B. M. YAVORSKY AND A. A. DETLAF

A Modern Handbook of PHYSICS

Б. М. ЯВОРСКИЙ, А. А. ДЕТЛАФ

СПРАВОЧНИК ПО ФИЗИКЕ

Москва «Наука»

B. M. YAVORSKY AND A. A. DETLAF

A Modern Handbook Of PHYSICS

Translated from the Russian by
NICHOLAS WEINSTEIN

На английском языке

Contents

PART TWO FUNDAMENTALS OF MOLECULAR PHYSICS AND THERMODYNAMICS

PART THREE ELECTRODYNAMICS

CHAPTER 14 ELECTRIC CHARGES. COULOMB'S LAW

CHAPTER 15 ELECTRIC FIELD STRENGTH AND DISPLACE-MENT

PART FOUR OSCILLATIONS AND WAVES

CHAPTER 31 ELECTROMAGNETIC WAVES

PART FIVE OPTICS

CHAPTER 32 INTERFERENCE OF LIGHT

CHAPTER 33 DIFFRACTION OF LIGHT

PART SIX ATOMIC AND MOLECULAR PHYSICS

PART SEVEN BASIC SOLID-STATE PHYSICS

PART EIGHT NUCLEAR PHYSICS AND ELEMENTARY PARTICLES

CHAPTER 42 STRUCTURE AND BASIC PROPERTIES OF ATOMIC NUCLEI

CHAPTER 43 ELEMENTARY PARTICLES

PART NINE APPENDICES

Preface

The basic sciences and physics, in particular, are of prime importance today in the training of engineers for the various branches of the national economy. This has led to fundamental changes in recent years in the teaching of physics in engineering institutes, and to the students of other educational institutions in which they do not major in physics. The scope and scientific level of physics courses have been substantially supplemented and cover the main trends in the development of modern physics. Consequently, the physics textbooks for engineering students have inevitably become three-volume editions of almost fifteen hundred pages. The need has arisen, in this connection, for a concise handbook on this subject. The aim of the authors was to fulfill this need.

In scope and depth this handbook includes all of the definitions, formulas and information covered in the most comprehensive and up-to-date physics courses of engineering institutes and the physics departments of universities and colleges. Physical laws are concisely formulated, all the necessary explanations are given and, in many cases, derivations as well. Though it plays a vital role in the teaching of physics, experimental material has been omitted. This is due only to a lack of space. All the units and symbols comply with the requirements of the SI. Units of physical quantities and systems of units are listed and dealt with in a short appendix.

This handbook is designed primarily for engineering students, as well as college and university students studying, but not majoring in physics. It can be used to advantage by engineers and graduate students, as well as by instructors and lecturers of intermediate schools and colleges.

Mathematical knowledge required in using the handbook is within the scope of the ordinary mathematics courses of engineering institutes. The detailed index and the numerous cross references, indicating the chapter, section and subsection, are of aid in finding any required information. The authors are sincerely grateful to the translator, N. Weinstein, and the science editor, T. Baranovskaya, for their comprehensive and expert work in the publication of the English edition of this handbook, as well as for their valuable suggestions, which should facilitate its use by the reader.

B. Yavorsky
A. Detlaf

Parts One, Four, Five and the appendices were written by Prof. A. A. Detlaf; Parts Two, Three, Six, Seven and Eight, by Prof. B. M. Yavorsky.

PART ONE MECHANICS

CHAPTER 1 KINEMATICS

1.1 Mechanical Motion. The Subject Matter of Mechanics

1.1.1 The simplest form of motion in nature is *mechanical motion*. It consists of changes in the relative positions of bodies or their parts in space in the course of time. The branch of physics dealing with the laws of mechanical motion is called *mechanics*. In a more narrow sense of the word, this is often understood to mean *classical mechanics*, which deals with the motion of macroscopic bodies travelling at velocities a great many times less than that of light in free space. Classical mechanics is based on Newton's laws and is therefore frequently referred to as *Newtonian mechanics*. The laws of motion of bodies at velocities close to that of light in free space are taken up in *relativistic mechanics* (5.1.1), and the laws of motion of microparticles (for example, the electrons in atoms, molecules, crystals, etc.) are the subject matter of *quantum mechanics* (38.1.1).
1.1.2 Classical mechanics comprises three basic branches: statics, kinematics and dynamics. *Statics* deals with the laws for the composition of forces and the conditions for the equilibrium of bodies. *Kinematics* provides a mathematical description of all possible kinds of mechanical motion, regardless of what causes each particular kind of motion. *Dynamics* investigates the effect of the interaction between bodies on their mechanical motion.
1.1.3 The mechanical properties of bodies are determined by their chemical nature, internal structure and state. These are beyond the scope of mechanics and are studied in other branches of physics. Hence, in accordance with the conditions of each specific problem on hand, various simplified models are employed to describe real bodies in motion. These include the particle, the perfectly rigid body, the perfectly elastic body, the perfectly inelastic body and others.

A *material point*, or *particle*, is a body whose shape and size are of no consequence under the conditions of the problem being considered. For example, in a first approximation, the motion of a ship from one point to another can be dealt with as the motion of a particle. If, however, it becomes necessary to take into account such details of this motion as the pitching and rolling of the ship by swells or heavy seas, the ship should be dealt with as an extended body of definite shape. In the literature the word particle is more frequently used than material point.

Any extended body or system of such bodies, making up the *mechanical system* being investigated, can be regarded as a *system of particles*. For this purpose, all the bodies of the system can be conceived of as being broken down into such a large number of parts that the size of each part is negligibly small compared to the sizes of the bodies themselves.

1.1.4 A *perfectly rigid body*, or simply *rigid body*, is one whose deformation can be neglected under the conditions of the given problem. The distance between any two points of a rigid body remains constant no matter what action the body is subjected to. A perfectly rigid body can be regarded as a system of particles rigidly connected to one another.

A *perfectly elastic body* is one whose deformation obeys Hooke's law (40.3.4). After the external forces applied to such a body have been removed, its initial size and shape are completely restored.

A *perfectly inelastic body* is one that completely retains its deformed state, caused by external forces applied to the body, when these forces have been removed.

1.2 Frames of Reference. Path, Path Length and Displacement Vector of a Particle

1.2.1 The position of a body in space can be specified only in reference to other bodies. It makes sense, for instance, to speak of the location of the planets with respect to the sun, that of an air liner or steamship with respect to the earth, but we cannot indicate their position in space "in general" without reference to some specific body. A perfectly rigid body in which a coordinate system is rigidly fixed, furnished with a "clock" and employed to determine the position of bodies and particles

being investigated in space at various instants of time, is called
a *frame of reference*. Sometimes the frame of reference is the
coordinate system itself if it is provided with a "clock" and
the solid in which the system is rigidly fixed is called the *body
of reference*. In each specific problem the frame of refer-
ence is selected so as to simplify the problem to the maximum
extent. Inertial frames of reference (2.1.2) are the most com-
monly used in physics.

1.2.2 Most extensively employed is the Cartesian rectangular
system of coordinates (Fig. 1.1) whose *orthonormal (normal or-
thogonal) basis* is formed by the
three vectors $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{k}$, which
are unit vectors with respect
to their magnitudes, mutually
orthogonal and pass through
the origin O of coordinates. The
position of an arbitrary point
M is specified by the radius
vector $\mathbf{r}$, which connects the ori-
gin O of coordinates to point M.
Vector $\mathbf{r}$ can be resolved along
the axes $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{k}$ of the basis
as follows:
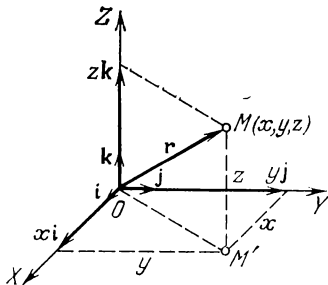
$$\mathbf{r} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k},$$

Fig. 1.1.

where $x\mathbf{i}$, $y\mathbf{j}$ and $z\mathbf{k}$ are the com-
ponents of vector $\mathbf{r}$ along the
coordinate axes. The coefficients of resolution, x, y and z, are
the Cartesian coordinates of point M and, due to the orthogo-
nality of the vectors of the basis, the projections of radius
vector $\mathbf{r}$ on the corresponding coordinate axes.
The motion of a particle is completely specified if three con-
tinuous and single-valued functions of the time t,

$$x = x\,(t), \quad y = y\,(t) \quad \text{and} \quad z = z\,(t),$$

are given to describe the changes in the coordinates of the
particle with time.
These equations are said to be the *kinematic equations of motion
of a particle*. They are equivalent to the single vector equation
$\mathbf{r} = \mathbf{r}\,(t)$, for a moving particle.

1.2.3 The line described by a moving particle in space is called its *path*. The kinematic equations of motion of a particle specify the equation of its path in parametric form (the parameter being the time t). Depending upon the shape of the path, distinction is made between *rectilinear* and *curvilinear* motion. The motion of a particle is said to be *plane* if its path lies completely in a single plane.

The mechanical motion of a body is relative, i.e. the nature of the motion and, in particular, the kind of path travelled by the points of the body depend upon the chosen frame of reference.

1.2.4 In the general case, the path of a moving particle is a spatial, rather than plane, curve. Introduced for such curves is the concept of an osculating plane. The *osculating plane* at the arbitrary point M of a curve is in the limiting position of a plane passing through any three points of the curve when these points approach point M without limit.
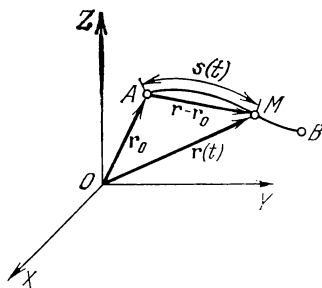
Fig. 1.2

An *osculating circle* at point M of a curve is the limit of a circle passing through three points of the curve being considered when these points approach point M without limit. The osculating circle lies in the osculating plane. The centre of the osculating circle and its radius are called the *centre of curvature* and *radius of curvature*, at point M, of the curve being considered. The straight line connecting point M to the centre of curvature is called the *principal normal* to the curve at point M. The tangent to the curve at point M is perpendicular to the principal normal and also lies in the osculating plane.

1.2.5 The *path length* is the sum of the lengths of all the portions of the path over which the particle passes in the given interval of time. The instant of time $t = t_0$, from which we begin to consider the motion of a particle, is called the *initial instant of time*, and the position of the particle at this instant (point A in Fig. 1.2) is called its *initial position*. Owing to the

arbitrary nature of the time reference point it is usually assumed that $t_0 = 0$. The path length s, travelled by the particle from its initial position, is a scalar function of time: $s = s(t)$. As is obvious from its definition, the path length of a particle, cannot be a negative quantity. If the particle travels along an arc of the path AB (Fig. 1.2), always in the same direction and is at point M at the instant of time t, then $s(t) = \overset{\frown}{AM}$. If the particle travels along the path in a more complicated way, for example if it travels from A to B, reaching B at the instant of time $t_1 < t$, and then reverses, returning to point M at instant t, then $s(t) = \overset{\frown}{AB} + \overset{\frown}{BM}$.

1.2.6 The *displacement vector* of a moving particle during the time interval from $t = t_1$ to $t = t_2$ is the vector from the position of the particle at the instant of time t_1 to its position at instant t_2. It is equal to the increment in the radius vector of the particle during the time interval being considered:

$$\mathbf{r}_2 - \mathbf{r}_1 = \mathbf{r}(t_2) - \mathbf{r}(t_1).$$

The displacement vector is always directed along the chord subtending the corresponding portion of the path.
Shown in Fig. 1.2 is the displacement vector of the particle for the time interval from t_0 to t. It is equal to $\mathbf{r} - \mathbf{r}_0 = \mathbf{r}(t) - \mathbf{r}(t_0)$.
The displacement vector of a moving particle for the time interval from t to $t + \Delta t$ equals

$$\Delta\mathbf{r} = \mathbf{r}(t + \Delta t) - \mathbf{r}(t) = \Delta x \times \mathbf{i} + \Delta y \times \mathbf{j} + \Delta z \times \mathbf{k},$$

where Δx, Δy and Δz are the increments (changes) in the coordinates of the particle during the time interval being considered.
1.2.7 A particle travelling freely in space can accomplish only three *independent motions*, i.e. motions that cannot be represented as combinations of the other two. As a matter of fact, the motion of a particle along each of the axes of a Cartesian rectangular coordinate system cannot be obtained by its motion along the other two axes. The number of independent motions that a mechanical system can accomplish is called the *number of degrees of freedom* of the system. Hence, a free particle has three degrees of freedom.

1.3 Velocity

1.3.1 The concept of velocity is introduced in mechanics to characterize the rate of motion of bodies. It is the time rate of change in the position of moving bodies. The *average velocity* of a moving particle in the time interval from t to $t + \Delta t$ is the vector $\mathbf{v}_{av}$, equal to the ratio of the increment Δr in the radius vector of the particle during this time interval to the length of the time interval Δt:

$$\mathbf{v}_{av} = \frac{\Delta \mathbf{r}}{\Delta t}.$$

Vector $\mathbf{v}_{av}$ has the same direction as Δr, i.e. along the chord subtending the corresponding portion of the particle's path.

1.3.2 The *velocity* (or *instantaneous velocity*) of a moving particle is the vector quantity $\mathbf{v}$, equal to the first time derivative of the radius vector $\mathbf{r}$ of the particle being considered:

$$\mathbf{v} = \frac{d\mathbf{r}}{dt}.$$

The velocity of the particle at the instant of time t is the limiting value that the average velocity $\mathbf{v}_{av}$ approaches for an infinitely small value of the time interval Δt. Thus

$$\mathbf{v} = \lim_{\Delta t \to 0} \frac{\Delta \mathbf{r}}{\Delta t} = \lim_{\Delta t \to 0} \mathbf{v}_{av},$$

The velocity vector $\mathbf{v}$ of the particle has the direction along a tangent to the path toward the motion in the same way as the vector $dr = \mathbf{v}\, dt$ of infinitely small displacement of the particle during an infinitely short time interval dt.

The path length ds travelled by the particle during the length of time dt is equal to the magnitude of the displacement vector: $ds = |\, d\mathbf{r}\,|$. Therefore the magnitude of the velocity vector of a moving particle is equal to the first time derivative of the path length:

$$v = |\, \mathbf{v}\,| = \frac{ds}{dt}.$$

1.3.3 The resolution of vector **v** along the basis of a Cartesian rectangular coordinate system is of the following form:

$$\mathbf{v} = v_x\mathbf{i} + v_y\mathbf{j} + v_z\mathbf{k}.$$

The projections of the velocity of a moving particle on the coordinate axes are equal to the first time derivative of the corresponding coordinates of the particle:

$$v_x = \frac{dx}{dt} \ , \ v_y = \frac{dy}{dt} \text{ and } v_z = \frac{dz}{dt},$$

and the magnitude of the velocity vector, sometimes called the *speed*, is

$$v = |\mathbf{v}| = \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2}.$$

1.3.4 In rectilinear motion of a particle, the direction of its velocity vector remains constant. The motion of a particle is said to be *uniform* if the magnitude of its velocity does not change with time: $v = ds/dt = \text{const}$. The path length s travelled by a uniformly moving particle is a linear function of time: $s = vt$ (provided that $t_0 = 0$, see Sect. 1.2.5).
If the magnitude of the particle's velocity, i.e. its speed, increases with time $(dv/dt > 0)$, the motion is said to be *accelerated*; if the speed decreases with time $(dv/dt < 0)$, the motion is said to be *decelerated*.
1.3.5 The *average velocity of nonuniform motion* of a particle over a given portion of its path is the scalar v_{av}, called the *travelling speed*, which is equal to the ratio of the length Δs of this portion to the length Δt of the time interval required to travel over this portion:

$$v_{av} = \frac{\Delta s}{\Delta t}.$$

It is equal to the speed of uniform motion in which the time required to travel the same path length Δs is the same as in the nonuniform motion being considered.
In curvilinear motion of a particle $|\Delta r| < \Delta s$. Therefore, in the general case, the average velocity (travelling speed) v_{av}

of the particle is not equal to the magnitude of its average velocity v_{av} over the same portion of its path (1.3.1): $v_{av} \geqslant$ $\geqslant |\mathbf{v}_{av}|$, where the equality sign corresponds to a rectilinear portion of the path.

1.3.6 In the case of plane motion of a particle M (1.2.3), it often proves convenient to use the polar coordinates r and φ, where r is the distance from pole O to particle M and φ is the vectorial angle measured from polar axis OA (Fig. 1.3). The velocity $\mathbf{v}$ of particle M can be resolved into two mutually perpendicular components: the *radial velocity* $\mathbf{v}_r$ and the *transverse velocity* $\mathbf{v}_\varphi$:
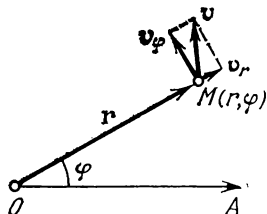
Fig. 1.3

$$\mathbf{v} = \mathbf{v}_r + \mathbf{v}_\varphi,$$

in which

$$\mathbf{v}_r = \frac{1}{r}\frac{dr}{dt}\mathbf{r} \text{ and } \mathbf{v}_\varphi = \frac{d\varphi}{dt}[\mathbf{kr}],$$

where $\mathbf{r}$ is the polar radius vector of particle M, and $\mathbf{k}$ is the unit vector directed perpendicular to the plane of motion of the particle so that the rotation of radius vector $\mathbf{r}$ upon an increase in angle φ is seen, from the head of unit vector $\mathbf{k}$, to be counterclockwise.

The magnitude v of the velocity vector of particle M, in plane motion, is

$$v = \sqrt{\left(\frac{dr}{dt}\right)^2 + \left(r\frac{d\varphi}{dt}\right)^2}.$$

During the infinitely short length of time dt, polar radius vector $\mathbf{r}$ of the particle in plane motion sweeps over a sector of a circle, the area of the sector being $dS = r^2 d\varphi/2$. Hence, the quantity

$$\sigma = \frac{dS}{dt} = \frac{1}{2}r^2\frac{d\varphi}{dt} = \frac{1}{2}rv_\varphi$$

is called the *areal velocity*.

1.4 Acceleration

1.4.1 The concept of acceleration is introduced in mechanics to characterize the rate of change of the velocity vector of a moving particle. The *average acceleration* in the time interval from t to $t + \Delta t$ is the vector a_{av}, which is equal to the ratio of the increment Δv of the velocity vector of the particle, during the time interval, to the length of the time interval Δt:

$$a_{av} = \frac{\Delta v}{\Delta t} .$$

1.4.2 *Acceleration* (or *instantaneous acceleration*) of a moving particle is the vector quantity a, equal to the first time derivative of the velocity v of the particle being considered or, otherwise, the second time derivative of the particle's radius vector r:

$$a = \frac{dv}{dt} = \frac{d^2 r}{dt^2} .$$

The acceleration of a moving particle at the instant of time t is the limiting value that the average acceleration a_{av} approaches for an infinitely small time interval Δt. Thus

$$a = \lim_{\Delta t \to 0} \frac{\Delta v}{\Delta t} = \lim_{\Delta t \to 0} a_{av} .$$

1.4.3 The resolution of vector a along the basis of a rectangular Cartesian coordinate system is of the form:

$$a = a_x i + a_y j + a_z k.$$

The projections of the acceleration vector of a moving particle on the coordinate axes are equal to the first time derivatives of the corresponding velocity projections or, otherwise, the second time derivatives of the corresponding coordinates of the particle:

$$a_x = \frac{dv_x}{dt} = \frac{d^2 x}{dt^2} , \quad a_y = \frac{dv_y}{dt} = \frac{d^2 y}{dt^2} \quad \text{and} \quad a_z = \frac{dv_z}{dt} = \frac{d^2 z}{dt^2} .$$

The magnitude of the acceleration vector is

$$a = |\mathbf{a}| = \sqrt{\left(\frac{dv_x}{dt}\right)^2 + \left(\frac{dv_y}{dt}\right)^2 + \left(\frac{dv_z}{dt}\right)^2}$$

$$= \sqrt{\left(\frac{d^2x}{dt^2}\right)^2 + \left(\frac{d^2y}{dt^2}\right)^2 + \left(\frac{d^2z}{dt^2}\right)^2}.$$

1.4.4 The acceleration vector of a moving particle lies in the osculating plane (1.2.4) passing through point M of the path being considered and is directed toward the concavity of the path BC (Fig. 1.4). In this plane the acceleration vector $\mathbf{a}$ can be resolved into two mutually perpendicular components $\mathbf{a}_\tau$ and $\mathbf{a}_n$:
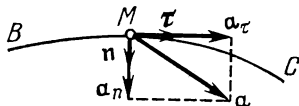
Fig. 1.4

$$\mathbf{a} = \mathbf{a}_\tau + \mathbf{a}_n.$$

1.4.5 Component $\mathbf{a}_\tau$ is called the *tangential acceleration* of the moving particle. It is directed along the tangent to the path and equals

$$\mathbf{a}_\tau = \frac{dv}{dt}\,\boldsymbol{\tau} \quad \text{and} \quad a_\tau = \frac{dv}{dt},$$

where $\boldsymbol{\tau} = \mathbf{v}/v$ is the unit vector of the tangent, from point M of the path in the direction of velocity $\mathbf{v}$ of the particle, and a_τ is the projection of the tangential acceleration on the direction of vector $\mathbf{v}$. The tangential acceleration characterizes the rate of change of the magnitude of the particle's velocity vector. Vectors $\mathbf{a}_\tau$ and $\mathbf{v}$ coincide in direction, i.e. $a_\tau > 0$, in accelerated motion (1.3.4); vectors $\mathbf{a}_\tau$ and $\mathbf{v}$ are opposite in direction i.e. $a_\tau < 0$, in decelerated motion of the particle; and $a_\tau = 0$ in its uniform motion. If $a_\tau = \text{const} \neq 0$, the motion is said to be *uniformly variable*. In uniformly variable motion the magnitude of the particle's velocity is linearly time-dependent:

$$v = v_0 + a_\tau t,$$

where $v_0 = v\,(0)$ is the magnitude of the initial velocity, i.e. the velocity at the initial instant of time $t = 0$. If $a_\tau = \text{const} > > 0$, the motion of the particle is said to be *uniformly acceler-*

ated, whereas if $a_\tau = $ const < 0, it is said to be *uniformly de-celerated*.

1.4.6 Component $\mathbf{a}_n$ of acceleration $\mathbf{a}$ of the moving particle is called its *normal acceleration*. It is directed along the principal normal to the path at point M being considered, toward the centre of curvature (1.2.4) of the path. Therefore, $\mathbf{a}_n$ is often called the *centripetal acceleration* of the particle. The normal acceleration equals

$$\mathbf{a}_n = \frac{v^2}{R}\,\mathbf{n},$$

where $\mathbf{n}$ is the unit vector of the principal normal, and R is the radius of curvature of the path. The normal acceleration charac-terizes the rate of change in direction of the particle's velocity vector.

If the particle travels in a straight line the normal acceleration $\mathbf{a}_n = 0$ and the acceleration is equal to the tangential accele-ration of the particle: $\mathbf{a} = \mathbf{a}_\tau$.

1.5 Translational and Rotary Motion of a Rigid Body *

1.5.1 *Translation* is the kind of motion of a rigid body in which any straight line rigidly fixed in the body (for instance, line AB in Fig. 1.5), remains paral-lel to its initial position (A_0B_0).
With respect to the earth, a lift cabin (elevator car), lathe tool, and many other items have translational motion.
In the translational motion of a rigid body all of its points are displaced in an identical way. During the infinitely short time dt the radius vectors of these points are changed by the same amount $d\mathbf{r}$. Hence the velocity of all points of the body is the

Fig. 1.5

same in each instant of time and is equal to $d\mathbf{r}/dt$ and, conse-quently, their acceleration is also the same. For this reason,
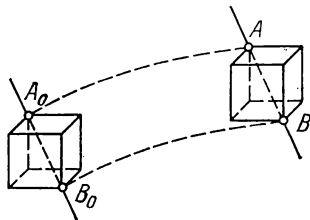
* Only perfectly rigid bodies are dealt with in this section; for the sake of brevity, they have been called simply "rigid bodies".

a kinematic discussion of the translational motion of a rigid
body reduces to the investigation of the motion of any one of
its points, or particles. Dynamics usually deals with the motion
of the centre of mass of the body (2.3.3). A rigid body travel-
ling freely in space has three translational degrees of freedom
(1.2.7), corresponding to its translational motions along the
three coordinate axes.

1.5.2 The motion of a rigid body in which two of its points, A
and B, remain fixed is called *rotation* (or *rotary motion*) *of the
body about a fixed axis*. The fixed straight line AB is the axis
of rotation. In the rotation of a rigid body about a fixed axis
all the points of the body describe circles whose centres lie on
the axis of rotation and whose planes are perpendicular to this
axis. The rotors of turbines, motors and generators, for exam-
ple, have rotary motion with respect to the earth if these ma-
chines are rigidly mounted.

A rigid body rotating about a fixed axis has only one degree of
freedom (1.2.7). Its position in space can be completely defined
by specifying the angle of rotation (position angle) φ from some
definite (initial) position.

1.5.3 The angular velocity serves to characterize the rate and
direction of rotation of a body about an axis. The *angular veloc-
ity* is the vector $\boldsymbol{\omega}$, which is equal in magnitude to the first
derivative of the angle of rotation φ with respect to time t
and is directed along the fixed axis of rotation in such a way
that from its head the rotation of the body is counterclockwise
(Fig. 1.6)*:

$$\boldsymbol{\omega} = \frac{d\varphi}{dt} \quad \text{and} \quad \omega = \frac{d\varphi}{dt}.$$

---

* The direction of vector $\boldsymbol{\omega}$ can also be determined by the
right-hand screw rule. It coincides with the direction of transla-
tional motion of a right-hand screw rotating together with the
body. Vectors like $\boldsymbol{\omega}$, whose direction is related to the direction
of rotation and is reversed in transforming from a right-handed
to a left-handed system of coordinates, are called *pseudovectors*,
or *axial vectors* (in distinction to ordinary, *polar vectors*, which
are not reversed in the above-mentioned transformation of coor-
dinates). For example, the vector product of two polar vectors
is a pseudovector, and that of a pseudovector and a polar vector
is a polar vector.

Here $d\varphi$ is the vector of an element of rotation of the body during the time dt. This vector is directed along the axis of rotation, also according to the right-hand screw rule (Fig. 1.6). Axial vectors $d\varphi$ and ω have no definite points of application: they can be laid off from any point on the axis of rotation. In Fig. 1.6 they are laid off from a certain point O of the fixed
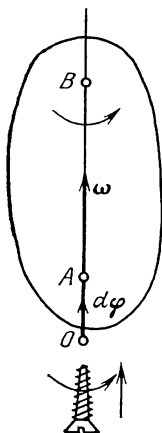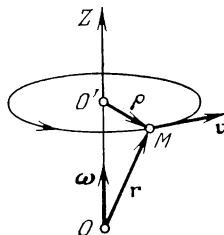
Fig. 1.6   Fig. 1.7

axis of rotation. This point is also taken as the origin of the reference frame.

Rotation of a body is said to be *uniform* if the magnitude of its angular velocity remains constant with time $\omega = $ const. In this case the angle of rotation of the body is linearly dependent on time: $\varphi = \omega t$.

1.5.4 An arbitrary point M of a rigid body rotating about fixed axis OZ with the angular velocity ω describes a circle of radius ρ with its centre at point O' (Fig. 1.7). As distinct from the angular velocity of the body, the velocity v of point M is often called the *linear velocity*. It is perpendicular both to the axis of rotation (i.e. to vector ω) and to radius vector ρ, from the

centre O' of the circle to point M, and is equal to their vector product:

$\mathbf{v} = [\boldsymbol{\omega}\boldsymbol{\rho}] = [\boldsymbol{\omega}\mathbf{r})$ and $v = \omega\rho$.

Here $\mathbf{r} = \mathbf{OO'} + \boldsymbol{\rho}$ and is the radius vector of point M, from point O of the axis of rotation. Point O is taken as the origin of coordinates.

1.5.5 The *period of revolution* is the time interval T required for a body, rotating uniformly at the angular velocity ω, to complete one revolution about its axis of rotation (i.e. to rotate through the angle $\varphi = 2\pi$): $T = 2\pi/\omega$.

The *rotational speed* $n = 1/T = \omega/2\pi$ is the number of revolutions made by a body in unit time in uniform rotation at the angular velocity ω.

1.5.6 The rotation of a rigid body in which one of its points remains fixed is said to be *rotation about a fixed point*. This point is usually taken as the origin of the fixed reference frame. In such rotation, all the points of the body travel along the surfaces of concentric spheres whose centre is at the fixed point. At each instant of time, this motion of the rigid body can be regarded as rotation about a certain axis that passes through the fixed point and is called the *instantaneous axis of rotation*. In the general case, the position of the instantaneous axis of rotation varies with respect to both a fixed reference frame and one rigidly fixed in the rotating body.

The velocity $\mathbf{v}$ of arbitrary point M of such a body is

$\mathbf{v} = [\boldsymbol{\omega}\mathbf{r}]$ and $v = \omega\rho$.

Here $\boldsymbol{\omega} = d\boldsymbol{\varphi}/dt$ is the angular velocity of the body and is directed along the instantaneous axis of rotation as is the vector $d\boldsymbol{\varphi}$ of an element of rotation of the body during the infinitely short time dt, $\mathbf{r}$ is the radius vector from fixed point O, about which the body rotates, to point M, and ρ is the distance from point M to the instantaneous axis of rotation. Such a body can have three independent motions: it can rotate about each of three mutually perpendicular axes passing through fixed point O. It consequently has three degrees of freedom (1.2.7).

1.5.7 The *angular acceleration* vector ε is introduced to indicate the rate of change of the angular velocity vector with time in nonuniform rotation of a body about a fixed axis or its rotation

about a fixed point. Vector ε is equal to the first derivative of the angular velocity ω of the body with respect to the time t:

$$\varepsilon = \frac{d\omega}{dt}.$$

If the body rotates about a fixed axis, vector ε is directed along this axis. It coincides in direction with ω in accelerated rotation $(d\omega/dt > 0)$ and is opposite to ω in direction in decelerated rotation $(d\omega/dt < 0)$. The projection of the angular acceleration on the fixed axis OZ of rotation is equal to

$$\varepsilon_z = \frac{d\omega_z}{dt},$$

where ω_z is the projection of vector ω on the same axis.

1.5.8 As distinct from the angular acceleration, the acceleration a of arbitrary point M of a rigid body rotating about fixed point O, or about a fixed axis passing through point O, is often called the *linear acceleration*. It equals

$$a = \frac{dv}{dt} = \frac{d}{dt}[\omega r] = a_r + a_{ax},$$

where $a_r = [\varepsilon r]$ is the *rotary acceleration* of the point, and $a_{ax} = [\omega [\omega r]]$ is its *axipetal acceleration*, which is directed toward the instantaneous axis of rotation.

If the body rotates about the fixed axis OZ (Fig. 1.7), the rotary acceleration of point M coincides with its tangential acceleration a_τ (1.4.5), and its axipetal acceleration coincides with its normal acceleration a_n (1.4.6):

$$a_\tau = [\varepsilon r] = [\varepsilon \rho] \quad \text{and} \quad a_n = -\omega^2 \rho.$$

1.5.9 Any complex motion of a rigid body can be broken down into two simple motions: translational motion at the velocity v_A of a certain arbitrarily chosen point A of the body and rotation about an instantaneous axis passing through this point. The angular velocity ω of rotation is independent of the choice of point A. The velocity of arbitrary point M of the body is

$$v = v_A + [\omega (r - r_A)],$$

where r and r_A are the radius vectors of points M and A.

In the dynamics of a rigid body it usually proves convenient to deal with its complex motion as the combination of two simultaneous motions: translational motion at the velocity of the body's centre of mass (2.3.3) and rotation about the centre of mass.

The simplest case of complex motion of a body is *plane*, or *plane-parallel*, *motion*, in which all the points of the body travel in parallel planes. This is the motion, for instance, of a homogeneous circular cylinder rolling down an inclined plane. In plane motion the direction of the instantaneous axis of rotation of the body about point A remains constant and vectors ω and v_A are perpendicular to each other.

CHAPTER 2   NEWTON'S LAWS

2.1 Newton's First Law. Inertial Frames of Reference

2.1.1 As the *first law of dynamics* Newton took a law that had already been established by Galileo: a particle persists in its state of rest or uniform motion in a straight line until the action of other bodies compel it to change that state.

