vectors $\Delta \boldsymbol{A}$ arrange themselves along a semicircle of length A_0, but with phase difference equal o 2π. (d) Constructing sequentially the vectors $\Delta \boldsymbol{A}$, when $\Delta = b \sin\varphi = 3\lambda/2$, we travel one and a half times around a circle of diameter $A_1 = 2A_0/3\pi$, which is the amplitude of the first maximum.

semicircle of length A_0 (Fig. 18.29b). Hence, the resultant amplitude is $2A_0/\pi$. When $\Delta = b \sin\varphi = \lambda$, the oscillations from the edges of the slit differ in phase by 2π. The corresponding vector diagram is shown in Fig. 18.29c. The vectors $\Delta \boldsymbol{A}$ arrange themselves along a circle of length A_0. The resultant amplitude is zero—the first minimum is obtained. The first maximum is obtained at $\Delta = b \sin\varphi = 3\lambda/2$. In this case, the oscillations from the edges of the slit differ in phase by 3π. Constructing sequentially the vectors $\Delta \boldsymbol{A}$, we travel one and a half times around a circle of diameter $A_1 = 2A_0/3\pi$ (Fig. 18.29d). It is exactly the diameter of this circle that is the amplitude of the first maximum. Thus, the intensity of the first maximum is $I_1 = (2/3\pi)^2 I_0 \approx 0.045 I_0$. We can find the relative intensity of the other maxima in a similar way. As a result, we get the following proportion:

$$I_0 : I_1 : I_2 : I_3 : \ldots = \left(\frac{2}{3\pi}\right)^2 : \left(\frac{2}{5\pi}\right)^2 : \left(\frac{2}{7\pi}\right)^2 : \ldots . \tag{18.30}$$

Thus, the central maximum considerably exceeds the remaining maxima in intensity; the main fraction of the light flux passing through the slit is concentrated in it.

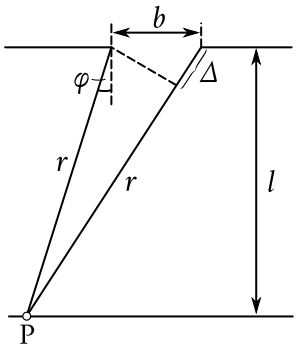


Fig. 18.30: Path difference of the rays from the edges of the slit to point P, to determine the kind of diffraction that will occur in each particular case.

When the width of the slit is very small in comparison with the distance from it to the screen, the rays travelling to point P from the edges of the slit will be virtually parallel even in the absence of a lens between the slit and the screen. Consequently, when a plane wave falls on a slit, Fraunhofer diffraction will be observed. All the equations obtained above will hold; by φ in them one should understand the angle between the direction from any edge of the slit to point P and a normal to the plane of the slit.

Let us establish a quantitative criterion permitting us to determine the kind of diffraction that will occur in each particular case. We shall find the path difference of the rays from the edges of the slit to point P (Fig. 18.30). We apply the cosine law to the triangle with the legs r, $r + \Delta$, and b:

$$(r + \Delta)^2 = r^2 + b^2 - 2rb \cos\left(\frac{\pi}{2} + \varphi\right).$$

Simple transformations yield

$$2r\Delta + \Delta^2 = b^2 + 2rb \sin\varphi. \tag{18.31}$$

We are interested in the case when the rays travelling from the edges of the slit to point P are almost parallel. When this condition is observed, $\Delta^2 \ll r\Delta$, and we can therefore ignore the addend Δ^2 in Eq. (18.31). In this approximation

$$\Delta = \frac{b^2}{2r} + b \sin\varphi. \tag{18.32}$$

In the limit at $r \to \infty$, we get a value of the path difference $\Delta_\infty = b \sin\varphi$ that coincides with the expression in Eq. (18.25).

At finite r's, the nature of the diffraction pattern will be determined by the relation between the difference $\Delta - \Delta_\infty$ and the wavelength λ. If

$$\Delta - \Delta_\infty = \ll \lambda, \tag{18.33}$$

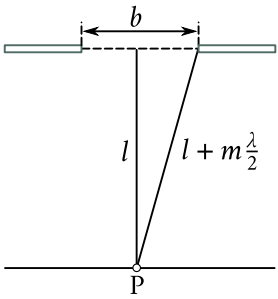


Fig. 18.31: Diagram to make a visual interpretation of the parameter (18.35). m is the number of Fresnel zone, λ the wavelength, b is the width of the slit and l the distance from the middle of the slit to the point P.

the diffraction pattern will be virtually the same as in Fraunhofer diffraction. At $\Delta - \Delta_\infty$ comparable with λ (*i.e.*, $\Delta - \Delta_\infty \sim \lambda$), Fresnel diffraction will take place. It follows from Eq. (18.32) that

$$\Delta - \Delta_\infty = \frac{b^2}{2r} \sim \frac{b^2}{l}$$

(here, l is the distance from the slit to the screen). Introduction of this expression into inequality (18.33) gives the condition $(b^2/l) \ll \lambda$ or

$$\frac{b^2}{l\lambda} \ll 1. \tag{18.34}$$

Thus, the nature of diffraction depends on the value of the dimensionless parameter

$$\frac{b^2}{l\lambda}. \tag{18.35}$$

If this parameter is much smaller than unity, Fraunhofer diffraction is observed, if it is of the order of unity, Fresnel diffraction is observed, and, finally, if this parameter is much greater than unity, the approximation of geometrical optics is applicable. For convenience of comparison, let us write what has been said above in the following form:

$$\frac{b^2}{l\lambda} \begin{cases} \ll 1 & \Rightarrow \text{ Fraunhofer diffraction,} \\ \sim 1 & \Rightarrow \text{ Fresnel diffraction,} \\ \gg 1 & \Rightarrow \text{ geometrical optics.} \end{cases} \tag{18.36}$$

Parameter (18.35) can be given a visual interpretation. Let us take point P opposite the middle of a slit (Fig. 18.31). For this point, the number m of Fresnel zones opened by the slit is determined by the expression

$$\left(l + m\frac{\lambda}{2}\right)^2 = l^2 + \left(\frac{b}{2}\right)^2.$$

Opening the parentheses and discarding the addend proportional to λ^2, we get[4]

$$m = \frac{b^2}{4l\lambda} \sim \frac{b^2}{l\lambda}. \tag{18.37}$$

Thus, parameter (18.35) is directly associated with the number of uncovered Fresnel zones (for a point opposite the middle of the slit).

If a slit opens a small fraction of the central Fresnel zone ($m \ll 1$), Fraunhofer diffraction is observed. The distribution of the intensity in this case is shown by the curve depicted in Fig. 18.28. If a slit uncovers a small number of Fresnel zones ($m \sim 1$), an image of the slit surrounded along its edges by clearly visible bright and dark fringes will be obtained on the screen. Finally, when a slit opens a large number of Fresnel zones ($m \gg 1$), a uniformly illuminated image of the slit is obtained on the screen. Only at the boundaries of the geometrical shadow are there very narrow alternating brighter and darker fringes virtually indistinguishable by the eye.

Let us see how the pattern changes when the screen is moved away from the slit. When the screen is near the slit ($m \gg 1$), the image corresponds to the laws of geometrical optics. Upon increasing the distance, we first obtain a Fresnel diffraction pattern which then transforms into a Fraunhofer pattern. The same sequence of changes is observed when we reduce the width of the slit b without changing the distance l.

It is clear from what has been said above that the value of the parameter (18.35) is the criterion of the applicability of geometrical optics (it must be much greater than unity) instead of the smallness of the wavelength in comparison with the characteristic dimension of the barrier (for example, the width of the slit). Assume, for instance, that both ratios b/λ and l/b equal 100. In this case, $\lambda \ll b$, but $b^2/(l\lambda) = 1$, and, therefore, a distinctly expressed Fresnel diffraction pattern will be observed.

18.6. Diffraction Grating

A diffraction grating is a collection of a large number of identical equispaced slits (Fig. 18.32). The distance d between the centres of adjacent slits is called the **period** of the grating.

Let us place a converging lens parallel to a grating and put a screen in the focal plane of the lens. We shall determine the nature of the diffraction pattern obtained on the screen when a plane light wave falls on the grating (we shall consider for

[4]We must note that the number of open zones will be larger for points greatly displaced to the region of the geometrical shadow.

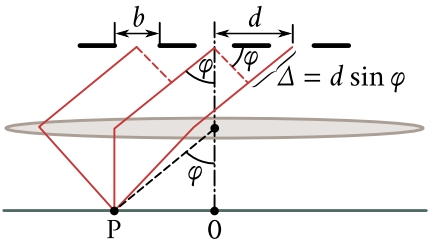


Fig. 18.32: Scheme of a diffraction grating with period d and slit width b. A converging lens is parallel to it focus a normal incident light and a screen in the focal plane of the lens serves to check the diffraction pattern similar to that of Fig. 18.28.

simplicity's sake that the wave falls normally on the grating). Each slit produces a pattern on the screen that is described by the curve depicted in Fig. 18.28. The patterns from all the slits will be at the same place on the screen (regardless of the position of the slit, the central maximum is opposite the centre of the lens). If the oscillations arriving at point P from different slits were incoherent, the resultant pattern produced by N slits would differ from the pattern produced by a single slit only in that all the intensities would grow N times. The oscillations from different slits are coherent to a greater or smaller extent, however. The resultant intensity will therefore differ from NI_φ [I_φ is the intensity produced by one slit; see Eq. (18.26)].

We shall assume in the following that the coherence radius of the incident wave is much greater than the length of the grating so that the oscillations from all the slits can be considered coherent relative to one another. In this case, the resultant oscillation at point P whose position is determined by the angle φ is the sum of N oscillations having the same amplitude A_φ shifted relative to one another in phase by the same amount δ. According to Eq. (17.47), the intensity in these conditions is

$$I_{\mathrm{gr}} = I_\varphi \frac{\sin^2(N\delta/2)}{\sin^2(\delta/2)} \tag{18.38}$$

(here I_φ plays the part of I_0).

A glance at Fig. 18.32 shows that the path difference from adjacent slits is $\delta = d \sin\varphi$. Hence, the phase difference is

$$\delta = 2\pi\frac{\Delta}{\lambda} = \frac{2\pi}{\lambda} d \sin\varphi, \tag{18.39}$$

where λ is the wavelength in the given medium.

Introducing into Eq. (18.38) Eqs. (18.26) and (18.39) for I_φ and δ, respectively, we get

$$I_{\mathrm{gr}} = I_0 \frac{\sin^2(\pi b \sin\varphi/\lambda)}{(\pi b \sin\varphi/\lambda)^2} \frac{\sin^2(N\pi d \sin\varphi/\lambda)}{sin^2(\pi d \sin\varphi/\lambda)} \tag{18.40}$$

(I_0 is the intensity produced by one slit opposite the centre of the lens).

The first multiplier of I_0 in Eq. (18.40) vanishes condition (18.25) is observed, *i.e.*,

$$b \sin \varphi = \pm k\lambda \quad (k = 1, 2, 3, \ldots).$$

At these points, the intensity produced by each slit individually equals zero.

The second multiplier of I_0 in Eq. (18.40) acquires the value N^2 for points satisfying the condition

$$d \sin \varphi = \pm m\lambda \quad (m = 0, 1, 2, \ldots) \tag{18.41}$$

[see Eq. (17.49)]. For the directions determined by this condition, the oscillations from individual slits mutually amplify one another. As a result, the amplitude of the oscillations at the corresponding point of the screen is

$$A_{\max} = NA_\varphi \tag{18.42}$$

(A_φ is the amplitude of the oscillation emitted by one slit at the angle φ).

Condition (18.41) determines the positions of the intensity maxima called the **principal** ones. The number m gives the order of the principal maximum. There is only one zero-order maximum, and there are two each of the maxima of the 1st, 2nd, etc. orders.

Squaring Eq. (18.42), we find that the intensity of the principal maxima $I_{\max}$ is N^2 times greater than the intensity I_φ, produced in the direction φ by a single slit:

$$I_{\max} = N^2 I_\varphi. \tag{18.43}$$

Apart from the minima determined by condition (18.25), there are $N - 1$ additional minima in each interval between adjacent principal maxima. These minima appear in the directions for which the oscillations from individual slits mutually destroy one another. In accordance with Eq. (17.50), the directions of the additional minima are determined by the condition

$$d \sin \varphi = \pm \frac{k'}{N}\lambda (k' = 1, 2, \ldots, N - 1, N + 1, \ldots, 2N - 1, 2N + 1, \ldots). \tag{18.44}$$

In Eq. (18.44), k' takes on all integral values except for $0, N, 2N, \ldots$, *i.e.*, except for those at which Eq. (18.44) transforms into Eq. (18.41).

It is easy to obtain condition (18.44) by the method of graphical addition of oscillations. The oscillations from the individual slits are depicted by vectors of the same length. According to Eq. (18.44), each of the following vectors is turned relative to the preceding one by the same angle

$$\delta = \frac{2\pi}{\lambda} d \sin \varphi = \frac{2\pi}{\lambda} k'.$$

Therefore, when k' is not an integral multiple of N, we put the tip of the following vector against the tail of the preceding one and obtain a closed broken line that completes k' (when $k' < N/2$) or $N - k'$ (when $k' > N/2$) revolutions before the tail of the N-th vector contacts the tip of the first one. The resultant amplitude

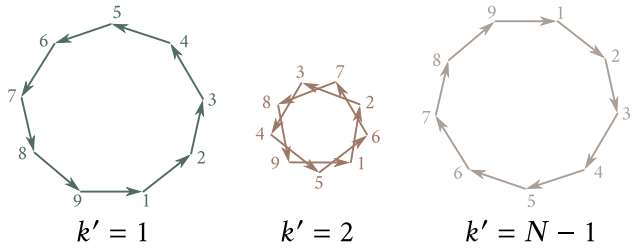


Fig. 18.33: Method of graphical addition of oscillations to arrive at Eq. (18.44). Sum of the vectors for $N = 9$ and for the values of k' equal to 1, 2, and $N - 1 = 8$.

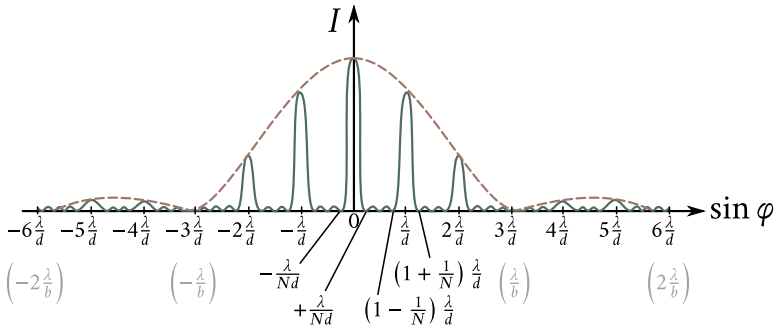


Fig. 18.34: Graph of Eq. (18.40) for $N = 4$ and $d/b = 3$. The dash curve shows the intensity produced by one slit multiplied by N^2 [Eq. (18.43)].

accordingly equals zero. The above is explained in Fig. 18.33 that shows the sum of the vectors for $N = 9$ and for the values of k' equal to 1, 2, and $N - 1 = 8$.

Between the additional minima, there are weak secondary maxima. The number of such maxima falling to an interval between adjacent principal maxima is $N - 2$. We showed in Sec. 17.6 that the intensity of the secondary maxima does not exceed 1/22nd of that of the closest principal maximum.

Figure 18.34 shows a graph of function (18.40) for $N = 4$ and $d/b = 3$. The dash curve passing through the peaks of the principal maxima shows the intensity produced by one slit multiplied by N^2 [see Eq. (18.43)]. At the ratio of the grating period to the slit width used in the figure ($d/b = 3$), the principal maxima of the third, sixth, etc. orders fall to the minima of intensity from one slit, owing to which these maxima vanish. In general, it can be seen from Eqs. (18.25) and (18.41) that the principal maximum of the m-th order falls to the k-th minimum from one slit if the equation $m/d = k/b$ or $m/k = d/b$ is satisfied. This is possible if d/b equals the ratio of two integers r and s (the ease when these integers are not great is of practical interest). Here, the principal maximum of the r-th order will be superposed on the s-th minimum from one slit, the maximum of the $2r$-th order will be superposed

on the $2s$-th minimum, etc. As a result, the maxima of orders r, $2r$, $3r$, etc. will be absent.

The number of principal maxima observed is determined by the ratio of the period of the grating d to the wavelength λ. The magnitude of $\sin\varphi$ cannot exceed unity. It therefore follows from Eq. (18.41) that

$$m \leqslant \frac{d}{\lambda}. \tag{18.45}$$

Let us determine the angular width of the central (zero) maximum. The position of the additional minima closest to it is determined by the condition $d\sin\varphi = \pm\lambda N$ [see Eq. (18.44)]. Hence, values of φ equal to $\pm\arcsin(\lambda/Nd)$ correspond to these minima. We thus obtain the following expression for the angular width of the central maximum:

$$\delta\varphi_0 = 2\arcsin\left(\frac{\lambda}{Nd}\right) \approx \frac{2\lambda}{Nd} \tag{18.46}$$

(we have taken advantage of the circumstance that $\lambda/ND \ll 1$).

The position of the additional minima closest to the principal maximum of the m-th order is determined by the condition $d\sin\varphi = (m \pm 1/N)\lambda$. Hence, for the angular width of the m-th maximum, we get the expression

$$\delta\varphi_m = 2\arcsin\left(m + \frac{1}{N}\right)\frac{\lambda}{d} - \arcsin\left(m - \frac{1}{N}\right)\frac{\lambda}{d}.$$

Introducing the notation $m\lambda/d = x$ and $\lambda/Nd = \Delta x$, we can write this equation in the form

$$\delta\varphi_m = \arcsin(x + \Delta x) - \arcsin(x - \Delta x). \tag{18.47}$$

With a great number of slits, the value of $\Delta x = \lambda/Nd$ will be very small. We can therefore assume that $\arcsin(x \pm \Delta x) \sim \arcsin x \pm (\arcsin x)'\Delta x$. The introduction of these values into Eq. (18.47) leads to the approximate expression

$$\delta\varphi_m = 2(\arcsin x)'\Delta x = \frac{2\Delta x}{\sqrt{1 - x^2}} = \frac{1}{\sqrt{1 - m^2(\lambda/d^2)^2}}\frac{\lambda}{Nd}. \tag{18.48}$$

When $m = 0$, this expression transforms into Eq. (18.46).

The product Nd gives the length of the diffraction grating. Consequently, the angular width of the principal maxima is inversely proportional to the length of the grating. The width $\delta\varphi_m$ grows with an increase in the order m of a maximum.

The position of the principal maxima depends on the wavelength λ. Therefore, when white light is passed through a grating, all the maxima except for the central one will expand into a spectrum whose violet end faces the centre of the diffraction pattern, and whose red end faces outward. Thus, a diffraction grating is a spectral instrument. We must note that whereas a glass prism deflects violet rays the greatest,

a diffraction grating, on the contrary, deflects red rays to a greater extent.

Figure 18.35 shows schematically the spectra of different orders produced by a grating when white light is passed through it. At the centre is a narrow zero-order maximum; only its edges are coloured [according to expression (18.46), $\delta\varphi_0$ depends on λ]. At both sides of the central maximum are two first-order spectra, then two second-order spectra, etc. The positions of the red end of the m-th order spectrum and the violet end of the $(m+1)$-th order one are determined by the relations

$$\sin\varphi_{\mathrm{r}} = m\frac{0.76}{d}, \quad \sin\varphi_{\mathrm{v}} = (m+1)\frac{0.40}{d},$$

where d has been taken in micrometres. When the condition is observed that

$$0.76m > 0.40(m+1),$$

the spectra of the m-th and $(m+1)$-th orders partly overlap. The inequality gives $m > 10/9$. Hence, partial overlapping begins from the spectra of the second and third orders (see Fig. 18.35, in which for illustration the spectra of different orders are displaced relative to one another vertically).

The main characteristics of a spectral instrument are its **dispersion** and **resolving power**. The dispersion determines the angular or linear distance between two spectral lines differing in wavelength by one unit (for example by 1 Å). The resolving power determines the minimum difference between wavelengths $\delta\lambda$ at which the two lines corresponding to them are perceived separately in the spectrum.

The **angular dispersion** is defined as the quantity

$$D = \frac{\delta\varphi}{\delta\lambda}, \tag{18.49}$$

where $\delta\varphi$ is the angular distance between spectral lines differing in wavelength by $\delta\lambda$.

To find the angular dispersion of a diffraction grating, let us differentiate condition (18.41) for the principal maximum at the left with respect to φ and at the right with respect to λ. Omitting the minus sign, we get

$$(d\cos\varphi)\delta\varphi = m\,\delta\lambda,$$

whence

$$D = \frac{\delta\varphi}{\delta\lambda} = \frac{m}{d\cos\varphi}. \tag{18.50}$$

Within the range of small angles; $\cos\varphi \approx 1$. We can therefore assume that

$$D \approx \frac{m}{d}. \tag{18.51}$$

It can be seen from expression (18.51) that the angular dispersion is inversely proportional to the grating period d. The higher the order m of a spectrum, the greater is the dispersion.

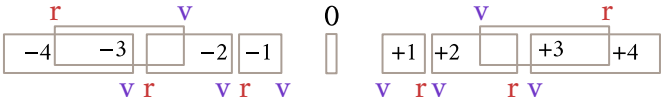


Fig. 18.35: Scheme of the spectra of different orders produced by a grating when white light is passed through it. At the centre there is a narrow zero-order maximum and only its edges are coloured [Eq. (18.46)]. At both sides of the central maximum are two first-order spectra, then two second-order spectra, etc.

Linear dispersion is defined as the quantity

$$D_{\text{lin}} = \frac{\delta l}{\delta \lambda}, \tag{18.52}$$

where δl is the linear distance on a screen or photographic plate between spectral lines differing in wavelength by $\delta\lambda$. A glance at Fig. 18.36 shows that for small values of the angle φ we can assume that $\delta l \approx f' \delta\varphi$, where f' is the focal length of the lens gathering the diffracted rays on a screen. Consequently, the linear dispersion is associated with the angular dispersion D by the relation

$$D_{\text{lin}} = f' D.$$

Taking expression (18.51) into consideration, we get the following equation for the linear dispersion of a diffraction grating (with small φ's):

$$D_{\text{lin}} = f' \frac{m}{d}. \tag{18.53}$$

The resolving power of a spectral instrument is defined as the dimensionless quantity

$$R = \frac{\lambda}{\delta \lambda}, \tag{18.54}$$

where $\delta\lambda$ is the minimum difference between the wavelengths of two spectral lines at which these lines are perceived separately.

The possibility of resolving (*i.e.*, perceiving separately) two close spectral lines depends not only on the distance between them (that is determined by the dispersion of the instrument), but also on the width of the spectral maximum. Figure 18.37 shows the resultant intensity (solid curves) observed in the superposition of two close maxima (the dash curves). In case (a), both maxima are perceived as a single one. In case (b), there is a minimum between the maxima. *Two close maxima are perceived by the eye separately if the intensity in the interval between them is not over* 80% *of the intensity of a maximum.* According to the criterion proposed by the British physicist John Rayleigh (1842-1919), such a ratio of the intensities occurs if the middle of one maximum coincides with the edge of another one (Fig. 18.37b). Such a mutual arrangement of the maxima is obtained at a definite (for the given instrument) value of $\delta\lambda$.

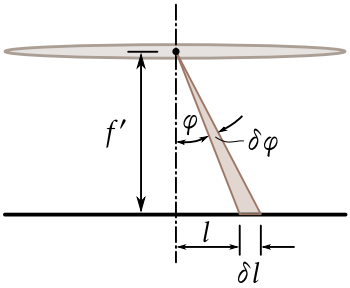


Fig. 18.36: For small values of the angle φ we can assume that $\delta l \approx f'\delta\varphi$, where f' is the focal length of the lens gathering the diffracted rays on a screen.

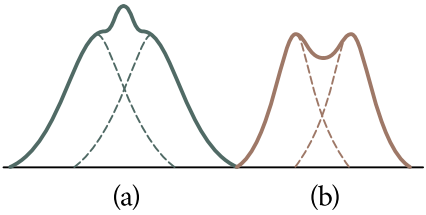


Fig. 18.37: Resultant intensity (solid curves) observed in the superposition of two close maxima (the dash curves). (a) Both maxima are perceived as a single one. (b) There is a minimum between the maxima.

Let us find the resolving power of a diffraction grating. The position of the middle of the m-th maximum for the wavelength $\lambda + \delta\lambda$ is riei;ermined by the condition

$$d \sin \varphi_{\max} = m(\lambda + \delta\lambda).$$

The edges of the m-th maximum for the wavelength λ are at angles complying with the condition

$$d \sin \varphi_{\min} = \left(m \pm \frac{1}{N}\right)\lambda.$$

The middle of the maximum for the wavelength $\lambda + \delta\lambda$ coincides with the edge of the maximum for the wavelength λ if

$$m(\lambda + \delta\lambda) = \left(m \pm \frac{1}{N}\right)\lambda,$$

whence

$$m\delta\lambda = \frac{\lambda}{N}.$$

Solving this equation relative to $\lambda/\delta\lambda$, we get an expression for the resolving power:

$$R = mN. \tag{18.55}$$

Thus, the resolving power of a diffraction grating is proportional to the order m of the spectrum and the number of slits N.

Figure 18.38 compares the diffraction patterns obtained for two spectral lines with the aid of gratings differing in the values of the dispersion D and the resolving power R. Gratings I and II have the same resolving power (they have the same number of slits N), but a different dispersion (in grating I, the period d is double and the dispersion D is half of the respective quantities of grating II). Gratings II

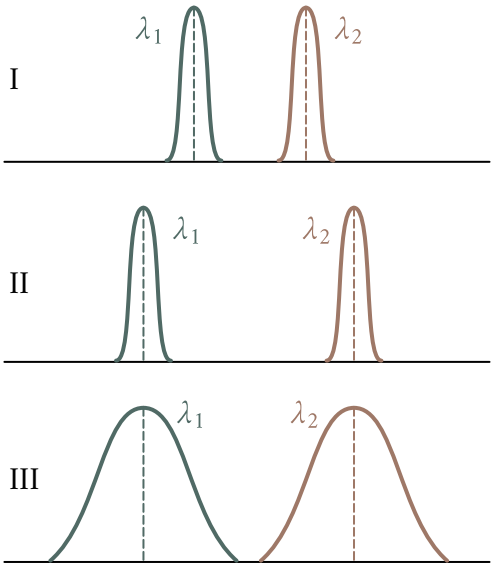


Fig. 18.38: Comparison of diffraction patterns obtained for two spectral lines with the aid of gratings differing in the values of the dispersion D and the resolving power R. Gratings I and II have the same resolving power, but in grating I the period is double and dispersion D is half of that in grating II. Gratings II and III have the same dispersion, but the resolving power of grating II doubles that of grating IIII.

and III have the same dispersion (they have the same d's), but a different resolving power (the number of slits N and the resolving power R of grating II are double the respective quantities of grating III).

Transmission and reflecting diffraction gratings are in use. Transmission gratings are made from glass or quartz plates on whose surface a special machine using a diamond cutter makes a number of parallel lines. The spaces between these lines are the slits.

Reflecting gratings are applied with the aid of a diamond cutter on the surface of a metal mirror. Light falls on a reflecting grating at an acute angle. A grating of period d functions in the same way as a transmission grating with the period $d \cos\theta$, where θ is the angle of incidence of the light, would function with the light falling normally. This makes it possible to observe a spectrum when light is reflected, for example, from a gramophone record having only a few lines (grooves) per millimetre if it is placed so that the angle of incidence is close to $\pi/2$. The American physicist Henry Row land (1848-1901) invented a concave reflecting grating which focuses the diffraction spectra by itself (without a lens).

The best gratings have up to 1200 lines per mm ($d \approx 0.8\,\mu$m). It can be seen

from Eq. (18.45) that no second-order spectra are observed in visible light with such a period. The total number of lines in such gratings reaches 200000 (they are about 200 mm long). With a focal length of the instrument $f' = 2$ m, the length of the visible first-order spectrum in this case is over 700 mm.

18.7. Diffraction of X-Rays

Let us place two diffraction gratings one after the other so that their lines are mutually perpendicular. The first grating (whose lines, say, are vertical) will produce a number of maxima in the horizontal direction. Their positions are determined by the condition

$$d_1 \sin \varphi_1 = \pm m_1 \lambda \quad (m_1 = 0, 1, 2 \ldots). \tag{18.56}$$

The second grating (with horizontal lines) will divide each of the beams formed in this way into vertically arranged maxima whose positions are determined by the condition

$$d_2 \sin \varphi_2 = \pm m_2 \lambda \quad (m_2 = 0, 1, 2 \ldots). \tag{18.57}$$

As a result, the diffraction pattern will have the form of regularly arranged spots, with two integral indices m_1 and m_2 corresponding to each of them (Fig. 18.39).

An identical diffraction pattern is obtained if instead of two separate gratings we take one transparent plate with two systems of mutually perpendicular lines applied on it. Such a plate is a two-dimensional periodic structure (a conventional grating is a one-dimensional structure). Having measured the angles φ_1 and φ_2 determining the positions of the maxima and knowing the wavelength λ, we can use Eqs. (18.56) and (18.57) to find the periods of the structure d_1 and d_2. If the directions in which a structure is periodic (for example, directions at right angles to the grating lines) make the angle α differing from $\pi/2$, the diffraction maxima will be at the apices of parallelograms instead of at the apices of rectangles (as in Fig. 18.39). In this case, the diffraction pattern can be used to determine not only the periods d_1 and d_2, but also the angle α.

Any two-dimensional periodic structures such as a system of small apertures or one of opaque tiny spheres produce a diffraction pattern similar to that shown in Fig. 18.39.

For diffraction maxima to appear, it is essential that the period of the structure d be greater than λ. Otherwise, conditions (18.56) and (18.57) can be satisfied only at values of m_1 and m_2 equal to zero (the magnitude of $\sin \varphi$ cannot exceed unity).

Diffraction is also observed in three-dimensional structures, *i.e.*, spatial formations displaying periodicity along three directions not in one plane. All crystalline bodies are such structures. Their period ($\sim 10^{-10}$ m), however, is too small for the

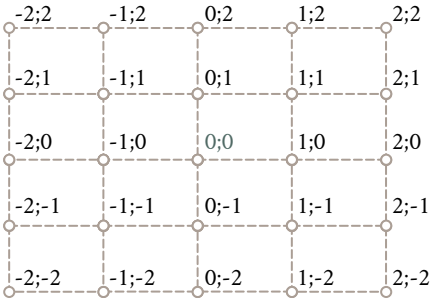


Fig. 18.39: Diffraction pattern with two integral indices m_1 and m_2 corresponding to two diffraction gratings placed one after the other so that their lines are mutually perpendicular.

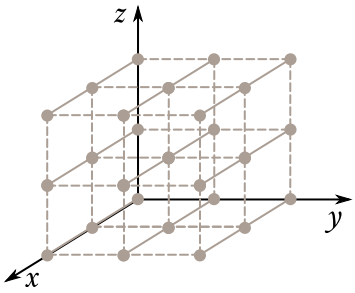


Fig. 18.40: Formation of diffraction maxima from a three-dimensional structure. The coordinate axes x, y and z are positioned in the directions along which the properties of the structure display periodicity.

observation of diffraction in visible light. The condition $d\lambda$ is observed for crystals only for X-rays. The diffraction of X-rays from crystals was first observed in 1913 in an experiment conducted by the German physicists Max von Laue (1879-1959), Walter Friedrich (1883-1968), and Paul Knipping (1883-1935). (The idea belonged to von Laue, while the other two authors ran the experiment.).

Let us find the conditions for the formation of diffraction maxima from a three-dimensional structure. We position the coordinate axes x, y, and z in the directions along which the properties of the structure display periodicity (Fig. 18.40). The structure can be represented as a collection of equally spaced parallel trains of structural elements arranged along one of the coordinate axes. We shall consider the action of an individual linear train parallel, for instance, to the x-axis (Fig. 18.41). Assume that a beam of parallel rays making the angle α_0 with the x-axis falls on the train. Every structural element is a source of secondary wavelets. An incident wave arrives at adjacent sources with a phase difference of $\delta_0 = 2\pi\Delta_0/\lambda$, where $\Delta_0 = d_1 \cos\alpha_0$ (here, d_1 is the period of the structure along the x-axis). Apart from this, the additional path difference $\Delta = d_1 \cos\alpha$ is produced between the secondary wavelets propagating in directions that make the angle α with the x-axis (all such directions lie along the generatrices of a cone whose axis is the x-axis). The oscillations from different structural elements will be mutually amplified for the directions for which

$$d_1(\cos\alpha - \cos\alpha_0) = \pm m_1\lambda \quad (m_1 = 0, 1, 2, \ldots). \tag{18.58}$$

There is a separate cone of directions for each value of m_1, and along these directions we get maxima of the intensity from one individually taken train parallel to the x-axis. The axis of this cone coincides with the x-axis.

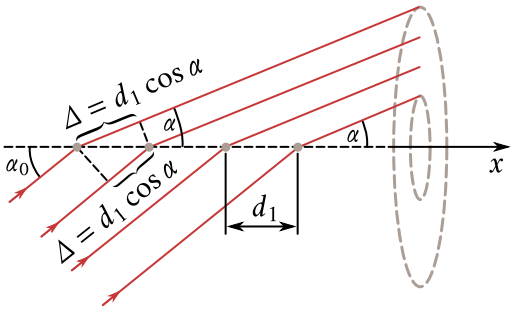


Fig. 18.41: Scheme of a collection of equally spaced parallel trains of structural elements arranged along the x-axis.

The condition of the maximum for a train parallel to the y-axis has the form

$$d_2(\cos\beta - \cos\beta_0) = \pm m_2\lambda \quad (m_2 = 0, 1, 2, \ldots), \tag{18.59}$$

where d_2 is the period of the structure in the direction of the y-axis, β_0 is the angle between the incident beam and the y-axis, and β is the angle between the y-axis and the directions along which diffraction maxima are obtained.

A cone of directions whose axis coincides with the y-axis corresponds to each value of m_2.

In directions satisfying conditions (18.58) and (18.59) simultaneously, mutual amplification of the oscillations from sources in the same plane perpendicular to the z-axis occur (these sources form a two-dimensional structure). The directions of the intensity maxima produced lie along the lines of intersection of the direction cones, of which one is determined by condition (18.58), and the second one by condition (18.59).

Finally, for the train parallel to the z-axis, the directions of the maxima are determined by the condition

$$d_3(\cos\gamma - \cos\gamma_0) = \pm m_3\lambda \quad (m_3 = 0, 1, 2, \ldots), \tag{18.60}$$

where d_3 is the period of the structure in the direction of the z-axis, γ_0 is the angle between the incident beam and the z-axis, and γ is the angle between the z-axis and the directions along which diffraction maxima are obtained.

As in the preceding cases, a cone of directions whose axis coincides with the z-axis corresponds to each value of m_3.

In the directions satisfying conditions (18.58), (18.59), and (18.60) simultaneously, mutual amplification of the oscillations from all the elements forming the three-dimensional structure occurs. As a result, diffraction maxima are produced by the three-dimensional structure. The directions of these maxima are on the lines of intersection of three cones whose axes are parallel to the coordinate axes.

The conditions

$$\begin{aligned} d_1(\cos\alpha - \cos\alpha_0) &= \pm m_1\lambda, \\ d_2(\cos\beta - \cos\beta_0) &= \pm m_2\lambda, \quad (m_i = 0, 1, 2, \ldots) \\ d_3(\cos\gamma - \cos\gamma_0) &= \pm m_3\lambda, \end{aligned} \tag{18.61}$$

which we have found are called **Laue's formulas**. Three integral numbers m_1, m_2, and m_3 correspond to each direction (α, β, γ) determined by these formulas. The greatest value of the magnitude of the difference between cosines is two. Hence, conditions (18.61) can be obeyed with values of the numbers m other than zero only provided that λ does not exceed $2d$.

The angles α, β and γ are not independent. For example, when a Cartesian system of coordinates is used, they are related by the expression

$$\cos^2\alpha + \cos^2\beta + \cos^2\gamma = 1. \tag{18.62}$$

Thus, when α_0, β_0 and γ_0 are given, the angles α, β and γ determining the directions of the maxima can be found by solving a system of four equations. If the number of equations exceeds the number of unknowns, a system of equations can be solved only when definite conditions are observed (only when these conditions are satisfied can the three cones intersect one another along a single line).

The system of Eqs. (18.61) and (18.62) can be solved only for certain quite definite wavelengths (λ can be considered as a fourth unknown whose values obtained from the solution of the system of equations are exactly the wavelengths for which maxima are observed). Generally speaking, only one maximum corresponds to each such value of λ. Several symmetrically arranged maxima may be obtained, however.

If the wavelength is fixed (monochromatic radiation), the system of equations can be made simultaneous by varying the values of α_0, β_0 and γ_0, *i.e.*, by turning the three-dimensional structure relative to the direction of the incident beam.

We have not treated the question of how rays travelling from different structural elements are made to converge to one point on a screen. A lens does this for visible light. A lens cannot be used for X-rays because the refractive index of these rays in all substances is virtually equal to unity. For this reason, the interference of the secondary wavelets is achieved by using very narrow beams of rays producing spots of a very small size on a screen (or a photographic plate) even without a lens.

The Russian scientist Yuri Vulf (1863-1925) and the British physicists William Henry Bragg (1862-1942) and his son William Lawrence Bragg (1890-1971) showed independently of each other that the diffraction pattern from a crystal lattice can be calculated in the following simple way. Let us draw parallel equispaced planes through the points of a crystal lattice (Fig. 18.42). We shall call these planes atomic layers. If the wave falling on the crystal is plane, the envelope of the secondary

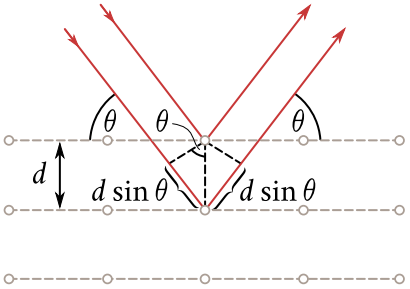


Fig. 18.42: Reflection of a plane wave upon parallel equispaced planes through the points of a crystal lattice (atomic planes). d is the period of identity of the crystal and θ is the angle supplementing the angle of incidence and called the glancing angle of the incident rays.

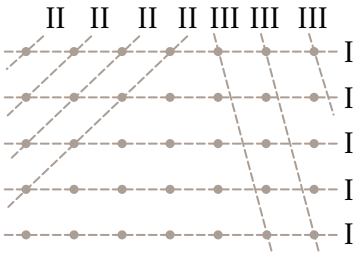


Fig. 18.43: The difference between the paths of two waves reflected from adjacent atomic layers is $2d \sin \theta$. (d and *theta* are described in the caption of Fig. 18.42).

waves set up by the atoms in such a layer will also be a plane. Thus, the summary action of the atoms in one layer can be represented in the form of a plane wave reflected from an atom-covered surface according to the usual law of reflection.

The plane secondary wavelets reflected from different atomic layers are coherent and will interfere with one another like the waves emitted in the given direction by different slits of a diffraction grating. As in the case of a grating, the secondary wavelets will virtually destroy one another in all directions except those for which the path difference between adjacent wavelets is a multiple of λ. Inspection of Fig. 18.42 shows that the difference between the paths of two waves reflected from adjacent atomic layers is $2d \sin \theta$, where d is the period of identity of the crystal in a direction at right angles to the layers being considered, and θ is the angle supplementing the angle of incidence and called the **glancing angle** of the incident rays. Consequently, the directions in which diffraction maxima are obtained are determined by the condition

$$2d \sin \theta = \pm m\lambda \quad (m = 0, 1, 2, \ldots). \tag{18.63}$$

This expression is known as the **Bragg-Vulf** formula.

The atomic layers in a crystal can be drawn in a multitude of ways (Fig. 18.43). Each system of layers can produce a diffraction maximum if condition (18.63) is observed for it. Only those maxima have an appreciable intensity, however, that are obtained as a result of reflections from layers sufficiently densely populated by atoms (for instance, from layers I and II in Fig. 18.43).

We must note that calculations by the Bragg-Vulf formula and by Laue's formulas [see Eqs. (18.61)] lead to coinciding results.

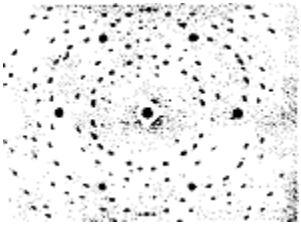


Fig. 18.44: Laue diffraction pattern of beryl (a mineral of the silicate group).

The diffraction of X-rays from crystals has two principal applications. It is used to investigate the spectral composition of X-radiation (**X-ray spectroscopy**) and to study the structure of crystals (**X-ray structure analysis**).

By determining the directions of the maxima obtained in the diffraction of the X-radiation being studied from crystals with a known structure, we can calculate the wavelengths. Originally, crystals of the cubic system were used to determine wavelengths, the spacing of the planes being determined from the density and relative molecular mass of the crystal.

In the method of structural analysis proposed by von Laue, a beam of X-rays is directed onto a stationary monocrystal. The radiation contains a wavelength at which condition (18.63) is satisfied for each system of layers sufficiently densely populated by atoms. Consequently, we obtain a collection of black spots on a photographic plate placed behind the crystal (after development). The mutual arrangement of the spots reflects the symmetry of the crystal. The distances between the spots and their intensities allow us to find the arrangement of the atoms in a crystal and their spacing. Figure 18.44 shows a Laue diffraction pattern of beryl (a mineral of the silicate group).

The method of structural analysis developed by the Dutch physicist Peter Debye and the Swiss physicist Paul Scherrer uses monochromatic X-radiation and polycrystalline specimens. The substance being studied is ground into a powder, and the latter is pressed into a wire-shaped specimen. The specimen is put along the axis of a cylindrical chamber on whose side surface a photographic film is placed (Fig. 18.45). Among the enormous number of chaotically oriented minute crystals, there will always be a multitude of such ones for which condition (18.63) will be observed, the diffracted ray being in the most diverse planes for different crystals. As a result, for each system of atomic layers and each value of m, we get not one direction of a maximum, but a cone of directions whose axis coincides with the direction of the incident beam (see Fig. 18.45). The pattern obtained on the film (a Debye powder pattern) has the form shown in Fig. 18.46. Each pair of symmetrically arranged lines corresponds to one of the diffraction maxima satisfying condition

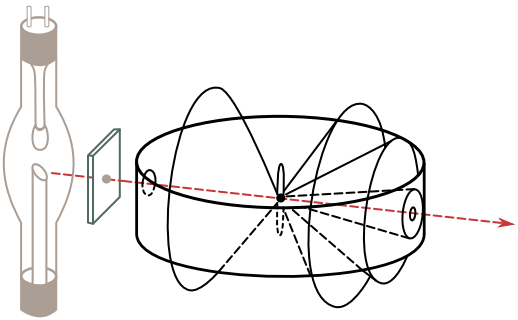


Fig. 18.45: Scheme of the instrument developed by Debye and Scherrer for structural analysis. It uses monochromatic X-radiation and polycrystalline specimens put along the axis of a cylindrical chamber on whose side surface a photographic film is placed.

Fig. 18.46: Pattern obtained on the film (a Debye powder pattern). Each pair of symmetrically arranged lines corresponds to one of the diffraction maxima satisfying condition (18.63) at a certain value of m.

(18.63) at a certain value of m. The structure of the crystal can be determined by decoding the X-ray pattern.

18.8. Resolving Power of an Objective

Assume that a plane light wave falls on an opaque screen with a round aperture of radius b cut out of it. The number of Fresnel zones opened by the aperture for point P opposite the centre of the aperture at the distance l from it can be found by Eq. (18.13) assuming that $a = \infty$, $r_0 = b$, and $b = l$. The result is

$$m = \frac{b^2}{l\lambda} \tag{18.64}$$

[compare with expression (18.37)].

In the same way as for a slit, depending on the value of parameter (18.64), we have to do either with the approximation of geometrical optics, or Fresnel diffraction, or, finally, Fraunhofer diffraction [see expressions (18.36)].

We can observe a Fraunhofer diffraction pattern from a round aperture on a screen in the focal plane of a lens placed behind the aperture by directing a plane light wave onto the aperture. This pattern has the form of a central bright spot

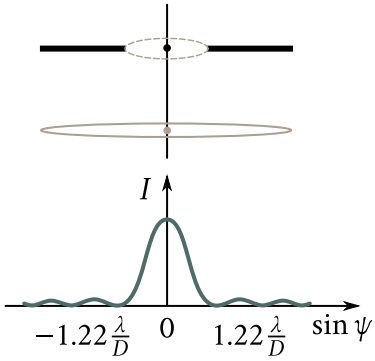


Fig. 18.47: Fraunhofer diffraction pattern from a round aperture on a screen in the focal plane of a lens placed behind the aperture by directing a plane light wave onto the aperture. This pattern has the form of a central bright spot surrounded by alternating dark and bright rings.

surrounded by alternating dark and bright rings (Fig. 18.47). The corresponding calculations show that the first minimum is at the angular distance from the centre of the diffraction pattern of

$$\varphi_{\min} = \arcsin\left(1.22\frac{\lambda}{D}\right), \tag{18.65}$$

where D is the diameter of the aperture [compare with Eq. (18.28)]. If $D \gg \lambda$, we may consider that

$$\varphi_{\min} = 1.22\frac{\lambda}{D}. \tag{18.66}$$

The major part (about 84%) of the light flux passing through the aperture gets into the region of the central bright spot. The intensity of the first bright ring is only 1.74%, and of the second, 0.41% of the intensity of the central spot. The intensity of the other bright rings is still smaller. For this reason, in a first approximation, we may consider that the diffraction pattern consists of only a single bright spot with an angular radius determined by Eq. (18.65). This spot is in essence the image of an infinitely remote point source of light (a plane light wave falls on the aperture).

The diffraction pattern does not depend on the distance between the aperture and the lens. In particular, it will be the same when the edges of the aperture are made to coincide with the edges of the lens. It thus follows that even a perfect lens cannot produce an ideal optical image. Owing to the wave nature of light, the image of a point produced by the lens has the form of a spot that is the central maximum of a diffraction pattern. The angular dimension of this spot diminishes with an increasing diameter of the lens mount D.

With a very small angular distance between two points, their images obtained

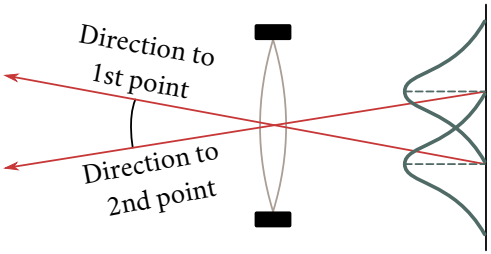


Fig. 18.48: Rayleigh criterion: two close points will still be resolved if the middle of the central diffraction maximum for one of them coincides with the edge of the central maximum for the second one. This occurs if the angular distance between the points $\delta\psi$ equals the angular radius given by Eq. (18.65).

with the aid of an optical instrument will be superposed and will produce a single luminous spot. Hence, two very close points will not be perceived by the instrument separately or, as we say, will not be resolved by the instrument. Consequently, no matter how great the image is in size, the corresponding details will not be seen on it.

Let $\delta\psi$ stand for the smallest angular distance between two points at which they can still be resolved by an optical instrument. The reciprocal of $\delta\psi$ is called the **resolving power of the instrument**:

$$R = \frac{1}{\delta\psi}. \tag{18.67}$$

Let us find the resolving power of the objective of a telescope or camera when very remote objects are being looked at or photographed. In this condition, the rays travelling into the objective from each point of the object may be considered parallel, and we can use formula (18.65). According to the Rayleigh criterion, two close points will still be resolved if the middle of the central diffraction maximum for one of them coincides with the edge of the central maximum (*i.e.*, with the first minimum) for the second one. A glance at Fig. 18.48 shows that this will occur if the angular distance between the points $\delta\psi$ will equal the angular radius given by Eq. (18.65). The diameter of the objective mount D is much greater than the wavelength λ. We may therefore consider that

$$\delta\psi = 1.22\frac{\lambda}{D}.$$

Hence,

$$R = \frac{D}{1.22\lambda}. \tag{18.68}$$

It can be seen from this formula that the resolving power of an objective grows with its diameter.

The diameter of the pupil of an eye at normal illumination is about 2 mm. Using this value in Eq. (18.68) and taking $\lambda = 0.5 \times 10^{-3}$ mm, we get

$$\delta\psi = 1.22 \times \frac{0.5 \times 10^{-3}}{2} = 0.305 \times 10^{-3} \text{ rad} \approx 1'.$$

Thus, the minimum angular distance between points at which the human eye still perceives them separately, equals one angular minute. It is interesting to note that the distance between adjacent light sensitive elements of the retina corresponds to this angular distance.

18.9. Holography

Holography (*i.e.*, "complete recording", from the Greek "bolos" meaning "the whole" and "grapho"-"write") is a special way of recording the structure of the light wave reflected by an object on a photographic plate. When this plate (a hologram) is illuminated with a beam of light, the wave recorded on it is reconstructed in practically its original form, so that when the eye perceives the reconstructed wave, the visual sensation is virtually the same as it would be if the object itself were observed.

Holography was invented in 1947 by the British physicist Dennis Gabor. The complete embodiment of Gabor's idea became possible, however, only after the appearance in 1960 of light sources having a high degree of coherence—lasers. Gabor's initial arrangement was improved by the American physicists Emmet Leith and Juris Upatnieks, who obtained the first laser holograms in 1963. The Soviet scientist Yuri Denisyuk in 1962 proposed an original method of recording holograms on a thick-layer emulsion. This method, unlike holograms on a thin-layer emulsion, produces a coloured image of the object.

We shall limit ourselves to an elementary consideration of the method of recording holograms on a thin-layer emulsion. Figure 18.49a contains a schematic view of an arrangement for recording holograms, and Fig. 18.49b a schematic view of reconstruction of the image. The light beam emitted by the laser, expanded by a system of lenses, is split into two parts. One part is reflected by the mirror to the photographic plate forming the so-called reference wave 1. The second part reaches the plate after being reflected from the object; it forms object beam 2. Both beams must be coherent. This requirement is satisfied because laser radiation has a high degree of spatial coherence (the light oscillations are coherent over the entire cross section of a laser beam). The reference and object beams superpose and form an interference pattern that is recorded by the photographic plate. A plate exposed in this way and developed is a **hologram**. Two beams of light participate in forming the hologram. In this connection, the arrangement described above is called

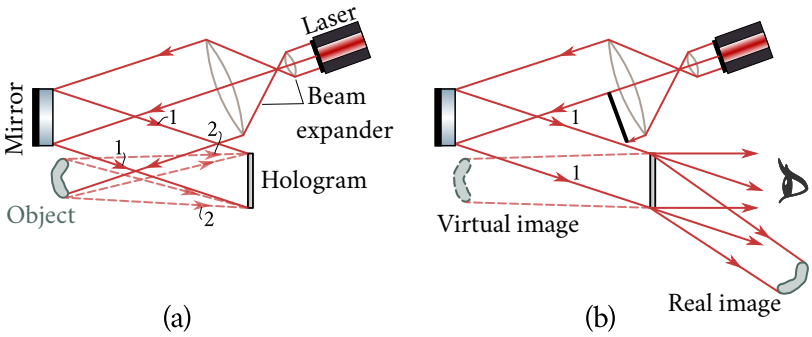


Fig. 18.49: Holograms on a thin-layer emulsion. (a) Schematic view of an arrangement for recording holograms. (b) Schematic view of reconstruction of the image.

two-beam or split-beam holography.

To reconstruct the image, the developed photographic plate is put in the same place where it was in recording the hologram, and is illuminated with the reference beam of light (the part of the laser beam that illuminated the object in recording the hologram is now stopped). The reference beam diffracts on the hologram, and as a result a wave is produced having exactly the same structure as the one reflected by the object. This wave produces a virtual image of the object that is seen by the observer. In addition to the wave forming the virtual image, another wave is produced that gives a real image of the object. This image is pseudoscopic; this means that it has a relief which is the opposite of the relief of the 2 object—the convex spots are replaced by concave ones, and vice versa.

Let us consider the nature of a hologram and the process of image reconstruction. Assume that two coherent parallel beams of light rays fall on the photographic plate, with the angle ψ between the beams (Fig. 18.50). Beam 1 is the reference one, and beam 2, the object one (the object in the given case is an infinitely remote point). We shall assume for simplicity that beam 1 is normal to the plate. All the results obtained below also hold when the reference beam falls at an angle, but the formulas will be more cumbersome.

Owing to the interference of the reference and object beams, a system of alternating straight maxima and minima of the intensity is formed on the plate. Let points A and B correspond to the middles of adjacent interference maxima. Hence, the path difference Δ' equals λ. Examination of Fig. 18.50 shows that $\Delta' = d \sin \psi$; hence,

$$d \sin \psi = \lambda. \tag{18.69}$$

Having recorded the interference pattern on the plate (by exposure and developing), we direct reference beam 1 at it. For this beam, the plate plays the part of

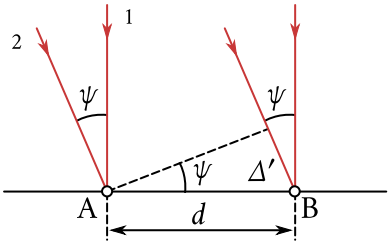


Fig. 18.50: Two coherent parallel beams of light rays fall on the photographic plate, with the angle ψ between the beams. Beam 1 is the reference one, and beam 2, the object one (the object is considered an infinitely remote point). We shall assume for simplicity that beam 1 is normal to the plate.

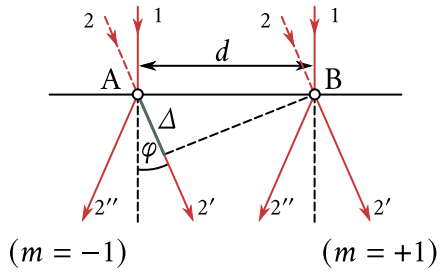


Fig. 18.51: Plate illuminated with a reference beam, produces a diffraction pattern whose maxima form the angles φ with a normal to the plate. The maximum corresponding to $m = 0$ is on the continuation of the reference beam. The maximum corresponding to $m = +1$ has the same direction as object beam 2 did during the exposure. In addition, a maximum corresponding to $m = -1$ appears.

a diffraction grating whose period d is determined by Eq. (18.69). A feature of this grating is the circumstance that its transmittance changes in a direction perpendicular to the "lines" according to a cosine law (in the gratings treated in Sec. 18.6 it changed in a jump: gap-dark-gap-dark, etc.). The result of this feature is that the intensity of all the diffraction maxima of orders higher than the first one virtually equals zero.