Newton's first law indicates that no external action is required to maintain a state of rest or of uniform motion in a straight line. This manifests a special dynamic property of bodies, called their *inertia*. Newton's first law is consequently called the *law of inertia*, and the motion of a body not subject to the action of other bodies is said to be *inertial motion*.

2.1.2 Mechanical motion is relative: its nature may differ for different frames of reference (1.2.1) that are moving in relation to each other. An astronaut, for instance, on board a man-made satellite orbiting the earth, may be at rest in a reference frame fixed in the satellite. At the same time, he is travelling together with the satellite with respect to the earth along an elliptic orbit, i.e. not with uniform motion and not in a straight line. It is thus obvious that Newton's first law does not hold for any frame of reference. A ball, for example, lying on the smooth floor of a stateroom in a ship travelling with uniform motion at sea in a straight line, may suddenly start moving along the floor without being subject to the action of any other bodies. This

will occur when the velocity of the ship, i.e. its speed and/or direction, begins to change.

A frame of reference with respect to which a particle, not subject to external action, remains at rest or moves uniformly in a straight line is called an *inertial frame of reference*. The content of Newton's first law reduces, in essence, to two statements: in the first place, that all material bodies possess the property of inertia, and secondly, that inertial reference frames exist.

2.1.3 Any two inertial reference frames can move only with translational motion with respect to each other, and only uniformly in a straight line. It has been experimentally established that the *heliocentric frame of reference* is practically inertial. The origin of its coordinate system is located at the centre of mass (2.3.3) of the solar system (approximately at the centre of the sun) and the three axes extend in the directions of three distant stars selected, for example, so that these coordinate axes are mutually perpendicular.

A *laboratory frame of reference*, whose coordinate axes are rigidly fixed in the earth, is noninertial chiefly owing to the earth's diurnal (daily) rotation. But the earth rotates so slowly that the maximum normal acceleration (1.4.6) of points on its surface due to diurnal rotation does not exceed 0.034 m/s². Therefore, in the great majority of practical problems, the laboratory reference frame can be regarded as being approximately inertial.

2.1.4 Inertial reference frames are of prime importance in all branches of physics, not only mechanics. This is due to the fact that according to Einstein's principle of relativity (5.1.2) the mathematical expression of any physical law should have the same form in all inertial reference frames. Hence, in the following we shall make use of only inertial frames of reference, without stipulation each time. The laws of motion of a particle with respect to a noninertial reference frame is dealt with in Chapter 7.

2.2 Force

2.2.1 *Force* is a vector quantity which is a measure of the mechanical action exerted on the body being considered by other bodies. Mechanical interaction can be accomplished either between bodies in direct contact (for instance, in friction, and in

the pressure of bodies against one another), or between remote
bodies. A special form of matter, connecting particles of matter
into unified systems and transmitting the action of certain par-
ticles on others at finite velocities, is called a *physical field*, or
simply *field*. Interaction between remote bodies is accomplished
by the gravitational or electromagnetic fields they set up (for
example, the attraction of the planets to the sun, and the inter-
action of charged bodies or current-carrying conductors). Me-
chanical action exerted on a given body by other bodies can be
manifested in two ways. In the first place, it can lead to a change
in the state of mechanical motion of the body being consid-
ered, and secondly, to its deformation. Both of these effects can
serve as the basis for measuring forces. For example, the mea-
surement of forces by means of a spring dynamometer is based
on Hooke's law (40.3.4) for axial tension.

Using the concept of force, we usually speak in mechanics of
the motion and deformation of a body by the action of the forces
applied to it. Here, of course, each force always corresponds
to a certain body exerting this force on the body being consid-
ered.

Force $\mathbf{F}$ is completely specified by giving its magnitude, direc-
tion in space and point of application. The straight line along
which a force is exerted is called its *line of action*.

A field acting on a particle with the force $\mathbf{F}$ is said to be *steady*,
or *stationary*, if it does not change in the course of time t, i.e.
if force $\mathbf{F}$ is explicitly independent of time at any point in the
field: $\partial\mathbf{F}/\partial t \equiv 0$. For a field to be steady it is necessary that
the bodies which set it up are at rest with respect to the ref-
erence frame used in considering the field.

2.2.2 The simultaneous action of several forces, $\mathbf{F}_1$, $\mathbf{F}_2$, . . ., $\mathbf{F}_n$
(Fig. 2.1a) on a particle M is equivalent to the action of a single
force, called the *resultant force* and equal to the vector sum of
the component forces. Thus

$$\mathbf{F} = \sum_{i=1}^{n} \mathbf{F}_i .$$

It is the closing line of the polygon of forces $\mathbf{F}_1$, $\mathbf{F}_2$, . . ., $\mathbf{F}_n$
(Fig. 2.1b).

If a body is perfectly rigid, the action of a force on it does not
change when the point of application is displaced along the
line of action, but within the limits of the body. In other words,

forces applied to perfectly rigid bodies can be regarded as non-localized vectors.

2.2.3 A body is said to be *free* if no limitations are imposed on its position and motion in space. An airplane, for example, flying in the air is a free body, as is a submarine travelling in the depths of the sea. Most bodies that we deal with are not
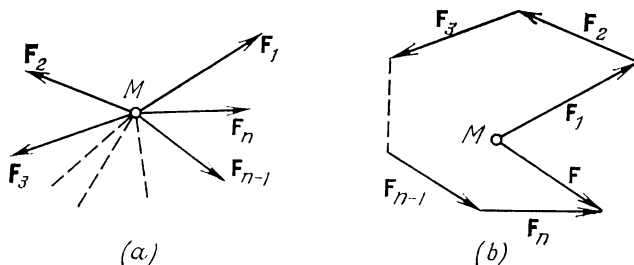
Fig. 2.1

free; certain limitations are imposed on their possible positions and motion. These limitations are called *constraints* in mechanics. A ball, for example, hung on a nonextensible thread, cannot move away from its point of suspension by more than the length of the thread; a streetcar (tram) must travel only along its tracks. Constraints are realized due to the action on the body being considered exerted by other bodies attached to or contacting this body (for instance, the thread on the ball tied to it, the track on the streetcar, etc.).

In investigating the behaviour of constrained bodies or systems of bodies in mechanics, the *principle of releasability* is applied. It states that a restrained body (or system of bodies) can be treated as a free body (or system) if the bodies imposing the constraints are replaced by the corresponding forces. These forces are called *reaction forces* of the constraints, or *constraint reactions*. All other forces exerted on the body are called *active forces*. Thus, the motion of the ball hung by the thread can be treated as the motion of a free ball subject, in addition to all

the active forces applied to it (for instance, gravity), to the
reaction of the thread as well.

In contrast to the active forces, which should be given in each
specific problem, the constraint reactions are not known before-
hand. They are to be determined in solving the problem. Their
values should be such that under the combined effects of the
active forces and constraint reactions, the motion of the "re-
leased" body completely agrees with the limitations imposed by
the constraints on the constrained body being considered. There
are no other distinctions between constraint reactions and active
forces.

2.2.4 Bodies not included in the mechanical system being in-
vestigated are called *external bodies*. Forces exerted on the
system by external bodies are called *external forces.* Corres-
pondingly, forces of interaction between the parts or particles
of the system being considered are called *internal forces.*

A mechanical system is said to be *closed*, or *isolated*, if it does
not interact with external bodies. No external forces act on
any of the bodies of a closed system.

2.3 Mass. Momentum

2.3.1 In classical (Newtonian) mechanics, the *mass of a particle*
(material point) is a positive scalar quantity which is a measure
of the inertia of this particle. Under the action of a force the
velocity of a particle changes *gradually*, not instantaneously,
i.e. it acquires an acceleration *finite* in magnitude and inversely
proportional to the mass of the particle. To compare the masses,
m_1 and m_2, of two particles, it is sufficient to measure the
magnitudes a_1 and a_2 of the accelerations acquired by these
particles when the same force is applied to them: $m_2/m_1 =
= a_1/a_2$. The mass of a body is ordinarily determined by weigh-
ing by beam scales.

Classical (Newtonian) mechanics contends that:

(a) the mass of a particle is independent of its state of motion,
being its unalterable characteristic;

(b) mass is an additive quantity, i.e. the mass of a system (for
instance, a body) is equal to the sum of the masses of all the
particles making up the system;

(c) the mass of an isolated (closed) system (2.2.4) does not

change with any processes occurring within the system (*law of conservation of mass*).

These propositions of Newtonian mechanics were reconsidered and revised in relativistic mechanics (5.6.1, 5.6.2, 5.7.3 and 5.7.6).

2.3.2 The density ρ of a body at a given point M is the ratio of the mass dm of an infinitely small element of the body, which includes point M, to the quantity dV, which is the volume of this element. Thus

$$\rho = \frac{dm}{dV} .$$

The size of the element being considered should be so small that any variation of density within its limits can be neglected. On the other hand, this size should be a great many times larger than intermolecular distances.

A body is said to be *homogeneous* if the density is constant at all of its points. The mass of a homogeneous body is the product of its density by its volume: $m = \rho V$.

The mass of a nonhomogeneous body is

$$m = \int_{(V)} \rho \, dV,$$

where ρ is a function of the coordinates, and integrating is carried out over the whole volume of the body. The average density $\langle \rho \rangle$ of a nonhomogeneous body is the ratio of its mass to its volume: $\langle \rho \rangle = m/V$.

2.3.3 The *centre of inertia*, or *centre of mass*, of a system of particles is point C, whose radius vector $\mathbf{r}_C$ equals

$$\mathbf{r}_C = \frac{1}{m} \sum_{i=1}^{n} m_i \mathbf{r}_i,$$

where m_i and $\mathbf{r}_i$ are the mass and radius vector of the ith particle, n is the total number of particles in the system, and $m = \sum_{i=1}^{n} m_i$ is the mass of the whole system,

The velocity of the centre of mass is

$$\mathbf{v}_C = \frac{d\mathbf{r}_C}{dt} = \frac{1}{m} \sum_{i=1}^{n} m_i \mathbf{v}_i .$$

2.3.4 The vector quantity $\mathbf{p}_i$, equal to the product of the mass m_i of a particle by its velocity $\mathbf{v}_i$, is called the *momentum of this particle*. The *momentum of a system* of particles is the vector $\mathbf{p}$, which is the geometric sum of the momenta of all the particles making up the system:

$$\mathbf{p} = \sum_{i=1}^{n} \mathbf{p}_i .$$

The momentum of a system is equal to the product of the mass of the whole system by the velocity of its centre of mass: $\mathbf{p} = m\mathbf{v}_C$.

2.4 Newton's Second Law

2.4.1 The basic law of the dynamics of a particle is Newton's second law, which discusses how the mechanical motion of a particle is changed by the effect of forces applied to it. *Newton's second law* states: the rate of change in the momentum $\mathbf{p}$ of a particle is equal to the force $\mathbf{F}$ acting on it. Thus

$$\frac{d\mathbf{p}}{dt} = \mathbf{F} \quad \text{or} \quad \frac{d}{dt}(m\mathbf{v}) = \mathbf{F},$$

where m and $\mathbf{v}$ are the mass and velocity of the particle.
If several forces act simultaneously on a particle, the force $\mathbf{F}$ in Newton's second law is understood to be the geometric sum of all acting forces, both active forces and constraint reactions (2.2.3), i.e. the resultant force (2.2.2).
2.4.2 The vector quantity $\mathbf{F}\, dt$ is called an element of *impulse of force* $\mathbf{F}$ during the infinitely short time dt of its action. The impulse of force $\mathbf{F}$ acting in the finite time interval from $t = t_1$

to $t = t_2$ is equal to the definite integral $\int\limits_{t_1}^{t_2} \mathbf{F}\, dt$, where $\mathbf{F}$, in the general case, depends upon the time t.

According to Newton's second law, the change in the momentum of a particle is equal to the impulse of the force acting on it:

$$d\mathbf{p} = \mathbf{F}\, dt \quad \text{and} \quad \Delta \mathbf{p} = \mathbf{p}_2 - \mathbf{p}_1 = \int\limits_{t_1}^{t_2} \mathbf{F}\, dt,$$

where $\mathbf{p}_2 = \mathbf{p}\,(t_2)$ and $\mathbf{p}_1 = \mathbf{p}\,(t_1)$ are the values of the momentum of a particle at the end $(t = t_2)$ and beginning $(t = t_1)$ of the time interval being considered.

2.4.3 Since in Newtonian mechanics the mass m of a particle is independent of its state of motion, $dm/dt = 0$. Hence, the mathematical expression of Newton's second law can also be presented in the form

$$\mathbf{a} = \frac{\mathbf{F}}{m},$$

where $\mathbf{a} = d\mathbf{v}/dt = d^2\mathbf{r}/dt^2$ is the acceleration of the particle and $\mathbf{r}$ is its radius vector. The corresponding statement of *Newton's second law* reads: the acceleration of a particle coincides in direction with the force acting on it and is equal to the ratio of this force to the mass of the particle.

The tangential and normal accelerations of a particle (1.4.4, 1.4.5 and 1.4.6) are determined by the corresponding components of force $\mathbf{F}$:

$$\mathbf{a}_\tau = \frac{\mathbf{F}_\tau}{m}, \quad a_\tau = \frac{dv}{dt} = \frac{F_\tau}{m},$$

and

$$\mathbf{a}_n = \frac{\mathbf{F}_n}{m}, \quad a_n = \frac{v^2}{R} = \frac{F_n}{m},$$

where v is the magnitude of the particle's velocity vector, and R is the radius of curvature of its path. The force $\mathbf{F}_n$, imparting normal acceleration to a particle, is directed toward the centre

of curvature (1.2.4) of the particle's path and is therefore called the *centripetal force*.

2.4.4 If a particle is subject to the simultaneous action of several forces, $\mathbf{F}_1$, $\mathbf{F}_2$, . . ., $\mathbf{F}_n$, its acceleration is

$$\mathbf{a} = \frac{1}{m} \sum_{l=1}^{n} \mathbf{F}_l = \sum_{l=1}^{n} \mathbf{a}_l,$$

where $\mathbf{a}_l = \mathbf{F}_l/m$. Consequently, each of the forces acting simultaneously on a particle imparts to it the acceleration it would have if there were no other forces (*principle of superposition*, or *independence*, *of forces*).
The equation

$$m \frac{d^2\mathbf{r}}{dt^2} = \mathbf{F} = \sum_{l=1}^{n} \mathbf{F}_l$$

is called the *differential equation of motion of a particle*.
In projections on the axes of a rectangular Cartesian coordinate system, this equation is of the form

$$m \frac{d^2x}{dt^2} = F_x, \ m \frac{d^2y}{dt^2} = F_y \text{ and } m \frac{d^2z}{dt^2} = F_z,$$

where x, y and z are the coordinates of the moving particle.

2.5 Newton's Third Law. Motion of the Centre of Mass

2.5.1 The mechanical action exerted by bodies on one another is of the nature of their *interaction*. This is what *Newton's third law* deals with: two particles exert forces on each other which are equal in magnitude and opposite in direction along a line connecting these particles.
If $\mathbf{F}_{ik}$ is a force exerted on the ith particle by the kth one, and $\mathbf{F}_{ki}$ is the force exerted on the kth particle by the ith one, then, according to Newton's third law,

$$\mathbf{F}_{ki} = -\mathbf{F}_{ik}.$$

Forces $\mathbf{F}_{ih}$ and $\mathbf{F}_{hi}$ are applied to different particles and can be mutually counterbalanced only when particles i and k belong to the same perfectly rigid body.

2.5.2 Newton's third law is an essential supplement to the first and second laws. It permits us to go over from the dynamics of a separate particle to that of an arbitrary mechanical system (system of particles). It follows from Newton's third law that in any mechanical system the geometric sum of all the internal forces (2.2.4) equals zero:

$$\sum_{i=1}^{n} \sum_{k=1}^{n} \mathbf{F}_{ih} = 0$$

where n is the number of particles in the system, and $\mathbf{F}_{ii} = 0$. The vector Fext, equal to the vector sum of all the external forces (2.2.4) acting on the system, is called the *principal vector of external forces*:

$$\mathbf{F}\text{ext} = \sum_{i=1}^{n} \mathbf{F}_{i}^{\text{ext}}$$

where $\mathbf{F}_{i}^{\text{ext}}$ is the resultant external force applied to the ith particle.

2.5.3 It follows from Newton's first and second laws that the first derivative with respect to the time t of the momentum $\mathbf{p}$ of a mechanical system (2.3.4) is equal to the principal vector of all external forces applied to the system. Thus

$$\frac{d\dot{\mathbf{p}}}{dt} = \mathbf{F}\text{ext}.$$

This equation expresses the *law of the change in the momentum of a system*.

Since $\mathbf{p} = m\mathbf{v}_C$, where m is the mass of a system and $\mathbf{v}_C$ is the velocity of its centre of inertia, or mass, the *law of motion of the centre of mass* of a mechanical system is of the form

$$\frac{d}{dt}(m\mathbf{v}_C) = \mathbf{F}\text{ext} \text{ or } m\mathbf{a}_C = \mathbf{F}\text{ext},$$

where $\mathbf{a}_C = d\mathbf{V}_C/dt$ is the acceleration of the centre of mass. Thus, the centre of mass (centre of inertia) of a mechanical

system travels like a particle whose mass is equal to the mass
of the whole system and which is subject to a force equal to
the principal vector of the external forces applied to the system.
If the system being considered is a rigid body that travels with
translational motion (1.5.1), the velocities $\mathbf{v}_i$ of all the points
of the body and $\mathbf{v}_C$ of its centre of mass are the same and are
equal to velocity $\mathbf{v}$ of the body. Correspondingly, the accelera-
tion of the body $\mathbf{a} = \mathbf{a}_C$ and the *basic equation in the dynamics
of translational motion of a rigid body* is of the form

$m\mathbf{a} = \mathbf{F}^{\text{ext}}.$

2.6 Motion of a Body of Variable Mass

2.6.1 In Newtonian mechanics the mass of a body can be
changed only if particles of matter are detached from or added
to it. An example of such a body is a rocket. In its flight the mass
of the rocket is gradually reduced because the gaseous products of
combustion in the rocket engine are ejected through the noz-
zles.
The equation of translational motion of a body of variable mass
(*Meshchersky equation*) is

$$m\frac{d\mathbf{v}}{dt} = \mathbf{F}^{\text{ext}} + (\mathbf{v}_1 - \mathbf{v})\,\frac{dm}{dt},$$

where m and $\mathbf{v}$ are the mass and velocity of the body at the in-
stant of time being considered, $\mathbf{F}^{\text{ext}}$ is the principal vector of
the external forces (2.5.2) acting on the body, and $\mathbf{v}_1$ is the
velocity of the detracted particles *after separation* from the
rocket (if $dm/dt < 0$) or the added particles *before being joined*
(if $dm/dt > 0$).
2.6.2 The second member on the right-hand side of the Me-
shchersky equation represents the additional force acting on a
body of variable mass. It is called the *reactive force*:

$$\mathbf{F}_r = (\mathbf{v}_1 - \mathbf{v})\frac{dm}{dt} = \mathbf{u}\frac{dm}{dt},$$

where $\mathbf{u} = \mathbf{v}_1 - \mathbf{v}$ is the relative velocity of the particles being
separated from or joined to the body, i.e. their velocity with
respect to a reference frame in translational motion together
with the body

The reactive force characterizes the mechanical action exerted on the body by the particles separated from or joined to it (for instance, the effect of the ejected exhaust gases on the rocket). 2.6.3 The equation of motion of a rocket not subject to the action of external forces is

$$m \ \frac{d\mathbf{v}}{dt} = \mathbf{u} \ \frac{dm}{dt} \ .$$

If the initial velocity of the rocket is zero the rocket flies in a straight line in the direction opposite to that of the relative velocity $\mathbf{u}$ of the exhaust gases ejected from the engine nozzle. Here

$$m \ \frac{dv}{dt} = - \ u \ \frac{dm}{dt}$$

and at $u = $ const the relation between the velocity of the rocket and its mass is expressed by *Tsiolkovsky's formula*

$$v = u \ \ln \frac{m_0}{m},$$

where m_0 is the initial, or launching, mass of the rocket. 2.6.4 The maximum velocity that a rocket can reach in the absence of external forces is called its *characteristic velocity*. This velocity is acquired at the moment the rocket engine stops because it has used up all the fuel and oxidizer that were on board,

$$v_{\max} = u \ \ln \ \frac{m_0}{m_0 - m_f}$$

where m_f is the initial mass of the fuel and oxidizer. The effects of the earth's gravity and air resistance lead to an appreciable reduction in the maximum velocity actually acquired by the rocket, during the operation of its engine, as compared to the characteristic velocity. 2.6.5 The *characteristic velocity of a multistage* (combination) *rocket* is

$$v_{\max} = \sum_{i=1}^{n} u_i \ \ln \ \frac{m_{0i}}{m_{0i} - m_{fi}} \ ,$$

where n is the total number of stages of the rocket, m_{fi} is the mass of the fuel and oxidizer intended for the operation of the ith stage engine, u_i is the exhaust gas velocity (nozzle velocity) of the ith stage engine, and m_{0i} is the launching mass of the multistage rocket, including all the stages from the ith to the nth. The velocity increase of a multistage rocket, compared to that of a single-stage one having the same launching mass and the same fuel and oxidizer capacity, is due to the reduction in the mass of the rocket as the first, second and subsequent stages are consecutively jettisoned when their fuel has been exhausted.

2.7 Law of Conservation of Momentum

2.7.1 The *law of conservation of momentum* is: the momentum **p** of a closed, or isolated, system does not change in the course of time, i.e.

$$\frac{d\mathbf{p}}{dt} \equiv 0 \text{ and } \mathbf{p} = \text{const.}$$

In contrast to Newton's laws, the law, or principle, of conservation of momentum is valid not only within the framework of classical mechanics. It belongs to the most fundamental of physical laws because it is concerned with a definite property of the symmetry of space: its homogeneity. The *homogeneity of space* is manifested in the fact that the physical properties of a closed system and the laws of its motion are independent of the choice of position of the origin of the inertial reference frame, i.e. these properties and laws are unchanged by a parallel translation of the closed system as a whole in space. According to modern concepts, fields can also have momentum, not only particles and bodies. Light, for example, exerts a pressure on the surface of a body that reflects or absorbs it because the electromagnetic field of the light wave has momentum.
2.7.2 In application to systems described by classical (Newtonian) mechanics, the principle of conservation of momentum can be regarded as a consequence of Newton's laws. For a closed mechanical system the principal vector of external forces

Fext $\equiv 0$ and the principle of conservation of momentum follows from (2.5.3):

$$p = \sum_{i=1}^{n} m_i \mathbf{v}_i = \text{const},$$

where m_i and $\mathbf{v}_i$ are the mass and velocity of the ith particle of the system, which consists of n particles.

Consequently, there is also no change in the projections of the momentum of a closed system onto the axes of Cartesian coordinates of an inertial reference frame:

$$p_x = \sum_{i=1}^{n} m_i v_{ix} = \text{const},$$

$$p_y = \sum_{i=1}^{n} m_i v_{iy} = \text{const},$$

$$p_z = \sum_{i=1}^{n} m_i v_{iz} = \text{const}.$$

The momentum of the system is $\mathbf{p} = m\mathbf{v}_C$, where m is the mass of the system and $\mathbf{v}_C$ is the velocity of its centre of mass (2.3.4). It follows, therefore, from the principle of conservation of momentum that in any processes occurring within a closed system the velocity of its centre of mass does not change: $\mathbf{v}_C = \text{const}$.

2.7.3 If a system is not closed, but the external forces acting on it are such that their principal vector is identically zero (Fext $\equiv 0$), then, according to Newton's laws (2.5.3), the momentum of the system does not change with time: $\mathbf{p} = \text{const}$.

As a rule, Fext $\not\equiv 0$ and $\mathbf{p} \neq \text{const}$. But if the projection of the principal vector of external forces onto any fixed axis is identically zero, the projection onto the same axis of the momentum vector of the system does not change with time. Thus, $p_x = \text{const}$ under the condition that $F_x^{\text{ext}} \equiv 0$. If, for instance, no other external forces, except gravity, are acting on a system, the horizontal component of the momentum of the system, perpendicular to the direction of the gravitational force, remains constant.

2.7.4 In some processes (e.g. collisions or in firing shells, etc.), the momenta of the parts of the system undergo great changes in relatively short time intervals. This is due to internal forces of interaction, of short duration, but of quite considerable magnitude, developed within the system. In comparison with these forces, all external forces acting continuously on the system (for instance, gravity) are small. In such processes, the action of the system of external forces can usually be ignored, i.e. it can be assumed that the momentum of the system as a whole is not changed in the process being considered.

2.8 Galilean Transformations. Mechanical Principle of Relativity

2.8.1 A *Galilean transformation* is a transformation of coordinates and time that is applied in Newtonian mechanics in going over from one inertial frame of reference K (x, y, z, t) to another K' (x', y', z', t') which travels with translational motion with respect to K at the constant velocity $\mathbf{V}$. The Galilean transformation is based on axioms concerning the absolute nature of time intervals and of lengths. The first axiom contends that the passage of time (time interval between any two events) is the same in all inertial frames of reference. According to the second axiom, the dimensions of a body are independent of its velocity of motion with respect to the frame of reference.

If the like axes of Cartesian coordinates of inertial reference frames K and K' are drawn parallel to one another pairwise and if at the initial instant of time ($t = t' = 0$) the origins O and O' of coordinates coincide with each other (Fig. 2.2), the Galilean transformations are of the form

$$x' = x - V_x t, \quad y' = y - V_y t, \quad z' = z - V_z t \quad \text{and} \quad t' = t$$

or

$$\mathbf{r}' = \mathbf{r} - \mathbf{V}t \quad \text{and} \quad t' = t,$$

where x, y, z and x', y', z' are the coordinates of point M in the frames of reference K (at the instant of time t) and K' (at the instant of time $t' = t$), $\mathbf{r}$ and $\mathbf{r}'$ are the radius vectors of point M in the same frames of reference, and V_x, V_y and V_z are the

projections of the velocity **V** of reference frame K' onto the coordinate axes of reference frame K.

The coordinate axes are usually arranged so that frame of reference K' travels along the positive direction of axis OX (Fig. 2.3). Under this condition the Galilean transformations are of their simplest form:

$x' = x - Vt, \ y' = y, \ z' = z$ and $t' = t$.

2.8.2 A consequence of the Galilean transformations is the following law for transforming the velocity of arbitrary point
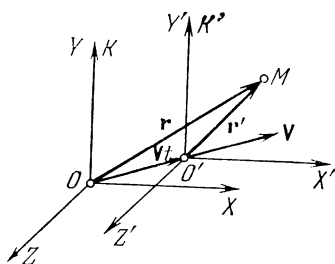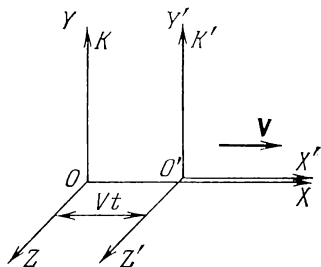
Fig. 2.2   Fig. 2.3

M (Fig. 2.2) in going over from one inertial frame of reference K (the velocity of the point is $\mathbf{v} = d\mathbf{r}/dt$) to another K' (in which the velocity of the same point is $\mathbf{v}' = d\mathbf{r}'/dt$):

$\mathbf{v}' = \mathbf{v} - \mathbf{V}$.

The projections of the velocity on like coordinate axes are transformed accordingly:

$v'_{x'} = v_x - V_x, \ v'_{y'} = v_y - V_y$ and $v'_{z'} = v_z - V_z$.

In particular, when reference frame K' travels along the positive direction of axis OX (Fig. 2.3)

$v'_{x'} = v_x - V, \ v'_{y'} = v_y$ and $v'_{z'} = v_z$.

4—0186

The accelerations of point M in frames of reference K ($\mathbf{a} = dv/dt$) and K' ($\mathbf{a}' = dv'/dt$) are the same: $\mathbf{a}' = \mathbf{a}$.

Hence, the acceleration of a particle is independent of the choice of the inertial frame of reference or, in other words, it is invariant with respect to Galilean transformations.

2.8.3 The forces of interaction between particles depend only upon their relative position and their velocity with respect to one another. The relative position of any two particles, 2 and 1, is represented by a vector equal to the difference between the radius vectors of the two particles, i.e. to the vector $\mathbf{r}_{21} = \mathbf{r}_2 - \mathbf{r}_1$ in reference frame K and to the vector $\mathbf{r}'_{21} = \mathbf{r}'_2 - \mathbf{r}'_1$ in reference frame K'. It follows from the Galilean transformations that $\mathbf{r}'_{21} = \mathbf{r}_{21}$. Therefore, the distance between particles 1 and 2 is the same in reference frames K and K':

$$r'_{21} = r_{21} \text{ or } (x'_2 - x'_1)^2 + (y'_2 - y'_1)^2 + (z'_2 - z'_1)^2$$
$$= (x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2.$$

The velocity of particle 2 with respect to particle 1 is equal to the difference of the velocities of these two particles: $\mathbf{v}_2 - \mathbf{v}_1$ (in reference frame K) and $\mathbf{v}'_2 - \mathbf{v}'_1$ (in reference frame K'). It follows from the Galilean transformations that $\mathbf{v}'_2 - \mathbf{v}'_1 = \mathbf{v}_2 - \mathbf{v}_1$.

Thus, the relative position and the velocity of relative motion of any two particles are independent of the choice of inertial reference frame; they are invariant under a Galilean transformation. Accordingly, the forces acting on a particle are also invariant under a Galilean transformation: $\mathbf{F}' = \mathbf{F}$.

2.8.4 Equations expressing Newton's laws (2.4.3 and 2.5.1) are invariant with respect to Galilean transformations, i.e. their form is unchanged in transforming the coordinates and time from one inertial frame of reference (K) to another (K'):

$$m\mathbf{a} = \mathbf{F} \text{ and } \mathbf{F}_{ki} = -\mathbf{F}_{ih} \text{ (in reference frame } K),$$

$$m'\mathbf{a}' = \mathbf{F}' \text{ and } \mathbf{F}'_{hi} = -\mathbf{F}'_{ih} \text{ (in reference frame } K'),$$

where $m' = m$ is the mass of the particle being considered and is the same in all frames of reference.

Thus, the *mechanical principle of relativity* (*Galileo's principle of relativity*) is valid in classical mechanics. It contends that the laws of mechanics are the same in all inertial frames of reference. This implies that under the same conditions all mechanical processes proceed in the same way in different frames of

reference. Consequently, it is impossible to establish, by means of any mechanical experiments conducted within a closed system of bodies, whether the system is at rest or is travelling at uniform velocity in a straight line (with respect to any inertial frame of reference whatsoever).

The mechanical principle of relativity points to the fact that all inertial frames of reference are equally valid in mechanics. It is impossible to single out any special, or "main", inertial frame of reference with respect to which the motion of bodies could be regarded as "absolute motion".

2.8.5 The principle of relativity was generalized to include all physical phenomena by Albert Einstein in his special theory of relativity (5.1.2). It was found that coordinates and time in various inertial frames of reference are related by the Lorentz transformations (5.3.2) rather than the Galilean transformations. But at low velocities of relative motion of the frames of reference (in comparison to the velocity of light in free space), the Lorentz transformations become the Galilean transformations.

CHAPTER 3   WORK AND MECHANICAL ENERGY

3.1 Energy, Work and Power

3.1.1 *Energy* is a scalar physical quantity, which is a common measure of the various forms of motion of matter dealt with in physics. The energy of a system quantitatively characterizes this system with respect to the conversions of motion that are possible within it. These conversions take place due to the interaction between parts of the system, both with one another and with external bodies (the environment). The concept of various forms of energy has been introduced in physics to analyze the qualitatively different forms of motion and the interactions corresponding to them. These include: mechanical (3.4.1), internal (9.1.2), electromagnetic (31.2.1), nuclear (42.2.2) and various other forms of energy.

3.1.2 The mechanical motion of a body is changed by forces exerted on the body by other bodies. Used in mechanics for a

4*

quantitative description of such a process of energy exchange between interacting bodies is the concept of the work done by the forces applied to the body being considered. An *element of work* done by force **F** in the infinitely small displacement $d\mathbf{r}$ is the scalar quantity

$$\delta W = \mathbf{F}\ d\mathbf{r} = \mathbf{F}\mathbf{v}\ dt,$$

where **r** and $\mathbf{v} = d\mathbf{r}/dt$ are the radius vector and velocity of the point of the body at which the force is applied, and dt is the infinitely short time interval during which force **F** performs the work δW (the meaning of the notation δW is given in 3.1.8). In rectangular Cartesian coordinates

$$\delta W = F_x\ dx + F_y\ dy + F_z\ dz = (F_x v_x + F_y v_y + F_z v_z)\ dt,$$

where x, y and z are the coordinates of the point of application of the force, and F_x, F_y, F_z and v_x, v_y, v_z are the projections of vectors **F** and **v** on the coordinate axes.