When the plate is illuminated with the reference beam (Fig. 18.51), a diffraction pattern appears whose maxima form the angles φ with a normal to the plate. These angles are determined by the condition

$$d \sin \varphi = m\lambda \quad (m = 0, \pm 1) \tag{18.70}$$

[compare with formula (18.41)]. The maximum corresponding to $m = 0$ is on the continuation of the reference beam. The maximum corresponding to $m = \pm 1$ has the same direction as object beam 2 did during the exposure [compare Eqs. (18.69) and (18.70)]. In addition, a maximum corresponding to $m = -1$ appears.

It can be shown that the result we have obtained also holds when object beam 2 consists of diverging rays instead of parallel ones. The maximum corresponding to $m = +1$ has the nature of diverging beam of rays 2′ (it produces a virtual image of the point from which rays 2 emerged during the exposure); the maximum corresponding to $m = -1$, on the other hand, has the nature of a converging beam of rays 2″ (it forms a real image of the point which rays 2 emerged from during the exposure).

In recording the hologram, the plate is illuminated by reference beam 1 and numerous diverging beams 2 reflected by different points of the object. An intricate interference pattern is formed on the plate as a result of superposition of the patterns produced by each of the beams 2 separately. When the hologram is illuminated with reference beam 1, all beams 2 are reconstructed, *i.e.*, the complete light wave reflected by the object ($m = +1$ corresponds to it). Two other waves appear in addition to it (corresponding to $m = 0$ and $m = -1$). But these waves propagate in other directions and do not hinder the perception of the wave producing a virtual image of the object (see Fig. 18.49).

The image of an object produced by a hologram is three-dimensional. It can be viewed from different positions. If in recording a hologram close objects concealed more remote ones, then by moving to a side we can look behind the closer object (more exactly, behind its image) and see the objects that had been concealed previously. The explanation is that when moving to a side we see the image reconstructed from the peripheral part of the hologram onto which the rays reflected from the concealed objects also fell during the exposure. When looking at the images of close and far objects, we have to accommodate our eyes as when looking at the objects themselves.

If a hologram is broken into several pieces, then each of them when illuminated will produce the same picture as the original hologram. But the smaller the part of the hologram used to reconstruct the image, the lower is its sharpness. This is easy to understand by taking into account that when the number of lines of a diffraction grating is reduced, its resolving power diminishes [see Eq. (18.55)].

The possible applications of holography are very diverse. A far from complete list of them includes holographic motion pictures and television, holographic microscopes, and control of the quality of processing articles. The statement can be encountered in publications on the subject that holography can be compared as regards its consequences with the setting up of radio communication.

Chapter 19

POLARIZATION OF LIGHT

19.1. Natural and Polarized Light

We remind our reader that light is called polarized if the directions of oscillations of the light vector in it are brought into order in some way or other (see Sec. 16.1). In natural light, oscillations in various directions rapidly and chaotically replace one another.

Let us consider two mutually perpendicular electrical oscillations occurring along the axes x and y and differing in phase by δ:

$$E_x = A_1 \cos(\omega t), \quad E_y = A_2 \cos(\omega t + \delta). \tag{19.1}$$

The resultant field strength $\boldsymbol{E}$ is the vector sum of the strengths $\boldsymbol{E}_x$ and $\boldsymbol{E}_y$ (Fig. 19.1). The angle φ between the directions of the vectors $\boldsymbol{E}$ and $\boldsymbol{E}_x$ is determined by the expression

$$\tan\varphi = \frac{E_x}{E_y} = \frac{A_2 \cos(\omega t + \delta)}{A_1 \cos(\omega t)}. \tag{19.2}$$

If the phase difference δ undergoes random chaotic changes, then the angle φ, *i.e.*, the direction of the light vector $\boldsymbol{E}$, will experience intermittent disordered changes too. Accordingly, natural light can be represented as the superposition of two incoherent electromagnetic waves polarized in mutually perpendicular planes and having the same intensity. Such a representation greatly simplifies the consideration of the transmission of natural light through polarizing devices.

Assume that the light waves E_x and E_y are coherent, with δ equal to zero or π. Hence, according to Eq. (19.2),

$$\tan\varphi = \pm\frac{A_2}{A_1} = \text{constant}.$$

Consequently, the resultant oscillation occurs in a fixed direction—the wave is plane-polarized.

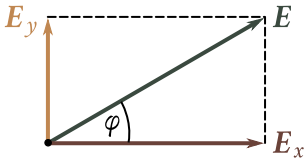


Fig. 19.1: The resultant field strength $\boldsymbol{E}$ is the vector sum of the strengths $\boldsymbol{E}_x$ and $\boldsymbol{E}_y$.

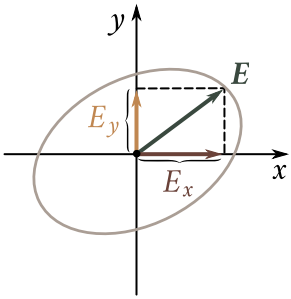


Fig. 19.2: We consider that quantities (19.1) are the coordinates of the tail of the resultant vector $\boldsymbol{E}$.

When $A_1 = A_2$, and $\delta = \pm\pi/2$, we have

$$\tan\varphi = \mp\tan(\omega t)$$

$[\cos(\omega t \pm \pi/2) = \mp\sin(\omega t)]$. It thus follows that the plane of oscillations rotates about the direction of the ray with an angular velocity equal to the frequency of oscillation ω. The light in this case will be circularly polarized.

To find the nature of the resultant oscillation with an arbitrary constant value of δ, let us take into account that quantities (19.1) are the coordinates of the tail of the resultant vector $\boldsymbol{E}$ (Fig. 19.2). We know from our treatment of oscillations (see Sec. 7.9 of Vol. I) that two mutually perpendicular harmonic oscillations of the same frequency produce motion along an ellipse when summated (in particular, motion along a straight line or a circle may be obtained). Similarly, a point with the coordinates determined by Eqs. (19.1), *i.e.*, the tail of vector $\boldsymbol{E}$, travels along an ellipse. Consequently, two coherent plane-polarized light waves whose planes of oscillations are mutually perpendicular produce an elliptically polarized light wave when superposed on each other. At a phase difference of zero or π, the ellipse degenerates into a straight line, and plane-polarized light is obtained. At $\delta = \pm\pi/2$ and equality of the amplitude of the waves being added, the ellipse transforms into a circle—circularly polarized light is obtained.

Depending on the direction of rotation of the vector $\boldsymbol{E}$, right and left elliptical and circular polarizations are distinguished. If with respect to the direction opposite that of the ray the vector $\boldsymbol{E}$ rotates clockwise, the polarization is called **right**, and in the opposite case it is **left**.

The plane in which the light vector oscillates in a plane-polarized wave will be called the **plane of oscillations**. For historical reasons, the term **plane of polarization** was applied not to the plane in which the vector $\boldsymbol{E}$ oscillates, but to the plane perpendicular to it.

Plane-polarized light can be obtained from natural light with the aid of devices called **polarizers**. These devices freely transmit oscillations parallel to the plane which we shall call the **polarizer plane** and completely or partly retain the oscillations perpendicular to this plane. We shall apply the adjective **imperfect** to a polarizer that only partly retains oscillations perpendicular to its plane. We shall apply the term "polarizer" for brevity to a perfect polarizer that completely retains the oscillations perpendicular to its plane and does not weaken the oscillations parallel to its plane.

Light is produced at the outlet from an imperfect polarizer in which the oscillations in one direction predominate over the oscillations in other directions. Such light is called **partly polarized**. It can be considered as a mixture of natural and plane-polarized light. Partly polarized light, like natural light, can be represented in the form of a superposition of two incoherent plane-polarized waves with mutually perpendicular planes of oscillations. The difference is that for natural light the intensity of these waves is the same, and for partly polarized light it is different.

If we pass partly polarized light through a polarizer, then when the device rotates about the direction of the ray, the intensity of the transmitted light will change within the limits from $I_{\max}$ to $I_{\min}$. The transition from one of these values to the other one will occur upon rotation through an angle of $\pi/2$ (during one complete revolution both the maximum and the minimum intensity will be reached twice). The expression

$$P = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}} \tag{19.3}$$

is known as the **degree of polarization**. For plane-polarized light, $I_{\min} = 0$, and $P = 1$. For natural light, $I_{\min} = I_{\max}$ and $P = 0$.

The concept of the degree of polarization cannot be applied to elliptically polarized light (in such light the oscillations are completely ordered, so that the degree of polarization always equals unity).

An oscillation of amplitude A occurring in a plane making the angle φ with the polarizer plane can be resolved into two oscillations having the amplitudes $A_{\parallel} = A\cos\varphi$ and $A_{\perp} = A\sin\varphi$ (Fig. 19.3; the ray is perpendicular to the plane of the drawing). The first oscillation will pass through the device, the second will be retained. The intensity of the transmitted wave is proportional to $A_{\parallel}^2 = A^2\cos^2\varphi$, *i.e.*, is $I\cos^2\varphi$, where I is the intensity of the oscillation of amplitude A. Consequently, an oscillation parallel to the plane of the polarizer carries along a fraction of the intensity equal to $\cos^2\varphi$. In natural light, all the values of φ are equally probable. Therefore, the fraction of the light transmitted through the polarizer will equal the average value of $\cos^2\varphi$, *i.e.*, one-half. When the polarizer is rotated

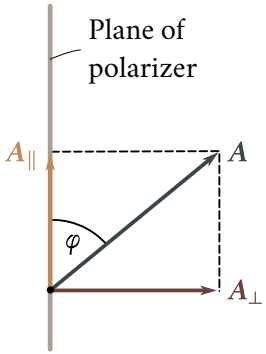


Fig. 19.3: An oscillation of amplitude A occurring in a plane making the angle φ with the polarizer plane can be resolved into two oscillations having the amplitudes parallel and perpendicular: $A_{\parallel} = A\cos\varphi$ and $A_{\perp} = A\sin\varphi$.

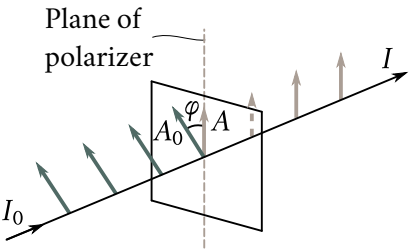


Fig. 19.4: Plane-polarized light of amplitude A_0 and intensity I_0 falling on a polarizer. φ is the angle between the plane of oscillations of the incident light and the plane of the polarizer. The component of the oscillation having the amplitude $A = A_0\cos\varphi$, will pass through the device.

about the direction of a natural ray, the intensity of the transmitted light remains the same. What changes is only the orientation of the plane of oscillations of the light leaving the device.

Assume that plane-polarized light of amplitude A_0 and intensity I_0 falls on a polarizer (Fig. 19.4). The component of the oscillation having the amplitude $A = A_0\cos\varphi$, where φ is the angle between the plane of oscillations of the incident light and the plane of the polarizer, will pass through the device. Hence, the intensity of the transmitted light I is determined by the expression

$$I = I_0\cos^2\varphi. \tag{19.4}$$

Relation (19.4) is known as **Malus's law**. It was first formulated by the French physicist Etienne Malus (1775-1812).

Let us put two polarizers whose planes make the angle φ in the path of a natural ray. Plane-polarized light whose intensity I_0 is half that of natural light will emerge from the first polarizer. According to Malus's law, light having an intensity of $I_0\cos^2\varphi$ will emerge from the second polarizer. The intensity of the light transmitted through both polarizers is

$$I = \frac{1}{2}I_{\text{nat}}\cos^2\varphi. \tag{19.5}$$

The maximum intensity equal to $I_{\text{nat}}/2$ is obtained at $\varphi = 0$ (the polarizers are parallel). At $\varphi = \pi/2$, the intensity is zero-crossed polarizers transmit no light.

Assume that elliptically polarized light falls on a polarizer. The device transmits

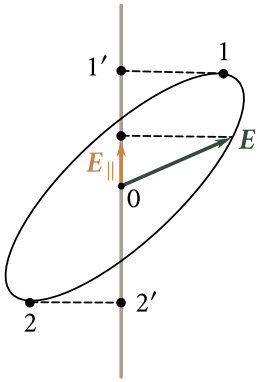


Fig. 19.5: For elliptically polarized light falling on a polarizer, the device transmits the component $\boldsymbol{E}_{\parallel}$ of the vector $\boldsymbol{E}$ in the direction of the plane of the polarizer. The maximum value of this component is reached at points 1 and 2.

the component $\boldsymbol{E}_{\parallel}$ of the vector $\boldsymbol{E}$ in the direction of the plane of the polarizer (Fig. 19.5). The maximum value of this component is reached at points 1 and 2. Hence, the amplitude of the plane-polarized light leaving the device equals the length of $01'$. Rotating the polarizer around the direction of the ray, we shall observe changes in the intensity ranging from $I_{\max}$ (obtained when the plane of the polarizer coincides with the semimajor axis of the ellipse) to $I_{\min}$ (obtained when the plane of the polarizer coincides with the semiminor axis of the ellipse). The intensity of light for partly polarized light will change in the same way upon rotation of the polarizer. For circularly polarized light, rotation of the polarizer is not attended (as for natural light) by a change in the intensity of the light transmitted through the device.

19.2. Polarization in Reflection and Refraction

If the angle of incidence of light on the interface between two dielectrics (for example, on the surface of a glass plate) differs from zero, the reflected and refracted rays will be partly polarized[1]. Oscillations perpendicular to the plane of incidence predominate in the reflected ray (in Fig. 19.6 these oscillations are denoted by points), and oscillations parallel to the plane of incidence predominate in the refracted ray (they are depicted in the figure by double-headed arrows). The degree of polarization depends on the angle of incidence. Let θ_{Br} stand for the angle satisfying the condition

$$\tan \theta_{\mathrm{Br}} = n_{12} \tag{19.6}$$

[1]Elliptically polarized light is obtained upon reflection from a conducting surface (for example, from the surface of a metal).

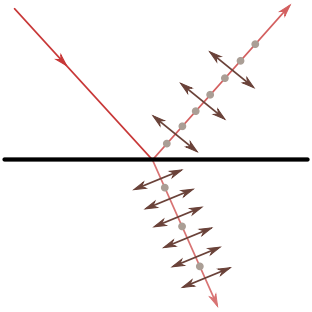


Fig. 19.6: Polarization in reflection and refraction. Oscillations perpendicular to the plane of incidence predominate in the reflected ray (dots) and oscillations parallel to the plane of incidence predominate in the refracted ray (double-headed arrows).

(n_{12} is the refractive index of the second medium relative to the first one). At an angle of incidence θ_1 equal to θ_{Br}, the Fig. 19.6 reflected ray is completely polarized (it contains only oscillations perpendicular to the plane of incidence). The degree of polarization of the refracted ray at an angle of incidence equal to θ_{Br} reaches its maximum value, but this ray remains polarized only partly.

Relation (19.6) is known as **Brewster's law**, in honour to its discoverer, the British physicist David Brewster (1781-1868), and the angle θ_{Br} is called **Brewster's angle**. It is easy to see that when light falls at Brewster's angle, the reflected and refracted rays are mutually perpendicular. The degree of polarization of the reflected and refracted rays for different angles of incidence can be obtained with the aid of Fresnel's formulas. The latter follow from the conditions imposed on an electromagnetic field at the interface between two dielectrics[2]. These conditions include the equality of the tangential components of the vectors $\boldsymbol{E}$ and $\boldsymbol{H}$, and also the equality of the normal components of the vectors $\boldsymbol{D}$ and $\boldsymbol{B}$ at both sides of the interface (for one side the sum of the relevant vectors for the incident and reflected waves must be taken, and for the other, the vector for the refracted wave).

Fresnel's formulas establish the relations between the complex amplitudes of the incident, reflected, and refracted waves. We remind our reader that by the complex amplitude $\hat{A}$ is meant the expression $Ae^{i\alpha}$, where A is the conventional amplitude, and α is the initial phase of the oscillations. Hence, the equality of two complex amplitudes signifies the equality of both the conventional amplitudes and the initial phases of the two oscillations:

$$\hat{A}_1 = \hat{A}_2 \Rightarrow A_1 = A_2 \text{ and } \alpha_1 = \alpha_2. \tag{19.7}$$

When the complex amplitudes differ in sign, the conventional ones are the same,

[2] Fresnel obtained these formulas on the basis of the notions of light as of elastic waves propagating in ether.

while the initial phases differ by π ($e^{i\pi} = -1$):

$$\hat{A}_1 = -\hat{A}_2 \Rightarrow A_1 = A_2 \text{ and } \alpha_1 = \alpha_2 + \pi. \tag{19.8}$$

Let us represent the incident wave in the form of a superposition of two incoherent waves in one of which the oscillations occur in the plane of incidence, and in the other, are perpendicular to this plane. Let us denote the complex amplitude of the first wave by $\hat{A}_{\parallel}$, and of the second by $\hat{A}_{\perp}$. We shall proceed similarly with the reflected and refracted waves. We shall use the same symbols for the amplitudes of the reflected waves, adding one prime, and the same symbols for the amplitudes of the refracted waves, adding two primes. Thus,

- $\hat{A}_{\parallel}$ and $\hat{A}_{\perp}$ = amplitudes of the incident waves,
- $\hat{A}'_{\parallel}$ and $\hat{A}'_{\perp}$ = amplitudes of the reflected waves,
- $\hat{A}''_{\parallel}$ and $\hat{A}''_{\perp}$ = amplitudes of the refracted waves.

Fresnel's formulas have the following form[3]:

$$\begin{cases} \hat{A}'_{\parallel} = \hat{A}_{\parallel} \dfrac{\tan(\theta_1 - \theta_2)}{\tan(\theta_1 + \theta_2)}, \quad \hat{A}'_{\perp} = \hat{A}_{\perp} \dfrac{\sin(\theta_1 - \theta_2)}{\sin(\theta_1 + \theta_2)} \\ \hat{A}''_{\parallel} = \hat{A}_{\parallel} \dfrac{2 \sin\theta_2 \cos\theta_1}{\sin(\theta_1 + \theta_2)\cos(\theta_1 - \theta_2)}, \quad \hat{A}''_{\perp} = \hat{A}_{\perp} \dfrac{2 \sin\theta_2 \cos\theta_1}{\sin(\theta_1 + \theta_2)} \end{cases} \tag{19.9}$$

(θ_1 is the angle of incidence, and θ_2 is the angle of refraction of the light wave). We must underline the fact that formulas (19.9) establish the relations between the complex amplitudes at the interface between dielectrics, *i.e.*, at the point of incidence of a ray on this interface.

It can be seen from the last two of formulas (19.9) that the signs of the complex amplitudes of the incident and refracted waves at any values of the angles θ_1 and θ_2 are the same (the sum of θ_1 and θ_2 cannot exceed π). This signifies that when penetrating into the second medium, the phase of the wave does not undergo a jump. In dealing with the phase relations between an incident and a reflected wave, we must take into account that for a wave polarized perpendicularly to the plane of incidence, the coincidence of the signs of $\hat{A}_{\perp}$ and $\hat{A}'_{\perp}$ corresponds to the absence of a jump in the phase in reflection (Fig. 19.7a). For a wave that is polarized in the plane of incidence, on the other hand, a jump in the phase is absent when the signs of $\hat{A}_{\parallel}$ and $\hat{A}'_{\parallel}$ are opposite (Fig. 19.7b).

The phase relations between the reflected and incident waves depend on the relation between the refractive indices n_1 and n_2 of the first and second media, and also on the relation between the angle of incidence θ_1 and Brewster's angle θ_{Br} (we remind our reader that when $\theta_1 = \theta_{\text{Br}}$, the sum of the angles θ_1 and θ_2, is $\pi/2$). Table

[3]Fresnel's formulas are customarily written without "caps" over the amplitudes. To underline the fact that we are dealing with complex amplitudes, however, we found it helpful to write the amplitudes with the "caps".

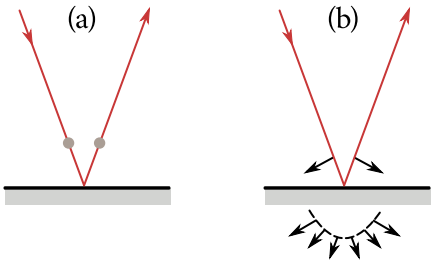


Fig. 19.7: Dealing with phase relations. (a) For a wave polarized perpendicularly to the plane of incidence, the coincidence of the signs of $\hat{A}_\perp$ and $\hat{A}'_\perp$ corresponds to the absence of a jump in the phase in reflection. (b) For a wave that is polarized in the plane of incidence, on the other hand, a jump in the phase is absent when the signs of $\hat{A}_\parallel$ and $\hat{A}'_\parallel$ are opposite.

19.1 gives the results following from the first two of formulas (19.9) in four possible cases. It follows from the table that for incidence at an angle less than Brewster's angle, reflection from an optically denser medium is attended by a jump in phase of π; reflection from an optically less dense medium occurs without a change in phase. This result for $\theta_1 = 0$ was obtained in Sec. 16.3. When $\theta_1 > \theta_{\text{Br}}$, the phase relations for both wave components are different.

We obtain from the first of formulas (19.9) that when $\theta_1 + \theta_2 = \pi/2$, *i.e.*, at $\theta_1 = \theta_{\text{Br}}$, the amplitude $\hat{A}'_\parallel$ vanishes. Consequently, only oscillations perpendicular to the plane of incidence are present in the reflected wave—the latter is completely polarized. Thus, *Brewster's law directly follows from Fresnel's formulas*.

At small angles of incidence, the sines and tangents in formulas (19.9) may be replaced by the angles themselves, and the cosines may be assumed equal to unity.

Table 19.1

	$\theta_1 < \theta_{\text{Br}}$ $(\theta_1 + \theta_2 < \pi/2)$	$\theta_1 > \theta_{\text{Br}}$ $(\theta_1 + \theta_2 > \pi/2)$
$n_2 > n_1$, $\theta_1 > \theta_2$	The signs of $\hat{A}'_\parallel$ and $\hat{A}_\parallel$ are the same (a phase jump by π). The sign of $\hat{A}'_\perp$ is opposite to that of $\hat{A}'_\perp$ (a phase jump by π).	The sign of $\hat{A}'_\parallel$ is opposite to that of $\hat{A}_\parallel$ (no phase jump). The sign of $\hat{A}'_\perp$ is opposite to that of $\hat{A}_\parallel$ (a phase jump by π).
$n_2 < n_1$, $\theta_1 < \theta_2$	The sign of $\hat{A}'_\parallel$ is opposite to that of $\hat{A}_\parallel$ (no phase jump). The signs of $\hat{A}'_\perp$ and $\hat{A}'_\perp$ are the same (no phase jump).	The signs of $\hat{A}'_\parallel$ and $\hat{A}_\parallel$ are the same (a phase jump by π). The signs of $\hat{A}'_\perp$ and $\hat{A}_\perp$ are the same (no phase jump).

In addition, in this case we may consider that $\theta_1 = n_{12}\theta_2$ (this follows from the law of refraction after the sines are replaced with the relevant angles). As a result, Fresnel's formulas for small angles of incidence acquire the form

$$\text{small } \theta_1 \Rightarrow \begin{cases} \hat{A}'_{\parallel} = \hat{A}_{\parallel}\dfrac{\theta_1 - \theta_2}{\theta_1 + \theta_2} = \hat{A}_{\parallel}\dfrac{n_{12} - 1}{n_{12} + 1} \\ \hat{A}'_{\perp} = \hat{A}_{\perp}\dfrac{\theta_1 - \theta_2}{\theta_1 + \theta_2} = -\hat{A}_{\perp}\dfrac{n_{12} - 1}{n_{12} + 1} \\ \hat{A}''_{\parallel} = \hat{A}_{\parallel}\dfrac{2\theta_2}{\theta_1 + \theta_2} = \hat{A}_{\parallel}\dfrac{2}{n_{12} + 1} \\ \hat{A}''_{\perp} = \hat{A}_{\perp}\dfrac{2\theta_2}{\theta_1 + \theta_2} = \hat{A}_{\perp}\dfrac{2}{n_{12} + 1}. \end{cases} \tag{19.10}$$

Squaring Eqs. (19.10) and multiplying the expressions obtained by the refractive index of the relevant medium, we get relations between the intensities of the incident, reflected, and refracted rays for small angles of incidence [see expression (19.6)]. Here, for example, the intensity of the reflected light I' can be calculated as the sum of the intensities of both components $I'_{\parallel}$ and $I'_{\perp}$ because these components are not coherent in natural light [the intensities instead of the amplitudes are summated for incoherent waves, see Eq. (17.1)]. As a result, we get

$$I' = I\left(\frac{n_{12} - 1}{n_{12} + 1}\right)^2,$$

From these formulas, we get Eqs. (16.33) and (16.34) for ρ and τ.

19.3. Polarization in Double Refraction

When light passes through all transparent crystals except for those belonging to the cubic system, a phenomenon is observed called **double refraction**[4]. It consists in that a ray falling on a crystal is split inside the latter into two rays propagating, generally speaking, with different velocities and in different directions.

Doubly refracting (or birefringent) crystals are divided into **uniaxial** and **biaxial** ones. In uniaxial crystals, one of the refracted rays obeys the conventional law of refraction, in particular, it is in the same plane as the incident ray and a normal to the refracting surface. This ray is called an **ordinary ray** and is designated by the symbol o. For the other ray, called an **extraordinary ray** (designated by e), the ratio of the sines of the angle of incidence and the angle of refraction does not remain constant when the angle of incidence varies. Even upon normal incidence of light Fig. 19.8 on a crystal, an extraordinary ray, generally speaking, deviates from

[4]Double refraction was first observed in 1669 by the Danish scientist Erasm Bartholin (1625-1698) for Iceland spar (a variety of calcium carbonate $CaCO_3$—crystals of the hexagonal system).

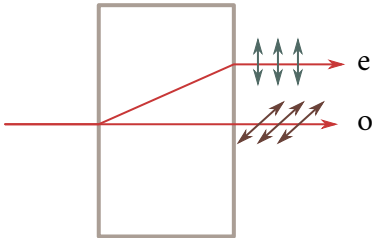


Fig. 19.8: Double refraction or birefringence. Doubly refracting (or birefringent) crystals are divided into uniaxial and biaxial ones. In uniaxial crystals, one of the refracted rays (ordinary rays "o") obeys the conventional law of refraction, in particular it is in the same plane as the incident ray and a normal to the refracting surface, while the extraordinary ray ("e") deviates from a normal.

a normal (Fig. 19.8). In addition, an extraordinary ray does not lie, as a rule, in the same plane as the incident ray and a normal to the refracting surface. Examples of uniaxial crystals are Iceland spar, quartz, and tourmaline. In biaxial crystals (mica, gypsum), both rays are extraordinary—the refractive indices for them depend on the direction in the crystal. In the following, we shall be concerned only with uniaxial crystals.

Uniaxial crystals have a direction along which ordinary and extraordinary rays propagate without separation and with the same velocity[5]. This direction is known as the **optical axis** of the crystal. It must be borne in mind that an optical axis is not a straight line passing through a point of a crystal, but a definite direction in the crystal. Any straight line parallel to the given direction is an optical axis of the crystal.

A plane passing through an optical axis is called a **principal section** or a **principal plane** of the crystal. Customarily, the principal section passing through the light ray is used.

Investigation of the ordinary and extraordinary rays shows that they are both completely polarized in mutually perpendicular directions (see Fig. 19.8). The plane of oscillations of the ordinary ray is perpendicular to a principal section of the crystal. In the extraordinary ray, the oscillations of the light vector occur in a plane coinciding with a principal section. When they emerge from the crystal, the two rays differ from each other only in the direction of polarization so that the terms "ordinary" and "extraordinary" have a meaning only inside the crystal.

In some crystals, one of the rays is absorbed to a greater extent than the other. This phenomenon is called **dichroism**. A crystal of tourmaline (a mineral of a complex composition) displays very great dichroism in visible rays. An ordinary

[5]Biaxial crystals have two such directions.

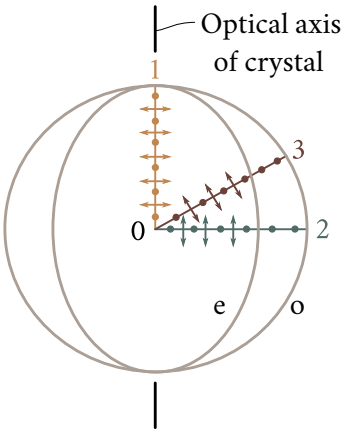


Fig. 19.9: In an ordinary ray, the oscillations of the light vector occur in a direction perpendicular to a principal section of the crystal (depicted by dots on the relevant ray). The oscillations in an extraordinary ray take place in a principal section. The directions of oscillations of the vector $\boldsymbol{E}$ are depicted by double-headed arrows, making different angles a with an optical axis.

ray is virtually completely absorbed in it over a distance of 1 mm. In crystals of iodoquinine sulphate, one of the rays is absorbed over a path of about 0.1 mm. This circumstance has been taken advantage of for manufacturing a polarizing device called a **polaroid**. It is a celluloid film into which a great number of identically oriented minute crystals of iodoquinine sulphate have been introduced.

Double refraction is explained by the anisotropy of crystals. In crystals of the noncubic system, the permittivity ε depends on the direction. In uniaxial crystals, ε in the direction of an optical axis and in directions perpendicular to it has different values $\varepsilon_{\parallel}$ and $\varepsilon_{\perp}$. In other directions, ε has intermediate values. According to Eq. (16.3) $n = \sqrt{\varepsilon}$. It thus follows from the anisotropy of ε, that *different values of the refractive index n correspond to electromagnetic waves with different directions of the oscillations of the vector* $\boldsymbol{E}$. Therefore, the velocity of the light waves depends on the direction of oscillations of the light vector $\boldsymbol{E}$.

In an ordinary ray, the oscillations of the light vector occur in a direction perpendicular to a principal section of the crystal (in Fig. 19.9 these oscillations are depicted by dots on the relevant ray). Therefore, with any direction of an ordinary ray (three directions 1, 2, and 3 are shown in the figure), the vector $\boldsymbol{E}$ makes a right angle with an optical axis of the crystal, and the velocity of the light wave will be the same, equal to $v_o = c/\sqrt{\varepsilon_{\perp}}$. Depicting the velocity of an ordinary ray in the form of lengths laid off in different directions, we shall get a spherical surface. Figure 19.9 shows the intersection of this surface with the plane of the drawing. A picture

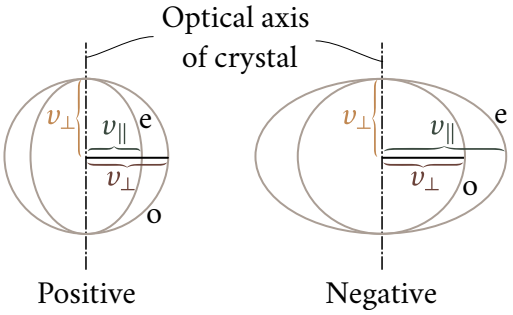


Fig. 19.10: Depending on which of the velocities, v_o or v_e, is greater, positive and negative uniaxial crystals are distinguished. Positive crystals, $v_e < v_o$ ($n_e > n_o$); negative crystals, $v_e > v_o$ ($n_e < n_o$).

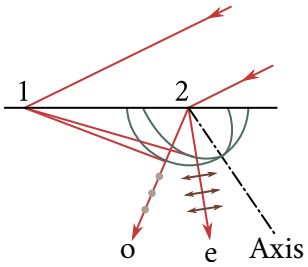


Fig. 19.11: Wave surfaces of an ordinary and extraordinary rays with their centre at point 2 on the surface of the crystal.

such as that in Fig. 19.9 is observed in any principal section, *i.e.*, in any plane passing through an optical axis. Let us imagine that a point source of light is placed at point 0 inside a crystal. Hence, the sphere which we have constructed will be the wave surface of ordinary rays.

The oscillations in an extraordinary ray take place in a principal section. Therefore, for different rays, the directions of oscillations of the vector $\boldsymbol{E}$ (in Fig. 19.9 these directions are depicted by double-headed arrows) make different angles a with an optical axis. For ray 1, the angle α is $\pi/2$, owing to which the velocity is $v_o = c/\sqrt{\varepsilon_\perp}$, for ray 2, the angle $\alpha = 0$, and the velocity is $v_e = c/\sqrt{\varepsilon_\parallel}$. For ray 3, the velocity has an intermediate value. We can show that the wave surface of extraordinary rays is an ellipsoid of revolution. At places of intersection with an optical axis of the crystal, this ellipsoid and the sphere constructed for the ordinary rays come into contact.

Uniaxial crystals are characterized by a **refractive index of an ordinary ray** equal to $n_o = c/v_o$, and a refractive index of an extraordinary ray perpendicular to an optical axis equal to $n_e = c/v_e$. The latter quantity is called simply the **refractive index of an extraordinary ray**.

Depending on which of the velocities, v_o or v_e, is greater, **positive** and **negative** uniaxial crystals are distinguished (Fig. 19.10). For positive crystals, $v_e < v_o$ (this means that $n_e > n_o$). For negative crystals, $v_e > v_o$ ($n_e < n_o$). It is simple to remember what crystals are called positive and what negative. For positive crystals, the ellipsoid of velocities is extended along an optical axis reminding one of the vertical line in the sign "+"; for negative crystals, the ellipsoid of velocities is extended in a direction perpendicular to an optical axis, reminding one of the sign "-".

The path of an ordinary and an extraordinary ray in a crystal can be determined

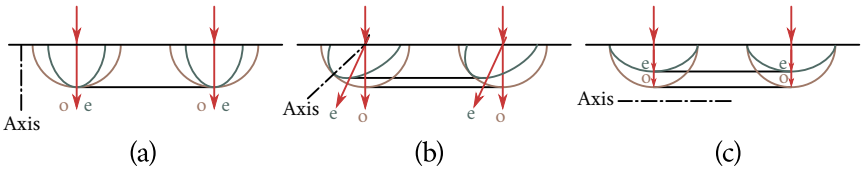


Fig. 19.12: Three cases of the normal incidence of light on the surface of a crystal differing in the direction of the optical axis. (a) The rays o and e propagate along an optical axis without separating. (b) An extraordinary ray may deviate from a normal to this surface. (c) The ordinary and extraordinary rays travel in the same direction, but propagate with different velocities.

with the aid of the Huygens principle. Figure 19.11 depicts wave surfaces of an ordinary and extraordinary rays with their centre at point 2 on the surface of the crystal. The construction is for the moment of time when the wavefront of the incident wave reaches point 1. The envelopes of all the secondary wavelets (the waves whose centres are in the interval between points 1 and 2 are not shown in the figure) for the ordinary and extraordinary rays are evidently planes. The refracted ray o or e emerging from point 2 passes through the point of contact of the envelope with the relevant wave surface.

We remind our reader that rays are defined as lines along which the energy of a light wave propagates (see Sec. 16.1). A glance at Fig. 19.11 shows that the ordinary ray o coincides with a normal to the relevant wave surface. The extraordinary ray e, on the other hand, appreciably deviates from a normal to the wave surface.

Figure 19.12 shows three cases of the normal incidence of light on the surface of a crystal differing in the direction of the optical axis. In case (a), the rays o and e propagate along an optical axis and therefore travel without separating. Inspection of Fig. 19.12b shows that even upon normal incidence of light on a refracting surface, an extraordinary ray may deviate from a normal to this surface (compare with Fig. 19.8). In Fig. 19.12c, the optical axis of the crystal is parallel to the refracting surface. In this case with normal incidence of the light, the ordinary and extraordinary rays travel in the same direction, but propagate with different velocities. The result is a constantly growing phase difference between them. The nature of polarization of the ordinary and extraordinary rays in Fig. 19.12 is not indicated. It is the same as for the rays depicted in Fig. 19.11.

19.4. Interference of Polarized Rays

When two coherent rays polarized in mutually perpendicular directions are superposed, no interference pattern with the characteristic alternation of maxima

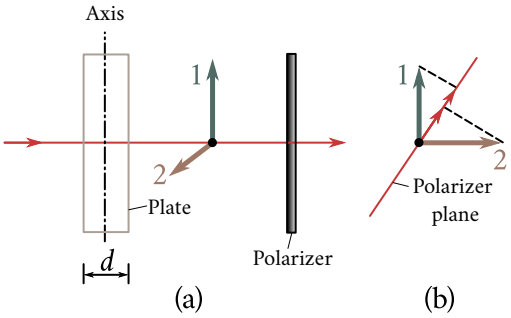


Fig. 19.13: Superposed ordinary and an extraordinary ray emerging from a crystal plate. (a) Light through a crystal plate cut out parallel to the optical axis. Rays 1 and 2 that are polarized in mutually perpendicular planes will emerge from the plate. (b) With a polarizer in the path of these rays, both rays will oscillate in one plane (polarizer plane) after passing through the polarizer.

and minima of the intensity can be obtained. Interference occurs only when the oscillations in the interacting rays occur along the same direction. The oscillations in two rays initially polarized in mutually perpendicular directions can be brought into one plane by passing these rays through a polarizer installed so that its plane does not coincide with the plane of oscillations of any of the rays.

Let us see what happens when an ordinary and an extraordinary ray emerging from a crystal plate are superposed. Assume that the plate has been cut out parallel to an optical axis (Fig. 19.13). With normal incidence of the light on the plate, the ordinary and extraordinary rays will propagate without separating, but with different velocities (see Fig. 19.12c). The following path difference appears between the rays while they pass through the plate:

$$\Delta = (n_o - n_e)d, \tag{19.11}$$

or the following phase difference:

$$\delta = \frac{(n_o - n_e)d}{\lambda_0} 2\pi \tag{19.12}$$

(d is the plate thickness, and λ_0 the wavelength in a vacuum).

Thus, if we pass natural light through a crystal plate cut out parallel to the optical axis (Fig. 19.13a), two rays 1 and 2 that are polarized in mutually perpendicular planes will emerge from the plate[6], and between them there will be a phase difference determined by Eq. (19.12). Let us place a polarizer in the path of these rays. Both rays after passing through the polarizer will oscillate in one plane. Their amplitudes

[6] In the crystal, ray 1 was extraordinary and could be designated by the symbol e, and ray 2 was ordinary (o). Upon emerging from the crystal, these rays lost their right to be called ordinary and extraordinary.

will equal the components of the amplitudes of rays 1 and 2 in the direction of the plane of the polarizer (Fig. 19.13b).

The rays emerging from the polarizer are produced as a result of division of the light obtained from a single source. Therefore, they ought to interfere. If rays 1 and 2 are produced as a result of natural light passing through the plate, however, they do not interfere. The explanation is very simple. Although the ordinary and extraordinary rays are produced by the same light source, they contain mainly oscillations belonging to different wave trains emitted by individual atoms. The oscillations in the ordinary ray are predominantly due to the trains whose oscillation planes are close to one direction in space, whereas those in the extraordinary ray are due to trains whose oscillation planes are close to another direction perpendicular to the first one. Since the individual trains are incoherent, the ordinary and extraordinary rays produced from natural light, and, consequently, rays 1 and 2 too, are also incoherent.

Matters are different if plane-polarized light falls on a crystal plate. In this case, the oscillations of each train are divided between the ordinary and extraordinary rays in the same proportion (depending on the orientation of an optical axis of the plate relative to the plane of oscillations in the incident ray). Consequently, rays o and e, and therefore rays 1 and 2 too, will be coherent and will interfere.

19.5. Passing of Plane-Polarized Light Through a Crystal Plate

Let us consider a crystal plate cut out parallel to an optical axis. We saw in the preceding section that when plane-polarized light falls on such a plate, the ordinary and extraordinary rays are coherent. At the entrance to the plate, the phase difference δ of these rays is zero, and at the exit from the plate

$$\delta = \frac{\Delta}{\lambda_0} 2\pi = \frac{(n_o - n_e)d}{\lambda_0} 2\pi \tag{19.13}$$

[see Eqs. (19.11) and (19.12); we assume that the light falls on the plate normally].

A plate cut out parallel to an optical axis for which

$$(n_o - n_e)d = m\lambda_0 + \frac{\lambda_0}{4}$$

(m is any integer or zero) is called a **quarter-wave plate**. An ordinary and an extraordinary rays passing through such a plate acquire a phase difference equal to $\pi/2$ (we remind our reader that the phase difference is determined with an accuracy to $2\pi m$). A plate for which

$$(n_o - n_e)d = m\lambda_0 + \frac{\lambda_0}{2}$$

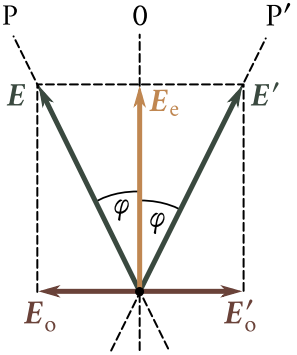


Fig. 19.14: A half-wave plate turns the plane of oscillations of the light passing through it through the angle 2φ (φ is the angle between the plane of oscillations in the incident ray and the axis of the plate). This means that passing through the plate the phase difference between the oscillations of $\boldsymbol{E}_\mathrm{o}$ and $\boldsymbol{E}_\mathrm{e}$ changes by π.

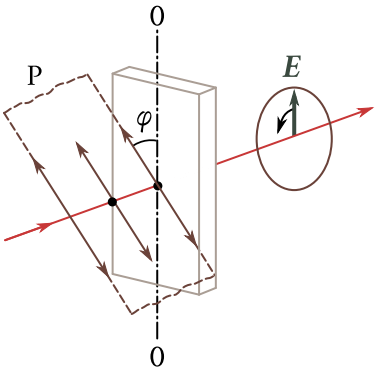


Fig. 19.15: For a plane-polarized light through a quarter-wave plate at $\varphi = 45°$, the amplitudes of both rays emerging from the plate will be the same (no dichroism), with phase shift of $\pi/2$, and the light will be circularly polarized. At a different value of the φ, the amplitudes of the rays emerging from the plate will be different. These rays when superposed form elliptically polarized light.

is called a **half-wave plate**, etc.

Let us see how plane-polarized light passes through a half-wave plate. The oscillation of $\boldsymbol{E}$ in the incident ray occurring in plane P produces the oscillation of $\boldsymbol{E}_\mathrm{o}$ of the ordinary ray and the oscillation of $\boldsymbol{E}_\mathrm{e}$ of the extraordinary ray when entering the crystal (Fig. 19.14). During the time spent in passing through the plate, the phase difference between the oscillations of $\boldsymbol{E}_\mathrm{o}$ and $\boldsymbol{E}_\mathrm{e}$ changes by π. Therefore, at the exit from the plate, the phase relation between the ordinary and extraordinary rays will correspond to the mutual arrangement of the vectors $\boldsymbol{E}_\mathrm{e}$ and $\boldsymbol{E}'_\mathrm{o}$ (at the entrance to the plate it corresponded to the mutual arrangement of the vectors $\boldsymbol{E}_\mathrm{e}$ and $\boldsymbol{E}_\mathrm{o}$). Consequently, the light emerging from the plate will be polarized in plane P′. Planes P and P′ are symmetrical relative to optical axis 0 of the plate. Thus, a half-wave plate turns the plane of oscillations of the light passing through it through the angle 2φ (φ is the angle between the plane of oscillations in the incident ray and the axis of the plate).

Now let us pass plane-polarized light through a quarter-wave plate (Fig. 19.15). If we arrange the plate so that the angle φ between plane of oscillations P in the incident ray and plate axis 0 is 45°, the amplitudes of both rays emerging from the plate will be the same (dichroism is assumed to be absent). The phase shift between the oscillations in these rays will be $\pi/2$. Hence, the light emerging from the plate

will be circularly polarized. At a different value of the angle φ, the amplitudes of the rays emerging from the plate will be different. Consequently, these rays when superposed form elliptically polarized light; one of the axes of the ellipse coincides with plate axis 0.

When plane-polarized light is passed through a plate with a fractional number of waves not coinciding with $m+1/4$ or $m+1/2$, two coherent light waves polarized in mutually perpendicular planes will emerge from the plate. Their phase difference is other than $\pi/2$ and other than π. Hence, with any relation between the amplitudes of these waves depending on the angle φ (see Fig. 19.15), elliptically polarized light will be produced at the exit from the plate, and none of the axes of the ellipse will coincide with plate axis 0. The orientation of the ellipse axes relative to axis 0 is determined by the phase difference δ, and also by the ratio of the amplitudes, *i.e.*, by the angle φ between the plane of oscillations in the incident wave and plate axis 0.

We must note that regardless of the plate thickness, when φ is zero or $\pi/2$, only one ray will propagate in the plate (in the first case an extraordinary ray, in the second case an ordinary one) so that at the plate exit the light remains plane-polarized with its plane of oscillations coinciding with P.

If we place a quarter-wave plate in the path of elliptically polarized light and arrange its optical axis along one of the ellipse axes, then the plate will introduce an additional phase difference equal to $\pi/2$. As a result, the phase difference between two plane-polarized waves whose sum is an elliptically polarized wave becomes equal to zero or π, so that the superposition of these waves produces a plane-polarized wave. Hence, a properly turned quarter-wave plate transforms elliptically polarized light into plane-polarized light. This underlies a method by means of which we can distinguish elliptically polarized light from partly polarized light, or circularly polarized light from natural light. The light being studied is passed through a quarter-wave plate and a polarizer placed after it. If the ray being studied is elliptically polarized (or circularly polarized), then by rotating the plate and the polarizer around the direction of the ray, we can achieve complete darkening of the field of vision. If the light is partly polarized (or natural), it is impossible to achieve extinction of the ray being studied with any position of the plate and polarizer.

19.6. A Crystal Plate Between Two Polarizers

Let us place a plate made from a uniaxial crystal cut out parallel to optical axis 0 between polarizers[7] P and P′ (Fig. 19.16). Plane-polarized light of intensity I will emerge from polarizer P. In passing through the plate, the light in the general case

[7]The second polarizer P′ in the direction of ray propagation is also called an analyzer.

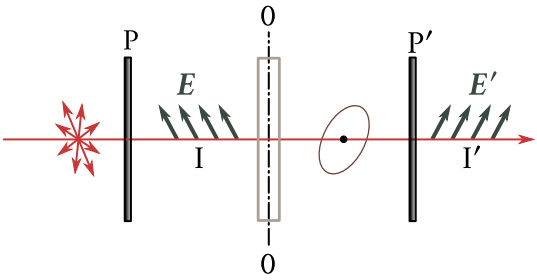


Fig. 19.16: Plate made from a uniaxial crystal cut out parallel to optical axis 0 between polarizers P and P′. Plane-polarized light of intensity I will emerge from polarizer P. In passing through the plate, the light in the general case will become elliptically polarized.

will become elliptically polarized. When it emerges from polarizer P′, the light will again be plane-polarized. Its intensity I' depends on the mutual orientation of the planes of polarizers P and P′ and an optical axis of the plate, and also on the phase difference δ acquired by the ordinary and extraordinary rays when they pass through the plate.

Assume that the angle φ between the plane of polarizer P and plate axis 0 is $\pi/4$. Let us consider two particular cases: the polarizers are parallel (Fig. 19.17a), and they are crossed (Fig. 19.17b). The light oscillation leaving polarizer P will be depicted by the vector $\boldsymbol{E}$ in plane P. At the entrance to the plate, the oscillation of $\boldsymbol{E}$ will produce two oscillations—the oscillation of $\boldsymbol{E}_\mathrm{o}$ (ordinary ray) perpendicular to the optical axis, and the oscillation of $\boldsymbol{E}_\mathrm{e}$ (extraordinary ray) parallel to the axis. These oscillations will be coherent; in passing through the plate, they acquire the phase difference δ that is determined by the plate thickness and the difference between the refractive indices of the ordinary and extraordinary rays. The amplitudes of these oscillations are the same and equal

$$E_\mathrm{o} = E_\mathrm{e} = E\cos\left(\frac{\pi}{4}\right) = \frac{E}{\sqrt{2}}, \tag{19.14}$$

where E is the amplitude of the wave emerging from the first polarizer.

The components of the oscillations of $\boldsymbol{E}_\mathrm{o}$ and $\boldsymbol{E}_\mathrm{e}$ will pass through the second polarizer in the direction of plane P′. The amplitudes of these components in both cases equal those given by Eq. (19.14) multiplied by $\cos(\pi/4)$, *i.e.*,

$$E'_\mathrm{o} = E'_\mathrm{e} = \frac{E}{2}. \tag{19.15}$$

For parallel polarizers (Fig. 19.17a), the phase difference of the waves emerging from polarizer P′ is δ, *i.e.*, the phase difference acquired when passing through the plate. For crossed polarizers (Fig. 19.17b), the projections of the vectors $\boldsymbol{E}_\mathrm{o}$ and $\boldsymbol{E}_\mathrm{e}$ onto the direction of P′ have different signs. This signifies that an additional phase

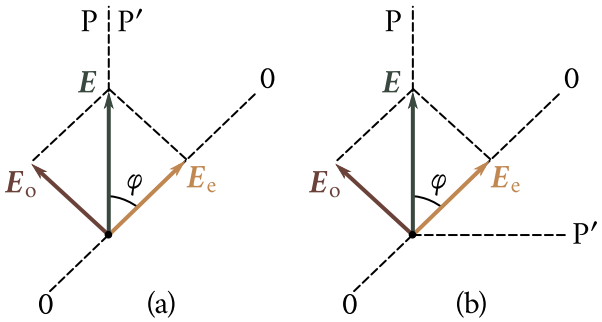


Fig. 19.17: Two particular cases for when the polarizers are parallel (a) and when are crossed (b). The light oscillation leaving polarizer P will be depicted by the vector $\boldsymbol{E}$ in plane P.

difference equal to π appears apart from the phase difference δ.

The waves leaving the second polarizer will interfere. The amplitude $E_{\parallel}$ of the resultant wave for parallel polarizers is determined by the relation

$$E_{\parallel}^2 = E_{\mathrm{o}}'^2 + E_{\mathrm{e}}'^2 + 2E_{\mathrm{o}}'E_{\mathrm{e}}' \cos\delta,$$

and for crossed polarizers by the relation

$$E_{\perp}^2 = E_{\mathrm{o}}'^2 + E_{\mathrm{e}}'^2 + 2E_{\mathrm{o}}'E_{\mathrm{e}}' \cos(\delta + \pi).$$

Taking Eq. (19.15) into consideration, we can write that

$$E_{\parallel}^2 = \frac{1}{4}E^2 + \frac{1}{4}E^2 + 2\frac{1}{4}E^2\cos\delta = \frac{1}{2}E^2(1+\cos\delta) = E^2\cos^2\left(\frac{\delta}{2}\right)$$

$$E_{\perp}^2 = \frac{1}{4}E^2 + \frac{1}{4}E^2 + 2\frac{1}{4}E^2\cos(\delta+\pi) = \frac{1}{2}E^2(1-\cos\delta) = E^2\sin^2\left(\frac{\delta}{2}\right).$$

The intensity is proportional to the square of the amplitude. Hence,

$$I_{\parallel}' = I\cos^2\left(\frac{\delta}{2}\right), \quad I_{\perp}' = I\sin^2\left(\frac{\delta}{2}\right), \tag{19.16}$$

where $I_{\parallel}'$ is the intensity of the light emerging from the second polarizer when the polarizers are parallel, $I_{\perp}'$ is the same intensity when the polarizers are crossed, and I is the intensity of the light that has passed through the first polarizer.

It follows from formulas (19.16) that the intensities $I_{\parallel}'$ and $I_{\perp}'$ are "complementary" —their sum gives the intensity I. In particular, when

$$\delta = 2m\pi \quad (m = 1, 2, \ldots), \tag{19.17}$$

the intensity $I_{\parallel}'$ will equal I, while the intensity $I_{\perp}'$ will vanish. At values of

$$\delta = (2m+1)\pi \quad (m = 0, 1, 2, \ldots), \tag{19.18}$$

on the other hand, the intensity $I_{\parallel}'$ will vanish, while the intensity $I_{\perp}'$ reaches the value I.

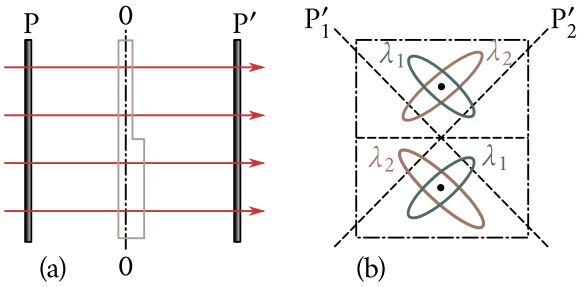


Fig. 19.18: (a) A plate placed between polarizers. The bottom half of the plate is thicker than the top one. The light passing through the plate contains radiation of only two wavelengths λ_1 and λ_2. (b) View from the side of polarizer P′. The light components will be elliptically polarized.

The difference between the refractive indices $n_o - n_e$ depends on the wavelength of the light λ_0. In addition, λ_0 directly enters expression (19.13) for δ. Assume that the light falling on polarizer P consists of radiation of two wavelengths λ_1 and λ_2 such that δ for λ_1 satisfies condition (19.17), and for λ_2 condition (19.18). In this case with parallel polarizers, light of wavelength λ_1 will pass without hindrance through the system depicted in Fig. 19.16, whereas light of wavelength λ_2 will be made completely extinct. With crossed polarizers, light of wavelength λ_2 will pass without hindrance, and light of wavelength λ_1 will be made completely extinct. Consequently, with one arrangement of the polarizers, the colour of the light transmitted through the system will correspond to the wavelength λ_1, and with the other arrangement, to the wavelength λ_2. Such two colours are called **complementary**. When one of the polarizers is rotated, the colour continuously changes, varying during each quarter of a revolution from one complementary colour to the other. A change in colour is also observed at φ differing from $\pi/4$ (but not equal to zero or $\pi/2$), the colours being less saturated, however.