3.1.3 The expression for an element of work can also be presented in the form

$$\delta W = F\ ds \cos \alpha = F_\tau\ ds,$$

where $ds = |\ d\mathbf{r}\ |$ is an element of the length of the path of the point of force application during the infinitely short time interval dt, α is the angle between vectors **F** and $d\mathbf{r}$, and $F_\tau = {} = F \cos \alpha$ is the projection of the force in the direction of the displacement $d\mathbf{r}$. A force normal to the path of its point of application performs no work.

Force **F** is called the *motive force* if $F_\tau > 0$, so that $\delta W > 0$. If $F_\tau < 0$ (so that $\delta W < 0$), force **F** is called the *retarding force (force of resistance)*.

3.1.4 If a mechanical system is subject to the simultaneous action of forces $\mathbf{F}_1$, $\mathbf{F}_2$, . . ., $\mathbf{F}_n$, the work δW done by them during the infinitely short time dt is equal to the algebraic sum of the work done during the same time dt by each of the forces separately. Thus

$$\delta W = \sum_{i=1}^{n} \delta W_i = \sum_{i=1}^{n} \mathbf{F}_i\ d\mathbf{r}_i = \sum_{i=1}^{n} \mathbf{F}_i \mathbf{v}_i\ dt,$$

where $\mathbf{r}_i$ and $\mathbf{v}_i$ are the radius vector and velocity of the point of application of force $\mathbf{F}_i$.

For example, for a particle $\mathbf{r}_i = \mathbf{r}$ is the radius vector of the particle and $\mathbf{v}_i = \mathbf{v}$ is its velocity. Accordingly, $\delta W = \mathbf{F} \, d\mathbf{r} =$
$= \mathbf{F}\mathbf{v} \, dt$, where $\mathbf{F} = \sum_{i=1}^{n} \mathbf{F}_i$ is the resultant force (2.2.2). It follows from Newton's second law (2.4.1) that for a particle

$$\delta W = \mathbf{v} \, d\mathbf{p},$$

where $\mathbf{p} = m\mathbf{v}$ is the momentum of the particle, and m is its mass.

In the translational motion of a perfectly rigid body $d\mathbf{r}_i =$
$= d\mathbf{r}_C$ and $\mathbf{v}_i = \mathbf{v}_C$, where $\mathbf{r}_C$ and $\mathbf{v}_C$ are the radius vector and velocity of the centre of mass (2.3.3) of the body.

The work done by internal forces in any motion of a perfectly rigid body equals zero. Hence, in the translational motion of such a body $\delta W = \mathbf{F}\text{ext} \, d\mathbf{r}_C = \mathbf{F}\text{ext}\mathbf{v}_C \, dt$, where $\mathbf{F}\text{ext}$ is the principal vector of external forces (2.5.2). It follows from the law of motion of the centre of mass (2.5.3) that

$$\delta W = \mathbf{v}_C \, d\mathbf{p},$$

where $\mathbf{p} = m\mathbf{v}_C$ is the momentum of a rigid body of mass m travelling at the velocity $\mathbf{v} = \mathbf{v}_C$.

3.1.5 The work W done by force $\mathbf{F}$ along a finite portion of the path L of its point of application is equal to the algebraic sum of the work done over all the infinitely small portions of this path, i.e. it is expressed by the line integral

$$W = \int\limits_{(L)} \mathbf{F} \, d\mathbf{r} = \int\limits_{0}^{s} F_\tau \, ds,$$

where s is the length of arc measured off along the path from the beginning of the portion of the path being considered, and F_τ is the projection of the force onto the direction of displacement $d\mathbf{r}$ of the point of force application. To calculate this integral it is necessary to know the dependence of F_τ on s along the given path L. If this dependence is represented graphically (Fig. 3.1), the required work W is proportional to the hatched area in Fig. 3,1.

3.1.6 *Potential-field (conservative)* forces are ones whose work depends only on the initial and final positions of their points

of application; it depends neither on the shape of the path of
these points nor on their laws of motion along this path.
For example, the forces of interaction between the parts (i.e.
particles) of a system are said to be potential (conservative)
forces if they depend only upon the configuration of the system,
i.e. on the relative position of all its particles. The work done
by these forces in moving the system from one arbitrary position
to another is independent of the kind of displacement, but is
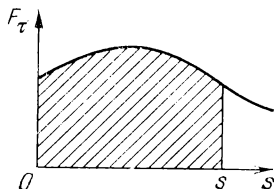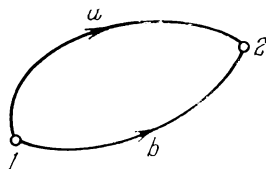
Fig. 3.1   Fig. 3.2

completely determined by the initial and final configurations
of the system. Examples are the forces of electrostatic and grav-
itational interaction.
A *steady (stationary) field* (2.2.1) is said to be *potential*, or con-
servative, if force **F**, which it exerts on a particle located in
the field, is a potential, or conservative force. This implies
that force **F** depends only upon the position of the particle in
the field, and that the amounts of work done by force **F** in
displacing the particle from one arbitrary position (*1*) to ano-
ther (*2*) (Fig. 3.2) along any two paths, for instance, *1a2* (work
W_{1a2}) and *1b2* (work W_{1b2}), are the same:

$$W_{1a2} = W_{1b2} = \int_{1}^{2} \mathbf{F}\, d\mathbf{r}.$$

Accordingly, the work done by a conservative force in displac-
ing its point of application along any closed path L (for exam-
ple, *1a2b1*) is equal to zero:

$$\oint_{(L)} \mathbf{F}\, d\mathbf{r} \equiv 0.$$

In the general case, the external bodies that set up the field being considered can be in motion with respect to an inertial frame of reference, so that their field is not stationary, i.e. force F depends explicitly on time: $\partial \mathbf{F}/\partial t \not\equiv 0$. A *nonstationary field* is a *conservative* one if the work done by force F upon *instantaneous* displacement of its point of application along any closed path L equals zero:

$$\oint_{(L)} \mathbf{F}\, d\mathbf{r} \equiv 0.$$

Here F depends, not only on the coordinates of the point of application, but on the time as well. To calculate this integral, however, time is assumed to be a fixed parameter.

3.1.7 Among the nonconservative forces are dissipative and gyroscopic forces. *Dissipative forces* are ones whose total work done in displacing a closed system in any way is always negative. Such, for instance, are the forces of sliding friction and the forces retarding the motion of bodies in liquids and gases. In contrast to conservative forces, dissipative forces depend, not only on the relative positions of the interacting bodies, but on their relative velocities as well.

Gyroscopic forces are ones that depend upon the velocity of the particle on which they act. They are directed perpendicular to this velocity. An example is the Lorentz force (24.1.1) exerted by a magnetic field on a charged particle travelling in this field. The work done by gyroscopic forces is always equal to zero, regardless of how the particle is displaced.

A mechanical system (system of particles) is said to be *conservative* if all the internal forces in it are conservative ones and all the external forces are both conservative and stationary. Systems that do not satisfy these conditions are said to be *nonconservative*.

3.1.8 An element of work done by force F, exerted on a particle by a stationary conservative field, can be represented in the form of an exact differential of a scalar function $\Phi\,(x,\,y,\,z)$ of the coordinates. This is called the *force function* of the field:

$$\mathbf{F}\, d\mathbf{r} = d\Phi \quad \text{or} \quad F_x\, dx + F_y\, dy + F_z\, dz$$
$$= \frac{\partial \Phi}{\partial x}\, dx + \frac{\partial \Phi}{\partial y}\, dy + \frac{\partial \Phi}{\partial z}\, dz\,.$$

Hence

$$F_x = \frac{\partial \Phi}{\partial x}, \quad F_y = \frac{\partial \Phi}{\partial y}, \quad F_z = \frac{\partial \Phi}{\partial z} \quad \text{and} \quad \mathbf{F} = \text{grad } \Phi.$$

These last relations are also valid for a nonstationary conservative field, whose force function depends, not only on the coordinates, but on time as well: $\Phi = \Phi (x, y, z, t)$. But here

$$\mathbf{F} \, d\mathbf{r} = d\Phi - \frac{\partial \Phi}{\partial t} \, dt.$$

An element of work done by a nonconservative force cannot be represented in the form of an exact differential of any function whatsoever of the coordinates. This is precisely why an element of work done by an arbitrary force is denoted by δW.

3.1.9 The concept of power is used in mechanics to characterize the amount of work done in unit time. *Power (instantaneous power)* is the scalar physical quantity P. It is equal to the ratio of the element of work δW to the infinitely short time interval dt during which it is performed:

$$P = \frac{\delta W}{dt}.$$

If $\mathbf{F}$ is the force doing the work δW, the power is the scalar product of force $\mathbf{F}$ by the velocity $\mathbf{v}$ of its point of application:

$$P = \mathbf{F}\mathbf{v} = F_\tau v.$$

In the general case, power may vary with time.

The *average power* in the time interval from t to $t + \Delta t$ is the physical quantity $\langle P \rangle$, which is equal to the ratio of the work W done during this time interval to the length of this interval Δt:

$$\langle P \rangle = \frac{W}{\Delta t}.$$

3.2 Kinetic Energy

3.2.1 The *kinetic energy* of a body is the energy of its mechanical motion. A change in the kinetic energy E_k of a particle due to the action of force $\mathbf{F}$ is equal to the work done by this force:

$$dE_k = \delta W = \mathbf{v} \, d\mathbf{p},$$

where $\mathbf{p} = m\mathbf{v}$ is the momentum of the particle, and m and $\mathbf{v}$ are its mass and velocity. In Newtonian mechanics $m = \text{const}$ and the expression for the kinetic energy of a particle is of the form

$$E_k = \frac{m\mathbf{v}^2}{2} = \frac{mv^2}{2}.$$

Kinetic energy in relativistic mechanics is discussed in Sect. 5.7.1.

3.2.2 The kinetic energy of a mechanical system is equal to the sum of the kinetic energies of all the parts of the system. For a system consisting of n particles, for instance,

$$E_k = \sum_{i=1}^{n} \frac{m_i \mathbf{v}_i^2}{2} = \sum_{i=1}^{n} \frac{m_i v_i^2}{2},$$

where m_i and $\mathbf{v}_i$ are the mass and velocity of the ith particle in the system.

The kinetic energy of a body is

$$E_k = \frac{1}{2} \int_{(V)} \rho \mathbf{v}^2 \, dV = \frac{1}{2} \int_{(V)} \rho v^2 \, dV,$$

where $\mathbf{v}$ is the velocity of points of the infinitely small element of volume dV of the body, whose density is ρ and mass is $dm = \rho dV$. Integrating is carried out throughout the whole volume V of the body. If a perfectly rigid body of mass m travels with translational motion at the velocity $\mathbf{v}$, its kinetic energy is $E_k = mv^2/2$. The kinetic energy of rotating bodies is discussed in Sects. 4.3.3 and 4.3.5.

3.2.3 The change in the kinetic energy of a mechanical system is equal to the algebraic sum of the amounts of work done by all the external and internal forces acting on the system (2.2.4). Thus

$$dE_k = \delta W^{\text{ext}} + \delta W^{\text{int}}.$$

For example, for a system consisting of n particles

$$dE_k = \sum_{i=1}^{n} \mathbf{F}_i^{\text{ext}} \, d\mathbf{r}_i + \sum_{i=1}^{n} \sum_{k=1}^{n} \mathbf{F}_{ik} \, d\mathbf{r}_i,$$

where $\mathbf{r}_i$ is the radius vector of the ith particle, $\mathbf{F}_i^{\text{ext}}$ is the resultant of the external forces acting on this particle, and $\mathbf{F}_{ii} = = 0$.

If the system is not deformed, then the work done by the internal forces

$$\delta W^{\text{int}} = 0 \quad \text{and} \quad dE_{\text{k}} = \delta W^{\text{ext}}.$$

For example, the change in the kinetic energy of a perfectly rigid body in translational motion is

$$dE_{\text{k}} = \mathbf{F}^{\text{ext}} d\mathbf{r},$$

where $\mathbf{F}^{\text{ext}}$ is the principal vector of external forces (2.5.2), and $d\mathbf{r}$ is the vector of an element of displacement of the body.

3.2.4 The kinetic energy of a mechanical system depends upon the choice of the frame of reference. If in the inertial reference frame K the kinetic energy of the system equals $E_{\mathbf{k}}$, and in the reference frame K', travelling with translational motion at the velocity $\mathbf{V}$ with respect to K, it is equal to $E'_{\mathbf{k}}$, then

$$E_{\mathbf{k}} = E'_{\mathbf{k}} + \frac{mV^2}{2} + \mathbf{p}'\mathbf{V},$$

where m is the mass of the system, $\mathbf{p}' = m\mathbf{v}'_C$ is the momentum of the system in its motion with respect to reference frame K', and $\mathbf{v}'_C$ is the velocity of the system's centre of mass with respect to K'. This equation is valid both for the case when $\mathbf{V} = = \text{const}$, i.e. when K' is an inertial frame of reference, as well as for $d\mathbf{V}/dt \neq 0$.

In particular, if the frame of reference K' travels with translational motion with respect to reference frame K at the velocity $\mathbf{v}_C$ of the centre of mass of the system, i.e. $\mathbf{V} = \mathbf{v}_C$, then $\mathbf{v}'_C = = 0$ and

$$E_{\mathbf{k}} = \frac{mv_C^2}{2} + E'_{\mathbf{k}}.$$

This equation is an expression of *Koenig's theorem*: the kinetic energy of a mechanical system is equal to the sum of the kinetic energy that a particle would have if its mass was that of the whole system and if it travelled at the velocity of the system's centre of mass, plus the kinetic energy of the same system in its motion with respect to a translational reference frame (i.e. one

travelling with translational motion) having its origin at the centre of mass.

It follows from Koenig's theorem that the kinetic energy of a perfectly rigid body is equal to the kinetic energy of translational motion of this body at the velocity of its centre of mass plus the kinetic energy of rotation of the body about its centre of mass.

3.3 Potential Energy

3.3.1 *Potential energy* is the part of the energy of a mechanical system that depends only on its *configuration*, i.e. on the relative arrangement of all the particles of the system and their positions in the external potential (conservative) field (3.1.6). The decrease in the potential energy when the system is moved from arbitrary position 1 to another arbitrary position 2 is measured by the work W_{12} done in this displacement by the conservative forces (internal and external) acting on the system. Thus

$$E_p(1) - E_p(2) = W_{12},$$

where $E_p(1)$ and $E_p(2)$ are the values of the potential energy of the system in its initial and final positions. Accordingly, the work done by conservative forces upon an infinitely small change in the configuration of the system is $\delta W = -dE_p$.

Note. It is assumed here that the external conservative forces are *stationary* (*steady*), i.e. they can change with time only as the consequence of a change of the position of the system being considered with respect to the frame of reference. Otherwise

$$dE_p = -\delta W + \frac{\partial E_p}{\partial t} dt.$$

In the simplest case, when the system is a particle located in a conservative field, the relation between force **F**, acting on the particle, and the potential energy E_p of this particle in the field is of the form

$$F_x = -\frac{\partial E_p}{\partial x}, \quad F_y = -\frac{\partial E_p}{\partial y}, \quad F_z = -\frac{\partial E_p}{\partial z}$$

and $\mathbf{F} = -\operatorname{grad} E_p.$

The potential energy E_p of a particle is related to the force function (3.1.8) of the corresponding conservative field by the equation

$$dE_p = -d\Phi \text{ or } E_p(x, y, z, t) = -\Phi(x, y, z, t) + C,$$

where C is the integration constant.

3.3.2 The equations in 3.3.1 enable the dependence of the potential energy of a system on its configuration to be determined only to an arbitrary constant term that has no effect on the *change* in energy. To obtain a unique dependence of the potential energy of the system on its configuration, a so-called *zero configuration* is chosen, in which the potential energy of the system is conditionally assumed to equal zero. Thus, the potential energy of a system in an arbitrary state is equal to the work done by all the conservative forces acting on the system in transforming it from the state being considered to the state corresponding to the zero configuration.

3.3.3 **Example 1.** *Potential energy of a particle in a uniform (homogeneous) force field.* Assume that force **F**, exerted on the particle by the field, is directed along axis OZ, i.e. $\mathbf{F} = F_z\mathbf{k}$, where **k** is the unit vector of axis OZ, and the projection F_z of force **F** onto axis OZ is independent of the coordinates of the particle. Then

$$dE_p = -\mathbf{F}\,dr = -F_z\,dz \text{ and } E_p(z) = -F_zz + E_p(0),$$

where $E_p(0)$ is the value of the potential energy of the particle at the level $z = 0$.

In particular, the potential energy of a particle of mass m located in a uniform gravity field at the earth's surface (axis OZ is directed vertically upward, $F_z = -mg$, where g is the free fall acceleration) is equal to

$$E_p(z) = mgz + E_p(0).$$

3.3.4 **Example 2.** *Potential energy of a particle in a central force field.* In a conservative *field of central forces* the particle is subject to the action of forces **F** which are directed everywhere along straight lines passing through a fixed point, called the *centre of force*, and depend only on the distance r to this point:

$$\mathbf{F} = F_r(r)\,\frac{\mathbf{r}}{r}\,.$$

Here r is the radius vector from the centre of force to the point being considered in the field, and F_r (r) is the projection of force F onto the direction of vector r; it depends only on the distance r. If the particle is attracted to the centre of force, F_r (r) = $-|$ F $|$ < 0, but if it is repelled from the centre of force, then F_r (r) = $|$ F $|$ > 0. An element of work done by force F is

$$\delta W = \mathbf{F} \, d\mathbf{r} = F_r \, (r) \, dr.$$

The potential energy of the particle is

$$E_p \, (r) = \int_r^\infty F_r \, (r) \, dr + E_p \, (\infty).$$

Usually the energy of a particle, located at an infinite distance from the centre of force, is taken as the reference point for potential energy, i.e. it is assumed that $E_p \, (\infty) = 0$. Thus

$$E_p \, (r) = \int_r^\infty F_r \, (r) \, dr.$$

Examples of a central force field, in which the force is inversely proportional to the square of the distance to the centre of force $(F_r \, (r) \propto r^{-2})$, are the gravitational fields of a particle or a homogeneous sphere, the electrostatic field of a point charge, as well as that of a

Fig. 3.3

hollow sphere and a solid sphere uniformly charged along the surface and throughout the volume, respectively.

3.3.5 **Example 3.** *Potential energy of a system of two particles* between which *central forces* act, i.e. forces depending upon the distance between the particles and directed along the straight line joining the particles. Shown in Fig. 3.3 are mutually repulsive forces $\mathbf{F}_{12}$ and $\mathbf{F}_{21} = -\mathbf{F}_{12}$:
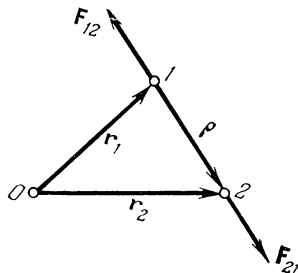
$$\mathbf{F}_{21} = F_\rho \, (\rho) \, \frac{\boldsymbol{\rho}}{\rho} \, ,$$

where $\rho = r_2 - r_1$ is the radius vector from particle 1 to particle 2, and $F\rho\ (\rho)$ is the projection of force F_{21} on the direction of vector ρ and depends only on the distance ρ between the particles. An infinitely small change in the potential energy of the system

Fig. 3.4

$$dE_p = -(F_{12}\ dr_1 + F_{21}\ dr_2)$$
$$= -F_{21}\ d\rho = -F_\rho\ (\rho)\ d\rho.$$

If we assume that $E_p \to 0$ as $\rho \to \infty$, then

$$E_p\ (\rho) = \int\limits_\rho^\infty F_\rho\ (\rho)\ d\rho.$$

This energy is often called the *mutual potential energy of two particles.*

3.3.6 Example 4. *Potential energy of an elastic body* (for instance, a spring) *in axial tension or compression.* When an elastic body is deformed, conservative internal forces (elastic forces) are developed in the body that oppose deformation. According to Hooke's law, the force F_{el} exerted by the deformed body A (Fig. 3.4) on body B, causing the deformation, is proportional to the amount of deformation (stretch or compression):

$$F_{el} = -kx i.$$

Here $x i$ is the displacement vector of body B that characterizes the deformation of body A (in the undeformed state $x = 0$; in compression $x > 0$ and in tension $x < 0$), and $k > 0$ is a factor characterizing the elastic properties of body A.

The potential energy of an elastically deformed body (in the absence of deformation, i.e. at $x = 0$, this energy is assumed equal to zero)

$$E_p = \frac{kx^2}{2}.$$
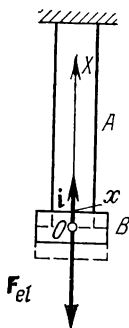
3.4 Law of Conservation of Mechanical Energy

3.4.1 The *mechanical energy*, or *total mechanical energy*, is the energy of mechanical motion and interaction. The mechanical energy E of a system of particles is equal to the sum of their kinetic energy E_k and the potential energy E_p of interaction of these particles with one another and with external bodies:

$$E = E_k + E_p.$$

An elementary increment of the mechanical energy of a system during the infinitely short time dt is equal to

$$dE = \delta W_{nonc} + \frac{\partial E_p}{\partial t}\, dt,$$

where δW_{nonc} is the algebraic sum of the elements of work done in the time interval dt by all the internal and external *nonconservative* forces. The term $(\partial E_p/\partial t)\, dt$ represents the change during the time interval dt in the potential energy of the system and, consequently, its total mechanical energy due to nonstationary external conservative forces (3.3.1).

3.4.2 If the system is *conservative* (3.1.7), $\delta W_{nonc} \equiv 0$ and $\partial E_p/\partial t \equiv 0$, because, in this case, there are no nonconservative forces and the external conservative forces are stationary. Consequently, the mechanical energy of such a system $E =$ const, i.e. the following law, called the *law of conservation of mechanical energy*, is valid: in the motion of a conservative system, its mechanical energy remains unchanged.

In particular, this law holds for closed conservative systems: the mechanical energy of a closed system does not change with time if all the internal forces acting in this system are conservative ones.

The law of conservation of mechanical energy is associated with the *uniformity of time*. This property of time is manifested in the fact that the laws of motion of a closed system (or a system located in a stationary external field) are independent of the choice of the zero time reference. For example, in the free fall of a body in a stationary conservative gravity field at the earth's surface, the velocity of the body and the length of path travelled depend only on the length of time in free fall and initial velocity of the body, but not on the specific instant of time the body began to fall.

3.4.3 The mechanical energy of a closed *nonconservative* system is changed owing to the work done by all the nonconservative internal forces:

$$dE = \delta W_{\text{nonc}}.$$

Gyroscopic forces (3.1.7) do no work and make no contribution to δW_{nonc}, i.e. the existence of such forces in a system does not lead to changes in its mechanical energy.

The action of dissipative forces (3.1.7), for instance, the force of friction, leads to a gradual reduction in the mechanical energy of a closed system. This process is called the *dissipation of energy*. Consequently, a system whose mechanical energy continually decreases with time is called a *dissipative system*. In the dissipation of energy, the mechanical energy of the system is converted into other forms of energy (for example, into the energy of chaotic motion of molecules). The conversion of mechanical energy is accomplished in full agreement with a universal law of nature: the law of conservation of energy (9.2.7).

According to this law, energy can be converted from one form to another and redistributed within the system, but its total amount in a closed system must remain constant. It follows from the law of conservation and conversion of energy that the changes in energy of an open system, occurring in interaction of the system with the external medium (external bodies and fields), must be numerically equal and opposite in sign to the changes in the energy of the external medium. In other words, the change in the energy of a system upon its interaction with the external medium should be equal to the energy the system receives or yields up in the process being considered.

3.4.4 Forces of resistance and friction act in all real mechanical systems. Hence, these systems are nonconservative. In certain cases, however, they can be assumed approximately conservative, and the law of conservation of mechanical energy can be applied to them. Such an approach is feasible if, in the process being considered, the work W_{nonc} done by all the nonconservative forces acting on the system is negligibly small in comparison to the mechanical energy E of the system, i.e. $|\ W_{\text{nonc}}/E\ | \ll \ll 1$, so that $|\ \Delta E/E\ | \ll 1$, where $\Delta E = W_{\text{nonc}}$ is the change in the mechanical energy of the system.

3.4.5 The state of *mechanical equilibrium of a system* is one that can be disturbed only as a result of external force action

applied to the system. In this equilibrium state all the particles
of the system are at rest so that the kinetic energy of the system
equals zero. The state of mechanical equilibrium is said to be
stable if a small external action applied to the system leads to
only a small change in its state. In this case, forces are devel-
oped in the system that tend to restore the equilibrium state.
A state of mechanical equilibrium is said to be *unstable* if it
can be disturbed by even an infinitely small external action
on the system, which does not return again to the equilibrium
state. In this case, forces are developed in the system that tend
to further disturb the equilibrium state.
The law of conservation of mechanical energy enables the condi-
tions of equilibrium of conservative systems to be formulated:
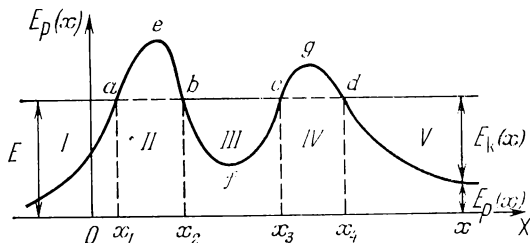
Fig. 3.5

in states of stable equilibrium the potential energy curve of
the system has minima, whereas in states of unstable equilib-
rium, it has maxima.
3.4.6 On the basis of the law of conservation of mechanical
energy, the region of possible configurations of a conservative
system (3.3.1) can be found. The kinetic energy of the system
$E_k \geqslant 0$. Hence, at a given value E of the mechanical energy
of the system, the latter can only be in states that conform to
the condition: $E_p \leqslant E$. Figure 3.5 illustrates the simplest case
in which a particle has one-dimensional motion along axis OX
in an external stationary conservative field. The potential
energy of the particle is a function of only a single coordinate x,
i.e. $E_p = E_p(x)$. The curve of this relation, shown in Fig. 3.5,

is called a *potential energy curve*. For a fixed value of the mechanical energy E of the particle, as shown in Fig. 3.5, the particle can remain, in its travel, in one of the following three regions: $x \leqslant x_1$ (region I), $x_2 \leqslant x \leqslant x_3$ (region III) and $x \geqslant x_4$ (region V). These regions are separated from one another by the regions II and IV, which are the so-called *potential barriers aeb* and *cgd*, within whose limits the particle cannot exist. At the boundaries of the potential barriers (at points a, b, c and d), the particle reverses its direction of motion; in region I the particle can travel without limit to the left of boundary a of the barrier, and in region V, without limit to the right of boundary d of the barrier. In region III, the particle oscillates between points b and c; it is in a so-called *potential well bfc*.

3.5 Perfectly Elastic and Inelastic Collisions

3.5.1 An impact in which a considerable change in the velocities of the colliding bodies occurs in an extremely short time interval is called a *collision*. Examples of collisions are a hammer striking a billet being forged that lies on the anvil, and a hammer striking the head of a nail being driven.
The *line of impact* is the common normal to the surfaces of the colliding bodies at their point of contact. If the centres of mass (2.3.3) of the colliding bodies lie on the line of impact at the instant of contact, the collision is said to be *central*. An example is the collision of two billiard balls. The collision, or impact, is said to be *head-on*, or *straight-line*, if the velocities of the centres of mass of the colliding bodies are parallel, before the collision, to the line of impact. Otherwise, it is called an *oblique impact*, or *glancing collision*.
3.5.2 In a collision the bodies are deformed and instantaneous, but quite substantial, forces are developed. They are called *impact forces*. For a system of colliding bodies these forces are internal ones*, i.e. they do not change the total momentum of the system. External forces, constantly exerted on the system

---

* It is assumed that the colliding bodies are either free (2.2.3) or the constraints imposed on them are such that no impact constraint reactions are developed.

(for instance, the gravity forces of the bodies) are usually very small compared to the impact forces. Therefore, even though the impulses of impact forces (2.4.2) in the duration τ of the collision are commensurable with the momenta (2.3.4) of the colliding bodies, the resultant impulse of all the constantly acting external forces during the same time interval τ is small compared to the momenta of the bodies. Accordingly, the work done on the system by the external forces during the time interval τ is small compared to the mechanical energy of the system. Hence, a system of bodies can be assumed, in the collision process, to be an approximately closed system (2.2.4). This means that in calculating the results of the collision, use can be made of the laws of the conservation of momentum (2.7.1), of angular momentum (4.4.1) and of energy (9.2.7). If the bodies deform in the collision like perfectly elastic ones, the impact forces are conservative and the law of conservation of mechanical energy (3.4.2) is valid in the system.

3.5.3 The collision of two bodies is said to be *perfectly inelastic* if, after impact, they travel together as a single body. Sufficiently close to a perfectly inelastic collision are, for example, the impact of the pile-driving hammer with the end of the pile being driven, and the firing of a bullet into a wagon filled with sand, when the bullet remains embedded in the sand. Various processes occur in the colliding bodies in an inelastic collision (plastic deformation of the bodies, friction, etc.). As a result, the kinetic energy of the system is partly transformed into its internal energy (9.1.2).

If two bodies of masses m_1 and m_2, travelling with translational motion at the velocities $\mathbf{v}_1$ and $\mathbf{v}_2$, undergo a *perfectly inelastic head-on central collision*, they continue their translational motion together at the velocity

$$\mathbf{u} = \frac{m_1\mathbf{v}_1 + m_2\mathbf{v}_2}{m_1 + m_2}.$$

Note. In the case of a perfectly inelastic collision that is neither head-on nor central, the preceding equation can be used to find the velocity of the centre of the mass of the bodies joined together in the collision. As the result of such a collision, however, the system may begin to rotate about its centre of mass in accordance with the law of conservation of angular momentum (4.4.1).

5*

3.5.4 The change in the kinetic energy of a system of two collid-ing bodies in a perfectly inelastic head-on central collision is

$$\Delta E_k = \frac{m_1 + m_2}{2}\, \mathbf{u}^2 - \frac{m_1}{2}\, \mathbf{v}_1^2 - \frac{m_2}{2}\, \mathbf{v}_2^2$$

$$= - \frac{m_1 m_2}{2\,(m_1 + m_2)}\, (\mathbf{v}_1 - \mathbf{v}_2)^2 < 0.$$

In particular, if the second body is at rest before the collision (for instance, a pile being driven by a pile driver, or a forging lying on the anvil), the relative reduction in the kinetic energy of the system in a perfectly inelastic head-on central collision is

$$-\frac{\Delta E_k}{E_{k_1}} = \frac{m_2}{m_1 + m_2}\,.$$

In engineering, a perfectly inelastic head-on central collision (impact) is made use of to change the shape of bodies (smith and die forging, rivetting, etc.) or to move a body in a medium with high resistance to motion (driving nails, piles, etc.). It proves expedient in the first case to have the ratio $-\Delta E_k/E_{k_1}$ as near as possible to unity, i.e. it is necessary that $m_2 \gg m_1$ (the mass of the forged item and the anvil should be many times greater than that of the hammer). In the second case, on the contrary, it is necessary that the loss in kinetic energy in the collision (impact) be as low as possible, i.e. that $m_1 \gg m_2$ (the mass of the hammer should be many times greater than that of the nail being driven).

3.5.5. A collision is said to be *perfectly elastic* if it does not lead to any change in the mechanical energy of the system, i.e. if the bodies are perfectly elastic.

Example 1. A *perfectly elastic head-on central collision* of two bodies (for instance, balls) of masses m_1 and m_2, which travel before the collision with translational motion at the velocities $\mathbf{v}_1$ and $\mathbf{v}_2$ along axis OX, passing through their centres of mass (Fig. 3.6a). The velocities of the bodies after the collision, $\mathbf{u}_1$ and $\mathbf{u}_2$ (Fig. 3.6b), can be found from the laws of conserva-tion of momentum and mechanical energy:

$$m_1\mathbf{u}_1 + m_2\mathbf{u}_2 = m_1\mathbf{v}_1 + m_2\mathbf{v}_2,\ \ m_1\mathbf{u}_1^2 + m_2\mathbf{u}_2^2 = m_1\mathbf{v}_1^2 + m_2\mathbf{v}_2^2.$$

Velocities u_1 and u_2 are directed along axis OX and their projections on this axis are

$$u_{1x} = \frac{(m_1 - m_2)\,v_{1x} + 2m_2 v_{2x}}{m_1 + m_2}$$

and $$u_{2x} = \frac{2m_1 v_{1x} + (m_2 - m_1)\,v_{2x}}{m_1 + m_2}\;.$$

In particular, if the masses of the bodies are the same, the bodies exchange velocities in a collision: $u_{1x} = v_{2x}$ and $u_{2x} = v_{1x}$.