The phase difference δ depends on the plate thickness. Hence, if a doubly refracting transparent plate placed between polarizers has a different thickness at different places, the latter when observed from the side of polarizer P′ will seem to be coloured differently. When polarizer P′ is rotated, these colours change, each of them transforming into its complementary colour. Let us explain this by the following example. Figure 19.18a shows a plate placed between polarizers. The bottom half of the plate is thicker than the top one. Assume that the light passing through the plate contains radiation of only two wavelengths λ_1 and λ_2. Figure 19.18b gives a "view" from the side of polarizer P′. At the exit from the crystal plate, each of the light components will, generally speaking, be elliptically polarized. The orientation and the eccentricity of the ellipses for the wavelengths λ_1 and λ_2 , and

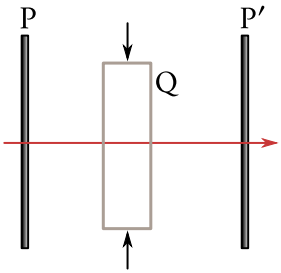


Fig. 19.19: Glass plate Q between crossed polarizers P and P′. Without glass deformation, there is no transmission of light. Compressing the plate light begins to pass through the system and the pattern observed in the transmitted rays being speckled with coloured fringes. Each fringe corresponds to identically deformed spots on the plate.

also for different halves of the plate, will be different. When the plane of polarizer P′ is placed in position P'_1, in the light transmitted through P′ the wavelength λ_1 will predominate in the top half of the plate and the wavelength λ_2 in the bottom half. Therefore, the two halves will be coloured differently. When polarizer P′ is placed in position P'_2, the colour of the top half will be determined by the light of wavelength λ_2, and of the bottom half by the light of wavelength λ_1. Thus, when polarizer P′ is turned through 90°, the two halves of the plate exchange colours, as it were. It is quite natural that this will occur only at a definite ratio of the thicknesses of the two halves of the plate.

19.7. Artificial Double Refraction

External action may cause double refraction to appear in transparent amorphous bodies, and also in crystals of the cubic system. This occurs, in particular, upon the mechanical deformations of bodies. The difference between the refractive indices of an ordinary and an extraordinary ray is a measure of the appearance of optical anisotropy. Experiments show that this difference is proportional to the stress σ at a given point of a body (*i.e.*, to the force per unit area; see Sec. 2.9 of Vol. I):

$$n_o - n_e = k\sigma \tag{19.19}$$

(k is a proportionality constant depending on the properties of the substance).

Let us place glass plate Q between crossed polarizers P and P′ (Fig. 19.19). As long as the glass is not deformed, such a system transmits no light. If the plate is subjected to compression, light begins to pass through the system, the pattern observed in the transmitted rays being speckled with coloured fringes. Each fringe corresponds to identically deformed spots on the plate. Consequently, the distribution of the fringes makes it possible to assess the distribution of the stresses inside the plate.

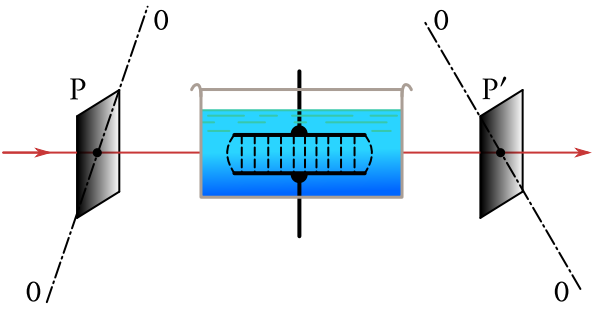


Fig. 19.20: Kerr effect in liquids. A Kerr cell is placed between crossed polarizers P and P′. The Kerr cell contains liquid into which capacitor plates have been introduced. When a voltage is applied across the plates, a virtually homogeneous electric field is set up between them. Thus, the liquid acquires the properties of a uniaxial crystal with an optical axis oriented along the field.

This underlies the optical method of studying stresses. A model of a component or structural member made from a transparent isotropic material (for example, from Plexiglas) is placed between crossed polarizers. The model is subjected to the action of loads similar to those which the article itself will experience. The pattern observed in transmitted white light makes it possible to determine the distribution of the stresses and also to estimate their magnitude.

The appearance of double refraction in liquids and amorphous solids under the action of an electric field was discovered by the Scotch physicist John Kerr (1824-1907) in 1875. This effect was named the **Kerr effect** after its discoverer. In 1930, it was also observed in gases. An arrangement for studying the Kerr effect in liquids is shown schematically in Fig. 19.20. It consists of a **Kerr cell** placed between crossed polarizers P and P′. A Kerr cell is a sealed vessel containing a liquid into which capacitor plates have been introduced. When a voltage is applied across the plates, a virtually homogeneous electric field is set up between them. Under its action, the liquid acquires the properties of a uniaxial crystal with an optical axis oriented along the field.

The resulting difference between the refractive indices n_o and n_e is proportional to the square of the field strength E:

$$n_o - n_e = kE^2. \tag{19.20}$$

The path difference

$$\Delta = (n_o - n_e)l = klE^2$$

appears between the ordinary and extraordinary rays along the path l. The corre-

sponding phase difference is

$$\delta = \frac{\Delta}{\lambda_0} 2\pi = 2\pi \frac{k}{\lambda_0} l E^2.$$

The latter expression is conventionally written in the form

$$\delta = 2\pi B l E^2, \tag{19.21}$$

where B is a quantity characteristic of a given substance and known as the **Kerr constant**.

The Kerr constant depends on the temperature of a substance and on the wavelength of the light. Among known liquids, nitrobenzene ($C_6H_5N_{02}$) has the highest Kerr constant.

The Kerr effect is explained by the different polarization of molecules in various directions. In the absence of a field, the molecules are oriented chaotically, therefore a liquid as a whole displays no anisotropy. Under the action of a field, the molecules turn so that either their electric dipole moments (in polar molecules) or their directions of maximum polarization (in non-polar molecules) are oriented in the direction of the field. As a result, the liquid becomes optically anisotropic. The thermal motion of the molecules counteracts the orienting action of the field. This explains the reduction in the Kerr constant with elevation of the temperature.

The time during which the prevailing orientation of the molecules sets in (when the field is switched on) or vanishes (when the field is switched off) is about 10^{-10} s. Therefore, a Kerr cell placed between crossed polarizers can be used as a virtually inertialess light shutter. In the absence of a voltage across the capacitor plates, the shutter will be closed. When the voltage is switched on, the shutter transmits a considerable part of the light falling on the first polarizer.

19.8. Rotation of Polarization Plane

Natural Rotation. Some substances known as optically active ones have the ability of causing rotation of the plane of polarization of plane-polarized light passing through them. Such substances include crystalline bodies (for example, quartz, cinnabar), pure liquids (turpentine, nicotine), and solutions of optically active substances in inactive solvents (aqueous solutions of sugar, tartaric acid, etc.).

Crystalline substances rotate the plane of polarization to the greatest extent when the light propagates along the optical axis of the crystal. The angle of rotation φ is proportional to the path l travelled by a ray in the crystal:

$$\varphi = \alpha l. \tag{19.22}$$

The coefficient α is called the **rotational constant**. It depends on the wavelength (dispersion of the ability to rotate).

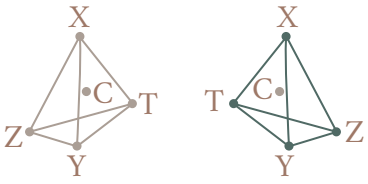


Fig. 19.21: Optically active substances exist in two varieties: right-hand and left-hand. The molecules or crystals of a right-hand substance are a mirror image of the molecules or crystals of the left-hand. The symbols C, X, Y, Z, and T stand for atoms or groups of atoms (radicals) differing from one another. Molecule (b) is a mirror image of molecule (a).

In solutions, the angle of rotation of the plane of polarization is proportional to the path of the light in the solution and to the concentration of the active substance, c:

$$\varphi = [\alpha] cl. \tag{19.23}$$

Here. $[\alpha]$ is a quantity called the **specific rotational constant**.

Depending on the direction of rotation of the polarization plane, optically active substances are divided into **right-hand** and **left-hand** ones. The direction of rotation (relative to a ray) does not depend on the direction of the ray. Consequently, if a ray that has passed through an optically active crystal along its optical axis is reflected by a mirror and made to pass through the crystal again in the opposite direction, then the initial position of the polarization plane is restored.

All optically active substances exist in two varieties—right-hand and left-hand. There exist right-hand and left-hand quartz, right-hand and left-hand sugar, etc. The molecules or crystals of one variety are a mirror image of the molecules or crystals of the other one (Fig. 19.21). The symbols C, X, Y, Z, and T stand for atoms or groups of atoms (radicals) differing from one another. Molecule (b) is a mirror image of molecule (a). If we look at the tetrahedron depicted in Fig. 19.21 along the direction CX, then in clockwise circumvention we shall encounter the sequence ZYTZ for molecule (a) and ZTYZ for molecule (b). The same is observed for any of the directions CY, CZ, and CT. The alternation of the radicals X, Y, Z, T in molecule (b) is the opposite of their alternation in molecule (a). Consequently, if, for example, a substance formed of molecules (a) is right-hand, then one formed of molecules (b) is left-hand.

If we place an optically active substance (a crystal of quartz, a transparent tray with a sugar solution, etc.) between two crossed polarizers, then the field of vision becomes bright. To get darkness again, one of the polarizers has to be rotated through the angle φ determined by expression (19.22) or (19.23). When a

solution is used, we can determine its concentration c by Eq. (19.23) if we know the specific rotational constant $[\alpha]$ of the given substance and the length l and have measured the angle of rotation φ. This way of determining the concentration is used in the production of various substances, in particular in the sugar industry (the corresponding instrument is called a saccharimeter).

Magnetic Rotation of the Polarization Plane. Optically inactive substances acquire the ability of rotating the plane of polarization under the action of a magnetic field. This phenomenon was discovered by Michael Faraday and is therefore sometimes called the **Faraday effect**. It is observed only when light propagates along the direction of magnetization. Therefore, to observe the Faraday effect, holes are drilled in the pole shoes of an electromagnet, and a light ray is passed through them. The substance being studied is placed between the poles of the electromagnet.

The angle of rotation of the polarization plane φ is proportional to the distance l travelled by the light in the substance and to the magnetization of the latter. The magnetization, in turn, is proportional to the magnetic field strength H [see Eq. (7.14)]. We can therefore write that

$$\varphi = VlH. \tag{19.24}$$

The coefficient V is known as the **Verdet constant** or the **specific magnetic rotation**. The constant V, like the rotational constant α, depends on the wavelength.

The direction of rotation is determined by the direction of the magnetic field. The sign of rotation does not depend on the direction of the ray. Therefore, if we reflect the ray from a mirror and make it pass through the magnetized substance again in the opposite direction, the rotation of the plane of polarization will double.

The magnetic rotation of the polarization plane is due to the precession of the electron orbits (see Sec. 7.7) produced under the action of the magnetic field.

Optically active substances when acted upon by a magnetic field acquire an additional ability of rotating the plane of polarization that is added to their natural ability.

Chapter 20

INTERACTION OF ELECTROMAGNETIC WAVES WITH A SUBSTANCE

20.1. Dispersion of Light

By the **dispersion of light** are meant phenomena due to the dependence of the refractive index of a substance on the length of the light wave. This dependence can be characterized by the function

$$n = f(\lambda_0), \tag{20.1}$$

where λ_0 is the length of a light wave in a vacuum.

The derivative of n with respect to λ_0 is called the **dispersion of a substance**.

Function (20.1) for all transparent colourless substances in the visible part of the spectrum has the nature shown in Fig. 20.1. Diminishing of the wavelength is attended by an increase in the refractive index at a constantly growing rate. Hence, the dispersion of a substance $dn/d\lambda_0$ is negative. Its absolute value increases when λ_0 decreases.

If a substance absorbs part of the rays, then the course of dispersion displays an anomaly in the region of absorption and near it (see Fig. 20.6). On a certain section, the dispersion of the substance $dn/d\lambda_0$ will be positive. Such a variation of n with λ_0 is called **anomalous dispersion**.

Media having the property of dispersion are known as **dispersing** ones. In these media, the speed of light waves depends on the wavelength λ_0 or the frequency ω.

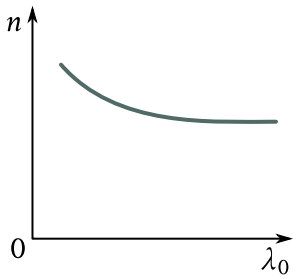


Fig. 20.1: Dispersion curve of a substance for all transparent colourless substances in the visible part of the spectrum.

20.2. Group Velocity

Strictly monochromatic light of the kind

$$E = A\cos(\omega t - kx + \alpha) \tag{20.2}$$

is an infinite sequence in time and space of "crests" and "valleys" propagating along the x-axis with the phase velocity

$$v = \frac{\omega}{k} \tag{20.3}$$

[see Eq. (19.4)]. We cannot use such a wave to transmit a signal because each following crest differs in no way from the preceding one. To transmit a signal, we must put a "mark" on the wave, say, interrupt it for a certain time Δt. In this case, however, the wave will no longer be described by Eq. (20.2).

It is the simplest to transmit a signal with the aid of a light pulse (Fig. 20.2). According to the Fourier theorem, such a pulse can be represented as the superposition of waves of the kind given by Eq. (20.2) having frequencies confined within a certain interval $\Delta\omega$. A superposition of waves differing only slightly from one another in frequency is called a **wave packet** or a **wave group**. The analytical expression for a wave packet has the form

$$E(x,t) = \int_{\omega_0-\Delta\omega/2}^{\omega_0+\Delta\omega/2} A_\omega \cos(\omega t - k_\omega x + \alpha_\omega)\,\mathrm{d}\omega \tag{20.4}$$

(the subscript ω used with A, k, and α indicates that these quantities differ for different frequencies). With a fixed value of t, a plot of function (20.4) has the form shown in Fig. 20.2. When t changes, the graph becomes displaced along the x-axis. Within the limits of a packet, plane waves amplify one another to a greater or smaller extent. Outside these limits, they virtually completely annihilate one another.

The relevant calculations show that the smaller the width of a packet Δx, the greater is the interval of frequencies $\Delta\omega$ or accordingly the greater is the interval

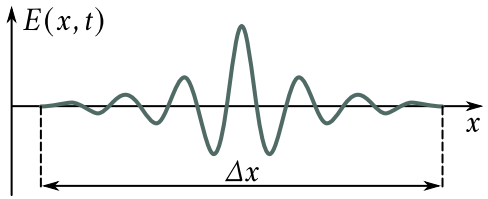


Fig. 20.2: Light pulse for transmitting a signal, Eq. (20.4), with a fixed value of t. When t changes, the graph becomes displaced along the x-axis. Within the limits of a packet, plane waves amplify one another to a greater or smaller extent. Outside these limits, they virtually completely annihilate one another.

of wave numbers Δk needed to describe a packet with the aid of Eq. (20.4). The following relation holds:

$$\Delta k \Delta x \approx 2\pi. \tag{20.5}$$

We must stress the fact that for the superposition of waves described by Eq. (20.4) to be considered a wave packet, the condition $\Delta\omega \ll \omega_0$ must be obeyed.

In a non-dispersing medium, all the plane waves forming a packet propagate with the same phase velocity v. It is evident that in this case the velocity of the packet coincides with v, and the shape of the packet does not change with time. It can he shown that a packet spreads in a dispersing medium with time—its width grows. If the dispersion is not great, spreading of the packet is not too fast. In this case, we can say that the packet travels with the velocity u, by which we mean the velocity of the centre of the packet, *i.e.*, of the point with the maximum value of E. This velocity is called the **group velocity**. In a dispersing medium, the group velocity u differs from the phase velocity v (here we mean the phase velocity of the harmonic component with the maximum amplitude, in other words, the phase velocity for the dominating frequency). We shall show below that when $\mathrm{d}n/\mathrm{d}\lambda_0 < 0$, the group velocity is smaller than the phase one ($u < v$); when $\mathrm{d}n/\mathrm{d}\lambda_0 > 0$, the group velocity is greater than the phase one ($u > v$).

Figure 20.3 shows "photographs" of a wave packet for three consecutive moments t_1, t_2, and t_3. The figure is for the case when $u < v$. Inspection of the figure shows that motion of the packet is attended by motion of the crests and valleys "inside" it. New crests constantly appear at the left-hand boundary of the packet. After travelling along the packet, they vanish at its right-hand boundary. Hence, whereas the packet as a whole travels with the velocity u, the individual crests and valleys travel with the velocity v.

When $u > v$, the directions of motion of the packet and of the crests inside it are opposite.

Let us explain what has been said above using the example of the superposition

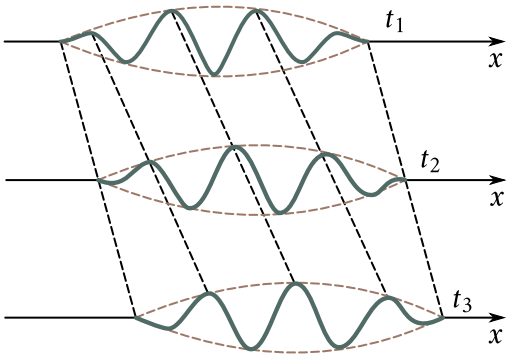


Fig. 20.3: "Photographs" of a wave packet for three consecutive moments t_1, t_2, and t_3, for $u < v$. The motion of the packet is attended by motion of the crests and valleys "inside" it. New crests constantly appear at the left-hand boundary of the packet. After travelling along the packet, they vanish at its right-hand boundary. Hence, whereas the packet as a whole travels with the velocity u, the individual crests and valleys travel with the velocity v.

of two plane waves of the same amplitude and of different wavelengths λ. Figure 20.4 gives an "instant photograph" of the waves. One of them is shown by a solid line, and the other by a dash line. The intensity is the greatest at point A where the phases of the two waves coincide at the given moment. At points B and C, the two waves are in counterphase, owing to which the intensity of the resultant wave is zero. Assume that both waves are propagating from left to right, the velocity of the "solid" wave being lower than that of the "dash" one (here $dn/d\lambda > 0$ and, consequently, $dn/d\lambda < 0$).

Thus, the place at which the waves amplify each other will move to the left with time relative to the waves. As a result, the group velocity will be lower than the phase value. If the velocity of the "solid" wave is greater than that of the "dash" one (*i.e.*, $dn/d\lambda > 0$), the place at which amplification of the waves occurs will move to the right so that the group velocity will be greater than the phase one.

Let us write the equations of the waves, assuming for simplicity that the initial phases equal zero:

$$E_1 = A\cos(\omega t - kx)$$

$$E_2 = A\cos[(\omega + \Delta\omega)t - (k + \Delta k)x].$$

Here $k = \omega/v_1$, and $(k+\Delta k) = (\omega+\Delta\omega)/v_2$. Assume that $\Delta\omega \ll \omega$, hence, $\Delta k \ll k$. Now, summating the oscillations and performing transformations according to the formula for the sum of cosines, we get

$$E = E_1 + E_2 = \left[2A\cos\left(\frac{\Delta\omega}{2}t - \frac{\Delta k}{2}x\right)\right]\cos(\omega t - kx) \qquad (20.6)$$

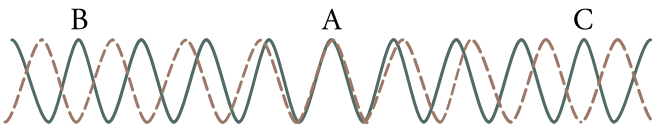


Fig. 20.4: "Instant photograph" of two waves. The intensity is the greatest at point A where the phases of the two waves coincide at the given moment. At points B and C, the two waves are in counterphase, owing to which the intensity of the resultant wave is zero.

(in the second multiplier, we have disregarded $\Delta\omega$ in comparison with ω and Δk in comparison with k).

The multiplier in brackets varies much more slowly with x and t than the second multiplier. We can therefore consider expression (20.6) as the equation of a plane wave whose amplitude varies according to the law[1]

$$\text{Amplitude} = \left|2A\cos\left(\frac{\Delta\omega}{2}t - \frac{\Delta k}{2}x\right)\right|.$$

In the given case, there is a number of identical amplitude maxima determined by the condition

$$\frac{\Delta\omega}{2}t - \frac{\Delta k}{2}x_{\max} = \pm m\pi \quad (m = 0, 1, 2, \ldots). \tag{20.7}$$

Each of these maxima can be considered as the centre of the relevant wave packet.

Solving Eq. (20.7) relative to $x_{\max}$ we get

$$x_{\max} = \frac{\Delta\omega}{\Delta k}t + \text{constant}.$$

It thus follows that the maxima travel with the velocity

$$u = \frac{\Delta\omega}{\Delta k}. \tag{20.8}$$

The expression obtained is the group velocity for a packet formed by two components.

Let us find the velocity with which the centre of a wave packet described by expression (20.5) travels. Passing over from cosines to exponents, we get

$$E(x,t) = \int_{\omega_0-\Delta\omega/2}^{\omega_0+\Delta\omega/2} \hat{A}_\omega \exp[i(\omega t - k_\omega x)]\,\mathrm{d}\omega \tag{20.9}$$

[$\hat{A}_\omega = A_\omega \exp(i\alpha_\omega)$ is the complex amplitude].

Let us expand the function $k_\omega = k(\omega)$ into a series in the vicinity of ω_0:

$$k_\omega = k_0 + \left(\frac{\mathrm{d}k}{\mathrm{d}\omega}\right)_0 (\omega - \omega_0) + \ldots. \tag{20.10}$$

[1] Compare with Eqs. (7.86) and (7.87) of Vol. I. The dependence of function (20.6) on x at a fixed value of t is depicted by a curve similar to the one in Fig. 7.11a of Vol. 1.

Here, $k_0 = k(\omega)_0$, and $(\mathrm{d}k/\mathrm{d}\omega)_0$ is the value of the derivative at point ω_0.

We shall introduce the variable $\xi = \omega - \omega_0$. Hence, $\omega = \omega_0 + \xi$ and $\mathrm{d}\omega = \mathrm{d}t$. Performing such a substitution in Eq. (20.9) and introducing the value of k_ω from Eq. (20.10), we can write

$$E(x,t) = \exp[i(\omega_0 t - k_0 x)] \int_{-\Delta\omega/2}^{+\Delta\omega/2} \hat{A}_\xi \exp\left\{ i \left[t - \left(\frac{\mathrm{d}k}{\mathrm{d}\omega} \right)_0 x \right] \xi \right\} \mathrm{d}\xi. \quad (20.11)$$

We have arrived at an equation of a plane wave of frequency ω_0, wave number k_0, and complex amplitude

$$\hat{A}(x,t) = \int_{-\Delta\omega/2}^{+\Delta\omega/2} \hat{A}_\xi \exp\left\{ i \left[t - \left(\frac{\mathrm{d}k}{\mathrm{d}\omega} \right)_0 x \right] \xi \right\} \mathrm{d}\xi. \quad (20.12)$$

It can be seen from Eq. (20.12) that the equation

$$t - \left(\frac{\mathrm{d}k}{\mathrm{d}\omega} \right)_0 x = \text{constant}, \quad (20.13)$$

relates the time t and the coordinate x of the plane in which the complex amplitude has a given fixed value, in particular including a value such that the magnitude of the complex amplitude, *i.e.*, the conventional amplitude $A(x,t)$, reaches a maximum.

Taking into account that $1/(\mathrm{d}k/\mathrm{d}\omega)_0 = (\mathrm{d}\omega/\mathrm{d}k)_0$, we can write Eq. (20.13) in the form

$$x_{\max} = \left(\frac{\mathrm{d}k}{\mathrm{d}\omega} \right)_0 t - \text{constant}' \qquad \left[\text{constant}' = \frac{\text{constant}}{(\mathrm{d}k/\mathrm{d}\omega)_0} \right]. \quad (20.14)$$

It follows from Eq. (20.14) that the place where the amplitude of a wave packet is maximum travels with the velocity $(\mathrm{d}\omega/\mathrm{d}k)_0$. We thus arrive at the following expression for the group velocity:

$$u = \frac{\mathrm{d}\omega}{\mathrm{d}k} \quad (20.15)$$

(the subscript 0 is no longer needed and has been omitted). We previously obtained a similar expression for a packet of two waves [see Eq. (20.8)]. We remind our reader that we have disregarded the terms of higher orders of smallness in expansion (20.8). In this approximation, the shape of the wave packet does not change with time. If we take into account the following terms of the expansion, then we get an expression for the amplitude from which it follows that the width of a packet grows with time—a wave packet broadens.

We can give a different form to the expression for the group velocity. Substituting vk for ω [see Eq. (20.3)], we can write Eq. (20.14) as follows:

$$u = \frac{\mathrm{d}(vk)}{\mathrm{d}k} = v + k\frac{\mathrm{d}v}{\mathrm{d}k}. \quad (20.16)$$

We shall further write

$$\frac{dv}{dk} = \frac{dv}{d\lambda}\frac{d\lambda}{dk}.$$

We find from the relation $\lambda = 2\pi/k$ that $d\lambda/dk = -2\pi/k^2 = -\lambda/k$. Accordingly, $dv/dk = -(dv/d\lambda)(\lambda/k)$. Using this value in Eq. (20.16), we get

$$u = v - \lambda\frac{dv}{d\lambda}. \tag{20.17}$$

A glance at this formula shows that the group velocity u can be either smaller or greater than the phase velocity v, depending on the sign of $dv/d\lambda$. In the absence of dispersion, $dv/d\lambda = 0$, and the group velocity coincides with the phase one.

The maximum of the intensity falls to the centre of a wave packet. Therefore, when the concept of group velocity has a meaning, the velocity of energy transfer by a wave equals the group velocity.

The concept of group velocity may be applied only provided that the absorption of the wave energy in the given medium is not great. With considerable attenuation of the waves, the concept of group velocity loses its meaning. This occurs in the region of anomalous dispersion. In this region, the absorption is very great, and the concept of group velocity cannot be applied.

20.3. Elementary Theory of Dispersion

The dispersion of light can be explained on the basis of the electromagnetic theory and the electron theory of a substance. For this purpose, we must consider the process of interaction of light with a substance. The motion of the electrons in an atom obeys the laws of quantum mechanics. In particular, the concept of the trajectory of an electron in an atom loses all meaning. As Lorentz showed, however, it is sufficient to restrict ourselves to the hypothesis on the existence of electrons bound quasi-elastically within atoms for a qualitative understanding of many optical phenomena. When brought out of their equilibrium position, such electrons will begin to oscillate, gradually losing the energy of oscillation on the emission of electromagnetic waves. As a result, the oscillations will be damped. The attenuation can be taken into account by introducing the "force of friction of emission" proportional to the velocity.

When an electromagnetic wave passes through a substance, every electron experiences the action of the Lorentz force

$$\boldsymbol{F} = -e\boldsymbol{E} - e(\boldsymbol{v}\times\boldsymbol{B}) = -e\boldsymbol{E} - e\mu_0(\boldsymbol{v}\times\boldsymbol{H}) \tag{20.18}$$

[see Eq. (6.35); the charge of an electron is $-e$]. According to Eq. (15.23), the ratio of the magnetic and electric field strengths in a wave is $H/E = \sqrt{\varepsilon_0/\mu_0}$. Hence, from

Eq. (20.18), we get the following value for the ratio of the magnetic and electric forces exerted on an electron

$$\frac{\mu_0 v H}{E} = \mu_0 v \left(\frac{\varepsilon_0}{\mu_0}\right)^{1/2} = v\sqrt{\varepsilon_0 \mu_0} = \frac{v}{c}.$$

Even if the amplitude a of electron oscillations reached a value of the order of 1 Å (10^{-10} m), *i.e.*, of the order of an atom's dimensions, the amplitude of the velocity of an electron $a\omega$ would be about $10^{-10} \times 3 \times 10^{15} = 3 \times 10^5$ m s^{-1} [according to Eq. (16.6), $\omega = 2\pi\nu$ equals about 3×10^{15} rad s^{-1}]. Thus, the ratio v/c is clearly less than 10^{-3} so that we may disregard the second addend in Eq. (20.18).

We can thus consider that when an electromagnetic wave passes through a substance, every electron experiences the force

$$F = -eE_0 \cos(\omega t + \alpha)$$

(α is a quantity determined by the coordinates of a given electron, and E_0 is the amplitude of the electric field strength of the wave).

To simplify our calculations, we shall first disregard the attenuation due to emission. We shall subsequently take the attenuation into account by introducing the relevant corrections into the formulas obtained. The equation of motion of an electron in this case has the form

$$\ddot{r} + \omega_0^2 r = -\frac{e}{m} E_0 \cos(\omega t + \alpha)$$

[see Eq. (7.13) of Vol. I,]; ω_0 is the natural frequency of oscillations of an electron). Let us add $-i(e/m)E_0 \sin(\omega t + \alpha)$ to the right-hand side of this equation and thus pass over to the complex functions $\hat{E}$ and $\hat{r}$:

$$\frac{\mathrm{d}^2 \hat{r}}{\mathrm{d}t^2} + \omega^2 \hat{r} = -\frac{e}{m} \hat{E}_0 \exp(i\omega t). \tag{20.19}$$

Here, $\hat{E}_0 = E_0 \exp(i\alpha)$ is the complex amplitude of the electric field of a wave.

We shall seek a solution of the equation in the form $\hat{r} = \hat{r}_0 \exp(i\omega t)$, where ω is the complex amplitude of oscillations of an electron. Accordingly, $\mathrm{d}^2\hat{r}/\mathrm{d}t^2 = -\omega^2 \hat{r}_0 \exp(i\omega t)$. Introducing these expressions into Eq. (20.19) and cancelling out the common factor $\exp(i\omega t)$, we arrive at the expression

$$-\omega^2 \hat{r}_0 + -\omega_0^2 \hat{r}_0 = -\frac{e}{m} \hat{E}_0,$$

whence

$$\hat{r}_0 = \frac{-(e/m)\hat{E}_0}{(\omega_0^2 - \omega^2}.$$

Multiplying the equation obtained by $\exp(i\omega t)$, we obtain

$$\hat{r}(t) = \frac{-(e/m)\hat{E}(t)}{(\omega_0^2 - \omega^2}.$$

Finally, taking the real parts of the complex functions $\hat{r}$ and $\hat{E}$, we find r as a function of t:

$$r(t) = \frac{-(e/m)E(t)}{(\omega_0^2 - \omega^2}. \tag{20.20}$$

To simplify the problem, we shall consider that the molecules are non-polar. In addition, since the masses of nuclei are great in comparison with the mass of an electron, we shall ignore the displacements of the nuclei from the equilibrium positions under the action of the wave field. In this approximation, the dipole electric moment of a molecule can be represented in the form

$$\begin{aligned}\boldsymbol{p}(t) &= \sum_l q_l \boldsymbol{R}_{0,l} + \sum_k e_k [\boldsymbol{r}_{0,k} + \boldsymbol{r}_k(t)] \\ &= \left\{ \sum_l q_l \boldsymbol{R}_{0,l} + \sum_k e_k \boldsymbol{r}_{0,k} \right\} + \sum_k e_k \boldsymbol{r}_k(t) \\ &= \boldsymbol{p}_0 + \sum_k e_k \boldsymbol{r}_k(t) = \sum_k e_k \boldsymbol{r}_k(t),\end{aligned}$$

where q_l and $R_{0,l}$ are the charges and position vectors of the equilibrium positions of the nuclei, e_k and $r_{0,k}$ are the charge and position vector of the equilibrium position of the k-th electron, $\boldsymbol{r}_k(t)$ is the displacement of the k-th electron from its equilibrium position under the action of the wave field, and $\boldsymbol{p}_0$ is the dipole moment of a molecule in the absence of a field, which is assumed to equal zero.

All the $\boldsymbol{r}_k(t)$'s are collinear with $\boldsymbol{E}(t)$. We therefore obtain the following expression for the projection of $\boldsymbol{p}(t)$ onto the direction of $\boldsymbol{E}(t)$:

$$p(t) = \sum_k e_k r_k(t) = \sum_k (-e) r_k(t)$$

(we have taken into account that e_k for all electrons is identical and equals $-e$). Let us introduce into this equation the value of $r(t)$ from Eq. (20.20), taking into consideration that the electrons in a molecule have different natural frequencies $\omega_{0,k}$. As a result, we get

$$p(t) = \sum_k \frac{e^2/m}{(\omega_{0,k}^2 - \omega^2)} E(t). \tag{20.21}$$

Let us denote the number of molecules in unit volume by the symbol N. The product $Np(t)$ gives the polarization $P(t)$ of a substance. According to Eqs. (2.5)

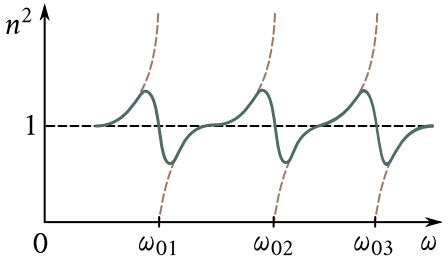


Fig. 20.5: Behaviour of function (20.22) when the friction of emission is disregarded (dashed line) and when is considered (solid line).

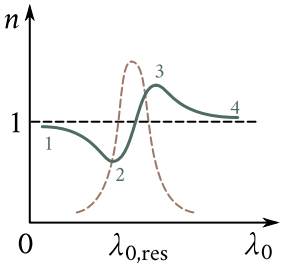


Fig. 20.6: Square root of Fig. 20.5 in terms of λ_0. The dash curve shows how the coefficient of absorption of light by a substance changes.

and (2.20), the permittivity is

$$\varepsilon = 1 + \chi = 1 + \frac{P(t)}{\varepsilon_0 E(t)} = 1 + \frac{N}{\varepsilon_0}\frac{p(t)}{E(t)}.$$

Using in this expression the ratio $p(t)/E(t)$ obtained from Eq. (20.21) and substituting n^2 for e [see Eq. (16.3)], we arrive at the formula

$$n^2 = 1 + \frac{N}{\varepsilon_0}\frac{e^2/m}{(\omega_{0,k}^2 - \omega^2)}. \tag{20.22}$$

At frequencies ω appreciably differing from all the natural frequencies $\omega_{0,k}$, the sum in Eq. (20.22) will be small in comparison with unity, so that $n^2 \approx 1$. Near each of the natural frequencies, function (20.22) is interrupted: when ω tends to $\omega_{0,k}$ from the left, it becomes equal to $+\infty$, and when it tends to $\omega_{0,k}$ from the right, the function becomes equal to $-\infty$ (see the dash curves in Fig. 20.5). Such a behaviour of function (20.22) is due to the fact that we have disregarded the friction of emission [we remind our reader that when friction is disregarded, the amplitude of the forced oscillations in resonance becomes equal to infinity; see Eq. (7.128) of Vol. I]. When the friction of emission is taken into consideration, we get the dependence of n^2 on ω depicted in Fig. 20.5 by the solid curve.

Passing over from n^2 to n and from ω to λ_0, we get the curve shown in Fig. 20.6 (the figure gives only a portion of the curve in the region of one of the resonance wavelengths). The dash curve in this figure shows how the coefficient of absorption of light by a substance changes (see the following section). Segment 3-4 is similar to the curve shown in Fig. 20.1. Segments 1-2 and 3-4 correspond to normal dispersion ($\mathrm{d}n/\mathrm{d}\lambda_0 < 0$). On segment 2-3, the dispersion is anomalous ($\mathrm{d}n/\mathrm{d}\lambda_0 > 0$). In region 1-2, the refractive index is less than unity, hence, the phase velocity of the wave exceeds c. This circumstance does not contradict the theory of relativity, which is based on the statement that the velocity of transmitting a signal cannot exceed c. In

the preceding section, we found that it is impossible to transmit a signal with the aid of an ideally monochromatic wave. Energy (*i.e.*, a signal) is transmitted with the aid of a not completely monochromatic wave (wave packet), however, with a velocity equal to the group velocity determined by Eq. (20.17). In the region of normal dispersion, $\mathrm{d}n/\mathrm{d}\lambda > 0$ ($\mathrm{d}n$ and $\mathrm{d}\lambda$ have different signs, while $\mathrm{d}n/\mathrm{d}\lambda < 0$), so that although $v > c$, the group velocity is less than c. In the region of anomalous dispersion, the concept of group velocity loses its meaning (the absorption is very great). Therefore, the value of u calculated by Eq. (20.17) will not characterize the rate of energy transmission. The relevant calculations give a value less than c for the velocity of energy transmission in this case too.

20.4. Absorption of Light

When a light wave passes through a substance, part of the wave energy is spent for producing oscillations of the electrons. This energy is partly returned to the radiation in the form of the secondary wavelets set up by the electrons; it is partly transformed, however, into the energy of motion of the atoms, *i.e.*, into the internal energy of the substance. This is the reason why the intensity of light transmitted through a substance diminishes—light is absorbed in the substance. The forced oscillations of the electrons and therefore the absorption of light become especially intensive at the resonance frequency (see the dash absorption curve in Fig. 20.6).

Experiments show that the intensity of light when it passes through a substance diminishes according to the exponential law

$$I = I_0 e^{-\varkappa l}. \tag{20.23}$$

Here, I_0 is the intensity of light at the entrance to the absorbing layer (on its boundary or at a certain place inside the substance), l is the thickness of the layer, and $\varkappa$ is the constant depending on the properties of the absorbing substance and called the **absorption coefficient**.

Equation (20.23) is known as **Bouguer's law**[2] [in honour of the French scientist Pierre Bouguer (1698-1758)].

Differentiation of Eq. (20.23) yields

$$\mathrm{d}I = -\varkappa I_0 e^{-\varkappa l}\,\mathrm{d}l = -\varkappa I\,\mathrm{d}l. \tag{20.24}$$

It follows from this expression that the decrement of the intensity along the path $\mathrm{d}l$ is proportional to the length of this path and to the value of the intensity itself. The absorption coefficient is the constant of proportionality.

Inspection of Eq. (20.23) shows that when $l = 1/\varkappa$, the intensity I is $1/e$-th of I_0.

[2] Also known as the Beer-Lambert law or Beer-Lambert-Bouguer law.

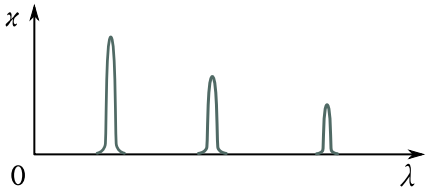


Fig. 20.7: The absorption coefficient of a substance whose atoms or molecules do not virtually act on one another (gases and metal vapours at a low pressure) is close to zero for most wavelengths. It displays sharp maxima (Fig. 20.7) only for very narrow spectral regions (having a width of several hundredths of an angstrom). These maxima correspond to the resonance frequencies of oscillations of the electrons inside the atoms. The molecular frequencies are in the infrared region of the spectrum.

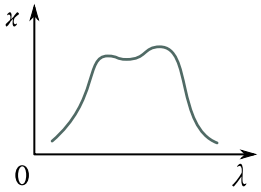


Fig. 20.8: Broad absorption bands of gases at high pressures (also liquids and solids). As the pressure of gases is increased, the absorption maxima expand and at high pressures the absorption spectrum of gases approaches those of liquids. This indicates that the expansion of the absorption bands is the result of the atoms interacting with one another.

Thus, the absorption coefficient is a quantity inversely proportional to the thickness of the layer that reduces the intensity of light passing through it to $1/e$-th of its initial value.

The absorption coefficient depends on the wavelength λ (or the frequency ω). The absorption coefficient of a substance whose atoms or molecules do not virtually act on one another (gases and metal vapours at a low pressure) is close to zero for most wavelengths. It displays sharp maxima (Fig. 20.7) only for very narrow spectral regions (having a width of several hundredths of an angstrom). These maxima correspond to the resonance frequencies of oscillations of the electrons inside the atoms. For polyatomic molecules, frequencies corresponding to the oscillations of the atoms inside the molecules are also detected. Since the masses of atoms are tens of thousands of times greater than the mass of an electron, the molecular frequencies are much smaller than the atomic ones—they are in the infrared region of the spectrum.

Gases at high pressures, and also liquids and solids produce broad absorption bands (Fig. 20.8). As the pressure of gases is increased, the absorption maxima, which are initially very narrow (see Fig. 20.7), expand more and more, and at high pressures the absorption spectrum of gases approaches those of liquids. This fact indicates that the expansion of the absorption bands is the result of the atoms interacting with one another.

Metals are virtually opaque for light ($\varkappa$ for them has a value of the order of 10^6 m^{-1}; for comparison we shall point out that for glass $\varkappa \approx 1$ m^{-1}). This is due to the presence of free electrons in metals. The action of the electric field of a light

wave causes the free electrons to come into motion—fast-varying currents attended by the liberation of Lenz-Joule heat are produced in the metal. As a result, the energy of the light wave rapidly diminishes and transforms into the internal energy of the metal.

20.5. Scattering of Light

From the classical viewpoint, the process of scattering of light consists in that light passing through a substance causes the electrons in the atoms to oscillate. The oscillating electrons produce secondary wavelets that propagate in all directions. This phenomenon should seem to result in the scattering of light in all conditions. The secondary wavelets, however, are coherent, so that their mutual interference must be taken into consideration.

The relevant calculations show that in a homogeneous medium the secondary wavelets completely destroy one another in all directions except for that of propagation of the primary wave. Therefore, no redistribution of the light by directions, *i.e.*, scattering of the light, occurs.

The secondary wavelets do not destroy one another in side directions only when light propagates in a non-homogeneous medium. The light waves become diffracted on the non-homogeneities of the medium and produce a diffraction pattern characterized by a quite uniform distribution of the intensity between all directions. Such diffraction on fine non-homogeneities is called the **scattering of light**.

Media having a clearly expressed optical non-homogeneity are known as **turbid media**. They include (1) smoke, *i.e.*, a suspension of very minute solid particles in a gas, (2) fogs and mists-suspensions of very minute liquid droplets in gases, (3) suspensions formed by solid particles in the bulk of a liquid, (4) emulsions, *i.e.*, suspensions of very minute droplets of one liquid in another one that does not dissolve the first liquid (an example of an emulsion is milk, which is a suspension of droplets of fat in water), and (5) solids such as mother-of-pearl, opals, and milk glass.

Light scattered on particles whose size is considerably smaller than the length of a light wave becomes partly polarized. The explanation is that the oscillations of the electrons produced by the scattered light beam occur in a plane at right angles to the beam (Fig. 20.9). The oscillations of the vector $\boldsymbol{E}$ in a secondary wavelet occur in a plane passing through the direction of oscillations of the charges (see Fig. 15.6). Therefore, the light scattered by the particles in directions normal to the beam will be completely polarized. The scattered light is polarized only partly in directions that make an angle other than a right one with the beam.

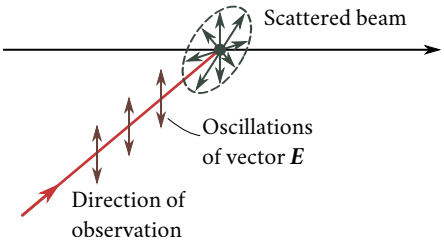


Fig. 20.9: The light scattered by the particles in directions normal to the beam will be completely polarized. The scattered light is polarized only partly in directions that make an angle other than a right one with the beam.

As a result of scattering of the light in side directions, the intensity in the direction of its propagation diminishes more rapidly than when only absorption occurs. Consequently, for a turbid substance, Eq. (20.23) must contain the coefficient $\varkappa'$ due to scattering in addition to the absorption coefficient $\varkappa$:

$$I = I_0 e^{-(\varkappa+\varkappa')l}. \tag{20.25}$$

The constant $\varkappa'$ is called the **extinction coefficient**.

If the dimensions of the non-homogeneities are small in comparison with the length of a light wave (not over $\sim 0.1\lambda$), then the intensity of the scattered light I is proportional to the fourth power of the frequency or is inversely proportional to the fourth power of the wavelength:

$$I \propto \omega^4 \propto \frac{1}{\lambda^4}. \tag{20.26}$$

This relation is known as **Rayleigh's law** after the British physicist John Rayleigh (1842-1919). It is easy to understand its origin if we take into account that the radiant power of an oscillating charge is proportional to the fourth power of the frequency and, consequently, is inversely proportional to the fourth power of the wavelength [see expression (15.46)].

If the dimensions of the non-homogeneities are comparable with the length of a wave, then the electrons at different spots on the non-homogeneities oscillate with an appreciable phase shift. This circumstance makes the phenomenon more complicated and leads to other regularities—the intensity of the scattered light becomes proportional to only the square of the frequency (inversely proportional to the square of the wavelength).

It is simple to observe the manifestation of law (20.26) by passing a beam of white light through a vessel with a turbid liquid (Fig. 20.10). Owing to scattering, the trace of the beam in the liquid is seen very well from a side. Since short light waves are scattered to a much greater extent than the long ones, the trace seems to be bluish. The beam passing through the liquid is enriched with long-wave radiation

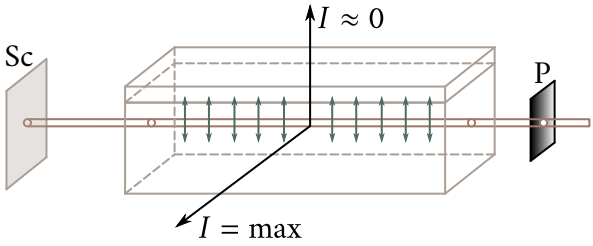


Fig. 20.10: A beam of white light passing through a vessel with a turbid liquid. The beam passing through the liquid is enriched with long-wave radiation and forms a reddish-yellow spot on screen Sc instead of a white one. With a polarizer P at the entrance of the beam to the vessel, we shall find that the intensity of the scattered light in different directions perpendicular to the initial beam is not the same.

and forms a reddish-yellow spot on screen Sc instead of a white one. If we put polarizer P at the entrance of the beam to the vessel, we shall find that the intensity of the scattered light in different directions perpendicular to the initial beam is not the same. The directivity of dipole emission (see Fig. 15.7) results in the fact that in the directions coinciding with the plane of oscillations of the primary beam, the intensity of the scattered light virtually equals zero, while in the directions perpendicular to the plane of the oscillations, the intensity of the scattered light is maximum. By turning the polarizer around the direction of the primary beam, we shall observe alternate amplification and attenuation of the light scattered in the given direction.

Even liquids and gases carefully purified of foreign admixtures and impurities scatter light to some extent. The Soviet physicist Leonid Mandelshtam (1879-1944) and the Polish physicist Marian Smoluchowski (1872-1917) established that the appearance of the optical non-homogeneities is due in this case to fluctuation of the density (*i.e.*, deviations of the density from its mean value observed within the confines of small volumes). These fluctuations are produced by chaotic motion of the molecules of the substance; therefore, the scattering of light due to them is called **molecular**.

Molecular scattering explains the light blue colour of the sky. The places of compression and rarefaction of the air continuously appearing in the atmosphere owing to the random motion of its molecules scatter sunlight. According to law (20.26), the light blue and blue rays are scattered to a greater extent than the yellow and red ones, the result being the light blue colour of the sky. When the Sun is low above the horizon, the rays propagating directly from it pass through a scattering medium of great thickness, and as a result they are enriched with waves of greater lengths. This is why the sky at sunrise and sunset has red tints.

There are especially favourable conditions for the appearance of considerable density fluctuations near the critical state of a substance (at the critical point $dp/dV = 0$; see Sec. 15.4 of Vol. I). These fluctuations result in intensive scattering of light such that a glass ampule with the substance seems to be absolutely black when looked through. This phenomenon is known as **critical opalescence**.

20.6. The Vavilov-Cerenkov Effect

In 1934, the Soviet physicist Pavel Cerenkov (born 1904), working under the supervision of Sergei Vavilov (1891-1951), discovered a special kind of glow of liquids under the action of radium gamma-rays. Vavilov advanced the correct assumption that the fast electrons produced by the gamma-rays are the source of the radiation. This phenomenon was named the **Vavilov-Cerenkov effect**. Its complete theoretical explanation was given in 1937 by the Soviet physicists Igor Tamm (1895-1971) and Ilya Frank (born 1908)[3].

According to the electromagnetic theory, a charge moving uniformly emits no electromagnetic waves (see Sec. 15.6). As Tamm and Frank showed, however, this holds only if the velocity v of a charged particle does not exceed the phase velocity c/n of electromagnetic waves in the medium in which the particle is moving. A particle emits electromagnetic waves even when travelling uniformly provided that $v > c/n$. The particle actually loses energy on radiation owing to which it travels with a negative acceleration. This acceleration is not the cause (as when $v < c/n$), but a consequence of radiation. If the loss of energy at the expense of radiation were replenished in some way or other, a particle travelling uniformly with the velocity $v > c/n$ would nevertheless be a source of radiation.

The Vavilov-Cerenkov effect was observed experimentally for electrons, protons, and mesons travelling in liquid and solid media.

Vavilov-Cerenkov radiation has a light blue colour because short waves predominate in it. The most characteristic feature of this radiation is the fact that it is emitted not in all directions, but only along the generatrices of a cone whose axis coincides with the direction of velocity of the relevant particle (Fig. 20.11). The angle θ between the directions of propagation of the radiation and the velocity vector of a particle is determined by the equation

$$\cos\theta = \frac{c/n}{v} = \frac{c}{nv}. \tag{20.27}$$

The Vavilov-Cerenkov effect finds widespread application in experimental equipment. In the so-called Cerenkov counters, a light pulse produced by a fast

[3] In 1958, Cerenkov, Tamm, and Frank were awarded a Nobel prize for their work.

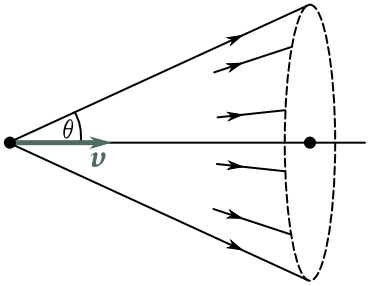


Fig. 20.11: Vavilov-Cerenkov radiation most characteristic feature is that it is emitted not in all directions, but only along the generatrices of a cone whose axis coincides with the direction of velocity of the relevant particle. The angle θ is formed between the directions of propagation of the radiation and the velocity vector of a particle.

charged particle is transformed with the aid of a photomultiplier[4] into a current pulse. To make such a counter function, the energy of a particle must exceed the threshold value determined by the condition $v = c/n$. Therefore, Cerenkov counters make it possible not only to register particles, but also to assess their energy. It is even possible to determine the angle θ between the direction of a flash and the velocity of the particle. This allows us to use Eq. (20.27) to calculate the velocity (and, consequently, also the energy) of a particle.

[4] By a photomultiplier is meant an electronic multiplier whose first electrode (a photocathode) is capable of emitting electrons under the action of light.

Chapter 21
MOVING-MEDIA OPTICS

21.1. The Speed of Light

The speed of light in a vacuum is one of the fundamental physical quantities. The establishment of the finite nature of the speed of light had a tremendous significance of principle. The finite nature of the speed of transmitting signals and of transmitting interactions underlies the theory of relativity.

In view of the fact that the numerical value of the speed of light is very high, the experimental determination of this speed is a very complicated task. The speed of light was first determined on the basis of astronomical observations. In 1676, the Danish astronomer Olaus Romer (1644-1710) determined the speed of light from observations of eclipses of Jupiter's satellites. He obtained a value of 215000 km s^{-1}.

The Earth's motion in orbit results in the visible position of stars on the celestial sphere changing. This phenomenon, called the **aberration of light**, was used in 1727 by the British astronomer James Bradley (1693-1762) to determine the speed of light.

Assume that the direction to a star seen in a telescope is perpendicular to the plane of the Earth's orbit. Hence, the angle between the direction toward the star and the vector of the Earth's velocity v will be $\pi/2$ during the entire year (Fig. 21.1). Let us point the axis of the telescope directly at the star. During the time τ needed for the light to cover the distance from the objective to the eyepiece, the telescope will move together with the Earth over the distance $v\tau$ in a direction at right angles to the light ray. As a result, the image of the star will be displaced from the centre of the eyepiece. For the image to be exactly at the centre of the eyepiece, the axis of the telescope must be turned in the direction of the vector $\boldsymbol{v}$ through the angle whose tangent is determined by the relation

$$\tan \alpha = \frac{v}{c} \tag{21.1}$$

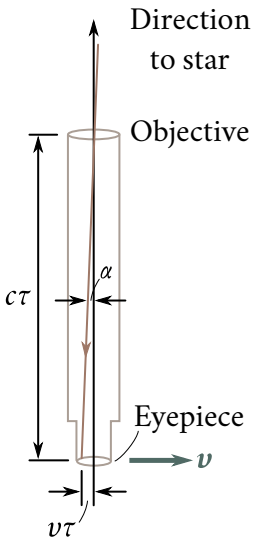


Fig. 21.1: Bradley experimental scheme to measure the speed of light. The direction to a star seen in a telescope is perpendicular to the plane of the Earth's orbit. The angle between the direction toward the star and the vector of the Earth's velocity v will be $\pi/2$ during the entire year.

(see Fig. 21.1). In exactly the same way, raindrops falling vertically will fly through a long tube placed on a moving cart only if the axis of the tube is inclined in the direction of motion of the cart.

Thus, the visible position of a star is displaced relative to the true one through the angle α. The Earth's velocity vector constantly turns in the plane of the orbit. Therefore, the telescope axis also turns, describing a cone about the true direction toward the star. Accordingly, the visible position of the star on the celestial sphere describes a circle whose angular diameter is 2α. If the direction toward the star makes an angle other than a right one with the plane of the Earth's orbit, the visible position of the star describes an ellipse whose major axis has the angular dimension 2α. For a star in the plane of the orbit, the ellipse degenerates into a straight line.

Bradley found from astronomical observations that $2\alpha = 40.9'$. The corresponding value of c obtained by Eq. (21.1) is 303000 km s^{-1}.

In terrestrial conditions, the speed of light was first measured by the French scientist Armand Fizeau (1819-1896) in 1849. The layout of his experiment is shown in Fig. 21.2. Light from source S fell on a half-silvered mirror. The light reflected from the mirror got onto the edge of a rapidly rotating toothed disk. Every time a space between the teeth was opposite the light beam, a light pulse was produced that reached mirror M and was reflected back. If at the moment when the light

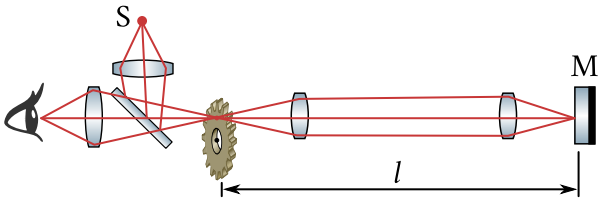


Fig. 21.2: Fizeau experimental setup to measure the speed of light. Light from source S falls on a half-silvered mirror. The light reflected from the mirror hits the edge of a rapidly rotating toothed disk. Every time a space between the teeth was opposite the light beam, a light pulse was produced that reached mirror M and was reflected back. If at the moment when the light returned to the disk a space was opposite the beam, the reflected pulse passed partly through the half-silvered mirror and reached the observer's eye. If a tooth of the disk was in the path of the reflected pulse, the observer saw no light.

returned to the disk a space was opposite the beam, the reflected pulse passed partly through the half-silvered mirror and reached the observer's eye. If a tooth of the disk was in the path of the reflected pulse, the observer saw no light.