If the mass of the second body is many times greater than that of the first body, then $u_{1x} \approx 2v_{2x} - v_{1x}$ and $u_{2x} \approx v_a$.

(a) Before collision   *(b) After collision*
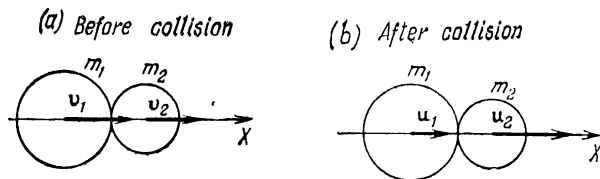
Fig. 3.6

3.5.6 Example 2. *Perfectly elastic oblique central collision.* If the bodies are smooth the impulse of friction forces in the collision can be neglected. In this case the tangential components of the velocities of the bodies are not changed, i.e. components perpendicular to the line of impact: $u_{1\tau} = v_{1\tau}$ and $u_{2\tau} = v_{2\tau}$. The normal components, directed along the line of impact, change in the same way as in a head-on collision:

$$u_{1n} = \frac{(m_1 - m_2)\,v_{1n} + 2m_2 v_{2n}}{m_1 + m_2}$$

and $$u_{2n} = \frac{2m_1 v_{1n} + (m_2 - m_1)\,v_{2n}}{m_1 + m_2}\;.$$

In particular, in the perfectly elastic oblique collision of a smooth ball with a fixed flat wall ($m_2 \gg m_1$ and $u_2 = v_2 = 0$):

$$u_{1\tau} = v_{1\tau} \text{ and } u_{1n} = -v_{1n},$$

i.e. the ball rebounds from the wall according to the law of mirror reflection: the angle of reflection is equal to the angle of incidence. The magnitude of the velocity is conserved: $u_1 = v_1$. The vector $\Delta\mathbf{p}_1$ of the change in momentum of the ball in the collision is directed perpendicular to the wall:

$$\Delta\mathbf{p}_1 = m_1 (\mathbf{u}_1 - \mathbf{v}_1) = -2m_1\mathbf{v}_{1n}.$$

The impulse of impact force acting on the wall is equal to $2m_1\mathbf{v}_{1n}$.

CHAPTER 4   DYNAMICS OF ROTARY MOTION

4.1 Moment of Force and Angular Momentum

4.1.1. The moment of force, or force moment is a concept introduced in mechanics to characterize the external mechanical action exerted on a body that changes its rotary motion. Distinction is made between the moment of force about a fixed point and about a fixed axis.

The *moment of the force* $\mathbf{F}$ *about a fixed point* O (the *pole*) is the vector quantity $\mathbf{M}$, equal to the vector product of the radius vector $\mathbf{r}$, from point O to point A of force application (Fig. 4.1), by the force vector $\mathbf{F}$:
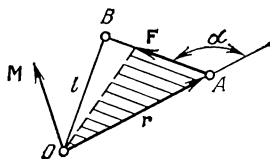
$$\mathbf{M} = [\mathbf{rF}].$$

The magnitude of the moment of force is $M = Fr \sin \alpha = Fl$, where α is the angle between vectors $\mathbf{r}$ and $\mathbf{F}$, and $l = r \sin \alpha$ is the length of the perpendicular OB (Fig. 4.1) dropped from point O to the line of action of the force. The quantity l is called the *arm of the force* with respect to point O. If the point of application of force $\mathbf{F}$ is displaced along its line of action, the moment $\mathbf{M}$ of this force about the same fixed point O is not changed. If the line of action of the force passes through point O, the moment of the force about point O equals zero.

4.1.2 The *principal,* or *resultant, moment* of a system of forces *about fixed point* O (*pole*) is the vector $\mathbf{M}$, equal to the geometric

Fig. 4.1

sum of the moments about point O of all n forces of the system:

$$M = \sum_{i=1}^{n} [\mathbf{r}_i \mathbf{F}_i],$$

where $\mathbf{r}_i$ is the radius vector from point O to the point of application of force $\mathbf{F}_i$.

It follows from Newton's third law (2.5.1) that the moments of the internal forces of interaction between the particles of a system about pole O are compensated for (cancelled out) pairwise: $[\mathbf{r}_i \mathbf{F}_{ih}] = - [\mathbf{r}_h \mathbf{F}_{hi}]$. Consequently, in calculating the resultant moment of forces, only the *external* forces exerted on the mechanical system being considered need be taken into account.

Fig. 4.2

4.1.3 The *moment of force* **F** *about a fixed axis a* is the scalar quantity M_a, equal to the projection on this axis of vector **M**, which is the moment of force **F** about arbitrary point O of axis a.

The value of moment M_a is independent of the choice of the position of point O along axis a.

Note. Sometimes the moment of a force about fixed axis a is understood to mean the vector quantity $\mathbf{M}_a = M_a \mathbf{i}_a$, where $\mathbf{i}_a$ is the unit vector of axis a. The vector $\mathbf{M}_a$ is the component of vector **M** of the moment of force about pole O, this component being directed along axis a.

If the line of action of the force intersects or is parallel to axis a, the moment of force about this axis is equal to zero.

Let A be the point of application of force **F**, and O_1 be the foot of the perpendicular dropped from point A to the axis OZ being considered (Fig. 4.2). It proves convenient to resolve force **F** into three mutually perpendicular components: axial component $\mathbf{F}_z$, parallel to axis OZ; radial component $\mathbf{F}_n$, directed along vector $\boldsymbol{\rho} = \overrightarrow{O_1 A}$, and tangential component $\mathbf{F}_\tau$, directed perpendicular to the axis and perpendicular to vector $\boldsymbol{\rho}$. The mo-

ment of force **F** about axis OZ is
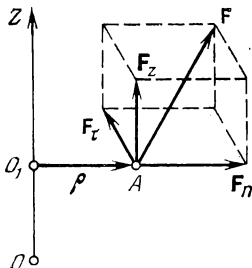
$M_z = [\rho \mathbf{F}_\tau]_z$ and $\mathbf{M}_z = [\rho \mathbf{F}_\tau]$.

Since vectors ρ and $\mathbf{F}_\tau$ are perpendicular to each other

$|\mathbf{M}_z| = |M_z| = \rho |\mathbf{F}_\tau|$.

The *principal*, or *resultant*, *moment* of a system of forces *about a fixed axis a* is equal to the algebraic sum of the moments of all the forces of the system about this axis.

4.1.4 The *angular momentum*, or *moment of momentum*, *of a particle about fixed point O (pole)* is the vector **L**. It is equal to the vector product of radius vector **r**, from point O to the location of the particle, by vector **p** of its momentum:

$\mathbf{L} = [\mathbf{rp}] = [\mathbf{r}m\mathbf{v}]$,

where m and **v** are the mass and velocity of the particle.
The *angular momentum of a system about fixed point O* is denoted by vector **L** and equals the vector sum of the angular momenta of all the particles of the system about the same point O:

$$\mathbf{L} = \sum_{i=1}^{n} [\mathbf{r}_i \mathbf{p}_i] = \sum_{i=1}^{n} [\mathbf{r}_i m_i \mathbf{v}_i],$$

where m_i, $\mathbf{r}_i$ and $\mathbf{v}_i$ are the mass, radius vector and velocity of the ith particle, and n is the total number of particles.
The *angular momentum of a system about fixed axis a* is the scalar quantity L_a, equal to the projection on axis a of vector **L**, the angular momentum of the system about some point O lying on this axis:

$$L_a = \sum_{i=1}^{n} [\mathbf{r}_i m_i \mathbf{v}_i]_a .$$

The numerical value L_a of the angular momentum is independent of the choice of the position of point O along axis a.
Note. Sometimes the angular momentum of a system about a fixed axis a is understood to mean the vector quantity $\mathbf{L}_a = L_a \mathbf{i}_a$, where $\mathbf{i}_a$ is the unit vector of axis a.

4.1.5 The *angular momentum of a body rotating about a fixed point O* at the angular velocity ω is equal to:

$$L = \int_{(m)} [\mathbf{r}\mathbf{v}] \, dm = \int_{(m)} [\mathbf{r}\,[\omega\mathbf{r}]] \, dm,$$

where $\mathbf{r}$ is the radius vector from point O to an infinitely small element of the body of mass dm, and $\mathbf{v} = [\omega\mathbf{r}]$ is the velocity of the element of the body. Since $[\mathbf{r}\,[\omega\mathbf{r}]] = r^2\omega - (\omega\mathbf{r})\,\mathbf{r}$, vectors L and ω do not coincide in direction in the general case:

$$L = \omega \int_{(m)} r^2 \, dm - \int_{(m)} (\omega\mathbf{r})\,\mathbf{r} \, dm.$$

The angular momentum of a body hinged at point O and its angular velocity coincide in direction if the body rotates about one of its principal axes of inertia (4.2.4) passing through point O

$$L = I\omega,$$

where I is the moment of inertia (4.2.1) of the body about this principal axis.

4.1.6 The resultant moments M and M^* of a system of forces about two different fixed points O and O^* are related by the equation

$$M \doteq M^* + [\mathbf{r}^*\mathbf{F}],$$

where $\mathbf{r}^*$ is the radius vector from origin O to point O^*, and F is the principal vector of the system of forces being considered. If $\mathbf{F} = 0$ the resultant moment of the system of forces is the same about any fixed point: $M^* = M$. This, precisely, is the property possessed by a *force couple*, or simply *couple*, i.e. a system of two forces, equal in magnitude and directed along two parallel straight lines in opposite directions. The shortest (perpendicular) distance d between the lines of action of the forces is called the *arm of the couple*. The moment, or torque, of a force couple is directed perpendicular to the plane of the forces, and its magnitude is equal to $M = Fd$, where F is the magnitude of each of the forces of the couple.

The resultant moment $\mathbf{M}_C$ of all the forces acting on a mechanical system about the centre of mass (2.3.3) of this system is related to the resultant moment $\mathbf{M}$ of the same system of forces about fixed point O by the equation

$$\mathbf{M} = \mathbf{M}_C + [\mathbf{r}_C\mathbf{F}],$$

where $\mathbf{r}_C$ is a radius vector from origin O to point C, and $\mathbf{F}$ is the principal vector of the system of forces.

4.1.7 The value of the angular momentum of a mechanical system about its centre of mass C is the same for the absolute motion of the particles at the velocities $\mathbf{v}_i$ (i.e. with respect to a fixed inertial frame of reference) and for the relative motion at the velocities $\mathbf{v}'_i = \mathbf{v}_i - \mathbf{v}_C$ (i.e. with respect to a frame of reference in translational motion with the origin of coordinates at point C):

$$\sum_{i=1}^{n} [\mathbf{r}'_i m_i \mathbf{v}_i] = \sum_{i=1}^{n} [\mathbf{r}'_i m_i \mathbf{v}'_i] = \mathbf{L}_C,$$

where $\mathbf{r}'_i = \mathbf{r}_i - \mathbf{r}_C$ is the radius vector of the ith particle in a frame of reference travelling together with the centre of mass.

The relation between the values of the angular momentum $\mathbf{L}$ of a mechanical system about fixed point O and $\mathbf{L}_C$ about the system's centre of mass is of the form

$$\mathbf{L} = \mathbf{L}_C + [\mathbf{r}_C\mathbf{p}],$$

where $\mathbf{p} = \sum_{i=1}^{n} m_i\mathbf{v}_i$ is the momentum of the system in its absolute motion.

4.2 Moment of Inertia

4.2.1 The *moment of inertia of a mechanical system* about fixed axis a is the physical quantity denoted by I_a and equal to the sum of the products of the masses of all n particles of the system by the squares of their normal distances from the axis:

$$I_a = \sum_{i=1}^{n} m_i \rho_i^2,$$

where m_i and ρ_i are the mass of the ith particle and its normal distance from the axis.

The *moment of inertia of a body* about fixed axis a is

$$I_a = \int\limits_{(m)} \rho^2 \, dm = \int\limits_{(V)} \rho^2 D \, dV,$$

where $dm = D \, dV$ is the mass of an infinitely small element of volume dV of the body, D is the density and ρ is the normal distance from the element dV to axis a.

If the body is homogeneous, i.e. its density is the same throughout, then

$$I_a = D \int\limits_{(V)} \rho^2 \, dV.$$

The moment of inertia I_a is a measure of inertness of a body in its rotary motion about fixed axis a (4.3.4), in the same way that its mass is a measure of inertness of the body in its translational motion.

4.2.2 The moment of inertia of a given body about some axis depends, not only on the mass, shape and size of the body, but also on the position of the body with respect to this axis. According to *Steiner's theorem of parallel axes*, the moment of inertia I of a body about an arbitrary axis is equal to the sum of its moment of inertia I_C about a parallel axis through the centre of mass (or of inertia) of the body plus the product of the mass m of the body by the square of the normal distance d between the two axes:

$$I = I_C + md^2.$$

4.2.3 The moments of inertia of certain homogeneous regularly shaped bodies about certain axes are listed in Table 4.1.

4.2.4 The *products of inertia of a body* about the axes of a rectangular Cartesian coordinate system are the quantities:

$$I_{xy} = \int\limits_{(m)} xy \, dm = \int\limits_{(V)} xy D \, dV,$$

TABLE 4.1

Body	Position of axis a	Moment of inertia I_a
Hollow thin-walled circular cylinder of radius R and mass m	Axis of the cylinder	mR^2
Solid circular cylinder (or disk) of radius R and mass m	Axis of the cylinder	$mR^2/2$
Solid sphere of radius R and mass m	Axis through the centre of the sphere	$2mR^2/5$
Hollow thin-walled sphere of radius R and mass m	Axis through the centre of the sphere	$2mR^2/3$
Thin straight rod of length l and mass m	Axis perpendicular to the rod and through its centre along its length	$ml^2/12$
The same rod	Axis perpendicular to the rod and through one end of it	$ml^2/3$

$$I_{xz} = \int\limits_{(m)} xz\, dm = \int\limits_{(V)} xzD\, dV,$$

$$I_{yz} = \int\limits_{(m)} yz\, dm = \int\limits_{(V)} yzD\, dV,$$

where x, y and z are the coordinates of an infinitely small element of the body of volume dV, density D and mass m.
Axis OX is called the *principal axis of inertia of the body* if the products of inertia I_{xy} and I_{xz} are simultaneously equal to zero. Three principal axes of inertia can be passed through each point of the body. The axes are mutually perpendicular.

The moments of inertia of a body about the three principal axes
of inertia, passed through an arbitrary point O of the body, are
called the *principal moments of inertia of the body*.
The principal axes of inertia, passing through the centre of mass
of the body, are called the *principal central axes of inertia of the
body*, and the moments of inertia about these axes are called
the *principal central moments of inertia of the body*. An axis of
symmetry of a homogeneous body is always one of its principal
central axes of inertia.

4.3 The Fundamental Law in the Dynamics of Rotary Motion

4.3.1 It follows from Newton's laws that the first derivative
with respect to time t of the angular momentum **L** of a mechan-
ical system about any fixed point O is equal to the resultant
moment $\mathbf{M}^{ext}$ of all the external forces, applied to the system,
about the same point O:

$$\frac{d\mathbf{L}}{dt} = \mathbf{M}^{ext}.$$

This equation expresses the *law of the variation of angular mo-
mentum of the system*. It is valid, in particular, for a rigid body,
hinged at point O and rotating about it. In this case, the equa-
tion expresses the *fundamental law of dynamics of a rigid body
rotating about a fixed point*.
In terms of the projections on the axes of a fixed rectangular
Cartesian coordinate system with its origin at point O, the law
for the variation of the angular momentum is written in the
form

$$\frac{dL_x}{dt} = M_x^{ext}, \qquad \frac{dL_y}{dt} = M_y^{ext} \quad \text{and} \quad \frac{dL_z}{dt} = M_z^{ext}.$$

Here L_x, L_y, L_z, M_x^{ext}, M_y^{ext} and M_z^{ext} are the angular mo-
menta of the system and the resultant moments of the external
forces about the corresponding coordinate axes.
4.3.2 **Example.** The *regular precession of a gyroscope* due to its
gravitational force. A *gyroscope (symmetrical gyroscope)* is a rigid
body spinning (rotating rapidly) about its axis of symmetry

that can change its position in space. A gyroscope has three degrees of freedom (1.5.6) if it is mounted at fixed point O, which lies on its axis and is called the *centre of suspension of the gyroscope*. If the centre of suspension coincides with the centre of gravity C (7.3.2) of the gyroscope, the gyroscope is said to be balanced: the force of gravity does not affect its state of rotation. Otherwise, it is said to be a *heavy gyroscope*, or top (Fig. 4.3). Owing to the action of the moment of gravitational force about point O
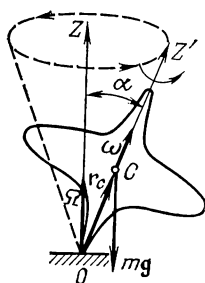
$$\mathbf{M}^{\text{ext}} = [\mathbf{r}_C m\mathbf{g}]$$

a heavy gyroscope rotates about this point so that its axis OZ' rotates uniformly about vertical axis OZ describing a conical surface (shown by dashed lines in Fig. 4.3). Such motion of the gyroscope is called *regular precession*. If the angular velocity of precession $\Omega \ll \omega$ (where ω is the spin angular velocity or, simply, spin velocity of the gyroscope about its axis of symmetry OZ'), it can be assumed with some approximation that the angular momentum $\mathbf{L}$ of the gyroscope about point O is directed along axis OZ' of the gyroscope and equals

$$\mathbf{L} = I\omega,$$

where I is the moment of inertia of the gyroscope about axis OZ'. Hence

$$\frac{d\mathbf{L}}{dt} = [\mathbf{r}_C m\mathbf{g}] = \left[\frac{\mathbf{r}_C}{I\omega_{z'}} \, \mathbf{L} m\mathbf{g}\right] = [\Omega\mathbf{L}],$$

Fig. 4.3

where $\Omega = -mr_C g/I\omega_{z'}$ is the angular velocity of precession, and $\omega_{z'} = \omega$ in the case shown in Fig. 4.3. The greater the spin angular velocity of the gyroscope the slower its precession.

4.3.3 The kinetic energy of a rigid body rotating about a fixed point at the angular velocity ω is

$$E_k = \frac{I\omega^2}{2} \, ,$$

where I is the moment of inertia of the body about the instantaneous axis of rotation (1.5.6).
An element of work done in the infinitely short time interval dt by force $\mathbf{F}$ acting on the body is

$$\delta W = \mathbf{M}\boldsymbol{\omega}\, dt = \mathbf{M}\, d\boldsymbol{\varphi} = M_{\omega} d\varphi,$$

where $\mathbf{M} = [\mathbf{rF}]$ is the moment of force $\mathbf{F}$ about point O ($\mathbf{r}$ is the radius vector from point O to the point of application of force $\mathbf{F}$), $d\varphi = \omega\, dt$ and $d\boldsymbol{\varphi} = \boldsymbol{\omega}\, dt$ are the angle of rotation and the vector of an element of rotation of the body in the time interval dt, and M_{ω} is the moment of force $\mathbf{F}$ about an instantaneous axis of rotation (M_{ω} is equal to the projection of vector $\mathbf{M}$ on the direction of vector $\boldsymbol{\omega}$).
The change in the kinetic energy of a rotating rigid body during the time interval dt is equal to the work done by the external forces:

$$dE_{\mathrm{k}} = M_{\omega}^{\mathrm{ext}} d\varphi,$$

where $M_{\omega}^{\mathrm{ext}}$ is the resultant moment (4.1.3) of external forces about an instantaneous axis of rotation of the body.
4.3.4 If a rigid body rotates about fixed axis OZ at the angular velocity $\boldsymbol{\omega}$, its angular momentum about this axis is

$$L_z = I_z\omega_z \quad \text{and} \quad \mathbf{L}_z = I_z\boldsymbol{\omega}.$$

Here I_z is the moment of inertia of the body about axis OZ and does not change with time ($I_z = \mathrm{const}$), and $|\omega_z| = \omega > 0$ ($\omega_z = \omega$ if vector $\boldsymbol{\omega}$ and the unit vector of axis OZ coincide in direction, otherwise $\omega_z = -\omega$).
The *fundamental law in the dynamics of a rigid body rotating about fixed axis OZ* is

$$I_z\frac{d\boldsymbol{\omega}}{dt} = \mathbf{M}_z^{\mathrm{ext}} \quad \text{or} \quad \boldsymbol{\varepsilon} = \frac{1}{I_z}\mathbf{M}_z^{\mathrm{ext}},$$

where $\boldsymbol{\varepsilon} = d\boldsymbol{\omega}/dt$ is the angular acceleration of the body.
It is evident from the preceding equation that the moment of inertia of a rigid body about some fixed axis is a measure of inertness of the body in its rotation about the given axis: the greater the moment of inertia of the body, the less the angular acceleration it acquires under the action of the same moment of external forces.

4.3.5 The kinetic energy of a rigid body rotating about fixed axis OZ at the angular velocity ω is

$$E_k = \frac{1}{2} I_z \omega^2.$$

An element of work done during the infinitely short time dt by force $\mathbf{F}$ applied to the body is

$$\delta W = \mathbf{M}_z \boldsymbol{\omega} \, dt = M_z \, d\varphi,$$

where M_z is the moment of force $\mathbf{F}$ about the axis OZ of rotation (the unit vector of axis OZ coincides in direction with vector $\boldsymbol{\omega}$).

The change in the kinetic energy of a rigid body during the time interval dt is equal to the work done by the external forces during this interval:

$$dE_k = M_z^{\text{ext}} d\varphi,$$

where M_z^{ext} is the resultant moment of external forces about the axis of rotation of the body.

4.3.6 The *motion of a free rigid body* satisfies the following two differential equations:

$$\frac{d}{dt}(m\mathbf{v}_C) = \mathbf{F}^{\text{ext}} \quad \text{and} \quad \frac{d\mathbf{L}_C}{dt} = \mathbf{M}_C^{\text{ext}}.$$

Here m is the mass of the body, $\mathbf{v}_C$ is the velocity of its centre of mass C, $\mathbf{F}^{\text{ext}}$ is the principal vector of external forces applied to the body (2.5.2), $\mathbf{M}_C^{\text{ext}}$ is the resultant moment of external forces about point C (4.1.6), and $\mathbf{L}_C$ is the angular momentum about the same point C (4.1.7).

The first equation describes translational motion of the free body at the velocity of its centre of mass (2.5.3). The second equation follows from the law of the variation of angular momentum (4.3.1) and describes the rotation of a rigid body about its centre of mass (1.5.9).

4.3.7 The kinetic energy of a free rigid body can be determined on the basis of Koenig's theorem (3.2.4):

$$E_k = \frac{m v_C^2}{2} + \frac{I_C \omega^2}{2},$$

where I_C is the moment of inertia of the body about an instantaneous axis of rotation passing through its centre of mass C, and ω is the angular velocity of the body. In the general case, the instantaneous axis is displaced in the body and the moment of inertia I_C varies with time. The value of I_C is constant if the body is in plane motion (1.5.9).

Example. The kinetic energy of a homogeneous circular cylinder rolling down an inclined plane without slipping. The cylinder has plane motion because all of its points travel in vertical planes that are parallel to one another. The cylinder travels with translational motion at the velocity $\mathbf{v}_C$ directed along the inclined plane, and rotates about its axis ($I_C = mR^2/2$, where m and R are the mass and radius of the cylinder) at the angular velocity ω. From the condition that the cylinder rolls without slipping it follows that the instantaneous velocities of the points of contact between the cylinder and inclined plane equal zero, i.e. $\omega = v_C/R$. Hence, the kinetic energy of the rolling cylinder is

$$E_k = \frac{mv_C^2}{2} + \frac{I_C\omega^2}{2} = \frac{3}{4}\,mv_C^2.$$

4.4 Law of Conservation of Angular Momentum

4.4.1 The *law of the conservation of angular momentum* can be formulated as follows: the angular momentum of a closed system (2.2.4) about any fixed point remains constant with time, i.e.

$$\frac{d\mathbf{L}}{dt} \equiv 0 \quad \text{and} \quad \mathbf{L} = \text{const.}$$

Accordingly, the angular momentum of a closed system about its centre of mass (4.1.7) remains constant with time:

$$\frac{d\mathbf{L}_C}{dt} \equiv 0 \quad \text{and} \quad \mathbf{L}_C = \text{const.}$$

Like the laws of conservation of momentum and energy, the law of conservation of angular momentum is of significance far beyond the scope of classical mechanics. It belongs to the most fundamental of physical laws because it is associated with a definite symmetry property of space: its isotropy. The *isotropy*

of space is manifested in the fact that the physical properties and laws of motion of a closed system are independent of the direction chosen for the coordinate axes in an inertial frame of reference. This means that these properties and laws do not change when the closed system is turned as a whole through any angle in space.

According to modern concepts, not only particles and bodies can have an angular momentum, but fields as well. Elementary particles and systems built from them (for instance, atomic nuclei) can have an angular momentum that is not related to their motion in space and is called their spin (Table 43.1).

4.4.2 With respect to systems that can be described by classical (Newtonian) mechanics, the law of conservation of angular momentum can be regarded as a consequence of Newton's laws. For a closed mechanical system, the resultant moment of external forces about any fixed point (as well as about the centre of mass of the system) is identically zero: $M^{ext} \equiv 0$ (consequently, $M_C^{ext} \equiv 0$, see 4.1.6, where $F = F^{ext} \equiv 0$) and the law of conservation of angular momentum follows from 4.3.1:

$$L = \sum_{i=1}^{n} [r_i m_i v_i] = \text{const},$$

where m_i, r_i and v_i are the mass, radius vector and velocity of the ith particle of the system, which consists of n such particles. Accordingly (see 4.1.7 and 2.5.3),

$$L_C = \sum_{i=1}^{n} [r_i' m_i v_i'] = \sum_{i=1}^{n} [r_i' m_i v_i] = \text{const},$$

where $r_i' = r_i - r_C$, $v_i' = v_i - v_C$, and r_C and v_C are the radius vector and velocity of the system's centre of mass.

4.4.3 If a system is not closed, but the external forces acting on it are such that their resultant moment about fixed point O is identically zero ($M^{ext} \equiv 0$), then, according to Newton's laws (4.3.1), the angular momentum of the system about the same point O remains constant with time: $L = \text{const}$. This condition is practically complied with, for example, by a balanced gyroscope (4.3.2) with three degrees of freedom, if the moment of the friction force in its suspension is sufficiently small.

If the stand of the gyroscope, holding its centre of suspension at rest, is turned in any way, the axis of the gyroscope retains its orientation with respect to a fixed inertial frame of reference.*
Usually $\mathbf{M}^{ext} \not\equiv 0$ and $\mathbf{L} \neq$ const. But if the resultant moment of external forces about any fixed axis passing through point O is identically zero, the angular momentum of the system about this axis remains constant with time. For instance, if $M_z^{ext} \equiv 0$, then $L_z =$ const.
If the system rotates about fixed axis OZ and the resultant moment of external forces about this axis $M_z^{ext} \equiv 0$, the angular momentum of the system about the axis of rotation remains constant with time:

$$I_z \omega = \text{const},$$

where ω and I_z are the angular velocity and moment of inertia of the system.
If the action of internal forces, as well as external forces that comply with the condition $M_z^{ext} \equiv 0$, leads to deformation of the system, changing its moment of inertia I_z, the angular velocity ω correspondingly increases or decreases.

4.4.4 *Free axes of a body* are ones about which a free rigid body (2.2.3) can rotate at constant angular velocity ω in the absence of all external action. Such rotation of a body is said to be *free*, or *inertial, rotation*. The free axes of a body coincide with its principal central axes of inertia (4.2.4). In the general case, the values I_1, I_2 and I_3 of the principal central moments of inertia (4.2.4) differ. The free rotation of such a body (for example, a homogeneous rectangular parallelepiped with edges of different length) is accomplished practically only about two free axes, corresponding to the extreme values of the principal central moments of inertia—the maximum and minimum values. The rotation of the body about its third principal central axis, corresponding to the intermediate value of the moment of inertia of the body, proves to be unstable: even very small

---

* It is assumed that vector $\mathbf{L}$ is directed along the axis of the gyroscope. Otherwise, a free gyroscope has a precessional motion: its axis describes a circular conical surface whose vertex is at the centre of suspension and whose axis is directed along the vector $\mathbf{L} =$ const.

external action is capable of leading to substantial deviation of the instantaneous axis of rotation from its initial direction in the body.

If the values of two principal central moments of inertia of a body are equal: $I_1 = I_2 \neq I_3$, stable free rotation of such a body (for instance, a homogeneous circular cylinder) is possible only about the free axis corresponding to the third, differing value I_3 of the moment of inertia. For a homogeneous circular cylinder such a free axis is its axis of symmetry. But if a long thin cylinder is rotated by means of a string secured to one end, stable rotation of the cylinder can be accomplished about the free axis corresponding to the maximum value of its moment of inertia. This free axis is perpendicular to the axis of symmetry of the cylinder.

CHAPTER 5   FUNDAMENTALS OF THE SPECIAL THEORY OF RELATIVITY

5.1 Postulates of the Special Theory of Relativity

5.1.1 The *special theory of relativity* (sometimes called the *restricted relativity theory*) is the modern physical theory of space and time. The special relativity theory and quantum mechanics (38.1.1) serve as the theoretical basis for all modern physics and engineering (for example, nuclear physics and nuclear technology). The special relativity theory is often called *relativistic theory*, and specific phenomena described by this theory are called *relativistic effects*. Relativistic effects are manifested, as a rule, at velocities of bodies close to that of light in free space ($c = 3 \times 10^8$ m/s) and called *relativistic velocities. Relativistic mechanics* refers to the mechanics of motion at relativistic velocities and is based on the special theory of relativity.

As in classical Newtonian mechanics it is assumed in special relativity theory that time is uniform (3.4.2) and that space is homogeneous (2.7.1) and isotropic (4.4.1).

5.1.2 The special theory of relativity is based on two main principles accepted as the initial postulates.

The *first postulate* is a generalization of Galileo's mechanical principle of relativity (2.8.4) to cover any physical processes.

This postulate, called the *principle of relativity*, or *Einstein's relativity principle*, states: all physical phenomena proceed in the same way, under the same conditions, in all inertial frames of reference (2.1.2). In other words, the principle of relativity contends that physical laws are independent (invariant) with respect to the chosen inertial frame of reference; equations expressing these laws have the same form in all inertial frames of reference. Consequently, it is impossible to ascertain, by means of any physical experiments whatsoever, conducted in a closed system of bodies (2.2.4), whether the system is at rest or is travelling at uniform velocity in a straight line with respect to some inertial frame of reference. All inertial frames of reference are equally justified in physics. It is impossible, on the basis of physical experiments, to select from the host of inertial reference frames any "preferred" or "absolute" frame of reference possessing any qualitative distinctions from other inertial reference frames.

5.1.3 The *second postulate* expresses the *principle of the invariance of the velocity of light*: the velocity of light in free space is independent of the motion of the source of light. It is the same in all directions and in all inertial frames of reference, being one of the most basic physical constants. Experiments show that c, the velocity of light in free space, is the limiting velocity in nature. The velocity of any particles and bodies, as well as the velocity of propagation of any interactions or signals, cannot exceed c.

These specific laws governing the process of light propagation in free space enable this real physical process to be made use of in establishing a procedure for the *chronometry of a frame of reference*, i.e. for the synchronization of clocks located at various points in space and travelling together with the frame of reference being considered (5.2.3).

5.1.4 The postulates of the special theory of relativity contradict the concepts of the properties of space and time that are accepted in classical mechanics and on which the Galilean transformations are based (2.8.1). In particular, this refers to the contention, considered to be "self-evident" in Newtonian mechanics, that the passage of time is identical in all inertial frames of reference. From this it follows that the time interval between any two events is of an absolute nature. For instance, if two events occur simultaneously according to the clock in one inertial frame of reference, then, according to classical

concepts, they also occur simultaneously according to the clock
of any other inertial frame.

The above-mentioned contradiction can be illustrated by the
following example (Fig. 5.1). We have two inertial frames of
reference: fixed frame K and frame K' which travels along the
OX axis at constant velocity $\mathbf{V}$. Assume that a flash of light is
produced at point O at the instant of the zero time reference
in both frames K and K' ($t = t' = 0$), when their origins O
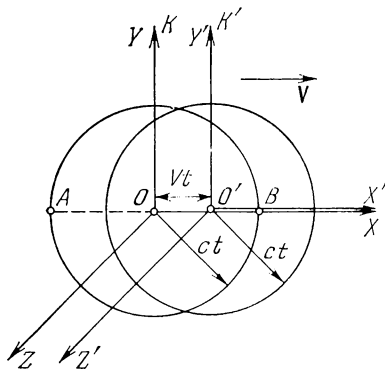
Fig. 5.1

and O' of coordinates coincide. By the instant of time $t > 0$,
light, propagating in free space at the velocity c, reaches
points at the surface of a hollow sphere in the frame K. This
sphere has its centre at point O and its radius equals ct.

It can be assumed that in frame K' the light flash occurred at
the instant of time $t' = 0$ at point O'. Hence, according to the
postulates of the special theory of relativity, by the instant
of time $t' = t$, light in frame K' reaches points at the surface
of a hollow sphere of the same radius ct as in the frame K but
with its centre at point O'. At this instant, however, point O'
is not at point O, but at a distance of Vt from this point. Hence,
any attempt to combine the postulates of the special theory of
relativity with classical conceptions of absolute time that elapses

identically in all frames of reference leads to an absurdity: the light flash must *simultaneously* reach points in space that belong to two different hollow spheres.

5.2 Simultaneity of Events. Synchronization of Clocks

5.2.1 The concept of the simultaneity of two or several events is extensively employed in making various physical measurements. For example, to determine the length l of a rod, arranged along the axis OX of reference frame K and travelling with respect to this frame, it is necessary to note the values of the coordinates $x_2(t)$ and $x_1(t)$ of the ends of the rod *simultaneously*, i.e. at the same instant of time t. Then

$$l = |\, x_2(t) - x_1(t)\, |.$$

The instant of time at which some event occurs (for instance, the launching or touchdown times of a spaceship) is determined by noting a clock's reading that is *simultaneous* with the event being considered. This can be simply done by means of a clock located at the place where the event occurs. Hence, there should be many clocks located at various points in space in each frame of reference. Obviously, these clocks should be in agreement with one another, i.e. run *synchronously*: their readings should be identical at each instant of time t.

5.2.2 The synchronism of clocks located adjacently, i.e. at the same place in space, can be checked by the coincidence of their readings at each arbitrary instant of time. The synchronism of clocks that are separated in space at the points A and B, distant from each other, could be checked in a similar manner if we had at our disposal standard time signals that could propagate *instantaneously* from A to B. Experiments indicate, however, that this method is unfeasible because the velocity of any signal cannot exceed the velocity of light in free space. We can, however, proceed in the following manner. We can convey the clock from B to A, check whether it is synchronous with the clock located at A, and then convey the clock carefully back to B. Then we can check whether the clock brought back to B runs at the same rate as the clock left at A by transmitting time signals from A to B after definite equal time intervals according to the clock at A. But it cannot be established by this procedure whether a shift in the zero time reference has

occurred in the clock reading while it was being conveyed back to B, i.e. whether the clock brought back to B is now fast or slow with respect to the clock at A by a constant amount Δt.

5.2.3 The problem of the synchronism of clocks separated in space at points A and B can be solved only by an unambiguous agreement (or definition) about when these clocks are to be regarded as synchronous. Einstein made use of a real physical process—the propagation of light in free space—as the basis for such a definition. He proceeded from the facts that the velocity of light in free space is the maximum possible velocity at which signals can be transmitted in nature and that it is identical in all directions and in all inertial frames of reference.

Assume that according to the clock at point A the light signal is transmitted from this point at the instant of time t_1 and, after being reflected at point B, returns to A at the instant of time t_3. Then, by definition, the clock at B is in synchronism with that at A if they run at the same rate and if, at the instant the light signal reaches B, the clock located there indicates the time $t_2 = (t_1 + t_3)/2$.

5.2.4 According to the special theory of relativity, the passage of time may differ in different inertial frames of reference. Consequently, the time interval between any two definite events is relative: it can change in going over from one inertial reference frame to another. In particular, the simultaneity of two events occurring at different points separated in space is also relative. Events occurring simultaneously in one inertial frame of reference are not at all simultaneous in other inertial frames that are travelling with respect to the first frame. In some frames of reference one of these events occurs before the other event; in other frames it occurs after the other event.

Thus, in the example illustrated in Fig. 5.1 (see 5.1.4), the arrival of the light flash at points A and B are events that occur simultaneously in the fixed frame of reference K. In the travelling frame of reference K' these events are not simultaneous. The light flash reaches A, receding from the light source O', later than it reaches B, approaching the source O'.

Events connected by a cause-and-effect relationship, i.e. by causality, cannot occur simultaneously in any frame of reference whatsoever, because any effect is due to some process initiated by the cause. At the same time, no process (physical, chemical, biological, etc.) can proceed instantaneously. Therefore, rela-

tivity does not contradict causality in any way. In any inertial frame of reference an event that is an effect always occurs later than the event causing this effect.

5.3 Lorentz's Transformations

5.3.1 It follows from the postulates of the special theory of relativity, as well as from the homogeneity (2.7.1) and isotropy (4.4.1) of space and the uniformity of time (3.4.2), that the relations between the coordinates and times of the same event in two inertial frames of reference are expressed by the *Lorentz transformations*, rather than the Galilean transformations (2.8.1) used in classical (Newtonian) mechanics. According to the principle of relativity and the above-mentioned properties of the symmetry of space and time, the Lorentz transformations should be linear.

5.3.2 The Lorentz transformations are in their simplest form if the like axes of the Cartesian coordinates of the fixed (K) and moving (K') inertial frames of reference are pairwise parallel to each other and if frame K' travels with respect to frame K at the constant velocity $\mathbf{V}$ along axis OX (Fig. 2.3, see 2.8.1). If, in addition, the zero time reference in both frames $(t = 0$ and $t' = 0)$ is taken as the instant when the origins O and O' of coordinates of the two systems coincide, then the Lorentz transformations are of the form

$$x' = \frac{x - Vt}{\sqrt{1 - V^2/c^2}}, \qquad x = \frac{x' + Vt'}{\sqrt{1 - V^2/c^2}},$$

$$y' = y, \qquad\qquad y = y',$$

$$z' = z, \qquad\qquad z = z',$$

$$t' = \frac{t - Vx/c^2}{\sqrt{1 - V^2/c^2}}, \qquad t = \frac{t' + Vx'/c^2}{\sqrt{1 - V^2/c^2}},$$

where c is the velocity of light in free space.

5.3.3 The Lorentz transformations show that not only the spatial coordinates of an event being considered are changed in going over from one inertial frame of reference to another, but the corresponding instants of time as well. But a definite interrelation exists between the spatial coordinates x', y' and z' of

an event and the time instant t' it occurs in an arbitrary inertial frame of reference K'. Hence, the quantity $[(x')^2 + (y')^2 + (z')^2 - c^2 (t')^2]$ is independent of the velocity V of frame K', i.e. it is identical in all inertial frames of reference:

$$(x')^2 + (y')^2 + (z')^2 - c^2 (t')^2 = x^2 + y^2 + z^2 - c^2 t^2.$$

The coordinate x' and time t' cannot be imaginary. It follows, therefore, from the Lorentz transformations that the velocity of relative motion of any two inertial frames of reference cannot exceed the velocity of light in free space ($V \leqslant c$).

5.3.4 According to Einstein's principle of relativity (5.1.2), physical laws must comply with the condition of *relativistic invariance* (*Lorentz invariance*). This requirement implies that equations expressing physical laws retain their form in going over from one inertial frame of reference to another in accordance with the Lorentz transformations.

The Lorentz transformations reduce to the Galilean transformations (2.8.1) at $V \ll c$ or, more precisely, in the limit as $V/c \to \to 0$, i.e. as $c \to \infty$. In other words, the Galilean transformations and classical (Newtonian) mechanics based on them are founded on the assumption of instantaneous propagation of interaction. Such an approximate approach is permissible only in considering the laws of mechanical motion of bodies at velocities a great many times less than the velocity of light in free space.

5.4 Relativity of Lengths and Time Intervals. Time Interval Between Two Events

5.4.1 It follows from the Lorentz transformations (5.3.2) that the linear dimension of a body travelling with respect to an inertial frame of reference is reduced in the direction of motion. This change in the longitudinal size of a body in motion is called the *Lorentz-FitzGerald contraction*. Assume l_0 to be the length of a rod at rest in reference frame K'. If the rod is located along axis $O'X'$ (Fig. 5.2), then $l_0 = x'_2 - x'_1$, where x'_2 and x'_1 are the coordinates of the ends of the rod. The length l of the same rod in reference frame K, with respect to which it travels along axis OX at the velocity V, is equal to the difference in the coor-

dinates of the ends of the rod, measured at the same instant of time t:

$$l = x_2(t) - x_1(t) = (x_2' - x_1') \sqrt{1 - V^2/c^2} = l_0 \sqrt{1 - V^2/c^2}.$$

Transverse dimensions of a body are independent of its velocity and are identical in all inertial frames of reference:

$$y_2 - y_1 = y_2' - y_1' \text{ and } z_2 - z_1 = z_2' - z_1'.$$

Thus, the linear dimensions of a body are relative. They are maximal in the inertial frame of reference with respect to which the body is at rest. These are said to be the body's *proper dimensions*.

5.4.2 The Lorentz-FitzGerald contraction is a kinematic effect of the special theory of relativity. It is not due to the effect of any kind of longitudinal forces compressing the body along its direction of motion. This contraction becomes appreciable only at velocities approaching that of light in free space. It
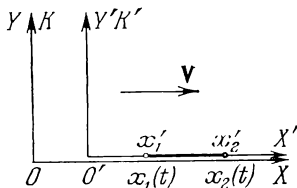
Fig. 5.2

follows from the contraction formula that bodies cannot travel at velocities $V \geqslant c$ because at $V = c$ the longitudinal dimension of the body (in the direction of motion) becomes zero, and at $V \cdot > c$ it would have to become imaginary.

5.4.3 It is evident from the Lorentz transformations that in the theory of relativity we can specify a definite "instant of time" only with respect to a single definite inertial frame of reference. Thus, for instance, a single "instant of time" in reference frame K (a single definite value of time t in this frame) corresponds to many values of time t' in frame K', depending upon the values of coordinate x:

$$t' = \frac{t - Vx/c^2}{\sqrt{1 - V^2/c^2}}.$$

Likewise, a single "instant of time" in frame K', i.e. a single definite value of time t', corresponds to many values of time t

in frame K, depending upon the values of coordinate x':

$$t = \frac{t' + Vx'/c^2}{\sqrt{1 - V^2/c^2}} \cdot$$

5.4.4 Another important consequence of the Lorentz transformations is the relativity of the time interval between any two events (for example, the beginning and end of some process), i.e. the dependence of this time interval on the choice of the inertial frame of reference. Assume that two events, 1 and 2, occur in the moving inertial frame of reference K' at the same point A, fixed in frame K' (i.e. $x_2' = x_1'$), at the instants of time t_1' and t_2'. Thus, the time interval between the events is $\tau_0 = = t_2' - t_1'$.

With respect to fixed inertial frame of reference K, point A travels with the same velocity $\mathbf{V}$ as does frame K'. Hence in frame K events 1 and 2 occur at different points in space, having the coordinates x_1 and x_2, and $x_2 - x_1 = V\tau$, where $\tau = t_2 - t_1$ is the time interval between events 1 and 2 according to the clock in frame K. It follows from the Lorentz transformations that

$$\tau = t_2 - t_1 = \frac{t_2' - t_1'}{\sqrt{1 - V^2/c^2}} = \frac{\tau_0}{\sqrt{1 - V^2/c^2}} \cdot$$

Thus, the time interval between two events is minimal in the inertial frame of reference with respect to which both events occur at the same point. The time measured by a clock travelling together with some object is said to be the *proper time* of the object.

5.4.5 The consequence of the Lorentz transformations that is dealt with in the preceding subsection is evidence of the existence of the relativistic effect of the slowing of time, called *time dilation*, in a moving inertial frame of reference as compared to a fixed one. A clock travelling at the velocity V with respect to a given inertial frame runs slower by a factor of $\sqrt{1 - V^2/c^2}$ than a fixed clock. Correspondingly, in agreement with the relativity principle, all physical processes in a moving frame of reference proceed at a slower rate than in a fixed frame.

The time dilation effect becomes appreciable only at extremely high velocities V, close to that of light in free space. This effect has been demonstrated, for example, by experiments conducted

with muons (43.3.3). The muon is an unstable elementary parti-cle. The mean proper lifetime of muons (according to a clock in the inertial frame of reference with respect to which they are at rest) $\tau_0 = 2.2 \times 10^{-6}$ s. Muons are created in the upper layers of the atmosphere by the action of primary cosmic rays, and they travel with respect to the earth at velocities V close to c. If there was no relativistic effect of time dilation, a muon would be capable of travelling an average distance, with respect to an observer on the earth, of $\tau_0 c = 660$ m through the at-mosphere. In other words, muons could not reach the earth's surface. Actually, they are registered by instruments installed at the surface of the earth, because the mean lifetime of a trav-elling muon, according to the clock of the observer on the earth, $\tau = \tau_0 / \sqrt{1 - V^2/c^2} \gg \tau_0$, and the path travelled by a muon during this time, $\tau V \gg 660$ m.

5.4.6 The relativistic effect of time dilation in a spaceship travelling with respect to the earth makes it feasible to under-take space flights of almost unlimited length and journeys "into the future". According to the relativity principle, all processes in a spaceship, including the aging process of the astronauts, proceed in agreement with the same laws as on the earth. But in the spaceship time must be measured by a clock travelling together with it at the velocity V with respect to the earth. If V is close to c the clock in the ship runs considerably slower than earth clocks (at the space-launch complex), by a factor of $1/\sqrt{1 - V^2/c^2}$. For example, at $\beta = V/c = 0.99999$ the rates of the spaceship and earth clocks differ by a factor of 224. Consequently, in a time interval of $\tau_0 = 10$ years in such a spaceship, one can make a space flight taking $\tau = 2240$ years according to earth clocks, aging only 10 years during the flight. In this flight the spaceship can travel a tremendous distance from the earth: $l = V\tau = \beta c \tau = 2239.98$ light years*. The closer V is to c, the greater the distance l travelled by the spaceship with respect to the earth in the same interval τ_0 of proper time in the spaceship, i.e. the longer the space flight that the astronauts can make during their lifetime.

If an astronaut completes a space flight at a velocity V close to c by returning to the earth, he will find that people remaining

---

* A light year is the distance travelled by light in free space in one year: 1 light year $= 9.4605 \times 10^{15}$ m.

on the earth (including his twin brother) has aged more than
he has during the flight. If V differs from c by a sufficiently
small amount, when $(1 - V^2/c^2)^{-1/2} \gg 1$, the astronaut can,
during the flight, outlive all his contemporaries and find him-
self, upon his return, among future generations.

5.4.7 It would seem, on the face of it, that, on the basis of the
relativity principle, we should arrive at a conclusion exactly
the opposite of that given in the preceding subsection. The clock
on the earth, travelling at the velocity $-V$ with respect to the
spaceship, should run slower than the clock in the ship. There-
fore, the time required for the flight should be longer for the
astronaut, and not for the people left on the earth. Hence, the
twin that made the flight should age more than his brother.

We have thus found that the difference between the readings
of the clocks at the space-launch complex and in the ship is
positive according to 5.4.6, while, on the other hand, it is
negative. This absurd result is called the *clock*, or *twin*, *paradox*.
Actually, no clock paradox exists. It arose due to improper
application of the relativity principle. This principle contends
that not all frames of reference are equally justified, but only
all *inertial* frames of reference. But the frame of reference of
the spaceship, in contrast to that of the earth, is not an inertial
one all throughout the flight. During launching, turning around
and in braking orbit when approaching the earth, the spaceship
travels with acceleration (or deceleration). Consequently, the
problem concerning the running of the clock at the launching
complex which is always at rest with respect to a single inertial
frame, and that of the clock in the spaceship, is, in principle,
not a symmetrical one. Hence, the earth and ship frames of
reference are not equally justified in the given problem. The
reasoning in Sect. 5.4.6 is the correct one because it is based
on the use of an inertial (the earth) frame of reference. Accord-
ingly, the reasoning given in Sect. 5.4.7, leading to the clock
paradox, is incorrect. In the second case, the general theory of
relativity should be applied, rather than the special theory.
When this is done, it is found that, from the point of view of
the astronaut as well, his clock runs slower than the clock at
the launching complex.

5.4.8 The *interval*, or *space-time interval*, between two events,
measured in the inertial frame of reference K', is the quantity

$$s'_{12} = \sqrt{c^2 (t'_{12})^2 - (l'_{12})^2},$$

where $t'_{12} = t'_2 - t'_1$ is the time interval between the two events being considered (according to the clock in frame K'), and $l'_{12} = \sqrt{(x'_2 - x'_1)^2 + (y'_2 - y'_1)^2 + (z'_2 - z'_1)^2}$ is the distance between the points at which events 1 and 2 occur, also measured in frame of reference K'.

It follows from the Lorentz transformations that the interval between two given events, 1 and 2, is invariant with respect to the choice of the inertial frame of reference, i.e. it does not change in transferring from moving inertial frame K' to fixed frame K:

$$s'_{12} = s_{12} = \text{inv}, \quad \text{where } s_{12} = \sqrt{c^2 t^2_{12} - l^2_{12}}.$$

If $s^2_{12} > 0$, i.e. if s_{12} is a real number, the interval s_{12} is said to be *timelike*.

The interval s_{12} is said to be *spacelike* if $s^2_{12} < 0$, i.e. s_{12} is an imaginary number.

5.4.9 It follows from the invariance of the interval with respect to the choice of the inertial frame K' that in all frames K' the values of t'_{12} and l'_{12} for the given two events, 1 and 2, satisfy the equation of a hyperbola:

$$c^2 (t'_{12})^2 - (l'_{12})^2 = s^2_{12}.$$

If $s^2_{12} > 0$ the relation between t'_{12} and l'_{12} in various inertial frames of reference K', travelling with respect to fixed frame K at any possible velocities ($0 \leqslant V < c$), can be represented graphically as the two branches *I* and *II* of a hyperbola (Fig.5.3). Consequently, the sign of the time interval between events 1 and 2, related by a timelike interval, is absolute. It is indepen-

Fig. 5.3

dent of the choice of the frame of reference: in all frames K' the second event always occurs either after the first, i.e. $t'_{12} > 0$ (branch *I*) or always before the first, i.e. $t'_{12} < 0$ (branch *II*). The distance l'_{12} is relative; an inertial frame K' can be chosen in which $l'_{12} = 0$, i.e. events 1 and 2 occur at the same place (points A and B on the branches *I* and *II* of the hyperbola).
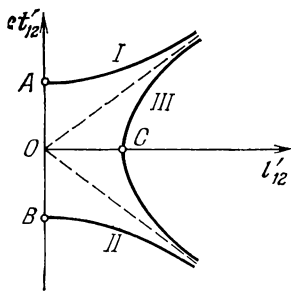
Two events related by causality should always correspond to
a timelike interval or, in any case, one equal to zero ($s_{12} = 0$).
This is due to the fact that the signal by means of which event 1
(the cause) leads to event 2 (the effect) cannot be propagated
in space at a velocity exceeding that of light in free space:
$l'_{12} \leqslant c\ (t'_2 - t'_1)$.
5.4.10 In the case of events related by a spacelike interval
($s^2_{12} < 0$), the sign of t'_{12} is relative: $t'_{12} > 0$ (upper part of hyper-
bola III in Fig. 5.3) in certain inertial frames K', whereas
in others $t'_{12} < 0$ (lower part of hyperbola III). Point C corres-
ponds to the frame K' in which $t'_{12} = 0$, i.e. events 1 and 2 occur
simultaneously.

5.5 Transformation of Velocities and Accelerations in Relativistic Kinematics

5.5.1 The velocities **v** and **v**′ of a particle in two inertial frames
of reference K and K' are

$$\mathbf{v} = \frac{d\mathbf{r}}{dt} = v_x \mathbf{i} + v_y \mathbf{j} + v_z \mathbf{k} \quad \text{and} \quad \mathbf{v}' = \frac{d\mathbf{r}'}{dt'} = v'_{x'} \mathbf{i}' + v'_{y'} \mathbf{j}' + v'_{z'} \mathbf{k}'$$

where $\mathbf{r} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$ and $\mathbf{r}' = x'\mathbf{i}' + y'\mathbf{j}' + z'\mathbf{k}'$ are the
radius vectors of the particle in frames K and K'. The projections
of velocities **v** and **v**′ on the axes of Cartesian coordinates of
frames K and K' are

$$v_x = \frac{dx}{dt}, \quad v_y = \frac{dy}{dt}, \quad v_z = \frac{dz}{dt} \ ; \quad v'_{x'} = \frac{dx'}{dt'}, \quad v'_{y'} = \frac{dy'}{dt'} \quad \text{and}$$

$$v'_{z'} = \frac{dz'}{dt'} \ .$$

If the like axes of the Cartesian coordinates of frames K' and
K are pairwise parallel, and frame K' travels with respect to
frame K at the constant velocity **V** along axis OX (Fig. 2.3,
see 2.8.1) and, in addition, if at the zero time references in
frames K and K' ($t = 0$ and $t' = 0$), the origins O and O' of
coordinates of these frames coincide, the Lorentz transforma-
tions in the form given in 5.3.2 are valid. It follows from these
transformations that the relations between the projections of

the velocities of the particle on the axes of the Cartesian coordinates of frames K and K' are of the form:

$$v'_{x'} = \frac{v_x - V}{1 - Vv_x/c^2}, \quad v_x = \frac{v'_{x'} + V}{1 + Vv'_{x'}/c^2},$$

$$v'_{y'} = \frac{v_y \sqrt{1 - V^2/c^2}}{1 - Vv_x/c^2}, \quad v_y = \frac{v'_{y'} \sqrt{1 - V^2/c^2}}{1 + Vv'_{x'}/c^2}$$

$$v'_{z'} = \frac{v_z \sqrt{1 - V^2/c^2}}{1 - Vv_x/c^2}, \quad v_z = \frac{v'_{z'} \sqrt{1 - V^2/c^2}}{1 + Vv'_{x'}/c^2}.$$

These equations express the *law of the composition of velocities in relativistic kinematics*. In the limit as $c \to \infty$ these equations lead to the ordinary law for the composition of velocities in classical mechanics (2.8.2):

$$v'_{x'} = v_x - V, \quad v'_{y'} = v_y, \quad v'_{z'} = v_z \text{ and } \mathbf{v'} = \mathbf{v} - \mathbf{V}.$$

5.5.2 The relations between the squares of the magnitudes of vectors $\mathbf{v}$ and $\mathbf{v'}$ are

$$v^2 = c^2 \left[1 - \frac{[1 - (v'/c)^2](1 - V^2/c^2)}{(1 + Vv'_{x'}/c^2)^2} \right] \quad \text{and}$$

$$(v')^2 = c^2 \left[1 - \frac{(1 - v^2/c^2)(1 - V^2/c^2)}{(1 - Vv_x/c^2)^2} \right].$$

In particular, if $v' = c$, then $v = c$ and vice versa. Hence, if the velocity of a particle with respect to some inertial frame of reference is equal to the velocity of light in free space, then it must have the same magnitude with respect to any other inertial frame regardless of the velocity of relative motion of these two frames. In other words, the sum of two velocities, of which one is equal to c, is always equal to c. This law, detected in the motion of such elementary particles as photons (37.1.4) and the neutrino (Table 43.1), demonstrates the limiting nature of the velocity of light in free space (5.1.3).

5.5.3 It is evident from the equations in the preceding subsection that a particle, travelling with respect to some inertial frame of reference at a velocity less than c, has a velocity with respect to any other inertial frame that is also less than c (for instance, if $v < c$, then $v' < c$, and vice versa). From this it follows, among other things, that no matter how close to c

the velocities of two particles are, their relative velocity is always less than c. Assume, for example, that two particles are travelling toward each other along the axis OX of frame K at the respective velocities: $\mathbf{v}_1 = 0.8c\mathbf{i}$ and $\mathbf{v}_2 = -0.8c\mathbf{i}$. The velocity $\mathbf{u}_{21}$ of the second particle with respect to the first is not equal, as is assumed in classical mechanics, to the vector difference $\mathbf{v}_2 - \mathbf{v}_1 = -1.6c\mathbf{i}$ if only because the magnitude of this velocity exceeds c. The required velocity is equal to the velocity of the second particle with respect to inertial frame K', travelling together with the first particle ($\mathbf{V}=0.8c\mathbf{i}$), i.e. $\mathbf{u}_{21}=\mathbf{v}_2'$. It follows from the equation in 5.5.1 that

$$v_{2x'}' = \frac{v_{2x} - V}{1 - Vv_{2x}/c^2} = -\frac{1.6c}{1+0.64} = -0.976c, \quad v_{2y'}' = v_{2z'}' = 0,$$

i.e. $\mathbf{u}_{21} = -0.976c\mathbf{i}'$ and $|\mathbf{u}_{21}| < c$.

5.5.4 The projections of the acceleration of a particle on the axes of the Cartesian coordinates of two inertial frames of reference K and K' (5.5.1) are related by the following equations;

$$a_{x'}' = \frac{dv_{x'}'}{dt'} = a_x \left(\frac{\sqrt{1-V^2/c^2}}{1-\dfrac{Vv_x}{c^2}} \right)^3,$$

$$a_{y'}' = \frac{dv_{y'}'}{dt'} = \left[\left(1-\frac{Vv_x}{c^2} \right) a_y + \frac{Vv_y}{c^2} a_x \right] \frac{1-V^2/c^2}{\left(1-\dfrac{Vv_x}{c^2} \right)^3},$$

$$a_{z'}' = \frac{dv_{z'}'}{dt'} = \left[\left(1-\frac{Vv_x}{c^2} \right) a_z + \frac{Vv_z}{c^2}.a_x \right] \frac{1-V^2/c^2}{\left(1-\dfrac{Vv_x}{c^2} \right)^3},$$

$$a_x = \frac{dv_x}{dt} = a_{x'}' \left(\frac{\sqrt{1-V^2/c^2}}{1+\dfrac{Vv_{x'}'}{c^2}} \right)^3,$$

$$a_y = \frac{dv_y}{dt} = \left[\left(1+\frac{Vv_{x'}'}{c^2} \right) a_{y'}' - \frac{Vv_{y'}'}{c^2} a_{x'}' \right] \frac{1-V^2/c^2}{\left(1+\dfrac{Vv_{x'}'}{c^2} \right)^3},$$

$$a_z = \frac{dv_z}{dt} = \left[\left(1+\frac{Vv_{x'}'}{c^2} \right) a_{z'}' - \frac{Vv_{z'}'}{c^2} a_{x'}' \right] \frac{1-V^2/c^2}{\left(1+\dfrac{Vv_{x'}'}{c^2} \right)^3}.$$

5.6 Basic Law of Relativistic Dynamics

5.6.1 In contrast to classical mechanics, the mass of a particle is not constant in relativistic mechanics. It depends upon the velocity $\mathbf{v}$ of the particle. Its value m differs in two inertial frames of reference moving with respect to each other. The *dependence of mass on velocity* is expressed by the equation:

$$m = \frac{m_0}{\sqrt{1 - v^2/c^2}} \,,$$

where m_0 is the *rest mass* of the particle, i.e. its mass measured in the inertial frame of reference with respect to which the particle is at rest, and c is the velocity of light in free space. The mass m is frequently called the *relativistic mass*.
The effect of the velocity of a particle on its mass becomes appreciable only at values of v close to c. For example, $m/m_0 = 1.005$ at $v/c = 0.1$ and $m/m_0 = 2.29$ at $v/c = 0.9$. It is evident from the law for the dependence of m on v that particles with a rest mass $m_0 \neq 0$ cannot travel at velocities greater than or equal to c (i.e. $v < c$). At the same time, particles whose rest mass equals zero [photons (37.1.4) and the neutrino (Table 43.1)] cannot have a velocity other than c.
5.6.2 The momentum $\mathbf{p} = m\mathbf{v}$ of a particle is a nonlinear function of its velocity:

$$\mathbf{p} = \frac{m_0 \mathbf{v}}{\sqrt{1 - v^2/c^2}} \,.$$

Vector $\mathbf{p}$ is sometimes called the *relativistic momentum* of the particle (to distinguish it from the value $m_0 \mathbf{v}$ of its momentum in classical mechanics). It is obvious that at $v \ll c$, the momentum $\mathbf{p} = m\mathbf{v} \approx m_0 \mathbf{v}$.
By virtue of the homogeneity of space (2.7.1), the *law of conservation of relativistic momentum* is valid in relativistic mechanics: the momentum of a closed system (2.2.4) does not change in the course of time.
Following from this law is the *law of conservation of relativistic mass*: in any processes that take place in a closed system, its total relativistic mass remains unchanged.

7*

5.6.3 The *basic law of relativistic dynamics* is: the rate of change in the momentum of a particle is equal to the force **F** acting on it, i.e.

$$\frac{d\mathbf{p}}{dt} = \mathbf{F}, \quad \text{or} \quad \frac{d}{dt} \left(\frac{m_0 \mathbf{v}}{\sqrt{1 - v^2/c^2}} \right) = \mathbf{F}.$$

Note. If the particle is subject to the simultaneous action of several forces, **F** is understood to be the resultant force (2.2.2).
5.6.4 An element of work done by force **F** in the infinitely small displacement $d\mathbf{r}$ of its point of application is

$$\delta W = (\mathbf{F}\, d\mathbf{r}) = (\mathbf{Fv})\, dt.$$

Combining the basic law of relativistic dynamics (5.6.3) and the dependence of mass on the velocity (5.6.1) we obtain

$$\mathbf{F} = m \frac{d\mathbf{v}}{dt} + \mathbf{v} \frac{dm}{dt} \quad \text{and} \quad \frac{dm}{dt} = \frac{mv}{c^2 - v^2} \frac{dv}{dt}.$$

Hence

$$\delta W = (\mathbf{Fv})\, dt = m\, (\mathbf{v}\, d\mathbf{v}) + v^2 dm = mv\, dv + v^2 dm = c^2 dm.$$

5.6.5 The acceleration imparted to a particle by the force **F** is

$$\mathbf{a} = \frac{d\mathbf{v}}{dt} = \frac{\mathbf{F}}{m} - \frac{\mathbf{v}}{m} \frac{dm}{dt} = \frac{1}{m} \left[\mathbf{F} - \frac{\mathbf{v}}{c^2} (\mathbf{Fv}) \right].$$

Hence, in contrast to classical mechanics, the acceleration of a particle in relativistic mechanics does not, in general, coincide in direction with the force imparting the acceleration. Vector **a** is collinear with force **F** in only two cases:
(a) when force **F** is perpendicular to the velocity **v** of the particle (transverse force) so that $(\mathbf{Fv}) = 0$ and

$$\mathbf{a} = \frac{\mathbf{F}}{m} = \frac{\mathbf{F}}{m_0} \sqrt{1 - \frac{v^2}{c^2}} \; ;$$

(b) when force **F** is parallel to the velocity vector **v** of the particle (longitudinal force) so that $\mathbf{v}\,(\mathbf{Fv}) = v^2 \mathbf{F}$ and

$$\mathbf{a} = \frac{\mathbf{F}}{m} \left(1 - \frac{v^2}{c^2} \right) = \frac{\mathbf{F}}{m_0} \left(1 - \frac{v^2}{c^2} \right)^{3/2} .$$

A longitudinal force imparts to a particle an acceleration less by a factor of $1 - v^2/c^2$ than that imparted by a transverse force of the same magnitude. This is due to the fact that a transverse force changes the velocity of a particle only in direction (the magnitude v of the velocity and the relativistic mass are not changed), whereas a longitudinal force changes the magnitude of the velocity of a particle as well as its mass.