During the time $\tau = 2l/c$ needed for the light to cover the distance to mirror M and back, the disk managed to turn through the angle $\Delta\omega = \omega\tau = 2l\omega/c$, where ω is the angular velocity of the disk. Assume that the number of disk teeth is N. Therefore, the angle between the centres of adjacent teeth is $\alpha = 2\pi/N$. The light did not return to the observer's eye at such disk velocities at which the disk in the time τ managed to turn through the angles $\alpha/2, 3\alpha/2, \ldots, (m - 1/2)\alpha$, etc. Hence, the condition for the m-th blackout has the form

$$\Delta\omega = \left(m - \frac{1}{2}\right)\alpha \quad \text{or} \quad \frac{2l\omega}{c} = \left(m - \frac{1}{2}\right)\frac{2\pi}{N}.$$

According to this formula, knowing l, N, and the angular velocity ω_m at which the m-th blackout is obtained, we can find c. In Fizeau's experiment, l was about 8.6 km. The value of 313000 km s^{-1} was obtained for c.

In 1928, Kerr cells (see Sec. 19.7) were used to measure the speed of light. They made it possible to interrupt a light beam with a much higher frequency (about 10^7 s^{-1}) than when a rotating toothed disk was used. This made measurements of c possible with l of the order of several metres.

Albert Michelson performed several measurements of the speed of light using the method of a rotating prism. In Michelson's experiment conducted in 1932, light propagated in a tube 1.6 km long from which the air was evacuated.

At present, the speed of light in a vacuum is taken equal to

$$c = 299792.5 \pm 0.1 \text{ km s}^{-1}. \tag{21.2}$$

We must note that in all the experiments in which light was interrupted, the group

velocity of the light waves was determined, and not the phase velocity. In air, these two velocities virtually coincide.

21.2. Fizeau's Experiment

Up to now, we assumed that the sources, receivers, and other bodies relative to which the propagation of light was considered are stationary. It is quite natural to be interested in how motion of a source of light waves affects the propagation of light. Here, it becomes necessary to indicate relative to what the motion takes place. We established in Sec. 14.11 that the motion of a source or a receiver of sound waves relative to the medium in which these waves are propagating affects the proceeding of acoustic phenomena (the Doppler effect), and, consequently, can be detected.

The wave theory initially treated light as elastic waves propagating in a hypothetic medium called universal ether. After Maxwell advanced his theory, elastic ether was replaced by an ether that was a carrier of electromagnetic waves and fields. By this ether was meant a special medium filling, like its elastic ether predecessor, the entire space of the universe and penetrating all bodies. Since ether was a certain medium, it would be possible to count on detecting the motion of bodies, for example light sources or receivers, with respect to this medium. In particular, the existence of an "ether wind" blowing around the Earth in its motion about the Sun ought to be expected.

Galileo's principle of relativity was established in mechanics. According to it, all inertial reference frames are equivalent in a mechanical respect. The detection of ether would make it possible to separate (with the aid of optical phenomena) a special (related to ether) predominant, absolute reference frame. Therefore, motion of the other frames could be considered relative to this absolute frame.

Thus, the establishment of how universal ether interacts with moving bodies, was a matter of principle. Three possibilities could be assumed: (1) ether is absolutely not disturbed by moving bodies, (2) ether is partly carried along by moving bodies, acquiring a velocity of αv, where v is the velocity of a body relative to the absolute reference frame, and α is a drag coefficient less than unity, and (3) ether is completely carried along by moving bodies, for example by the Earth, in the same way as a body in its motion carries along the layers of gas adjoining its surface. The last possibility, however, is disproved by the existence of the phenomenon of light aberration. We established in the preceding section that the change in the visible position of stars can be explained by the motion of the telescope relative to the reference frame (medium) in which the light wave is propagating.

To find out whether ether is carried along by moving bodies, Fizeau conducted the following experiment in 1851. A parallel beam of light from source S was split by

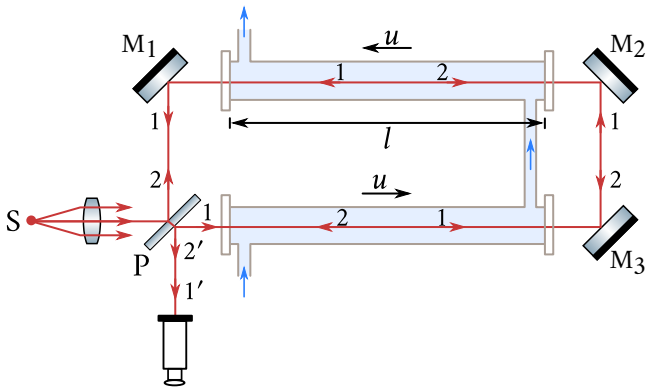


Fig. 21.3: Fizeau's interferometer experiment to determine the role of the ether in the motion bodies in it. A parallel beam of light from source S was split by half-silvered plate P into two beams 1 and 2.

half-silvered plate P into two beams 1 and 2 (Fig. 21.3). As a result of reflection from mirrors M_1, M_2 and M_3, the beams, after completing the same total path L, again reached plate P. Beam 1 partly passed through P, while beam 2 was partly reflected. As a result, two coherent beams $1'$ and $2'$ were set up. They produced an interference pattern in the form of fringes in the focal plane of a telescope. Two tubes along which water could be passed with the velocity u in the directions indicated by the arrows were installed in the paths of beams 1 and 2. Ray 2 propagated in both tubes opposite to the flow of the water, and ray 1 with the flow.

When the water was stationary, beams 1 and 2 covered the path L in the same time. If water in its motion even partly carries along ether, then when the flow of the water was switched on, ray 2, which propagates opposite to the flow, would spend more time to cover the path L than ray 1 travelling in the direction of flow. As a result, a certain path difference will appear between the rays, and the interference pattern will be displaced.

The path difference we are interested in appears only in the path of the rays in the water. This path has the length $2l$. Let the velocity of light in the water relative to the ether be v. When ether is not carried along by the water, the speed of light relative to the arrangement will coincide with v. Let us assume that the water in its motion partly carries along the ether, imparting to it the velocity αu relative to the arrangement (u is the velocity of the water, and α is the drag coefficient). Hence, the velocity of light relative to the arrangement will be $v + \alpha u$ for ray 1 and $v - \alpha u$ for ray 2. Ray 1 covers the path $2l$ during the time $t_1 = 2l/(v + \alpha u)$, and ray 2 during the time $t_2 = 2l/(v - \alpha u)$. It can be seen from Eq. (16.54) that the optical length of a path to cover which the time t is required equals ct. Hence, the path difference of

rays 1 and 2 is $\delta = c(t_2 - t_1)$. Dividing δ by Δ by λ_0, we get the number of fringes by which the interference pattern will be displaced when the flow of water is switched on:

$$\Delta N = \frac{c(t_2 - t_1)}{\lambda_0} = \frac{c}{\lambda_0}\left(\frac{2l}{v - \alpha u} - \frac{2l}{v + \alpha u}\right) = \frac{4cl\alpha u}{\lambda_0(v^2 - \alpha^2 u^2)}.$$

Fizeau discovered that the interference fringes are indeed displaced. The value of the drag coefficient corresponding to this displacement was

$$\alpha = 1 - \frac{1}{n^2}, \tag{21.3}$$

where n is the refractive index of water. Thus, Fizeau's experiment showed that ether (if it exists) is carried along by moving water only partly.

It is easy to see that the result of Fizeau's experiment is explained by the relativistic law of velocity addition. According to the first of equations (8.27) in Vol. I, the velocities v_x and v'_x of a body in frames K and K′ are related by the expression

$$v_x = \frac{v'_x + v_0}{1 + v_0 v'_x/c^2} \tag{21.4}$$

(v_0 is the velocity of the frame K′ relative to the frame K).

Let us relate the reference frame K to Fizeau's instrument, and the frame K′ to the moving water. Now, the part of v_0 will be played by the velocity of the water u, that of v'_x by the velocity of the light relative to the water equal to c/n, and, finally, the part of v_x will be played by the velocity of the light relative to the instrument v_{inst}. Introduction of these values into Eq. (21.4) yields

$$v_{\text{inst}} = \frac{c/n + u}{1 + u(c/n)c^2} = \frac{(c/n) + u}{1 + u/(cn)}.$$

The velocity of the water u is much smaller than c. The expression obtained can therefore be simplified as follows:

$$v_{\text{inst}} = \frac{(c/n) + u}{1 + u/(cn)} = \approx \left(\frac{c}{n} + u\right)\left(1 - \frac{u}{cn}\right) \approx \frac{c}{n} + u\left(1 - \frac{1}{n^2}\right) \tag{21.5}$$

[we have disregarded the term $u^2/(cn)$].

According to classical notions, the velocity of light relative to the instrument v_{inst} equals the sum of the velocity of light relative to ether, *i.e.*, c/n, and of the velocity of ether relative to the instrument, *i.e.*, αu:

$$v_{\text{inst}} = \frac{c}{n} + \alpha u.$$

A comparison with Eq. (21.5) gives the value obtained by Fizeau for the drag coefficient α [see Eq. (21.3)].

It must be borne in mind that only the velocity of light in a vacuum is the same in all reference frames. Its velocity in a substance differs in different reference

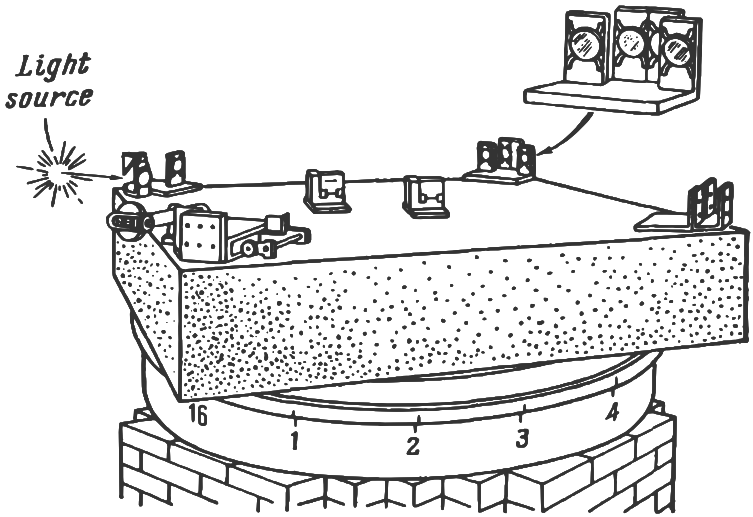


Fig. 21.4: Michelson and Morley experiment. A brick foundation supported an annular iron trough with mercury. A wooden float having the shape of the bottom half of a longitudinally cut doughnut floated on the mercury. The float carried a massive square stone slab. This design made it possible to smoothly turn the slab about the vertical axis of the arrangement.

frames. It has the value c/n in the frame associated with the medium in which the light is propagating.

21.3. Michelson's Experiment

In 1881, Michelson carried out his famous experiment by means of which he counted on detecting the motion of the Earth relative to ether (the ether wind). In 1887, he repeated his experiment together with Morley on an improved instrument. The arrangement used by Michelson and Morley is shown in Fig. 21.4. A brick foundation supported an annular iron trough with mercury. A wooden float having the shape of the bottom half of a longitudinally cut doughnut floated on the mercury. The float carried a massive square stone slab. This design made it possible to smoothly turn the slab about the vertical axis of the arrangement. A Michelson interferometer (see Fig. 17.6) was installed on the slab. The interferometer was modified so that both rays before returning to the half-silvered plate cover a distance coinciding with the diagonal of the slab several times. A diagram of the path of the rays is shown in Fig. 21.5. The symbols in this figure correspond to those used in Fig. 17.16.

The experiment was based on the following reasoning. Let us assume that interferometer arm PM_2 (Fig. 21.6) coincides with the direction of motion of the Earth relative to ether. Consequently, the time needed for ray 1 to cover the path to

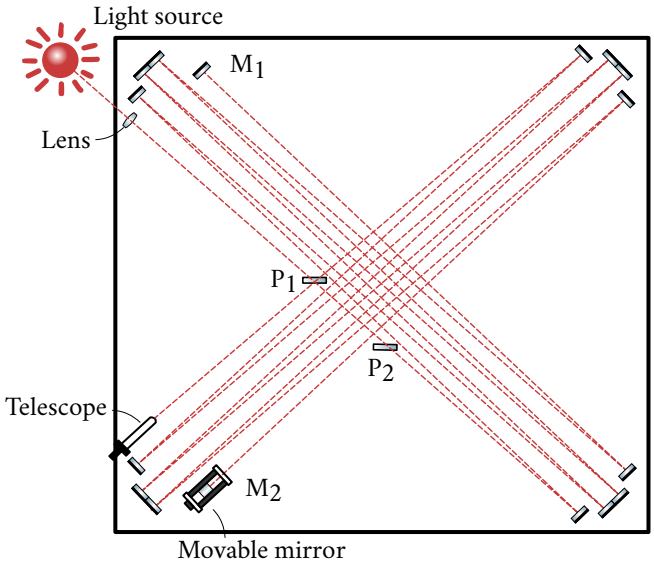


Fig. 21.5: Modified Michelson interferometer. The interferometer was modified so that both rays before returning to the half-silvered plate cover a distance coinciding with the diagonal of the slab several times. The symbols in this figure correspond to those used in Fig. 17.16.

mirror M_1 and back will differ from the time needed for ray 2 to cover path PM_2P. As a result, even when the lengths of both arms are equal, rays 1 and 2 will acquire a certain path difference. If we turn the arrangement through 90°, the arms will exchange places, and the path difference will change its sign. This should result in displacement of the interference pattern whose magnitude, as shown by calculations performed by Michelson, could be detected quite readily.

To calculate the expected displacement of the interference pattern, let us find the time spent by rays 1 and 2 to cover the relevant paths. Assume that the Earth's velocity relative to the ether is v. If the ether is not carried along by the Earth and the velocity of light relative to the ether is c (the refractive index of air is practically equal to unity), then the velocity of light relative to the instrument will be $c - v$ for direction PM_2 and $c + v$ for direction M_2P. Hence, the time needed for ray 2 is determined by the expression

$$t_2 = \frac{l}{c-v} + \frac{l}{c+v} = \frac{2lc}{c^2-v^2} = \frac{2l}{c}\frac{1}{(1-v^2/c^2)} \approx \frac{2l}{c}\left(1+\frac{v^2}{c^2}\right) \qquad (21.6)$$

(the Earth's velocity along its orbit is $30\,\mathrm{km\,s^{-1}}$, therefore, $v^2/c^2 = 10^{-8} \ll 1$).

Before commencing to calculate the time t_1, let us consider the following example from mechanics. Suppose that a launch developing the velocity c relative to water has to cross a river with a current velocity of v in a direction strictly perpen-

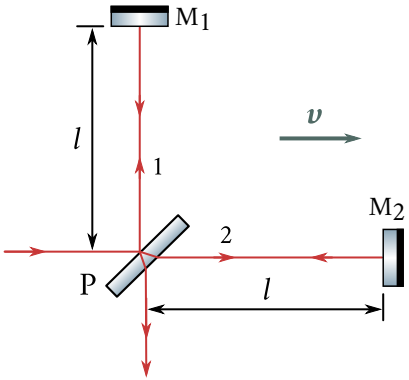


Fig. 21.6: Reasoning of Michelson and Morley experiment, assuming that the interferometer arm PM_2 coincides with the direction of motion of the Earth relative to ether. Then, the time needed for ray 1 to cover the path to mirror M_1 and back will differ from the time needed for ray 2 to cover path PM_2P. As a result, even when the lengths of both arms are equal, rays 1 and 2 will acquire a certain path difference. Turning the arrangement 90°, the arms will exchange places, and the path difference will change its sign.

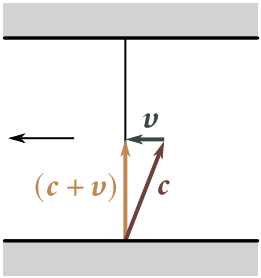


Fig. 21.7: Considerations to calculate time t_1. Suppose that a launch developing the velocity c relative to water has to cross a river with a current velocity of v in a direction strictly perpendicular to its banks. For the launch to travel in the required direction, its velocity c relative to the water must be directed as shown here.

dicular to its banks (Fig. 21.7). For the launch to travel in the required direction, its velocity c relative to the water must be directed as shown in the figure. Therefore, the velocity of the launch relative to the banks will be $|c + v| = \sqrt{c^2 - v^2}$. The velocity of ray 1 relative to the arrangement (as assumed by Michelson) will be the same. Consequently, the time taken by ray 1 is[1]

$$t_1 = \frac{2l}{\sqrt{c^2 - v^2}} = \frac{2l}{c}\frac{1}{\sqrt{1 - v^2/c^2}} \approx \frac{2l}{c}\left(1 + \frac{1}{2}\frac{v^2}{c^2}\right). \tag{21.7}$$

Substituting for t_2 and t_1 in the expression $\Delta = c(t_2 - t_1)$ their values from expressions (21.6) and (21.7), we get the path difference for rays 1 and 2:

$$\Delta = 2l\left[\left(1 + \frac{v^2}{c^2}\right) - \left(1 + \frac{1}{2}\frac{v^2}{c^2}\right)\right] = l\frac{v^2}{c^2}.$$

When the arrangement is turned through 90°, the path difference changes its sign. Consequently, the number of fringes by which the interference pattern will be displaced is

$$\Delta N = \frac{2\Delta}{\lambda_0} = 2\frac{l}{\lambda_0}\frac{v^2}{c}. \tag{21.8}$$

[1]We have used the formulas $\sqrt{1 - x} \approx 1 - x/2$ and $1/(1 - x) \approx 1 + x$, for small values of x.

The arm length l (taking into account multifold reflections) was 11 m. The wavelength of the light used by Michelson and Morley was 0.59 μm. The use of these values in Eq. (21.8) gives

$$\Delta N = \frac{2 \times 11}{0.59 \times 10^{-6}} \times 10^{-8} = 0.37 \approx 0.4 \text{ fringe.}$$

The arrangement made it possible to detect a displacement of the order of 0.01 fringe. But no displacement of the interference pattern was detected. The experiment was repeated during different times of the day to exclude the possibility of the horizon plane being perpendicular to the vector of the Earth's orbital velocity at the moment of measurements. Subsequently, the experiment was repeated many times during different seasons of the year (during a year, the vector of the Earth's orbital velocity turns in space through 360°), and negative results were constantly obtained. The attempt to detect an ether wind was not successful Universal ether remained elusive.

Several attempts were made to explain the negative result of Michelson's experiment without refuting the hypothesis of the existence of universal ether. But all these attempts were groundless. An exhaustive non-contradictory explanation of all the experimental facts including the results of Michelson's experiment was given by Albert Einstein in 1905. He arrived at the conclusion that universal ether, *i.e.*, a special medium that could serve as an absolute reference frame, does not exist. Accordingly, Einstein extended the mechanical principle of relativity to all physical phenomena without any exception. He further postulated in accordance with experimental data that the speed of light in a vacuum is the same in all inertial reference frames and does not depend on the motion of the light sources and receivers.

The principle of relativity and the principle of the constancy of the speed of light form the foundation of the special theory of relativity developed by Einstein (see Chapter 8 of Vol. I).

21.4. The Doppler Effect

In acoustics, the change in frequency due to the Doppler effect is determined by the velocities of the source and the receiver relative to the medium that is the carrier of the sound waves [see Eq. (14.78)]. The Doppler effect also exists for light waves. But there is no special medium that would serve as the carrier of electromagnetic waves. Therefore, the Doppler displacement of the frequency of light waves is determined only by the relative velocity of the source and the receiver.

Let us associate the origin of coordinates of the frame K with a light source and the origin of coordinates of the frame K′ with a receiver (Fig. 21.8). We shall direct the axes x and x', as usual, along the velocity vector v with which the frame K′ (*i.e.*, the receiver) is moving relative to the frame K (*i.e.*, the source). The equation of a

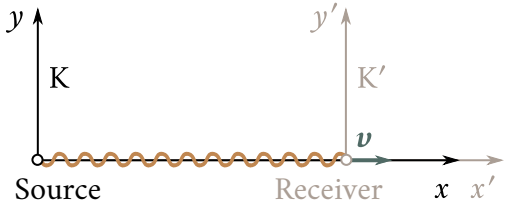


Fig. 21.8: The Doppler effect for light waves. Let us associate the origin of coordinates of the frame K with a light source and the origin of coordinates of the frame K′ with a receiver. We shall direct the axes x and x', along the velocity vector v with which the frame K′ (the receiver) is moving relative to the frame K (the source).

plane light wave emitted by the source in the direction of the receiver will have the following form in the frame K:

$$E(x,t) = A\cos\left[\omega\left(t - \frac{x}{c}\right) + \alpha\right]. \tag{21.9}$$

Here, ω is the frequency of a wave registered in the reference frame associated with the source, *i.e.*, the frequency of oscillations of the source. We assume that the light wave is propagating in a vacuum; therefore, the phase velocity is c.

According to the principle of relativity, the laws of nature have the same form in all inertial reference frames. Hence, in the frame K′, the wave given by Eq. (21.9) will be described by the equation

$$E(x',t') = A'\cos\left[\omega'\left(t' - \frac{x'}{c}\right) + \alpha'\right], \tag{21.10}$$

where ω' is the frequency registered in the reference frame K′, *i.e.*, the frequency picked up by the receiver. We have provided all the quantities except c, which is the same in all reference frames, with primes.

We can obtain an equation of a wave in the frame K′ from an equation in the frame K by passing over from x and t to x' and t' with the aid of the Lorentz transformations. Introducing instead of x and t in Eq. (21.9) their values in accordance with Eqs. (8.17) of Vol. I, we get

$$E(x',t') = A\cos\left[\omega\left(\frac{t' + (v/c^2)x'}{\sqrt{1 - v^2/c^2}} - \frac{x' + vt'}{c\sqrt{1 - v^2/c^2}}\right) + \alpha\right]$$

(the part of v_0 is played by v). The latter expression is easily transformed into the following one:

$$E(x',t') = A\cos\left[\omega\left(\frac{1 - v/c}{\sqrt{1 - v^2/c^2}}\right)\left(t' - \frac{x'}{c}\right) + \alpha\right]. \tag{21.11}$$

Equation (21.11) describes the same wave in the frame K′ as Eq. (21.10). Therefore,

the following relation must be observed:

$$\omega' = \omega \frac{1 - v/c}{\sqrt{1 - v^2/c^2}} = \omega \left(\frac{1 - v/c}{1 + v/c} \right)^{1/2}.$$

Let us change our notation: we shall denote the frequency ω of the source by ω_0, and the frequency ω' of the receiver by ω. The preceding equation will thus become

$$\omega = \omega_0 \left(\frac{1 - v/c}{1 + v/c} \right)^{1/2}. \tag{21.12}$$

Passing over from the angular frequency to the ordinary one, we have

$$\nu = \nu_0 \left(\frac{1 - v/c}{1 + v/c} \right)^{1/2}. \tag{21.13}$$

The velocity v of the receiver relative to the source in Eqs. (21.12) and (21.13) is an algebraic quantity. When the receiver moves away from the source, $v > 0$, and by Eq. (21.12), $\omega < \omega_0$; when the receiver approaches the source, $v < 0$, so that $\omega > \omega_0$. When $v \ll c$, Eq. (21.12) can be written approximately as follows:

$$\omega \approx \omega_0 \left[\frac{1 - (1/2)(v/c)}{1 + (1/2)(v/c)} \right] \approx \omega_0 \left(1 - \frac{1}{2} \frac{v}{c} \right) \left(1 - \frac{1}{2} \frac{v}{c} \right).$$

Hence, after Taylor expansion around $v = 0$ and limiting ourselves to terms of the order of v/c, we get

$$\omega = \omega_0 \left(1 - \frac{v}{c} \right). \tag{21.14}$$

From this formula, we can find the relative change in the frequency:

$$\frac{\Delta\omega}{\omega} = -\frac{v}{c} \tag{21.15}$$

($\Delta\omega$ stands for $\omega - \omega_o$).

We can show that a **transverse Doppler effect** exists for light waves in addition to the longitudinal effect we have considered. It consists in a reduction in the frequency picked up by the receiver observed when the vector of the relative velocity is directed at right angles to the straight line passing through the receiver and the source[2] (when, for example, the source travels along a circle at whose centre the receiver is). In this case, the frequency ω_0 in the frame of the source is associated with the frequency ω in the frame of the receiver by the relation

$$\omega = \omega_0 \left(1 - \frac{v^2}{c^2} \right)^{1/2} \approx \omega_0 \left(1 - \frac{1}{2} \frac{v^2}{c^2} \right). \tag{21.16}$$

[2]We remind our reader that the transverse Doppler effect does not exist for sound waves.

The relative change in frequency in the transverse Doppler effect

$$\frac{\Delta\omega}{\omega} = -\frac{1}{2}\frac{v^2}{c^2}, \tag{21.17}$$

is proportional to the square of the ratio v/c and, consequently, is considerably smaller than in the longitudinal effect for which the relative change in the frequency is proportional to the first power of v/c.

The existence of the transverse Doppler effect was proved experimentally by the American physicist Herbert Ives (1882-1953) in 1938. He determined the change in the frequency of emission of hydrogen atoms in canal rays (see the last paragraph of Sec. 12.6).

The velocity of the atoms was about 2×10^8 m s^{-1}. These experiments were a direct experimental confirmation of the correctness of Lorentz transformations.

In the general case, the vector of the relative velocity can be resolved into two components of which one is directed along the ray, and the other at right angles to it. The first component gives rise to the longitudinal, and the second to the transverse Doppler effect.

The longitudinal Doppler effect is used to determine the radial velocity of stars. By measuring the relative shift of the lines in the spectra of stars, we can use Eq. (21.12) to determine v.

The thermal motion of the molecules of a luminous gas, owing to the Doppler effect, leads to broadening of the spectral lines. As a result of the chaotic nature of the thermal motion, all the directions of the molecules' velocities relative to a spectrograph are equally probable. Therefore, the radiation registered by the instrument contains all the frequencies in the interval from $\omega_0(1-v/c)$ to $\omega_o(1+v/c)$, where ω_0 is the frequency emitted by the molecules, and v is the velocity of thermal motion [see Eq. (21.14)]. The registered width of a spectral line is thus $2\omega_0 v/c$. The quantity

$$\delta\omega_{\mathrm{D}} = 2\omega_0\frac{v}{c}, \tag{21.18}$$

is called the **Doppler width of a spectral line** (v stands for the most probable velocity of the molecules). The magnitude of the Doppler broadening of spectral lines makes it possible to assess the velocity of thermal motion of the molecules and, consequently, the temperature of a luminous gas.