5.7 Mass-Energy Relation

5.7.1 The increment of the kinetic energy E_h of a particle is equal to the work done by the force F acting on the particle (5.6.4):

$$dE_k = \delta W = c^2 \, dm,$$

where dm is the corresponding increment of relativistic mass of the particle. It follows that

$$E_k = (m - m_0) \, c^2 = m_0 c^2 \left[\frac{1}{\sqrt{1 - v^2/c^2}} - 1 \right],$$

where m_0 is the rest mass of the particle. Expanding $(1 - v^2/c^2)^{-1/2}$ into a Maclaurin series we obtain

$$E_k = m_0 c^2 \left[\frac{1}{2} \left(\frac{v}{c} \right)^2 + \frac{3}{8} \left(\frac{v}{c} \right)^4 + \ldots \right].$$

At $v \ll c$ this equation reduces to the ordinary expression for the kinetic energy in classical mechanics:

$$E_k = \frac{m_0 v^2}{2} = \frac{m v^2}{2} .$$

5.7.2 It follows from the first equation in the preceding subsection that an increase in the kinetic energy of a body should be accompanied by a corresponding increase in its relativistic mass m:

$$dm = \frac{1}{c^2} \, dE_k.$$

Changes in other kinds of energy of a body are also associated with an increase in its relativistic mass. If, for example, in

heating a body at rest its internal energy (9.1.2) is increased
by the amount dU, the mass m of the body, equal to its   rest
mass m_0, is increased by

$$dm = dm_0 = \frac{1}{c^2}\ dU.$$

In the general case, a change in the total energy E of the body
by the amount dE is accompanied by a change in its relativistic
mass by the amount

$$dm = \frac{1}{c^2}\ dE.$$

Accordingly, the following universal relation exists between E
and m:

$$E = mc^2 = \frac{m_0 c^2}{\sqrt{1 - v^2/c^2}}\ ;$$

this is known as the *mass-energy relation*: the total energy of
a body (or system) is equal to the product of the relativistic
mass of the body (or system) by the square of the velocity of
light in free space.
By virtue of the uniformity of time (3.4.2), the *law of the   con-
servation of energy* is valid in relativistic mechanics as it is in
classical mechanics: the total energy of a closed system (2.2.4)
does not change in the course of time.
It follows from the mass-energy relation that the laws of the
conservation of relativistic mass (5.6.2) and of the total energy
are not independent laws.
5.7.3 The total energy of a particle or a system of particles (for
instance, of an atomic nucleus, atom, molecule or body) at
rest is equal to

$$E_0 = m_0 c^2,$$

where m_0 is the rest mass; E_0 is called the *rest energy* of the
particle or system. The values of m_0 and E_0 are independent
of the choice of the inertial frame of reference. These quantities
are constant characteristics of structureless (elementary) par-
ticles, similar, for example, to their electric charge and spin
(Table 43.1). The rest mass and rest energy of a system of par-
ticles depend upon the composition of the system and its inter-
nal state. For example, the rest mass of an excited nucleus (or

atom) is greater than the rest mass of the same nucleus (or atom) in the ground state.

5.7.4 The total energy E of a particle and its momentum $\mathbf{p}$ are related by the equations:

$$\mathbf{p} = \frac{E}{c^2}\,\mathbf{v} \quad \text{or} \quad \frac{E^2}{c^2} - p^2 = m_0^2 c^2 \quad \text{or} \quad E = \sqrt{p^2 c^2 + m_0^2 c^4}\,.$$

The values of the total energy, relativistic mass and momentum of a given particle, in contrast to its rest mass m_0, are relative, i.e. they are different in two inertial frames of reference K (E, m and $\mathbf{p}$) and K' (E', m' and $\mathbf{p}'$). But the difference between the square of the total energy of a particle divided by c^2 and the square of the momentum of the particle, like the interval between two events (5.4.8), does not depend upon the choice of the inertial frame of reference:

$$\frac{(E')^2}{c^2} - (p')^2 = \frac{E^2}{c^2} - p^2 = m_0^2 c^2.$$

5.7.5 In going over from one inertial frame K to another K' that is travelling with respect to the first at the velocity $\mathbf{V} = $ $=$ const along the OX axis (Fig. 2.3, see 2.8.1), the projections of the momentum of the particle on the coordinate axes and its total energy are transformed as follows:

$$p_{x'}' = \frac{p_x - \dfrac{VE}{c^2}}{\sqrt{1 - V^2/c^2}}\,, \qquad p_x = \frac{p_{x'}' + \dfrac{VE'}{c^2}}{\sqrt{1 - V^2/c^2}}\,,$$

$$p_{y'}' = p_y, \qquad\qquad p_y = p_{y'}',$$

$$p_{z'}' = p_z, \qquad\qquad p_z = p_{z'}',$$

$$E' = \frac{E - Vp_x}{\sqrt{1 - V^2/c^2}}\,, \qquad E = \frac{E' + Vp_{x'}'}{\sqrt{1 - V^2/c^2}}\,.$$

5.7.6 It does not at all follow from the laws of conservation of relativistic mass (5.6.2) and of total energy (5.7.2) that the rest mass and rest energy of a closed system cannot change. For instance, the sum of the rest masses of free protons and neutrons is always greater than the rest mass of the atomic nucleus they form.

A quantity introduced to characterize systems held together with a margin of safety (for example, atomic nuclei, atoms and molecules) is the concept of the binding energy. The *binding energy of a system* is measured as the least amount of work that must be done to separate the system into its component parts (for instance, an atom into its nucleus and electrons). The binding energy of a system is

$$E_b = \sum_{i=1}^{n} m_{0i} c^2 - M_0 c^2,$$

where M_0 is the rest mass of a system of n particles, and m_{0i} is the rest mass of the ith particle in the free state. The quantity

$$\Delta m = \sum_{i=1}^{n} m_{0i} - M_0 = \frac{E_b}{c^2}$$

is sometimes called the *mass defect of the system*.

CHAPTER 6   GRAVITATION

6.1 Law of Universal Gravitation

6.1.1 *Isaac Newton's law of universal gravitation* states that the force of mutual attraction between any two particles is directly proportional to the product of their masses and inversely proportional to the square of the distance between them. Such forces are called *gravitational forces*. If m_1 and m_2 are the masses of the particles being considered and $\mathbf{r}_1$ and $\mathbf{r}_2$ are the radius vectors of these particles, then, according to the law of universal gravitation, the force $\mathbf{F}_{12}$ attracting the first particle toward the second particle, and the force $\mathbf{F}_{21}$ attracting the second particle toward the first particle are equal in magnitude and opposite in direction; they are equal to:

$$\mathbf{F}_{12} = -G\,\frac{m_1 m_2}{r_{12}^3}\,\mathbf{r}_{12} \quad \text{and} \quad \mathbf{F}_{21} = -G\,\frac{m_1 m_2}{r_{21}^3}\,\mathbf{r}_{21}.$$

Here $\mathbf{r}_{12} = \mathbf{r}_1 - \mathbf{r}_2$ and $\mathbf{r}_{21} = \mathbf{r}_2 - \mathbf{r}_1$ are radius vectors from the second particle to the first and from the first to the second, respectively, and $r_{12} = r_{21} = |\mathbf{r}_{12}| = |\mathbf{r}_{21}|$ is the distance between the particles. The proportionality factor G is called the *constant of universal gravitation*, or the *gravitational constant*. In value the gravitational constant is equal to the force of mutual attraction of two particles of unit mass with a distance
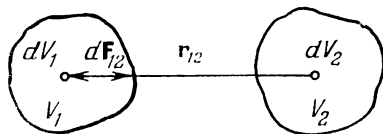
Fig. 6.1

of unit length between them. It has been found from experiments that

$$G = (6.6720 \pm 0.0041) \times 10^{-11} \ (\text{N-m}^2)/\text{kg}^2.$$

6.1.2 The gravitational interaction of two bodies of arbitrary size and shape (Fig. 6.1) is described by the equation

$$\mathbf{F}_{12} = -G \int\limits_{(V_1)} \rho_1 \, dV_1 \int\limits_{(V_2)} \frac{\rho_2}{r_{12}^3} \mathbf{r}_{12} \, dV_2$$

where $\mathbf{r}_{12}$ is the radius vector from the infinitely small element dV_2 of volume of the second body to the infinitely small element dV_1 of volume of the first body, ρ_1 and ρ_2 are the densities of the two elements, and integration is carried out throughout the whole volume of both bodies.
The calculations of force $\mathbf{F}_{12}$ are substantially simplified in two cases:
(a) when there is a spherically symmetrical distribution of the masses in the two bodies, i.e. both bodies are of spherical shape and the density of each body depends only upon the distance to its centre (in particular, the bodies can be homogeneous);
(b) when one of the bodies is of negligibly small size with re-

spect to the other body in which the distribution of its mass is
spherically symmetrical.
In these cases

$$\mathbf{F}_{12} = -G\,\frac{m_1 m_2}{r_{12}^3}\,\mathbf{r}_{12},$$

where m_1 and m_2 are the masses of the bodies, and $\mathbf{r}_{12}$ is the ra-
dius vector connecting the centres of mass (2.3.3) of the second
and first bodies.
6.1.3 As a first approximation, it can be assumed that the earth
is of spherical shape and the distribution of its mass is spherical-
ly symmetrical. Hence the gravitational force $\mathbf{F}$, exerted by the
earth on a body of mass m, is directed toward the centre of the
earth and its magnitude is

$$F = G\,\frac{m M_{\mathrm{E}}}{r^2}$$

where M_{E} is the mass of the earth and r is the distance from
the body to the earth's centre (the size of any body on earth
is negligibly small compared to the radius of the earth).
6.1.4 In application to such microscopic objects as elementary
particles, gravitational interaction is practically of no import-
ance whatsoever. It is found to be extremely weak in comparison
to all other kinds of interaction: strong, electromagnetic and
weak interaction (43.2.5 through 43.2.8). For example, the
electric force of mutual repulsion of two electrons exceeds their
gravitational force by over 10^{42} times! Even for ordinary macro-
scopic objects on the earth the forces of gravitational inter-
action are extremely small. Two homogeneous spheres, for
instance, with a mass of 1000 kg each and a distance of one
metre between their centres are attracted to each other with
a force of only 7×10^{-5} N.
But, at the same time, gravitational forces are the determining
factors in the motion of objects that are investigated in astron-
omy and astronautics (spaceships, planets and their satellites,
planetary systems, stars, etc.). This is due, primarily, to the
huge size of astronomical bodies and, in the second place, to
the small electromagnetic forces of interaction of such bodies,
which, as a whole, are practically electroneutral.

6.2 Gravitational Field

6.2.1 Gravitational interaction between bodies is due to the *gravitational field* set up by the bodies. A distinctive feature of a gravitational field is that a particle placed in this field is subject to a force proportional to the mass of the particle.
The force characteristic of a gravitational field is its *intensity*, the vector quantity g_f, which is equal to the ratio of the force $\mathbf{F}$, exerted on the particle by the field in which it is placed, to the mass m of the particle:

$$g_f = \frac{\mathbf{F}}{m}.$$

The intensity of a gravitational field is independent of the mass m of the particle. It is a function of the coordinates (x, y, z) of points of the field being considered. In the case of an unsteady field the intensity also depends upon the time t.
A gravitational field is steady, or stationary (2.2.1), if the body that sets it up is fixed in the frame of reference chosen to describe the field. The intensity of a steady gravitational field depends only upon the coordinates: $g_f = g_f(x, y, z)$.
It follows from Newton's second law (2.4.3) that the forces of a gravitational field impart an acceleration $\mathbf{a}$ to a free particle in the field. This acceleration equals the intensity of the field:

$$\mathbf{a} = \frac{\mathbf{F}}{m} = g_f.$$

6.2.2 It follows from the law of universal gravitation (6.1.1) that the intensity of the gravitational field of a fixed particle of mass M, located at the origin of coordinates, equals

$$g_f = -G\frac{M}{r^3}\,\mathbf{r},$$

where $\mathbf{r}$ is the radius vector of the point being considered in the field.
A gravitational field is a conservative, or potential, one (3.1.6) because the force acting on a particle of mass m brought into the field is a central force (3.3.4);

$$\mathbf{F} = m g_f = -G\frac{mM}{r^2}\,\frac{\mathbf{r}}{r}.$$

Correspondingly, the potential energy of a particle in such a field is equal to (3.3.4):*

$$E_p = \int\limits_r^\infty F_r \, dr = -GmM \int\limits_r^\infty \frac{dr}{r^2} = -G\,\frac{mM}{r}\,.$$

It is equally correct to regard the quantity E_p as the potential energy of a particle of mass M in a gravitational field set up by a particle of mass m or, finally, as the *mutual potential energy of two particles* due to their gravitational interaction.

6.2.3 A gravitational field complies with the *principle of superposition of fields*: in superposing several (n) gravitational fields their intensities at each point in space are added vectorially, i.e. the intensity of the resultant field is

$$\mathbf{g}_f = \sum_{i=1}^{n} \mathbf{g}_{fi},$$

where $\mathbf{g}_{fi}$ is the intensity of only the ith field at the point being considered in space.

The intensity of the gravitational field of an arbitrary system consisting of n fixed particles is

$$\mathbf{g}_f = -G \sum_{i=1}^{n} \frac{m_i}{\rho_i^3}\,\boldsymbol{\rho}_i,$$

where $\boldsymbol{\rho}_i = \mathbf{r} - \mathbf{r}_i$ is the radius vector from the ith particle, whose radius vector is equal to $\mathbf{r}_i$, to the point in the field specified by radius vector $\mathbf{r}$. Consequently, the potential energy of a particle of mass m in this gravitational field is

$$E_p = -Gm \sum_{i=1}^{n} \frac{m_i}{\rho_i}\,,$$

----

* Here and throughout Sects. 6.2 and 6.3 an infinitely distant point is chosen as the reference point for the potential energy, i.e. it is assumed that $E_p\,(\infty) = 0$.

In particular, if the gravitational field is set up by a body whose mass M is distributed in a spherically symmetrical manner (6.1.2), then outside this body

$$g_f = -G\frac{M}{r^3}\,\mathbf{r} \quad \text{and} \quad E_p = -\frac{GmM}{r}\,,$$

where $\mathbf{r}$ is a radius vector from the centre of the body to the point being considered in the field. These equations are valid, for example, for the gravitational field of the earth.

6.2.4 By virtue of the conservative nature (potentiality) of the gravitational field (3.1.6), its energy characteristic, the potential, can be introduced. The *potential of a gravitational field* is the scalar φ, equal to the ratio of the potential energy E_p of a particle, placed at the point being considered in the field to the mass m of the particle:

$$\varphi = \frac{E_p}{m}\,.$$

The potential φ is independent of the mass m of the particle; it is a function of the coordinates of points of the gravitational field. For example, the potential of a gravitational field set up by a fixed particle of mass M is

$$\varphi = -\frac{GM}{r}\,,$$

where r is the distance from the source of the field to the point being considered.

The potential of a gravitational field set up by an arbitrary system of n fixed particles is

$$\varphi = -\sum_{i=1}^{n} G\,\frac{m_i}{\rho_i}\,,$$

where ρ_i is the distance from a particle of mass m_i to the point being considered in the field. Thus, in superposing gravitational fields, their potentials are added together algebraically, i.e. the potential φ at any point of a resultant field is equal to

the algebraic sum of the potentials at the same point for all
the superposed fields taken separately;

$$\varphi = \sum_{i=1}^{n} \varphi_i.$$

Note. In applying this equation it is necessary that the same
reference point be chosen for the potentials φ_i of all the fields
being superposed: $\varphi_i (\infty) = 0$ (see the footnote to Sec. 6.2.2).
6.2.5 An element of work done by the forces of a gravitational
field in an infinitely small displacement dr of a particle of mass m
in this field is:

$$dW = (\mathbf{F} \, d\mathbf{r}) = m \, (\mathbf{g}_f \, d\mathbf{r}).$$

But, on the other hand, the work dW is equal to the decrease
in the potential energy of the particle in the gravitational
field:

$$dW = -dE_\mathrm{p} = -m \, d\varphi.$$

Consequently, the potential and intensity of a gravitational
field are related by the equation:

$$d\varphi = -(\mathbf{g}_f \, d\mathbf{r}) = - (g_{fx} \, dx + g_{fy} \, dy + g_{fz} \, dz),$$

where g_{fx}, g_{fy} and g_{fz} are the projections of vector $\mathbf{g}_f$ on the
axes of rectangular Cartesian coordinates. Since

$$d\varphi = \frac{\partial \varphi}{\partial x} \, dx + \frac{\partial \varphi}{\partial y} \, dy + \frac{\partial \varphi}{\partial z} \, dz,$$

then

$$\frac{\partial \varphi}{\partial x} = -g_{fx}, \quad \frac{\partial \varphi}{\partial y} = -g_{fy} \quad \text{and} \quad \frac{\partial \varphi}{\partial z} = -g_{fz}.$$

Hence

$$\mathbf{g}_f = - \left(\frac{\partial \varphi}{\partial x} \, \mathbf{i} + \frac{\partial \varphi}{\partial y} \, \mathbf{j} + \frac{\partial \varphi}{\partial z} \, \mathbf{k} \right) = - \mathrm{grad} \, \varphi,$$

i.e. the intensity of a gravitational field is numerically equal
and opposite in direction to the gradient of the potential of
the field.

The relation between φ and g_f can also be written in the form

$$d\varphi = -g_f \, dl \cos \alpha = -g_{fl} \, dl \quad \text{or} \quad g_{fl} = -\frac{d\varphi}{dl} \, ,$$

where α is the angle between vectors $\mathbf{g}_f$ and $d\mathbf{r}$, $dl = |\, d\mathbf{r}\,|$, and g_{fl} is the projection of vector $\mathbf{g}_f$ onto the direction of vector $d\mathbf{r}$. Thus, the projection of the intensity vector of a gravitational field onto some direction is numerically equal and opposite in sign to the change in the field potential per unit length in the same direction.

6.2.6 The nonrelativistic theory of gravitation discussed above and based on Newton's law of universal gravitation is only approximate. It describes with sufficient accuracy only comparatively weak gravitational fields, having a potential $|\,\varphi\,| \ll c^2$, where $c^2 = 9 \times 10^{16}$ m²/s² is the square of the velocity of light in free space. It is, in particular, applicable to the gravitational fields of the earth and sun because the absolute values of the potentials of these fields at the surfaces of the earth and sun are equal to 6.3×10^7 and 1.9×10^{11} m²/s².

6.2.7 The *modern* (relativistic) *theory of gravitation*, which is a unified theory of space, time and gravitation, was formulated by Albert Einstein. He called it the *general theory of relativity*. The existence of a close relationship between space and time had already been indicated in his special theory of relativity. This relationship is evident in the Lorentz transformations (5.3.2) and the invariance of the interval between two events (5.4.8). It was found necessary, in describing physical processes, to employ a four-dimensional space-time continuum, in which the position of a point is specified by three space coordinates and the time coordinate ct.

According to the relativistic theory of gravitation, the geometrical properties (metric) of the space-time continuum depend upon the distribution of gravitating masses and their motion in space. Bodies setting up a gravitational field "distort" real three-dimensional space and differently change the passage of time at various points in this space. This means that the bodies cause a deviation of the metric of space-time from that of the "flat" space-time described by Euclidean geometry and dealt with in the special theory of relativity. Therefore, the motion of a body in a gravitational field can be regarded as inertial motion, but in a "distorted" (non-Euclidean) space-

time continuum. Accordingly, a particle, subject to the action of a gravitational field, travels neither in a straight line nor at uniform velocity in a real three-dimensional space.

It was shown in the relativistic theory of gravitation that the principle of superposition (6.2.3) does not hold for arbitrary gravitational fields. This principle, as well as the whole non-relativistic theory of gravitation, is sufficiently accurate only in weak fields ($|\varphi| \ll c^2$) and for motion in these fields at low velocities $v \ll c$.

6.3 Kepler's Laws. Space Velocities

6.3.1 The motion of the planets of the solar system along their orbits around the sun complies with the three laws of Kepler. These laws can be derived from Newton's law of universal gravitation, dealing, as a first approximation, with the sun and planets as particles. In the central force field of gravitation of the sun, a planet of mass m is subject to the action of the gravitational force

$$\mathbf{F} = -G\,\frac{mM_S}{r^3}\,\mathbf{r},$$

where M_S is the mass of the sun and $\mathbf{r}$ is the radius vector of the planet from the centre of forces O, which is taken as the origin of coordinates.

The moment of force $\mathbf{F}$ about the centre of forces is $\mathbf{M} = [\mathbf{rF}] \equiv 0$. Hence, the angular momentum $\mathbf{L}$ of the planet about the same point O does not change with time (4.3.1):

$\mathbf{L} = [\mathbf{r}m\mathbf{v}] = \text{const}$.

Consequently, the planet travels along a plane path (orbit), whose plane is perpendicular to vector $\mathbf{L}$. According to Sect. 1.3.6, $\mathbf{L} = [\mathbf{r}m\mathbf{v}_\varphi]$, where $\mathbf{v}_\varphi$ is the transverse velocity of the planet. Therefore, the orbital motion of the planet satisfies the condition

$$r^2\,\frac{d\varphi}{dt} = \frac{L}{m} = \text{const},$$

where r and φ are the polar coordinates of the planet.

A second condition is imposed by the law of conservation of mechanical energy: $E_k + E_p = E = \text{const}$. According to 1.3.6 and 6.2.2

$$E_k = \frac{mv^2}{2} = \frac{m}{2}\left[\left(\frac{dr}{dt}\right)^2 + \left(r\,\frac{d\varphi}{dt}\right)^2\right]$$

$$= \frac{m}{2}\left[\left(\frac{dr}{dt}\right)^2 + \left(\frac{L}{mr}\right)^2\right]$$

and

$$E_p = -G\,\frac{mM_S}{r}\,,$$

so that the second condition is of the form

$$\left(\frac{dr}{dt}\right)^2 + \left(\frac{L}{mr}\right)^2 - \frac{2GM_S}{r} = \frac{2E}{m}\,.$$

6.3.2 The equation of the path of the planet (in the polar coordinates r and φ) is

$$r = \frac{p}{1 + e\cos\varphi}\,,$$

where $p = L^2/Gm^2M_S$ and $e = \sqrt{(2EL^2/G^2m^3M_S^2) + 1}$. The total mechanical energy of the planet $E < 0$, so that $e < 1$ and the path (orbit) of the planet is an ellipse.

Kepler's first law states that all the planets of the solar system travel along elliptic orbits of which the sun occupies one focus. It follows from the first condition (6.3.1) that the areal velocity (1.3.6) of the planet is constant:

$$\sigma = \frac{1}{2}\,r^2\,\frac{d\varphi}{dt} = \frac{L}{2m} = \text{const}$$

Kepler's second law states that the radius vector of a planet sweeps out equal areas in equal lengths of time.

6.3.3 According to Kepler's second law, the orbital period, or period of revolution, T of a planet about the sun is equal to

the ratio of the area A of the orbit to the areal velocity σ:

$$T = \frac{A}{\sigma} = \frac{\pi a b}{\sigma},$$

where $a = p/(1 - e^2)$ and $b = a\sqrt{1 - e^2}$ are the major and minor semiaxes of the elliptic orbit. Thus

$$T^2 = \frac{\pi^2 p}{L^2/4m^2}\, a^3 = \frac{4\pi^2}{GM_S}\, a^3.$$

This equation is an expression of *Kepler's third law*: the squares of the periods of revolution of the planets about the sun are proportional to the cubes of the major semiaxes of the elliptic orbits of these planets.

6.3.4 The *orbital velocity* is the minimum velocity that it is necessary to impart to a body to convert it into an artificial satellite of the earth. It is also called the *circular orbital velocity* because it is equal to the velocity of an artificial satellite in circular orbit about the earth in the absence of air friction of the atmosphere. The orbital velocity is

$$v_1 = \sqrt{\frac{GM_E}{r}},$$

where M_E is the mass of the earth, and r is the radius of the circular orbit. At the earth's surface $v_1 = 7.9$ km/s.

6.3.5 The *escape velocity* is the minimum velocity that it is necessary to impart to a body to enable it, without any additional forces, to overcome the earth's gravitational attraction and become an artificial satellite of the sun. It is also called the *parabolic velocity* because it launches the body on a parabolic orbit in the earth's gravitational field (in the absence of air friction of the atmosphere). The escape velocity

$$v_2 = \sqrt{\frac{2GM_E}{r}},$$

where r is the radius from the launching site of the body to the centre of the earth. At the earth's surface $v_2 = 11.2$ km/s.

6.3.6. The *solar escape velocity* is the minimum velocity that it is necessary to impart to a spacecraft, launched at the earth's surface, to overcome the sun's gravitational attraction and leave the solar system. This velocity is $v_3 = 16.7$ km/s.

CHAPTER 7   MOTION IN NONINERTIAL FRAMES OF REFERENCE

7.1 Kinematics of Relative Motion

7.1.1 It is assumed in classical (Newtonian) mechanics that distances and time intervals remain unchanged in going over from one frame of reference to any other frame that is in motion in even the most arbitrary manner with respect to the first frame. For example, assume that K is an inertial frame with its origin at point O^* and S is a *noninertial* frame with its origin at point O (Fig. 7.1). In the general case, the motion of frame S with respect to frame K can be regarded as the sum of two motions: translational motion of point O at the velocity $\mathbf{v}_0$ and rotation about this point at the
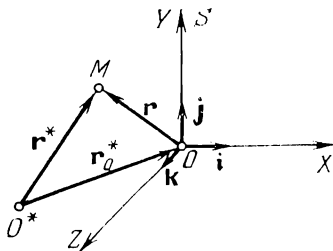
Fig. 7.1

angular velocity Ω. Radius vectors $\mathbf{r}^*$ and $\mathbf{r}$ of arbitrary particle M, measured in frames K and S, are related by the equation

$$\mathbf{r}^* = \mathbf{r}_0^* + \mathbf{r},$$

where $\mathbf{r}_0^*$ is the radius vector of point O measured in frame of reference K.

7.1.2 The motion of particle M with respect to any inertial frame of reference K, conditionally assumed to be fixed, is called the *absolute motion* of particle M. The motion of the

same particle with respect to the noninertial frame of reference S is said to be *relative motion.*

The *relative velocity* $\mathbf{v}_r$ of particle M, i.e. its velocity with respect to frame S, equals

$$\mathbf{v}_r = \frac{dx}{dt}\,\mathbf{i} + \frac{dy}{dt}\,\mathbf{j} + \frac{dz}{dt}\,\mathbf{k},$$

where x, y and z are the Cartesian coordinates of particle M, and $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{k}$ are the unit vectors of the coordinate axes in frame S.

The *absolute velocity* of particle M, i.e. its velocity $\mathbf{v}$ in the frame of reference K, equals

$$\mathbf{v} = \frac{d\mathbf{r}^*}{dt} = \mathbf{v}_0 + x\,\frac{d\mathbf{i}}{dt} + y\,\frac{d\mathbf{j}}{dt} + z\,\frac{d\mathbf{k}}{dt} + \mathbf{v}_r,$$

where $\mathbf{v}_0 = d\mathbf{r}_0^*/dt$ is the absolute velocity of point O. Since the unit vectors of moving frame S can change in frame of reference K only due to the rotation of frame S about point O at the angular velocity $\boldsymbol{\Omega}$,

$$\frac{d\mathbf{i}}{dt} = [\boldsymbol{\Omega}\mathbf{i}], \quad \frac{d\mathbf{j}}{dt} = [\boldsymbol{\Omega}\mathbf{j}] \quad \text{and} \quad \frac{d\mathbf{k}}{dt} = [\boldsymbol{\Omega}\mathbf{k}]$$

and

$$\mathbf{v} = \mathbf{v}_t + \mathbf{v}_r,$$

where $\mathbf{v}_t = \mathbf{v}_0 + [\boldsymbol{\Omega}\mathbf{r}]$ is the *velocity of transportation* of particle M. It is also called the *bulk*, or *drift*, *velocity* and is equal to the absolute velocity of the point of moving frame of reference S (i.e. rigidly fixed in this frame) in which particle M is located at the given instant of time.

7.1.3 The *relative acceleration* $\mathbf{a}_r$ of particle M (its acceleration with respect to frame S) equals

$$\mathbf{a}_r = \frac{d^2x}{dt^2}\,\mathbf{i} + \frac{d^2y}{dt^2}\,\mathbf{j} + \frac{d^2z}{dt^2}\,\mathbf{k}.$$

The *absolute acceleration* of particle M, i.e. its acceleration with respect to frame K, equals

$$\mathbf{a} = \frac{d\mathbf{v}}{dt} = \mathbf{a}_t + \mathbf{a}_C + \mathbf{a}_r.$$

Here

$$\mathbf{a}_t = \frac{d\mathbf{v}_0}{dt} + \left[\frac{d\boldsymbol{\Omega}}{dt}\ \mathbf{r}\right] + [\boldsymbol{\Omega}\,[\boldsymbol{\Omega}\mathbf{r}]]$$

is the *acceleration of transportation* of particle M. It is equal to the absolute acceleration of the point in moving frame S in which particle M is located at the given instant of time, and

$$\mathbf{a}_C = 2\,[\boldsymbol{\Omega}\mathbf{v}_r]$$

is the *Coriolis acceleration* of particle M. The Coriolis acceleration has its maximum value if the relative velocity $\mathbf{v}_r$ of the particle is perpendicular to the vector $\boldsymbol{\Omega}$ of angular velocity of the moving frame of reference. It is equal to zero if the angle between vectors $\mathbf{v}_r$ and $\boldsymbol{\Omega}$ equals zero or π, or if even one of these vectors equals zero.

7.2 Inertial Forces

7.2.1 Newton's laws of motion are not valid in noninertial frames of reference. In particular, a particle can change the state of its motion with respect to noninertial frame S without any action being exerted on the particle by other bodies. A ball, for instance, suspended by a string from the ceiling of a railway coach of a train travelling at uniform velocity in a straight line, deviates backward when the motion of the train is accelerated and frontward when the motion is decelerated, i.e. it begins to move with respect to a noninertial frame of reference attached to the coach. But no horizontal forces act on the ball when it deviates.

7.2.2 The fundamental law of dynamics for a particle in noninertial frames of reference can be derived from Newton's second law and the relation between the absolute and relative accelerations of a particle. It follows from Sect. 7.1.3 that the product of the mass m of a particle by its relative acceleration equals

$$m\mathbf{a}_r = m\mathbf{a} - m\mathbf{a}_t - m\mathbf{a}_C.$$

According to Newton's second law, formulated to apply to the absolute motion of a particle, i.e. its motion with respect to inertial frame K,

$$m\mathbf{a} = \mathbf{F},$$

where **F** is the vector sum of all forces acting on the particle. Hence, the *fundamental equation of dynamics of a particle's relative motion* is of the form:

$$m\mathbf{a}_r = \mathbf{F} - m\mathbf{a}_t - m\mathbf{a}_C.$$

This equation can be reduced to a form similar to the fundamental law of dynamics for the absolute motion of a particle:

$$m\mathbf{a}_r = \mathbf{F} + \mathbf{I}_t + \mathbf{I}_C.$$

The vector quantities $\mathbf{I}_t = -m\mathbf{a}_t$ and $\mathbf{I}_C = -m\mathbf{a}_C$ have the dimensionality of force and are called the *inertial force of transportation* and the *Coriolis inertial force*, respectively.

7.2.3 It follows from Sect. 7.1.3 that in the general case the inertial force of transportation is the sum of three terms:

$$\mathbf{I}_t = -m\,\frac{d\mathbf{v}_0}{dt} - m\left[\frac{d\mathbf{\Omega}}{dt}\,\mathbf{r}\right] - m\,[\mathbf{\Omega}\,[\mathbf{\Omega r}]].$$

The last term on the right-hand side of this equation

$$\mathbf{I}_{cf} = -m\,[\mathbf{\Omega}\,[\mathbf{\Omega r}]]$$

is called the *centrifugal force of inertia*, or simply the *centrifugal force*, because this vector is perpendicular to the instantaneous axis of rotation (to vector $\mathbf{\Omega}$) of the noninertial frame S and its direction is away from this axis. Numerically the centrifugal force equals

$$I_{cf} = m\Omega^2\rho,$$

where ρ is the distance from a particle of mass m to the instantaneous axis of rotation of the frame of reference.

The inertial force of transportation coincides with the centrifugal force when the noninertial frame of reference travels with translational motion at constant velocity ($\mathbf{v}_0 = $ const) and rotates at constant angular velocity ($\mathbf{\Omega} = $ const).

7.2.4 The Coriolis inertial force is

$$\mathbf{I}_C = 2m\,[\mathbf{v}_r\mathbf{\Omega}].$$

This force acts on the particle only when the noninertial frame of reference rotates and the particle moves with respect to this frame. Thus, for example, particles of water in rivers of the Northern Hemisphere that flow meridionally are subject

to the action of a Coriolis inertial force. This force is perpendicular to the velocity of flow of the river and leads to undermining the right bank in the direction of flow.

The Coriolis inertial force does no work in the relative motion of a particle because the force is perpendicular to the velocity of relative motion of the particle. Hence, the Coriolis inertial force is an example of the gyroscopic forces (3.1.7).

7.2.5 The action of inertial forces on a particle in a noninertial frame of reference are real and can be measured in this frame by, for example, a spring dynamometer. But, in contrast to ordinary forces of interaction between bodies, it is impossible to indicate for inertial forces the action of what specific bodies on the particle being considered they express. This feature of inertial forces is due to the fact that the very occurrence of the vector quantities I_t and I_C in the fundamental equation of the dynamics of relative motion results only from the noninertial nature of the frame of reference used to describe the relative motion of the particle. The addition of the inertial forces I_t and I_C to the force F, which characterizes the action on the particle of all the other bodies, enables the fundamental equation of the dynamics of relative motion to be written in a form that resembles the expression for Newton's second law in an inertial frame of reference.

There can be no closed systems of bodies in noninertial frames of reference because the inertial forces are always external forces for any of the bodies in the system. Therefore, the laws of conservation of momentum (2.7.1), of angular momentum (4.4.1) and of energy (3.4.2) are not valid in noninertial frames of reference.

7.3 Relative Motion in a Frame of Reference Fixed to the Earth. Gravity Force and Weight of a Body

7.3.1 A frame of reference fixed in the earth is noninertial for two reasons: firstly, due to the diurnal, or daily, rotation of the earth at constant angular velocity Ω (where $\Omega = 2\pi$ rad per 24 h $= 7.3 \times 10^{-5}$ rad/s) and, in the second place, due to the effects of the gravitational field of the sun, moon, planets and other astronomical bodies on the earth. This gravitational field is practically uniform within the limits of the earth and imparts the same acceleration of translational motion

to the terrestrial frame of reference and all bodies moving with respect to this frame. This acceleration is $\mathbf{a}_0 = d\mathbf{v}_0/dt = \mathbf{g}_f$, where $\mathbf{g}_f$ is the intensity of the field (6.2.1).

It follows from the equations in Sects. 7.2.2 and 7.2.3 that the equation of relative motion of a particle of mass m in a frame of reference fixed to the earth is of the form

Fig. 7.2

$$m\mathbf{a}_r = \mathbf{F} + \mathbf{F}_{gr} + \mathbf{I}_{cf} + \mathbf{I}_C,$$

where $\mathbf{I}_{cf}$ and $\mathbf{I}_C$ are the centrifugal and Coriolis inertial forces, $\mathbf{F}_{gr}$ is the gravitational attraction of the earth for the particle (6.1.3), and $\mathbf{F}$ is the sum of all the other forces acting on the particle, with the exception of the gravitational forces.

7.3.2 The *gravity force of a body* is the force $\mathbf{P}$, applied to the body and equal to the vector sum of the force $\mathbf{F}_{gr}$ of gravitational attraction of the earth for the body (6.1.3) and the centrifugal inertial force $\mathbf{I}_{cf}$ due to the diurnal rotation of the earth (Fig. 7.2):

$$\mathbf{P} = \mathbf{F}_{gr} + \mathbf{I}_{cf},$$

i.e.

$$\mathbf{P} = -G\,\frac{mM_E}{r^3}\,\mathbf{r} - m\,[\mathbf{\Omega}\,[\mathbf{\Omega r}]]\cdot$$

Here m and M_E are the masses of the body and the earth, $\mathbf{r}$ is the radius vector drawn from the centre of the earth to the place where the body is located, $\mathbf{\Omega}$ is the angular velocity of diurnal rotation of the earth, and G is the gravitational constant (6.1.1).

The gravity force causes an unsupported or unfastened body to fall to the earth. It equals the force with which the body presses against a horizontal support (or pulls on a vertical suspension) due to the gravitation attraction of the earth. It can be measured in the terrestrial frame of reference by
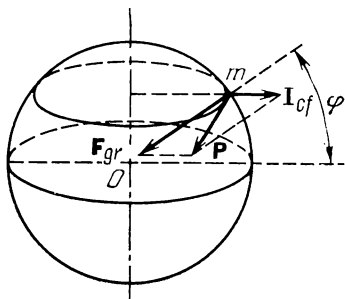
means of a spring dynamometer. The point of application
of the gravity force acting on a body, i.e. the point of appli-
cation of the resultant of all the gravity forces acting on all
the particles of the body is called the *centre of gravity of the
body*. It coincides with its centre of mass (2.3.3).
7.3.3 The gravity force of a body is independent of its velocity
of relative motion. It is proportional to the mass m of the body
and can be presented in the form

$$\mathbf{P} = m\mathbf{g},$$

where $\mathbf{g}$ is the *acceleration due to gravity*, or *free fall acceleration*
(see 7.3.5). At a given place on earth, vector $\mathbf{g}$ is the same for
all bodies and changes for other places.
The gravity force of a body coincides with the gravitational
attraction of the earth for the body only at the earth's North
and South poles because there the centrifugal inertial force
$\mathbf{I}_{cf} = 0$. The greatest difference between the gravity force and
the gravitational attraction is observed at the equator, where
the force $\mathbf{I}_{cf}$ reaches its maximum value and is opposite in
direction to $\mathbf{F}_{gr}$. Even at the equator, however, the gravity
force and the force of gravitational attraction differ by only
0.35 per cent. At all points of the earth's surface, except the
equator and the poles, forces $\mathbf{P}$ and $\mathbf{F}_{gr}$ fail to coincide in direc-
tion as well (Fig. 7.2), but the maximum angle between them
does not exceed 6 minutes of arc. The gravity force decreases
with the altitude. Near the earth's surface this decrease is
approximately 0.034 per cent per kilometre above the surface.
7.3.4 The acceleration g near the earth's surface varies from
9.78 m/s² at the equator to 9.83 m/s² at the North and South
poles. This is due, in the first place, to the dependence of the
centrifugal inertial force on the geographical latitude of the
location and, in the second, to the nonspherical shape of the
earth, which is slightly flattened at the poles of its axis of
rotation, having the shape of an ellipsoid of rotation, or sphe-
roid (the polar and equatorial radii of the earth are $R_p =$
$= 6357$ km and $R_e = 6378$ km). The *standard value of free
fall acceleration*, accepted in establishing a system of units
and for barometric computation, is 9.80665 m/s².
7.3.5 The *free fall of a body* is its motion due to the action of
the gravitational field alone.
The free fall acceleration of a body falling to the earth, mea-
sured in a noninertial frame of reference rotating together

with the earth, can be found by the motion equation (7.3.1)
by putting $F = 0$, $F_{gr} + I_{cf} = mg$ and $I_C = 2m\,[v_r\Omega]$. Then

$$a_r = g + 2\,[v_r\Omega].$$

If $v_r = 0$, $a_r = g$.

Consequently, vector g is equal to the free fall acceleration
of the body, measured with respect to the terrestrial frame of
reference at the instant when the relative velocity of the body
equals zero. For this reason, vector g is called the *free fall
acceleration*.

If the relative velocity of a freely falling body $v_r \neq 0$, its
acceleration with respect to the earth is not g: $g = a_r + 2\,[\Omega v_r]$.
At velocities $v_r < 680$ m/s, however, the values of g and a_r
differ by less than one per cent. It can therefore be assumed
in many cases that with respect to an observer on the earth
free fall of a body is caused only by the gravity force acting
on and imparting the acceleration g to the body. In a similar
manner the action of the Coriolis inertial force on a freely falling
body can be treated as a comparatively small perturbation.
Thus, for instance, a freely falling body is deviated by the
Coriolis force to the east of the direction assumed by a plumb
line, i.e. from the direction of the vector $P = mg$. This devia-
tion for a body falling freely without any initial velocity from
the height h is equal, at the latitude φ, to

$$s = \frac{2}{3}\,\Omega h \sqrt{\frac{2h}{g}}\,\cos\varphi.$$

If, for example, $h = 160$ m and $\varphi = 45°$, $s = 1.55$ cm.

7.3.6 The *weight of a body* is the force Q with which it acts,
due to the earth's gravity, on a support or suspension that
prevents it from falling freely. It is assumed that the body
and its support (or suspension) are fixed with respect to the
frame of reference in which the weight is determined. The
force $-Q$ is exerted by the support or suspension on the body.
It follows from the fundamental equation of dynamics of rela-
tive motion (7.2.2), where $a_r = a_C = 0$ and $F = F_{gr} - Q$,
that

$$Q = F_{gr} + I_t,$$

where $\mathbf{F}_{gr}$ is the force of gravitational attraction of the earth
for the body, and $\mathbf{I}_t$ is the inertial force of transportation due
to the noninertial nature of the frame of reference.

Example 1. The weight of a body in a frame of reference fixed
to the earth is equal to the gravity force of the body (7.3.2):

$$Q = \mathbf{F}_{gr} + \mathbf{I}_{cf} = \mathbf{P}.$$

Example 2. The weight of a body in a frame of reference at-
tached to a lift (elevator) which travels with translational motion
with respect to the earth with the acceleration $\mathbf{a}_0$:

$$Q = \mathbf{P} - m\mathbf{a}_0.$$

If the lift falls freely, $\mathbf{a}_0 = \mathbf{g}$, the free fall acceleration, and
the weight of the body in the lift is $Q = 0$, i.e. the body is
in the state of weightlessness.

7.3.7 *Weightlessness*, or zero gravity, is the state of a mechan-
ical system in which the external gravitational field acting
on the system does not cause any pressure of the parts against
one another or their deformation. Such a state occurs in a me-
chanical system complying with the following three conditions:
(a) no other external forces, except those of the gravitational
field, act on the system; (b) the size of the system is such that
within its limits the external gravitational field can be regarded
as being uniform; and (c) the system is in translational motion.
The state of weightlessness is characteristic, for instance, of
bodies in a spaceship because, during the main part of its
path in a gravitational field, it is in free flight, i.e. coasting
with the engines turned off.

7.4 Principle of Equivalence

7.4.1 The inertial forces acting on bodies in a noninertial
frame of reference are proportional to the masses of the bodies
and, other conditions being equal, impart the same relative
accelerations (7.1.3) to them. In other words, all bodies free
of external action travel in an "inertial force field", i.e. with
respect to a noninertial frame of reference, in absolutely the
same way, provided that the initial conditions of their motion
are also the same. A similar behaviour is observed for bodies
travelling with respect to an inertial frame of reference and

subject to the action of the forces of a gravitational field. At each point in the field these forces, like the inertial forces, are proportional to the masses of the bodies and impart the same free fall acceleration to them. This acceleration is equal to the field intensity (6.2.1) at the point being considered.

In a noninertial frame of reference, for example, fixed to a lift (elevator) travelling with uniform acceleration vertically upward at the acceleration of transportation $a_0 = \text{const}$, all free bodies fall in the absence of a gravitational field with the same relative acceleration $a_r = -a_0$. Free bodies behave in exactly the same way in the same lift when it travels at uniform velocity in a uniform gravitational field of the intensity $g_f = -a_0$. Thus, on the basis of experiments conducted on the free fall of bodies within a tightly closed lift, it is impossible to establish whether the lift is travelling at uniform velocity in a gravitational field of the intensity $g_f = a_r$ (in particular, the lift may also be at rest in this field) or whether it is travelling with constant acceleration of transportation $a_t = -a_r$ in the absence of a gravitational field.

7.4.2 The *local equivalence principle* states that the gravitational field in a limited region of space is equivalent to an "inertial force field" in a properly chosen noninertial frame of reference. The region of space should be small enough so that the gravitational field in it can be assumed to be uniform.

The equivalence principle is not to be understood as a statement of the identity of inertial forces and Newton's gravitational forces that act between bodies. As a matter of fact, the intensity of a true gravitational field, set up by bodies, decreases as the bodies recede farther and farther away, and approaches zero at infinity. Gravitational fields "equivalent" to inertial forces do not comply with this condition. For example, the intensity of a gravitational field, "equivalent" to the centrifugal inertial forces in a rotating frame of reference increases without limit at points farther and farther away from the axis of rotation. The intensity of a field "equivalent" to the inertial forces of transportation in a frame of reference having translational motion is the same at all points of the field.

7.4.3 A true gravitational field, in contrast to one "equivalent" to inertial forces, exists in both noninertial and inertial frames of reference. No possible choice of a noninertial frame of reference can completely eliminate a true gravitational field, i.e. compensate for it throughout space by means of a "field

of inertial forces". This follows in any case, even if only from the different behaviour of "fields of inertial forces" and true gravitational fields at infinity. A true gravitational field can be eliminated only locally, i.e. for a small region of space within whose limits the field can be assumed uniform, and for a short time interval during which the field can be assumed constant. The noninertial frame of reference corresponding to this operation should travel with an acceleration of transportation equal to the free fall acceleration of bodies in the region of the true gravitational field being considered. Thus, in a space ship in free flight in a gravitational field, the gravitational forces of attraction are compensated for by the inertial forces of transportation and do not lead to the relative motion of bodies in the ship.

PART TWO FUNDAMENTALS OF MOLECULAR PHYSICS AND THERMODYNAMICS

CHAPTER 8 IDEAL GASES

8.1 Subject Matter of Molecular Physics. Thermal Motion

8.1.1 The branch of physics called *molecular physics* deals with the dependence of the structure and physical properties of bodies on the nature of the motion of and the interaction between the particles making up these bodies.

Molecular physics is based on the *molecular-kinetic theory* of the structure of matter. According to this theory all bodies consist of submicroscopic particles—atoms, molecules or ions—in continuous chaotic motion, which is called *thermal motion*. Experimental confirmations of the molecular-kinetic theory are: the Brownian movement (11.7.1), the transport phenomenon in various aggregate states of matter and other phenomena.

8.1.2 The molecular-kinetic theory of the structure of matter is used to advantage in various branches of physics. A great variety of physical phenomena that proceed due to the interaction and motion of particles of matter are dealt with in this theory from a unified point of view. This theory, for example, enables the mechanism of elasticity in solids (40.3.1) to be understood, reveals the cause of internal friction in gases (10.8.4) and liquids, explains the difference between real gases (12.1.2) and ideal gases (8.4.1), and elucidates the mechanism of electrical conduction in conductors of electric current of various kinds, and the electrical and magnetic properties of matter.

8.1.3 The thermal motion of particles of matter in various states of aggregation is not the same. It depends upon the forces of attraction or repulsion between the atoms, molecules and ions.

There are practically no attractive forces between the atoms and molecules in sufficiently dilute gases. This is due to the fact that the particles of such gases are located at distances

from one another exceeding the range of molecular action (12.2.3). The particles of such gases travel at uniform velocity and in a straight line until they collide with one another or with the walls of the container. These collisions are of a random nature. Each molecule in the volume of the gas may collide with any of the nearby particles, changing the direction of its motion in any arbitrary manner. The collision of a molecule (or atom) of the gas with a wall may occur at any angle θ (where $-\dfrac{\pi}{2} \leqslant \theta \leqslant \dfrac{\pi}{2}$). As a result, the thermal motion of molecules of gas is a purely random one and, on the average, the same number of molecules are travelling at any instant of time in any arbitrary direction within the gas.

8.1.4 Solid crystalline bodies are characterized by considerable forces of interaction between their particles (atoms, molecules and ions). The joint action of the forces of attraction and repulsion between the particles (12.1.3) leads to the vibration of the particles of solids about their equilibrium positions, which are called the *sites*, or *points*, *of the crystal lattice* (40.1.1). Intermolecular interaction and violations of periodicity in crystals (40.1.1) make these vibrations anharmonic (28.1.3).

8.1.5 The thermal motion of molecules of a liquid is of a kind intermediate between the two preceding types of motion (8.1.3 and 8.1.4). For a definite length of time the molecules of a liquid vibrate about a certain equilibrium position and are in a temporarily settled state. At the end of this time the equilibrium position of the molecules is displaced and a new settled arrangement is formed. Simultaneously, a slow displacement of the molecules and their vibrations occur within small volumes of the liquid (see also Sect. 13.2.3).

8.2 Statistical and Thermodynamic Methods of Investigation

8.2.1 There is a huge number of atoms (or molecules) in any body. For instance, in 1 cm³ of a gas whose properties are close to those of an ideal gas there are 2.7×10^{19} molecules at standard conditions. In the condensed states, liquid and solid, the amount is of the order of 10^{22} particles/cm³. If we assume

that each atom (or molecule) obeys Newton's second law (2.4.1) it is obviously quite impossible, not only to solve the differential equations of motion of the separate particles, but even to write these equations. Therefore, the behaviour of a separate molecule (or atom) of a body, for example, its path and the sequence of changes in its state, cannot be investigated by the methods used in classical mechanics.

8.2.2 The macroscopic properties of systems consisting of a very great number of particles are studied by the *statistical method*. It is based on the application of probability theory to definite models of the structure of the systems being investigated. The branch of theoretical physics in which the physical properties of systems are studied by the statistical method is called *statistical physics (statistical mechanics)*. Special laws, known as *statistical laws* are manifested in the joint behaviour of large numbers of particles. Certain *average values of the physical quantities* can be found for a system consisting of a great many particles. These values characterize the set of particles as a whole. In gases, for instance, there are average values of the velocity of thermal motion of the molecules (10.3.6) and of their energy (10.2.4). In a solid there is an average energy per degree of freedom of the vibrational motion of its particles (41.7.2), etc. All properties of a system of particles are due, not only to the individual properties of the particles themselves, but to the features of their joint motion and average values of their dynamic characteristics (average velocities, average energies, etc.).

In addition to statistical laws, there are also *dynamical laws*, describing the motion of separate particles. The relationship between dynamical and statistical laws is manifested in the fact that the laws of motion of individual particles have an effect of the description of a system of particles, which is being investigated by the statistical method.

8.2.3 Another procedure for studying physical phenomena is the *thermodynamic method*, which is not concerned with either the internal structure of the substances making up the bodies (or systems) being investigated or with the nature of the motion of the separate particles. The thermodynamic method is based on the study of the various conversions of energy that occur in a system. The conditions of these conversions and the relations between the various kinds of energy enable the physical properties of systems to be studied in

a great variety of processes in which the systems participate. The branch of physics in which research in the physical properties of systems is conducted by means of the thermodynamic method is called *thermodynamics (phenomenological thermodynamics)*. Thermodynamics is based on two experimentally established principles, or laws, of thermodynamics (9.3.1 and 11.3.2), as well as the Nernst heat theorem, or third law of thermodynamics (11.8.2).

8.3 Thermodynamic Variables. Equations of State. Thermodynamic Processes

8.3.1 Thermodynamics deals with *thermodynamic systems*. There are macroscopic items (bodies and fields) that can interchange energy with one another as well as with the external medium, i.e. bodies and fields that are external with respect to the given system.

8.3.2 The physical quantities introduced to describe the state of a thermodynamic system are called *thermodynamic variables*, or *parameters (state variables)*, of the system. Usually the pressure, specific volume and temperature are taken as the thermodynamic variables.

The *pressure p* is a physical quantity equal to the force acting on unit area of a body's surface in the direction normal to the surface:

$$p = \frac{dF_\mathrm{n}}{dA},$$

where dF_n is the numerical value of the normal force acting on an infinitely small part of the surface of the body, the area of this part being dA.

The *specific volume v* is the reciprocal of the density ρ of a body: $v = 1/\rho$. The specific volume of a homogeneous body is equal to the volume of unit mass of the body.

8.3.3 The concept of temperature is of significance for equilibrium states of a thermodynamic system (8.3.4). An *equilibrium state (state of thermodynamic equilibrium)* is a state of a system which does not change with time *(steady state)*, the steady nature of the state not being due to processes occurring in the external medium. An equilibrium state is set up in

a system subject to constant external conditions and is main-
tained in the system for an arbitrary length of time. The tem-
perature is the same in all parts of a thermodynamic system
that is in a state of thermodynamic equilibrium. If two bodies
with different temperatures come into contact, internal ener-
gy (9.1.2) is transferred from the body at the higher temperature
to that at the lower temperature. The process ceases when the
temperatures of the two bodies become equal.

8.3.4 The temperature of an equilibrium system is a measure
of the intensity of thermal motion of its molecules (or atoms
or ions). In an equilibrium system of particles that obeys the
laws of classical statistical mechanics (8.2.2), the average
kinetic energy of the thermal motion of the particles is pro-
portional to the absolute temperature (10.2.5).

Temperature can only be measured indirectly, making use
of the fact that many physical properties of bodies, lending
themselves to direct measurement, depend on the temperature.
A change in the temperature of a body leads to changes in its
length and volume, density, electric resistance, elastic prop-
erties, etc. A change in any of these properties can be used
to measure temperature. For this purpose, it is necessary for
a certain body, said to be *temperature-indicating*, that the
functional dependence of the given property on the temperature
be known. Temperature scales based on temperature-indicating
bodies are said to be *empirical*.

In the *international centigrade (Celsius) scale*, temperature is
measured in °C and is denoted by t (degree centigrade or degree
Celsius). It is assumed that at a standard pressure of 1.01325×10^5 Pa (Appendix I) the melting point of ice and the boiling
point of water are equal to 0°C and 100°C.

In the *thermodynamic (absolute) temperature scale*, the tem-
perature is measured in kelvins (K) and is denoted by T.

The relation between the absolute temperature and the tem-
perature measured by the Celsius scale is: $T = 273.15 + t$.

The temperature $T = 0$ (or $t = -273.15$°C) is called *absolute
zero temperature* (the impossibility of reaching absolute zero
is discussed in Sect. 11.8.4).

8.3.5 The state variables of a system are classified as being
either external or internal. *External state variables* are physical
quantities depending upon the position in space and various
properties of bodies that are external with respect to the given
system. For a gas, for instance, the volume V of the vessel in

which the gas is contained is an external variable because the volume depends upon the position of external bodies: the walls of the vessel. For a dielectric in an electric field, the intensity of the field, set up by certain external sources, is an external variable. If a liquid is held in an open vessel, the atmospheric pressure, for example, is an external variable.

The *internal variables of a system* are physical quantities that depend both on the position of bodies external with respect to the system, and on the coordinates and velocities of the particles that make up the system. Internal variables of a gas, for instance, are its pressure and energy because they depend upon the coordinates and velocities of the moving molecules and on the density of the gas.

8.3.6 The state variables of a system that is in equilibrium (8.3.3) are not independent. The internal variables of such a system depend only on its external variables and on the temperature. The equilibrium state of a simple system* of given chemical composition and mass M is determined by specifying two variables: the volume V and the temperature T. The *equation of state (thermal equation of state)* of a simple system is the functional dependence of the equilibrium pressure p in the system on the volume and temperature

$$p = f(V, \ T).$$

The equation of state is obtained experimentally in thermodynamics. In statistical physics (8.2.2), it is derived theoretically. This constitutes the interrelation between the statistical (8.2.2) and thermodynamical (8.2.3) methods of investigation.

8.3.7 If any of the external variables of the system is changed, the state of a thermodynamic system is changed. This is called a *thermodynamic process*. A thermodynamic process is said to be equilibrium (an *equilibrium, quasi-static process*) if the system passes at an infinitely slow rate through a continuous succession of infinitely close thermodynamic equilibrium states (8.3.3). All processes that do not comply with these

---

* An example of a simple system is a gas in the absence of external fields. Mixtures of chemically homogeneous gases of constant composition, chemically pure liquids, etc. are also simple systems.

conditions are said to be *nonequilibrium* ones. Real processes
are nonequilibrium ones because they proceed at a finite rate.
But the slower they proceed, the closer they are to equilibrium
processes.
Isoprocesses are thermodynamic processes that proceed in
a system of constant mass with a constant value of one of the
state variables.
An *isothermal process* proceeds at constant temperature ($T =$
$=$ const).
An *isochoric (isovolumic) process* proceeds at constant volume
($V =$ const).
An *isobaric process* proceeds at constant pressure ($p =$ const).
An *adiabatic process* is a thermodynamic process that occurs
in a system without heat exchange (9.2.4) with external bodies
(see also Sect. 9.5.10).
8.3.8 *State functions* are physical quantities characterizing the
state of a system. Variation in the state functions during ther-
modynamic processes is independent of the kind of process
that occurs. State functions are uniquely determined by the
values of the variables of the initial and final states of the
system. The simplest state functions of a system are its inter-
nal energy U (9.1.2) and its entropy S (11.4.2).

8.4 Equation of State of an Ideal Gas

8.4.1 An *ideal*, or *perfect*, *gas* is one whose molecules do not
interact with one another at a distance and are of vanishingly
small size. In real gases (12.1.2) the molecules are subject to
the forces of intermolecular interaction (12.1.3).
In collisions with one another and in striking the walls of
the vessel the molecules of an ideal gas behave as perfectly
elastic spheres (1.1.4) of diameter d *(effective diameter of the
molecule)*, which depends upon the chemical nature of the gas.
The existence of an effective diameter d (where $d \approx 10^{-1}$ m)
signifies that forces of mutual repulsion (12.1.3) act between
the molecules. Intermolecular forces of attraction (12.1.3)
decrease drastically with an increase in the distance r between
molecules and are practically negligible at $R_m > 10^{-9}$ m.
In real gases at low densities the average distance $\langle r \rangle$ between
molecules exceeds R_m and, to a good approximation, they can

be regarded as ideal. Hydrogen, helium, oxygen and nitrogen are assumed to be ideal gases at densities corresponding to standard conditions.

8.4.2 For a given mass of an ideal gas, the ratio of the product of the numerical values of pressure and volume to the absolute temperature is a constant value (*Clapeyron equation*):

$$\frac{pV}{T} = C = \text{const.}$$

The numerical value of the *gas constant* C depends upon the choice of the units of measurement of p, V and T.

If v is the specific volume of the gas (8.3.2) and M is its mass, then $V = Mv$ and the Clapeyron equation assumes the form

$$pv = \frac{C}{M} T = BT,$$

where $B = C/M$ is the *specific gas constant* referred to unit mass.

8.4.3 From the definition of the mole (Appendix I), it is evident that moles of different gases contain the same number of molecules. This is called *Avogadro's number* N_A (Appendix II).

The *molar mass* μ of a gas or any body is the physical quantity equal to the ratio of the mass of the gas (or any body) to the number of molès N it contains: $\mu = M/N$.

The molar mass is proportional to the relative mass of the molecules of a gas: $\mu = 10^{-3} m/m_0$, where m is the mass of the molecules of the given gas and m_0 is the atomic mass unit based on the carbon scale (Appendix II). The coefficient 10^{-3} is required because molar mass is measured in kg/mol in the SI (International System of Units).

The *molar volume* V_μ is the physical quantity equal to the ratio of the volume V of a gas to the number of moles N it contains: $V_\mu = V/N$. The mass of a mole is numerically equal to μ and therefore $V_\mu = \mu v$, where v is the specific volume (8.3.2).

8.4.4 The equation of state for a mole of ideal gas is

$$\frac{pV_\mu}{T} = \mu B = R \quad \text{or} \quad pV_\mu = RT,$$

Here R is the *universal gas constant*, being the gas constan-
per mole of gas. The universal nature of R follows from *Avot
gadro's law*, which states that moles of all ideal gases occupy
the same volume at the same pressure and the same tempera-
ture. Under standard conditions ($T = 273.15$ K and $p =$
$= 1.0132 \times 10^5$ Pa $= 1$ atm $= 760$ mmHg) a mole of any
gas has a volume of $V_\mu = 22.415 \times 10^{-3}$ m^3. From this we can
calculate the numerical values of R for various systems of
units (Appendix II).
If a volume V of gas has the mass M kg, i.e. if it contains M/μ
moles, then $V = M/\mu \times V_\mu$ and the equation of state of the
gas assumes the form called the *Mendeleev-Clapeyron equation*:

$$pV = \frac{M}{\mu} RT.$$

8.4.5 *Boltzmann's constant k* is a physical quantity equal
to the ratio of the universal gas constant R to Avogadro's
number N_A, i.e. $k = R/N_A$. The values of k in various systems
of units are listed in Appendix II.
The equation of state of an ideal gas, expressed by means of
Boltzmann's constant, is of the form

$$p = \frac{kN_A}{V_\mu} = kn_0 T,$$

where $n_0 = N_A/V_\mu$ is the number of gas molecules in unit
volume (molecule concentration).
At constant temperature, the pressure of a gas is proportional
to the concentration of its molecules (or to the density of the
gas).

CHAPTER 9   FIRST LAW OF THERMODYNAMICS

9.1 Total and Internal Energy of a System

9.1.1 An arbitrary thermodynamic system (8.3.1), in any
thermodynamic state (8.3.3), has a *total energy E*, which is
the sum of:
(a) the kinetic energy E_k^{mech} of mechanical motion of the system
as a whole (or of its macroscopic parts);

(b) the potential energy E_p^{ext} (3.3.1) of the system in all external force fields (for instance, electromagnetic or gravitational fields); and
(c) the internal energy U:

$$E = E_k^{mech} + E_p^{ext} + U.$$

9.1.2 The *internal energy* of a body or of a thermodynamic system (8.3.1) is the energy depending only upon the thermodynamic state of the body (or system). For a fixed (stationary) system, not located in any external force fields, the internal egergy coincides with the total energy. The internal energy also coincides with the rest energy of a body (or system) (5.7.3) and includes the energy of all the types of internal motion of a body (or system) and the energy of interaction of all the particles (atoms, molecules, ions, etc.) making up the body (or system).
For example, the internal energy of a gas consisting of polyatomic molecules (ammonia, carbon dioxide, etc.) consists of:
(a) kinetic energy of thermal translational and rotary motion of the molecules;
(b) kinetic and potential energy of vibration of the atoms and molecules;
(c) potential energy due to intermolecular interactions;
(d) energy of the electron shells of the atoms and ions;
(e) kinetic energy and potential energy of interaction of the nucleons (42.1.1) in the nuclei of the atoms.
Terms (d) and (e) do not ordinarily change in processes that proceed at not very high temperatures so that ionization and excitation play no essential role. Under these conditions, terms (d) and (e) are not taken into account in the internal energy balance. Term (c) is also neglected for an ideal gas (8.4.1).
9.1.3 The internal energy is a single-valued function of the thermodynamic state of a system. The value of the internal energy in any state is independent of the process by means of which the system reached this state. The change in internal energy in the transition of a system from state 1 to state 2 equals $\Delta U = U_2 - U_1$ and is independent of the kind of transition $1 \to 2$ that occurs. If the system accomplishes a cycle (11.1.1)

the total change in its internal energy is equal to zero: $\oint dU = $ $= 0$.

As is well known, such a relation mathematically signifies that an element of change in the internal energy dU is a complete (exact) differential. This property is also possessed by the entropy (11.4.2) and other state functions (8.3.8) (cf. Sect. 9.2.5).
9.1.4 In a system that is in a state of thermodynamic equilibrium the internal energy depends only upon the temperature and the external variables (8.3.5). In particular, for a simple system (8.3.6) of constant mass M, the internal energy is a function of the temperature T and volume V of the system (*calorific equation of state* of a simple system):

$$U = \varphi\ (V,\ T).$$

Example 1. The internal energy of an ideal gas (8.4.1) depends only on its absolute temperature and is proportional to the mass M of the gas:

$$U = \int\limits_0^T C_V\,dT + U_0 = M\ \left(\int\limits_0^T c_V\,dT + u_0 \right)\ ,$$

where C_V and $c_V = C_V/M$ are the heat capacity (9.5.1) and specific heat (9.5.2) of the gas at constant volume (in an isochoric process) (8.3.7), $u_0 = U_0/M$ is the internal energy of unit mass of the gas at $T = 0$ K. For monatomic gases at ordinary temperatures C_V is independent of T, and $U = C_V T + U_0$.
Example 2. The internal energy of a van der Waals gas (12.2.1) is:

$$U = \int\limits_0^T C_V\,dT - \frac{M^2}{\mu^2}\ \frac{a}{V} + U_0,$$

where M is the mass of the gas, μ is its molar mass (8.4.3), and a is the van der Waals' coefficient (12.2.3).
9.1.5 In thermodynamics, the internal energy is determined to an accuracy of the constant term U_0, whose value depends upon the choice of the reference point for the quantity U,

i.e. from a state with zero internal energy. The quantity U_0 plays practically no role in thermodynamical calculations in which the changes ΔU of internal energy, independent of U_0, are determined (see also Sect. 8.1.2).