APPENDICES

A.1. List of Symbols

A	amplitude; gas amplification; work
$\boldsymbol{A}$	vector potential of magnetic field†
a	amplitude
$\boldsymbol{a}$	acceleration; vector
B	Kerr constant
$\boldsymbol{B}$	magnetic induction
C	capacitance; circulation of vector; Curie constant
c	electromagnetic constant; speed of light
D	angular dispersion
$\boldsymbol{D}$	electric displacement
d	period of diffraction grating; separation distance of capacitor plates
$\div$	divergence
E	illuminance; Young's modulus
$\boldsymbol{E}$	electric field strength
$\boldsymbol{E}^*$	strength of extraneous force field
$\mathscr{E}$	electromotive force (e.m.f.)
e	base of natural logarithms; positive elementary charge
$\hat{\boldsymbol{e}}$	unit vector
F	Faraday constant
$\boldsymbol{F}$	force
f	focal length
G	shear modulus

† The magnitude of a vector is denoted by the same symbol as the vector itself, but in ordinary italic (sloping) type.

$\boldsymbol{H}$ magnetic field strength
$\hbar$ Planck's constant h divided by 2π
I current; luminous intensity; sound intensity
i imaginary unity ($i = \sqrt{-1}$)
$\boldsymbol{j}$ current density; density of energy flux
$\boldsymbol{K}$ momentum
k constant of proportionality; wavenumber
$\boldsymbol{k}$ wavevector
L inductance; loudness level; luminance; optical path
$\boldsymbol{L}$ angular momentum
l length; mean free path
$\boldsymbol{l}$ displacement
M luminous emittance; magnification; mass of mole
$\boldsymbol{M}$ magnetization; moment of force
m mass
N demagnetization factor; number
N_{A} Avogadro constant
n number; refractive index
P degree of polarization; optical power; power; probability; radiated power
$\boldsymbol{P}$ force of gravity; polarization
p pressure
$\boldsymbol{p}$ dipole moment; electric moment
Q amount of heat; quality of oscillator circuit
q charge
R molar gas constant; radius; resistance; resolving power
R_{H} Hall coefficient
$\Re$ real number
r distance; radius
$\boldsymbol{r}$ position vector
S area
$\boldsymbol{S}$ Poynting vector
T absolute temperature; period
$\boldsymbol{T}$ torque
t time
U voltage
u mobility of ion
$\boldsymbol{u}$ velocity
V Verdet constant; visibility
$\boldsymbol{v}$ velocity

W energy
w energy density
X reactance
x coordinate
y coordinate
Z atomic number of element; impedance
z coordinate; valence
α angle; drag coefficient; initial phase of oscillations; rotational constant
β angle; polarizability of molecule; relative velocity
γ angle; attenuation coefficient
Δ difference in optical path; increment; Laplacian operator
δ density of metal; fraction of energy; phase difference
ε relative permittivity; strain
ε_0 electric constant
θ angle; polar angle; polar coordinate
$\varkappa$ thermal conductivity; wave absorption coefficient
$\varkappa'$ extinction coefficient
λ linear charge density; logarithmic decrement; wavelength
μ permeability
μ_B Bohr magneton
μ_0 magnetic constant
ν frequency
ξ displacement of wave point
π ratio of circumference to diameter
ρ coherence radius; density; reflection coefficient; resistivity; volume density of charge
σ conductivity; cross-sectional area; stress; surface charge density
τ retardation time; time; time constant of a circuit; transmission coefficient
Φ flux
φ angle; azimuthal angle; potential
χ electric susceptibility
χ_m magnetic susceptibility
Ψ flux linkage; total magnetic flux
ψ angle
Ω solid angle
ω angular frequency
$\boldsymbol{\omega}$ angular velocity
∇ del (Hamiltonian) operator

A.2. Units of Electrical and Magnetic Quantities in the International System (SI) and in the Gaussian System

The electric constant

$$\varepsilon_0 = \frac{1}{4\pi(2.997925)^2 \times 10^9}\,\mathrm{F\,m^{-1}} \approx \frac{1}{4\pi \times 9 \times 10^9}\,\mathrm{F\,m^{-1}}.$$

The magnetic constant

$$\mu_0 = 4\pi \times 10^{-7}\,\mathrm{H\,m^{-1}}.$$

The electromagnetic constant

$$c = \frac{1}{\sqrt{\varepsilon_0 \mu_0}} = 2.997925 \times 10^8\,\mathrm{m\,s^{-1}} \approx 3 \times 10^8\,\mathrm{m\,s^{-1}}.$$

The relations between the units are given approximately. To obtain the exact values, substitute 2.997925 for 3 and $(2.997925)^2$ for 9.

Table A.1

Quantity and aymbol	Unit and symbol		Relation
	SI	**Gaussian**	
Force F	newton (N)	dyne (dyn)	$1\ \mathrm{N} = 10^5$ dyn
Work A and energy W	joule (J)	erg (erg)	$1\ \mathrm{J} = 10^7$ erg
Charge q	coulomb (C)	cgse_q	$1\ \mathrm{C} = 3 \times 10^9\ \mathrm{cgse}_q$
Electric field strength E	volt per metre ($\mathrm{V\,m^{-1}}$)	cgse_E	$1\ \mathrm{C} = 3 \times 10^4\ \mathrm{m^{-1}}$
Potential φ, voltage U, e.m.f. $\mathscr{E}$	volt (V)	$\mathrm{cgse}_{\varphi,U,\mathscr{E}}$	$1\ \mathrm{cgse}_{\varphi,U,\mathscr{E}} = 300\,\mathrm{V}$
Electric moment of dipole p	coulomb-metre (C m)	cgse_p	$1\ \mathrm{C\,m} = 3 \times 10^{11}\ \mathrm{cgse}_q$
Polarization P	coulomb per metre squared ($\mathrm{C\,m^{-2}}$)	cgse_P	$1\ \mathrm{C\,m^{-2}} = 3 \times 10^5\ \mathrm{cgse}_P$
Electric susceptibility χ	SI_χ	cgse_χ	$1\ \mathrm{cgse}_\chi = 4\pi\ \mathrm{SI}_\chi$
Electric displacement D	coulomb per metre squared ($\mathrm{C\,m^{-2}}$)	cgse_D	$1\ \mathrm{C\,m^{-2}} = 4\pi 3 \times 10^5\ \mathrm{cgse}_D$
Flux of electric displacement Φ	coulomb (C)	cgse_Φ	$1\ \mathrm{C} = 4\pi 3 \times 10^9\ \mathrm{cgse}_\Phi$
Capacitance C	farad (F)	centimetre (cm)	$1\ \mathrm{F} = 10^{11}$ cm
Current I	ampere (A)	cgse_I	$1\ \mathrm{A} = 3 \times 10^9\ \mathrm{cgse}_I$
Current density j	ampere per metre squared ($\mathrm{A\,m^{-2}}$)	cgse_j	$1\ \mathrm{A\,m^{-2}} = 3 \times 10^5\ \mathrm{cgse}_j$
Resistance R	ohm (Ω)	cgse_R	$1\ \mathrm{cgse}_R = 9 \times 10^{11}\,\Omega$
Resistivity ρ	ohm-metre (Ω m)	cgse_ρ	$1\ \mathrm{cgse}_\rho = 9 \times 10^9\,\Omega\,\mathrm{m}$
Conductivity σ	siemens per metre ($\mathrm{S\,m^{-1}}$)	cgse_σ	$1\ \mathrm{S\,m^{-1}} = 9 \times 10^9\ \mathrm{cgse}_\sigma$
Magnetic induction B	tesla (T)	gauss (Gs)	$1\ \mathrm{T} = 10^4$ Gs
Flux of magnetic induction Φ	weber (Wb)	maxwell (Mx)	$1\ \mathrm{Wb} = 10^8$ Mx
Flux linkage Ψ	weber (Wb)	maxwell (Mx)	$1\ \mathrm{Wb} = 10^8$ Mx
Magnetic moment p_m	ampere-metre squared ($\mathrm{A\,m^2}$)	$\mathrm{cgsm}_{p_\mathrm{m}}$	$1\ \mathrm{A\,m^2} = 10^3\ \mathrm{cgsm}_{p_\mathrm{m}}$
Magnetization M	ampere per metre ($\mathrm{A\,m^{-1}}$)	cgsm_M	$1\ \mathrm{cgsm}_M = 10^3\,\mathrm{A\,m^{-1}}$
Magnetic field strength H	ampere per metre ($\mathrm{A\,m^{-1}}$)	oersted (Oe)	$1\,\mathrm{A\,m^{-1}} = 4\pi 10^{-3}$ Oe
Magnetic susceptibility χ_m	$\mathrm{SI}_{\chi_\mathrm{m}}$	$\mathrm{cgsm}_{\chi_\mathrm{m}}$	$1\ \mathrm{cgsm}_{\chi_\mathrm{m}} = 4\pi\ \mathrm{SI}_{\chi_\mathrm{m}}$
Inductance L	henry (H)	centimetre (cm)	$1\ \mathrm{H} = 10^9$ cm
Mutual inductance L_{12}	henry (H)	centimetre (cm)	$1\ \mathrm{H} = 10^9$ cm

A.3. Basic Formulas of Electricity and Magnetism in the SI and in the Gaussian System

1. Coulomb's law:

$$F = \frac{1}{4\pi\varepsilon_0}\frac{q_1 q_2}{r^2} \text{ (SI)} \qquad F = \frac{q_1 q_2}{r^2} \text{ (GS)}.$$

2. Electric field strength (definition):

$$\boldsymbol{E} = \frac{\boldsymbol{F}}{q}.$$

3. Field strength of point charge:

$$E = \frac{1}{4\pi\varepsilon_0}\frac{q}{\varepsilon r^2} \text{ (SI)} \qquad E = \frac{q}{\varepsilon r^2} \text{ (GS)}.$$

4. Field strength between charged planes and near surface of charged conductor:

$$E = \frac{\sigma}{\varepsilon_0 \varepsilon} \text{ (SI)} \qquad E = \frac{4\pi\sigma}{\varepsilon} \text{ (GS)}.$$

5. Potential (definition):

$$\varphi = \frac{W_\mathrm{p}}{q}.$$

6. Potential of field of point charge:

$$\varphi = \frac{1}{4\pi\varepsilon_0}\frac{q}{\varepsilon r} \text{ (SI)} \qquad \varphi = \frac{q}{\varepsilon r} \text{ (GS)}.$$

7. Work of field forces on charge:

$$A = q(\varphi_1 - \varphi_2).$$

8. Relation between $\boldsymbol{E}$ and φ:

$$\boldsymbol{E} = -\nabla\varphi.$$

9. Relation between φ and $\boldsymbol{E}$:

$$\varphi_1 - \varphi_2 = \int_1^2 \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{l}.$$

10. Curl of vector $\boldsymbol{E}$ for electrostatic field:

$$\nabla \times \boldsymbol{E} = 0.$$

11. Circulation of vector $\boldsymbol{E}$ for electrostatic field:

$$\oint \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{l} = 0.$$

12. Electric moment of dipole:

$$p = ql.$$

13. Torque acting on dipole in electric field:

$$\boldsymbol{T} = \boldsymbol{p} \times \boldsymbol{E}.$$

14. Energy of dipole in field:

$$W = -\boldsymbol{p} \times \boldsymbol{E}.$$

15. Dipole moment of "elastic" molecule:

$$\boldsymbol{p} = \beta\varepsilon_0 \boldsymbol{E} \ \text{(SI)} \qquad \boldsymbol{p} = \beta \boldsymbol{E} \ \text{(GS)}.$$

16. Polarization (definition):

$$\boldsymbol{P} = \frac{1}{\Delta V} \sum \boldsymbol{p}.$$

17. Relation between $\boldsymbol{P}$ and $\boldsymbol{E}$:

$$\boldsymbol{P} = \chi\varepsilon_0 \boldsymbol{E} \ \text{(SI)} \qquad \boldsymbol{P} = \chi \boldsymbol{E} \ \text{(GS)}.$$

18. Relation between $\boldsymbol{P}$ and volume density of bound charges:

$$\rho' = -\nabla \cdot \boldsymbol{P}.$$

19. Relation between $\boldsymbol{P}$ and surface density of bound charges:

$$\sigma' = P_{\mathrm{n}}.$$

20. Electric displacement (definition):

$$\boldsymbol{D} = \varepsilon_0 \boldsymbol{E} + \boldsymbol{P} \ \text{(SI)} \qquad \boldsymbol{D} = \boldsymbol{E} + 4\pi \boldsymbol{P} \ \text{(GS)}.$$

21. Divergence of vector $\boldsymbol{D}$:

$$\nabla \cdot \boldsymbol{D} = \rho \ \text{(SI)} \qquad \nabla \cdot \boldsymbol{D} = 4\pi\rho \ \text{(GS)}.$$

22. Gauss's theorem for $\boldsymbol{D}$:

$$\oint \boldsymbol{D} \cdot \mathrm{d}\boldsymbol{S} = \sum q \ \text{(SI)} \qquad \oint \boldsymbol{D} \cdot \mathrm{d}\boldsymbol{S} = 4\pi \sum q \ \text{(GS)}.$$

23. Relation between permittivity ε and electric susceptibility χ:

$$\varepsilon = 1 + \chi \ \text{(SI)} \qquad \varepsilon = 1 + 4\pi\chi \ \text{(GS)}.$$

24. Relation between values of χ in the SI and in the Gaussian system:

$$\chi_{\mathrm{SI}} = 4\pi\chi_{\mathrm{GS}}.$$

25. Relation between $\boldsymbol{D}$ and $\boldsymbol{E}$:

$$\boldsymbol{D} = \varepsilon\varepsilon_0 \boldsymbol{E} \ \text{(SI)} \qquad \boldsymbol{D} = \varepsilon \boldsymbol{E} \ \text{(GS)}.$$

26. Relation between $\boldsymbol{D}$ and $\boldsymbol{E}$ in a vacuum:

$$\boldsymbol{D} = \varepsilon_0 \boldsymbol{E} \ \text{(SI)} \qquad \boldsymbol{D} = \boldsymbol{E} \ \text{(GS)}.$$

27. $\boldsymbol{D}$ of point charge field:

$$D = \frac{1}{4\pi} \frac{q}{r^2} \ \text{(SI)} \qquad D = \frac{q}{r^2} \ \text{(GS)}.$$

28. Capacitance of capacitor (definition):

$$C = \frac{q}{U}.$$

29. Capacitance of plane capacitor:

$$C = \frac{\varepsilon_0 \varepsilon S}{d} \text{ (SI)} \qquad C = \frac{\varepsilon_0 \varepsilon S}{4\pi d} \text{ (GS)}.$$

30. Energy of system of charges:

$$W = \frac{1}{2} \sum q\varphi.$$

31. Energy of charged capacitor:

$$W = \frac{CU^2}{2}.$$

32. Density of electric field energy:

$$w = \frac{\varepsilon_0 \varepsilon E^2}{2} \text{ (SI)} \qquad w = \frac{\varepsilon E^2}{8\pi} \text{ (GS)}.$$

33. Current (definition):

$$I = \frac{\mathrm{d}q}{\mathrm{d}t}.$$

34. Current density (definition):

$$j = \frac{\mathrm{d}I}{\mathrm{d}S_\perp}.$$

35. Continuity equation:

$$\nabla \cdot \boldsymbol{j} = -\frac{\partial \rho}{\partial t}.$$

36. Voltage (definition):

$$U = \varphi_1 - \varphi_2 + \mathscr{E}_{12}.$$

37. Ohm's law:

$$I = \frac{U}{R}.$$

38. Ohm's law in differential form

$$\boldsymbol{j} = \frac{1}{\rho}\boldsymbol{E} = \sigma \boldsymbol{E}.$$

39. Joule-Lenz law:

$$Q = \int_0^t RI^2 \, \mathrm{d}t.$$

40. Joule-Lenz law in differential form:

$$w = \rho j^2.$$

41. Force of interaction of two parallel currents in a vacuum (per unit length):

$$F = \frac{\mu_0}{4\pi}\frac{2I_1I_2}{b} \quad \text{(SI)} \qquad F = \frac{1}{c^2}\frac{2I_1I_2}{b} \quad \text{(GS)}.$$

42. Field of freely moving charge:

$$\boldsymbol{B} = \frac{\mu_0}{4\pi}\frac{q(\boldsymbol{v}\times\boldsymbol{r})}{r^3} \quad \text{(SI)} \qquad \boldsymbol{B} = \frac{1}{c}\frac{q(\boldsymbol{v}\times\boldsymbol{r})}{r^3}.$$

43. Biot-Savart law:

$$\mathrm{d}\boldsymbol{B} = \frac{\mu_0}{4\pi}\frac{I(\mathrm{d}\boldsymbol{l}\times\boldsymbol{r})}{r^3} \quad \text{(SI)} \qquad \mathrm{d}\boldsymbol{B} = \frac{1}{c}\frac{I(\mathrm{d}\boldsymbol{l}\times\boldsymbol{r})}{r^3} \quad \text{(GS)}.$$

44. Lorentz force:

$$\boldsymbol{F} = q\boldsymbol{E} + q(\boldsymbol{v}\times\boldsymbol{B}) \quad \text{(SI)} \qquad \boldsymbol{F} = q\boldsymbol{E} + \frac{q}{c}(\boldsymbol{v}\times\boldsymbol{B}) \quad \text{(GS)}.$$

45. Ampere's law:

$$\mathrm{d}\boldsymbol{F} = I(\mathrm{d}\boldsymbol{l}\times\boldsymbol{B}) \quad \text{(SI)} \qquad \mathrm{d}\boldsymbol{F} = \frac{1}{c}I(\mathrm{d}\boldsymbol{l}\times\boldsymbol{B}) \quad \text{(GS)}.$$

46. Magnetic moment of loop with current:

$$p_\mathrm{m} = IS \quad \text{(SI)} \qquad p_\mathrm{m} = \frac{1}{c}IS \quad \text{(GS)}.$$

47. Angular momentum exerted on magnetic moment in a magnetic field:

$$\boldsymbol{L} = \boldsymbol{p}_\mathrm{m}\times\boldsymbol{B}.$$

48. "Mechanical" energy of magnetic moment in a magnetic field:

$$W = -\boldsymbol{p}_\mathrm{m}\cdot\boldsymbol{B}.$$

49. Divergence of vector $\boldsymbol{B}$:

$$\nabla\cdot\boldsymbol{B} = 0.$$

50. Gauss's theorem for $\boldsymbol{B}$:

$$\oint \boldsymbol{B}\cdot\mathrm{d}\boldsymbol{S} = 0.$$

51. Magnetization (definition):

$$\boldsymbol{M} = \frac{1}{\Delta V}\sum \boldsymbol{p}_\mathrm{m}.$$

52. Magnetic field strength (definition):

$$\boldsymbol{H} = \frac{1}{\mu_0}\boldsymbol{B} - \boldsymbol{M} \quad \text{(SI)} \qquad \boldsymbol{H} = \boldsymbol{B} - 4\pi\boldsymbol{M} \quad \text{(GS)}.$$

53. Relation between $\boldsymbol{M}$ and $\boldsymbol{H}$:

$$\boldsymbol{M} = \chi_\mathrm{m}\boldsymbol{H}.$$

54. Relation between permeability μ and magnetic susceptibility χ_m:

$$\mu = 1 + \chi_\mathrm{m} \quad \text{(SI)} \qquad \mu = 1 + 4\pi\chi_\mathrm{m} \quad \text{(GS)}.$$

55. Relation between values of χ_{m} in the SI and in the Gaussian system:

$$\chi_{\mathrm{m,SI}} = 4\pi\chi_{\mathrm{m,GS}}.$$

56. Relation between $\boldsymbol{B}$ and $\boldsymbol{H}$:

$$\boldsymbol{B} = \mu\mu_0\boldsymbol{H} \ \text{(SI)} \qquad \boldsymbol{B} = \mu\boldsymbol{H} \ \text{(GS)}.$$

57. Relation between $\boldsymbol{B}$ and $\boldsymbol{H}$ a vacuum:

$$\boldsymbol{B} = \mu_0\boldsymbol{H} \ \text{(SI)} \qquad \boldsymbol{B} = \boldsymbol{H} \ \text{(GS)}.$$

58. Curl of vector $\boldsymbol{H}$ for a stationary field:

$$\nabla \times \boldsymbol{H} = \boldsymbol{j} \ \text{(SI)} \qquad \nabla \times \boldsymbol{H} = \frac{4\pi}{c}\boldsymbol{j} \ \text{(GS)}.$$

59. Circulation of vector $\boldsymbol{H}$ for a stationary field:

$$\oint \boldsymbol{H} \cdot \mathrm{d}\boldsymbol{l} = \sum I \ \text{(SI)} \qquad \oint \boldsymbol{H} \cdot \mathrm{d}\boldsymbol{l} = \frac{4\pi}{c}\sum I \ \text{(GS)}.$$

60. Magnetic field strength of straight current:

$$H = \frac{1}{4\pi}\frac{2I}{b} \ \text{(SI)} \qquad H = \frac{1}{c}\frac{2I}{b} \ \text{(GS)}.$$

61. Magnetic field strength at centre of ring current:

$$H = \frac{I}{2R} \ \text{(SI)} \qquad H = \frac{1}{c}\frac{2\pi I}{R} \ \text{(GS)}.$$

62. Field strength of solenoid:

$$H = nI \ \text{(SI)} \qquad H = \frac{4\pi}{c}nI \ \text{(GS)}.$$

63. Flux of magnetic induction (definition):

$$\Phi = \int_S \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{S}.$$

64. Work done on loop with current when it is moved in a magnetic field:

$$A = I\Delta\Phi \ \text{(SI)} \qquad A = \frac{1}{c}I\Delta\Phi \ \text{(GS)}.$$

65. Flux linkage or total magnetic flux (definition):

$$\Psi = \sum \Phi.$$

66. Induced e.m.f.:

$$\mathscr{E}_{\mathrm{i}} = -\frac{\mathrm{d}\Psi}{\mathrm{d}t} \ \text{(SI)} \qquad \mathscr{E}_{\mathrm{i}} = -\frac{1}{c}\frac{\mathrm{d}\Psi}{\mathrm{d}t} \ \text{(GS)}.$$

67. Inductance (definition):

$$L = \frac{\Psi}{I} \ \text{(SI)} \qquad L = c\frac{\Psi}{I} \ \text{(GS)}.$$

68. Inductance of solenoid:

$$L = \mu_\mu n^2 lS \ \text{(SI)} \qquad L = 4\pi\mu_\mu n^2 lS \ \text{(GS)}.$$

69. E.m.f. of self-induction (in absence of ferromagnetics):

$$\mathscr{E}_s = -L\frac{\mathrm{d}I}{\mathrm{d}t} \text{ (SI)} \qquad \mathscr{E}_s = -\frac{1}{c^2}L\frac{\mathrm{d}I}{\mathrm{d}t} \text{ (GS)}.$$

70. Energy of magnetic fteld of current:

$$W = \frac{LI^2}{2} \text{ (SI)} \qquad W = \frac{1}{c^2}\frac{LI^2}{2} \text{ (GS)}.$$

71. Density of energy of magnetic field:

$$w = \frac{\mu_0\mu H^2}{2} \text{ (SI)} \qquad w = \frac{\mu H^2}{8\pi} \text{ (GS)}.$$

72. Energy of linked loops with current:

$$W = \frac{1}{2}\sum_{i,k} L_{ik}I_iI_k \text{ (SI)} \qquad W = \frac{1}{2c^2}\sum_{i,k} L_{ik}I_iI_k \text{ (GS)}.$$

73. Density of displacement current:

$$\boldsymbol{j}_{\text{dis}} = \dot{\boldsymbol{D}} \text{ (SI)} \qquad \boldsymbol{j}_{\text{dis}} = \frac{1}{4\pi}\dot{\boldsymbol{D}} \text{ (GS)}.$$

74. Maxwell's equations in differential form:

$$\nabla \times \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t} \text{ (SI)} \qquad \nabla \times \boldsymbol{E} = -\frac{1}{c}\frac{\partial \boldsymbol{B}}{\partial t} \text{ (SI)}$$

$$\nabla \cdot \boldsymbol{B} = 0 \text{ (SI)} \qquad \nabla \cdot \boldsymbol{B} = 0 \text{ (GS)}$$

$$\nabla \times \boldsymbol{H} = \boldsymbol{j} + \frac{\partial \boldsymbol{D}}{\partial t} \text{ (SI)} \qquad \nabla \times \boldsymbol{H} = \frac{4\pi}{c}\boldsymbol{j} + \frac{1}{c}\frac{\partial \boldsymbol{D}}{\partial t} \text{ (GS)}$$

$$\nabla \cdot \boldsymbol{D} = \rho \text{ (SI)} \qquad \nabla \cdot \boldsymbol{D} = 4\pi\rho \text{ (GS)}.$$

75. Maxwell s equations in integral form:

$$\oint_\Gamma \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{l} = -\int_S \frac{\partial \boldsymbol{B}}{\partial t} \cdot \mathrm{d}\boldsymbol{S} \text{ (SI)} \qquad \oint_\Gamma \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{l} = -\frac{1}{c}\int_S \frac{\partial \boldsymbol{B}}{\partial t} \cdot \mathrm{d}\boldsymbol{S} \text{ (GS)}$$

$$\oint_S \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{S} = 0 \text{ (SI)} \qquad \oint_S \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{S} = 0 \text{ (GS)}$$

$$\oint_\Gamma \boldsymbol{H} \cdot \mathrm{d}\boldsymbol{l} = \int_S \boldsymbol{j} \cdot \mathrm{d}\boldsymbol{S} + \int_S \frac{\partial \boldsymbol{D}}{\partial t} \cdot \mathrm{d}\boldsymbol{S} \text{ (SI)}$$

$$\oint_\Gamma \boldsymbol{H} \cdot \mathrm{d}\boldsymbol{l} = \frac{4\pi}{c}\int_S \boldsymbol{j} \cdot \mathrm{d}\boldsymbol{S} + \frac{1}{c}\int_S \frac{\partial \boldsymbol{D}}{\partial t} \cdot \mathrm{d}\boldsymbol{S} \text{ (GS)}$$

$$\oint_S \boldsymbol{D} \cdot \mathrm{d}\boldsymbol{S} = \int_V \rho\, \mathrm{d}V \text{ (SI)} \qquad \oint_S \boldsymbol{D} \cdot \mathrm{d}\boldsymbol{S} = 4\pi\int_V \rho\, \mathrm{d}V \text{ (GS)}.$$

76. Velocity of electromagnetic waves:

$$v = \frac{c}{\sqrt{\varepsilon\mu}}.$$

77. Relation between amplitudes of vectors $\boldsymbol{E}$ and $\boldsymbol{H}$ in an electromagnetic wave:

$$E_{\mathrm{m}} = \sqrt{\varepsilon_0 \varepsilon}\, H_{\mathrm{m}} \sqrt{\mu_0 \mu} \quad \text{(SI)} \qquad E_{\mathrm{m}} = \sqrt{\varepsilon}\, H_{\mathrm{m}} \sqrt{\mu} \quad \text{(GS)}.$$

78. Poynting vector:

$$\boldsymbol{S} = \boldsymbol{E} \times \boldsymbol{H} \quad \text{(SI)} \qquad \boldsymbol{S} = \frac{1}{4\pi c} \boldsymbol{E} \times \boldsymbol{H} \quad \text{(GS)}.$$

79. Density of electromagnetic field momentum:

$$\boldsymbol{K} = \frac{1}{c^2} \boldsymbol{E} \times \boldsymbol{H} \quad \text{(SI)} \qquad \boldsymbol{K} = \frac{1}{4\pi c^2} \boldsymbol{E} \times \boldsymbol{H} \quad \text{(GS)}.$$