9.2 Heat and Work

9.2.1 An energy exchange between a thermodynamic system and external bodies takes place in two ways: either by doing work or by means of heat exchange.* The amount of energy transferred to the system by external bodies in force inter-action between the system and the bodies is called *work* done on the system. The amount of energy transferred to the system by external bodies by heat transfer is called the *quantity of heat* imparted to the system.**

9.2.2 If the thermodynamic system is fixed (stationary), a necessary condition to do work is displacement of the external bodies interacting with it, i.e. it is necessary to change the external state variables of the system (8.3.5). In the absence of external force fields, energy exchange between the fixed system and the external medium may take place by doing work when the volume and shape of the system are changed. According to the law of conservation of energy, the work W' done on the system by external forces is numerically equal and opposite in sign to the work W that the system does on the external medium, i.e. against the external forces:

$$W' = -W.$$

9.2.3 *Expansion work* is the work done by the system in over-coming the external pressure. An element of expansion work $\delta W = p_{ext} \, dV$, where p_{ext} is the uniformly distributed exter-nal pressure, and dV is an element of change of the system's volume. If the expansion process is an equilibrium (quasi-

---

* A discussion of the third way of energy exchange—mass trans-fer—is beyond the scope of the present handbook.
** Sometimes the term "heat" is employed in place of the term "quantity of heat" if this leads to no confusion. It should be indicated, however, that heat, generally speaking, is to be under-stood as a form of energy transfer (9.2.4).

static) one (8.3.7), $p_{\text{ext}} = p$, where p is the pressure in the system. Then $\delta W = p \, dV$.

The significance of the different ways used in denoting the elements of change δW and dU is explained in Sects. 9.4.3 and 9.1.3. The work done in equilibrium expansion of system from volume V_1 to volume V_2 is

$$W = \int_{V_1}^{V_2} p \, dV.$$

Example. The work done by a gas in a vessel with a weightless movable piston (Fig. 9.1). The pressure of the gas $p > 0$ and, therefore, in its expansion $(dV > 0)$ the gas performs positive work $(\delta W > 0)$. In compressing the gas $(dV < 0$ and $\delta W < 0)$ the positive work is done on the gas by the forces of external pressure. Here the gas does negative work.
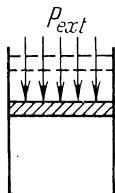
The graphical representation of work is discussed in 9.4.2.

9.2.4 Heat exchange takes place between bodies (or parts of bodies) heated to different temperatures. There are three types of heat exchange: convection, conduction and radiation (radiative heat exchange).

Fig. 9.1   *Convective heat exchange* is the transfer of energy in the form of heat between the unequally heated parts of gases or liquids or between gases, liquids and solids. Convective heat exchange in liquids takes place by motion of parts of the liquid with respect to one another or with respect to solids. For instance, in a radiator of a hot-water heating system, energy of the hot water flowing through the radiator is transferred by convection to the less heated walls of the radiator.

The phenomenon of *heat conduction* consists in the transfer of energy in the form of heat from one part of a nonuniformly heated body to another part. Thus, for example, energy is transferred through the walls of a heating radiator from the more heated inner surfaces to the less heated outer surfaces.

Radiative heat exchange takes place without direct contact of the bodies exchanging energy and consists in the emission and

absorption of the energy of an electromagnetic field by the bodies. An immense amount of energy is delivered from the sun to the earth's surface by radiative heat exchange.

9.2.5 Work and heat are the energy characteristics of processes for changing the state of thermodynamic systems and are meaningful only when such processes occur. Depending upon the kind of process the system undergoes in its transition from state 1 to state 2, it is necessary to perform different amounts of work and to deliver different quantities of heat to the system. A comparison with Sect. 8.1.3 (on the change in the internal energy in processes) indicates that work and heat are not forms of energy and it is incorrect to speak of a "stockpile of work" or a "reserve of heat" in a body. For the same reasons, an element of the quantity of heat δQ and an element of work δW are not complete differentials (cf. Sect. 9.1.3).

9.2.6 The performance of work on a system can change any type of its energy. For example, when a gas is rapidly compressed in a vessel by means of a moving piston (9.2.3), the work done by external forces on the gas increases the internal energy of the gas. In the inelastic collision of two bodies (3.5.3), a part of the work is expended to change the kinetic energy of the bodies (3.2.2) and a part to change their internal energy. If energy is delivered to a system in the form of heat, it is used only to increase the internal energy of the system. This is due to the fact that in any kind of heat exchange (9.2.4), energy is transferred directly between chaotically moving particles of the bodies. This changes their internal energy. In the process of heat conduction, for instance, in a nonuniformly heated rigid body the particles of the body located in its more heated parts transmit some of their energy to particles located in its less heated parts. The result is an equalization of the internal energies of the various parts of the body, an equalization of their temperatures and the discontinuation of the heat conduction process.

It follows from the aforesaid that there is a qualitative difference and lack of equivalence between work and heat as a form of energy transmission. Frequently, these two forms of transmitting energy exist simultaneously. For example, in heating a gas in a vessel with a moving piston, the volume of the gas is increased and work is done against the external pressure simultaneously.

9.2.7 A thermodynamical system is said to be *closed* (or *isolated*)

if there is no exchange of energy in any form between it and the external medium. Such systems comply with the *law of conservation of energy*: the total energy of the system remains unchanged regardless of the processes that occur within it.

9.2.8 A thermodynamical system is said to be *adiabatically isolated* (i.e. *thermally isolated*) if no heat exchange takes place between it and the external medium. Such a system can perform work on external bodies. Moreover, external forces can do work on the system. An example of such a system is a cylinder with a moving piston, filled with gas and surrounded all over by a dense layer of heatproof felt. The absence of heat exchange with the external medium does not exclude the possibility of the gas doing expansion work (9.2.3), or the external pressure doing compression work on the gas.

The properties of a system approximate those of an adiabatically isolated one if the change in its state is so rapid that there is no time, during the process, for heat exchange with the external medium. (For example, rapid expansion of a compressed gas released from a gas cylinder when the outlet valve is opened for a short time.)

9.3 First Law of Thermodynamics

9.3.1 The *first law of thermodynamics* states that the change in the internal energy ΔU_{1-2} of a system that occurs in the process $1 \to 2$ of transition of the system from state 1 to state 2 is equal to the sum of the work W'_{1-2} done by external forces on the system and the quantity of heat Q_{1-2} transmitted to the system:

$$\Delta U_{1-2} = W'_{1-2} + Q_{1-2},$$

where $W'_{1-2} = -W_{1-2}$, and W_{1-2} is the work done by the system to overcome external forces in the process $1 \to 2$. Hence

$$Q_{1-2} = \Delta U_{1-2} + W_{1-2}.$$

The quantity of heat transferred to a system is expended in changing the internal energy of the system and in the work done by the system to overcome the external forces.

For an element of the quantity of heat δQ, an element of work δW and an infinitely small change dU of the internal energy, the first law of thermodynamics assumes the form

$\delta Q = dU + \delta W$.

9.3.2 If $\delta Q > 0$, then heat is being transferred to the system; if $\delta Q < 0$, then heat is being removed from the system. In the finite process $1 \rightarrow 2$, elements of the quantity of heat can have both signs, and the total quantity of heat Q_{1-2} in the process $1 \rightarrow 2$ is equal to the algebraic sum of the quantities of heat transferred during all the parts of this process:

$$Q_{1-2} = \int_{1}^{2} \delta Q.$$

It is customary to consider $\delta W > 0$ when the system performs work on external bodies, and $\delta W < 0$ when work is performed by external bodies on the system. The work W_{1-2} done by the system in the finite process $1 \rightarrow 2$ is equal to the algebraic sum of the amounts of work δW performed by the system during all the parts of this process:

$$W_{1-2} = \int_{1}^{2} \delta W.$$

9.3.3 If the system for instance, the working medium (11.1.1) in a periodic-action engine completes a cycle $1 \rightarrow 1$ (11.1.1), then $\Delta U_{1-1} = 0$ and $W_{1-1} = Q_{1-1}$. It is impossible to build an operating engine that could perform more work than the energy transferred to the system from outside. Such an engine is called a *perpetual motion machine of the first kind*. The impossibility of constructing a perpetual motion engine of the first kind is also a statement of the first law of thermodynamics.

9.4 Graphical Representation of Thermodynamic Processes and Work

9.4.1 The equation of state (8.3.6) of a thermodynamic system enables the value of any state variable, for example p, to be determined from any values of two other variables, such as V

and T. Therefore, thermodynamic processes can be graphically
represented in various two-dimensional coordinate systems.
In addition to the most extensively used pV diagrams, pT
and VT diagrams are also used.

In Fig. 9.2, the thermodynamic process in the pV diagram
is represented by the curve $C_1 C_2$, and the points C_1 $(p_1,\ V_1)$
and C_2 $(p_2,\ V_2)$ indicate the initial and final states of the
thermodynamic system.

Only equilibrium processes (8.3.7) can be represented graphi-
cally. State variables for the whole body (or system) cannot
be specified for nonequilibrium processes (8.3.7) because they
differ at different parts of the body (or system). Hence such
graphical representation of nonequilibrium processes is im-
possible.

9.4.2 An element of work δW done by a system (9.2.3) in an
equilibrium process (8.3.7) can be measured by the area of the
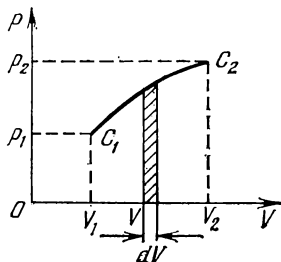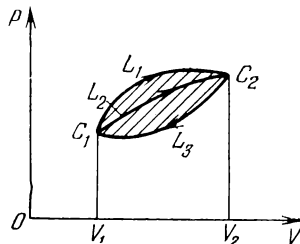

Fig. 9.2   Fig. 9.3

trapezoid having one curvilinear side, hatched in Fig. 9.2.
The work W_{1-2} done by the system in the process $C_1 C_2$ is equal to

$$W_{1-2} = \int_{V_1}^{V_2} p\, dV$$

and is measured by the area that is bounded by the curve re-
presenting process $C_1 C_2$, the axis of abscissas and the ordinates
p_1 and p_2 of points C_1 and C_2. The amount of work W_{1-2} done

depends on how the system passes from state C_1 to state C_2, i.e. upon the kind of process C_1C_2 the system undergoes in changing its state. In the pV diagram of Fig. 9.3 the amounts of work done by the system in the processes $C_1L_1C_2$, $C_1L_2C_2$ and $C_1L_3C_2$, and equal to W_{L_1}, W_{L_2} and W_{L_3}, respectively, are measured by three areas that differ in size: $W_{L_1} > W_{L_2} > > W_{L_3}$.

After the system has completed the cycle (11.1.1) $C_1L_1C_2L_3C_1$ the total amount of work $W_{C_1-C_1}$ is not equal to zero. The positive expansion work in process $C_1L_1C_2$ exceeds the negative work done in the compression process $C_2L_3C_1$. The resultant positive work is measured by the hatched area in Fig. 9.3.

9.4.3 The work W and the quantity of heat Q are not state functions (8.3.8). In various processes $1 \rightarrow 2$, in which the state of a system is changed, various quantities of heat are transferred to the system, and various amounts of work are done. Elements of the changes in the physical quantities δQ and δW are not complete differentials (9.1.3).

9.5 Heat Capacity of Matter. Applying the First Law of Thermodynamics to Isoprocesses in an Ideal Gas

9.5.1 The *heat capacity* C of a body is a physical quantity, numerically equal to the ratio of the quantity of heat δQ transferred to the body to the change dT in the temperature of the body in the thermodynamic process being considered:

$$C = \frac{\delta Q}{dT}.$$

The value of C depends upon the mass of the body, its chemical composition, thermodynamic state and the process in which the heat δQ is transferred to it.

9.5.2 The *specific heat* c is the heat capacity per unit mass of a substance. For a homogeneous body $c = C/M$, where M is the mass of the body.

The *molar heat capacity* C_μ is the heat capacity of a single mole (Appendix I) of the substance: $C_\mu = \mu c$, where μ is the molar mass (8.4.3) of the substance.

9.5.3 The element of the quantity of heat δQ that must be transferred to a body to change its temperature from T to $T + \Delta T$ is

$$\delta Q = C \, dT.$$

For a homogeneous body $\delta Q = Mc \, dT = (M/\mu) \, C_\mu \, dT$, where M is the mass of the body, μ is its molar mass and M/μ is the number of moles in the body.
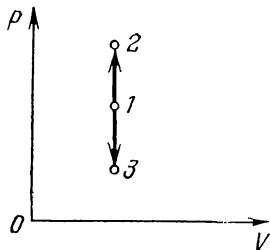
9.5.4 For equilibrium (8.3.7) isoprocesses in gases (8.3.7), the first law of thermodynamics (9.3.1) is of the form

$$\frac{M}{\mu} \, C_\mu \, dT = dU + p \, dV.$$

In the isochoric process (8.3.7) of heating or cooling a gas (straight lines *1-2* and *1-3* in Fig. 9.4), no element of work $\delta W = p \, dV$ is done (because $dV = 0$). The whole quantity of heat δQ transferred to the gas is expended in changing its internal energy: $\delta Q = dU$. If C_{V_μ} is the molar heat capacity at constant volume,

$$dU = \frac{M}{\mu} \, C_{V_\mu} \, dT. \tag{*}$$

Fig. 9.4

Within a definite temperature range it can be assumed that $C_{V_\mu} \approx$ const (for greater detail see Sect. 10.7.4) and that the change in the internal energy of the gas ΔU_{1-2}, as its temperature is changed from T_1 to T_2, is equal to $\Delta U_{1-2} = U_2 - U_1 = (M/\mu) \, C_{V_\mu} \, (T_2 - T_1)$ and is due to the quantity of heat Q_{1-2} transferred to the gas in the isochoric process:

$$Q_{1-2} = \frac{M}{\mu} \, C_{V_\mu} \, (T_2 - T_1). \tag{**}$$

At $T_2 > T_1$, $Q_{1-2} > 0$ and a definite quantity of heat is transferred to gas; at $T_2 < T_1$, $Q_{1-2} < 0$ and a definite quantity of heat is conducted away from the gas.

9.5.5 For an ideal gas, equations (*) and (**) express the change
in its internal energy in *any process* for changing the state of
the gas in the temperature range $(T_2 - T_1)$. The internal energy
of an ideal gas depends only
on its chemical composition,
mass and temperature.

In real gases (12.1.2) the inter-
nal energy includes the poten-
tial energy of interaction be-
tween the molecules (9.1.2),
which depends upon the dis-
tance between them. This part
of the internal energy of a
real gas is changed with changes
in its volume. Therefore,
equations (*) and (**) express
the change in the internal

Fig. 9.5

energy of a real gas only when it is heated or cooled in an
isochoric (constant-volume) process.

9.5.6 In an arbitrary equilibrium process (8.3.7) that takes
place in an ideal gas, the first law of thermodynamics assumes
the form

$$\frac{M}{\mu}\, C_\mu\, dT = \frac{M}{\mu}\, C_{V\mu}\, dT + p\, dV,$$

where C_μ is the molar heat capacity of an ideal gas in the given
process. The isobaric process (8.3.7) of heating (straight line
1-2 in Fig. 9.5) or cooling (straight line 1-3 in Fig. 9.5) a gas
can be carried out, for instance, in a vessel (cylinder) closed by
a movable piston subject to a constant external pressure. An
element of work δW done by an ideal gas in an isobaric process is

$$\delta W = p\, dV = \frac{M}{\mu}\, R\, dT,$$

in which use is made of the expression $dV = \dfrac{M}{\mu}\, \dfrac{R}{p}\, dT$ from
the Mendeleev-Clapeyron equation (8.4.4) at $p = \text{const.}$
The universal gas constant R (8.4.4) is numerically equal
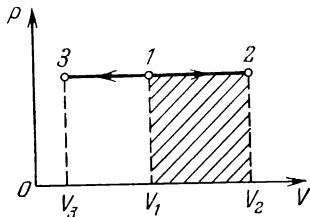to the work done by one mole of an ideal gas when it is heated

isobarically by one degree:

$$R = \frac{\delta W}{(M/\mu)\, dT}\,.$$

The work W_{1-2} done by the gas in isobaric expansion 1-2 is

$$W_{1-2} = \int_{V_1}^{V_2} p\, dV = p\,(V_2 - V_1).$$

It is represented by the hatched area in Fig. 9.5. For an ideal gas the work W_{1-2} is also equal to

$$W_{1-2} = \frac{M}{\mu}\, R\,(T_2 - T_1).$$

9.5.7 The element of the quantity of heat δQ transferred to the gas in the isobaric process is

$$\delta Q = \frac{M}{\mu}\, C_{p\mu}\, dT,$$

where $C_{p\mu}$ is the molar heat capacity at constant pressure. If the quantity $C_{p\mu}$ can be assumed constant in the temperature range $(T_2\text{-}T_1)$, then the quantity of heat Q_{1-2} transferred to (or conducted away from) the gas is

$$Q_{1-2} = \frac{M}{\mu}\, C_{p\mu}\,(T_2 - T_1).$$

9.5.8 The molar heat capacities $C_{p\mu}$ and $C_{V\mu}$ of a gas are related by the *Mayer equation*

$$C_{p\mu} - C_{V\mu} = R.$$

For specific heats c_p and c_V, the Mayer relation is of the form

$$c_p - c_V = R/\mu,$$

where μ is the molar mass (8.4.3) of the gas.
For the heat capacities C_p and C_V

$$C_p - C_V = \frac{M}{\mu}\, R,$$

where M is the mass of the gas and M/μ is the number of moles it contains.

The Mayer relation indicates that in heating a gas isobarically by one degree a greater quantity of heat must be transferred to the gas than in isochoric heating by the same amount. This difference between the quantities of heat should be equal to the work done by the gas in its isobaric expansion.

9.5.9 The isothermal process (8.3.7) of expanding (or compressing) a gas takes place in heat exchange (9.2.4) between the gas and the external medium at a constant temperature difference. For this to be feasible, the heat capacity (9.5.1) of the external medium should be sufficiently great and the process of expansion (or
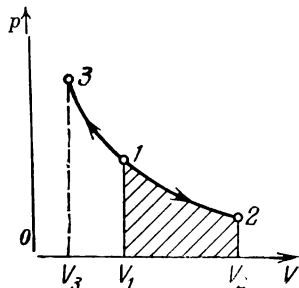

Fig. 9.6

compression) should proceed extremely slowly. First-order phase transitions (12.3.3), such as boiling and condensation, are isothermal processes if they proceed at constant external pressure.

In an isothermal process in an ideal gas the internal energy of the gas remains unchanged (9.5.5) and the whole quantity of heat Q_{1-2} transferred to the gas is expended in the work W_{1-2} done by the gas to overcome the external forces:

$$Q_{1-2} = W_{1-2} = \int_{V_1}^{V_2} p\, dV = \frac{M}{\mu} RT \int_{V_1}^{V_2} \frac{dV}{V} = \frac{M}{\mu} RT \ln \frac{V_2}{V_1}.$$

Here M/μ is the number of moles contained in mass M of the gas, T is the constant temperature of the gas, and V_1 and V_2 are the initial and final volumes of the gas. If the gas isothermally expands ($V_2 > V_1$), heat is being transferred to it ($Q_{1-2} > 0$) and it is doing positive work ($W_{1-2} > 0$) represented by the hatched area in Fig. 9.6. In isothermal compression of the gas (process 1-3 in Fig. 9.6), the work done by the gas is negative ($W_{1-3} < 0$). Positive work ($W'_{1-3} = -W_{1-3} > 0$) is

being done by the external forces to compress the gas. In compression, a certain quantity of heat is conducted away from the gas ($Q_{1-3} < 0$). The heat capacity of a substance in an isothermal process equals infinity ($dT = 0$ and $\delta Q \neq 0$).

9.5.10 An adiabatic process (8.3.7) proceeds under the condition that $\delta Q = 0$. It is essential that the condition $Q = 0$ does not define this process, because it does not specify the absence of heat exchange with the external medium. It only indicates that the algebraic sum of the quantities of heat transferred to and conducted away from the gas in various parts of the process equals zero. In an adiabatic process, work is done by an ideal gas at the expense of its internal energy:

$$\delta W = -dU = \frac{-M}{\mu}\, C_{V\mu}\, dT,$$

where $C_{V\mu}$ is the molar heat capacity of the gas at constant volume, M/μ is the number of moles contained in mass M of the gas, and dT is an element of change in the temperature of the gas. If the gas adiabatically expands ($\delta W = p\, dV > 0$), it is cooled ($dT < 0$). In adiabatic compression the gas is heated: $\delta W = p\, dV < 0$ and $dT > 0$.

9.5.11 The *adiabatic*, or *Poisson*, *equation*: $pV^{\varkappa} = $ const, is valid for an equilibrium adiabatic process (8.3.7).

Making use of the Mendeleev-Clapeyron equation (8.4.4), the relation between p and T, and also between V and T, in an adiabatic process can be found from the adiabatic equation:

$$pT^{\frac{\varkappa}{\varkappa-1}} = \text{const} \quad \text{and} \quad VT^{\frac{1}{\varkappa-1}} = \text{const}.$$

In these equations the dimensionless quantity $\varkappa = C_{p\mu}/C_{V\mu} = c_p/c_V > 1$ is the *adiabatic exponent*. The full curve in Fig. 9.7, called an *adiabat* or *adiabatic curve*, represents an adiabatic process in a pV diagram. The dashed curve on this diagram, called an *isotherm*, represents an isothermic process at the temperature corresponding to the initial state *1* of the gas. In an adiabatic process the pressure varies with the volume more rapidly than in an isothermic process (the curve is steeper). In adiabatic expansion the temperature of the gas and its pressure decrease faster than in the corresponding isothermal

expansion. In adiabatic compression of the gas its pressure increases at a higher rate than in the corresponding isothermal compression. The reason is that the increased pressure is due both to the decrease in the volume of the gas and to the increase in its temperature.

9.5.12 The work W_{1-2} done by the gas in the adiabatic process $1 \rightarrow 2$ is measured by the hatched area in Fig. 9.7.
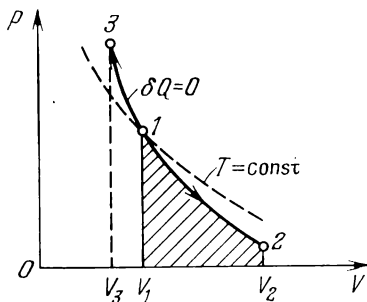

Fig. 9.7

The amount of work W_{1-2} done in an adiabatic process is expressed by

$$W_{1-2} = \frac{M}{\mu} C_{V\mu} (T_1 - T_2), \quad W_{1-2} = \frac{M}{\mu} \frac{R}{\varkappa - 1} (T_1 - T_2),$$

$$W_{1-2} = \frac{p_1 V_1}{\varkappa - 1} \left(1 - \frac{T_2}{T_1} \right), \quad W_{1-2} = \frac{p_1 V_1}{\varkappa - 1} \left[1 - \left(\frac{V_1}{V_2} \right)^{(\varkappa - 1)} \right].$$

In these equations M/μ is the number of moles contained in mass M of the gas, $C_{V\mu}$ is the molar heat capacity at constant volume, $\varkappa$ is the adiabatic exponent, and p, V and T are the state variables of the gas in the states *1* and *2* as indicated.

9.5.13 Table 9.1 lists summarized information on the isoprocesses in gases.

TABLE 9.1

Thermodynamic Process	Isochoric	Isobaric	Isothermal	Adiabatic
Basic condition	$V = \text{const}$	$p = \text{const}$	$T = \text{const}$	$\delta Q = 0$
Relationships between state variables	$\dfrac{p}{T} = \text{const}$	$\dfrac{V}{T} = \text{const}$	$pV = \text{const}$	$pV^{\varkappa} = \text{const}$ $pT^{\varkappa/(\varkappa-1)} = \text{const}$ $VT^{1/(\varkappa-1)} = \text{const}$
Work done in the process	$\delta W = 0$ $W = 0$	$\delta W = p\,dV$ $W = p(V_2 - V_1)$	$\delta W = p\,dV$ $W = \dfrac{M}{\mu}\,RT \ln \dfrac{V_2}{V_1}$	$\delta W = p\,dV = -dU$ $W = -\Delta U$ $= C_V(T_1 - T_2)$
Quantity of heat transferred in the process	$\delta Q = C_V dT$ $Q = C_V(T_2 - T_1)$	$\delta Q = C_p dT$ $Q = C_p(T_2 - T_1)$	$\delta Q = \delta W$ $Q = W$	$\delta Q = 0$ $Q = 0$
Change in internal energy	$dU = \delta Q$ $U = Q$	$dU = C_V dT$ $U = C_V(T_2 - T_1)$	$dU = 0$ $\Delta U = 0$	$dU = -\delta W = C_V dT$ $\Delta U = -W$ $= C_V(T_2 - T_1)$
Heat capacity	$C_V = \dfrac{M}{\mu}\,\dfrac{R}{\varkappa - 1}$	$C_p = \dfrac{M}{\mu}\,\dfrac{\varkappa R}{\varkappa - 1}$	$C_T = \infty$	$C_{ad} = 0$

CHAPTER 10   KINETIC THEORY OF GASES

10.1 Certain Information on Classical Statistical Physics

10.1.1 The *kinetic theory of gases* deals with the structure and physical properties of gases and is based on the statistical method of investigation (8.2.2). Besides those mentioned in Sect. (8.2.1), the following propositions form the foundation for *classical statistical physics.*

10.1.2 A system of particles complies with the laws of conservation of energy (9.2.7), momentum (2.7.1) and of angular momentum (4.4.1). A system of charged particles complies with the law of conservation of electric charge (14.1.3).

10.1.3 All physical processes in a system of particles proceed *continuously* in space and time. A space-time description of any physical phenomena in classical mechanics and classical statistical physics assumes the possibility of continuous variation of all physical quantities that characterize the state of the system. For instance, the velocity and energy of any particle can vary continuously due to the action of various forces.

10.1.4 Each particle in a system is regarded as being "tagged". It can be distinguished from all other such particles (*distinguishability of identical particles in classical statistical physics*).

10.1.5 Any particle of a system can have arbitrary values for its coordinates and momenta (or velocities) independently of the values of these quantities for other particles of the system. If an infinitely small volume $dx\,dy\,dz$ is chosen in the coordinate system XYZ, any particle can be located within this volume regardless of the presence in this volume of any arbitrary number of other particles. Likewise, any particle may be located within an arbitrary element of "volume" $dp_x\,dp_y\,dp_z$ of momentum space (or $du_x\,du_y\,du_z$ of velocity space) regardless of the presence in these "volumes" of any arbitrary number of other particles. This implies that any particle can have momentum components along the coordinate axes within the limits from p_x to $p_x + dp_x$, from p_y to $p_y + dp_y$ and from p_z to $p_z + dp_z$ (and, respectively, velocity components of the particle along the coordinate axes within the limits from u_x to $u_x + du_x$, u_y to $u_y + du_y$ and u_z to $u_z + du_z$). The minimum

values of the volumes $dx\ dy\ dz$ and $dp_x\ dp_y\ dp_z$ are not restricted in any manner.

Sections 10.1.2 and 10.1.3 apply not only to classical statistical physics, they are characteristic of all of classical physics. Sections 10.1.4 and 10.1.5 apply only to classical statistical physics; they do not hold for quantum statistics (41.1.1).

10.2 Basic Equation of the Kinetic Theory of Gases

10.2.1 The pressure of a gas (8.3.2) in a vessel is the result of the collisions of the gas molecules with the walls of the vessel. The pressure of a gas is the macroscopic effect of the thermal motion of molecules (8.1.1). As the molecules of the gas strike the walls of the vessel (8.4.1) they impart their momentum (2.3.4) to the walls. This change in momentum of the molecules exerts the pressure on the walls. For an ideal gas the collisions between the molecules within the vessel has no effect on the pressure exerted by the gas on the walls. Owing to the chaotic nature of thermal motion of the molecules (8.1.3) the pressure of the gas is the same on all the walls and, by definition (8.3.2), is the average force exerted normal to unit area of the wall surface.

10.2.2 The *basic equation of the kinetic theory of gases* is

$$pV = \frac{2}{3}\ E_k,$$

where p is the pressure of the gas, V is its volume and $E_k = \sum_{i=1}^{N} \frac{m_i u_i^2}{2}$ is the total kinetic energy of translational motion of all the N molecules of gas in the vessel.

10.2.3 In a homogeneous gas $m_i = m$, i.e. the mass of each molecule is the same, but their velocities u_i differ (10.3.1), and

$$E_k = \frac{m}{2} \sum_{i=1}^{N} u_i^2.$$

It proves expedient to introduce the root-mean-square velocity v_{rms} of translational motion of gases:

$$v_{rms} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} u_i^2},$$

where N is the total number of molecules in volume V of the gas. The root-mean-square velocity pertains to the whole
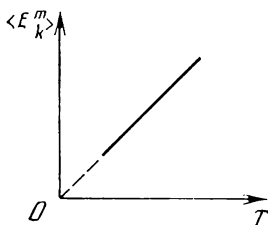

Fig. 10.1

population of molecules, but is meaningless in application to a single molecule or to a small number of molecules. Making use of v_{rms} the expression for E_k assumes the form

$$E_k = \frac{1}{2} N m v_{rms}^2 \quad \text{and} \quad pV = \frac{1}{3} N m v_{rms}^2 = \frac{1}{3} M v_{rms}^2,$$

where $M = Nm$ is the mass of the gas.
The basic equation for the pressure of a gas is

$$p = \frac{2}{3} E_{k_0} = \frac{1}{3} n_0 m v_{rms}^2 = \frac{1}{3} \rho v_{rms}^2,$$

where $E_{k_0} = E_k/V$, $n_0 = N/V$ is the number of molecules in unit volume, and $\rho = n_0 m$ is the density of the gas,

10.2.4 From a comparison of the Mendeleev-Clapeyron equation (8.4.4) with the basic equation (10.2.3) $RT = \mu v_{\text{rms}}^2/3$ it follows that

$$v_{\text{rms}} = \sqrt{\frac{3RT}{\mu}} = \sqrt{\frac{3RT}{mN_{\text{A}}}} = \sqrt{\frac{3kT}{m}} = 1.73 \sqrt{pv},$$

where k is Boltzmann's constant (8.4.5), m is the mass of a molecule, N_{A} is Avogadro's number (Appendix II), p is the pressure of the gas, and v is its specific volume.

10.2.5 The *average kinetic energy of translational motion of the molecules in an ideal gas* is

$$\langle E_{\text{k}}^{\text{m}} \rangle = \frac{E_{\text{k}}}{N} = \frac{mv_{\text{rms}}^2}{2} = \frac{3}{2} kT.$$

The average kinetic energy $\langle E_{\text{k}}^{m} \rangle$ is proportional to the absolute temperature and is independent of any other variables (Fig. 10.1). The absolute, or thermodynamic, temperature is a measure of the average kinetic energy of translational motion of the molecules in an ideal gas. This consequence of the kinetic theory of gases is not valid in the region of ultralow temperatures, close to absolute zero (8.3.4).

10.3 Maxwell's Molecular Velocity and Energy Distribution Law (Maxwell's Distribution Law)

10.3.1 The *law of the distribution of molecular velocities for an ideal gas,* theoretically established by Maxwell, determines the number dn of molecules of a homogeneous monatomic ideal gas, out of the total number n_0 of its molecules in unit volume, that has a velocity, at a given temperature, lying within the range from u to $u + du$. This law is valid for gases in a state of thermodynamic equilibrium (8.3.3). The velocity distribution of the molecules of such a gas is a steady-state value. The Maxwellian distribution is based on pairwise collisions of the randomly moving gas molecules. This assumes uniform distribution of the molecules throughout the volume of the vessel, i.e. constant density of the gas,

10.3.2 In its most customary form, the law of the distribution of molecular speeds (magnitudes of the molecular velocities) is presented as

$$dn = n_0 \left(\frac{m}{2\pi kT} \right)^{3/2} e^{-\frac{mu^2}{2kT}} 4\pi u^2 \, du.$$

Here u is the magnitude of the molecular velocity (speed of a molecule), m is the mass of a molecule, k is Boltzmann's
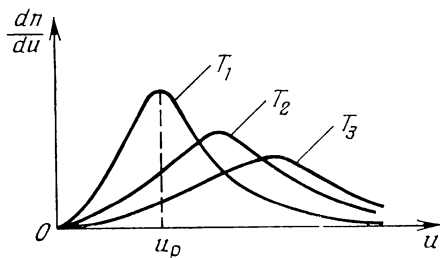

Fig. 10.2

constant, and T is the absolute (thermodynamic) temperature. Molecular velocity distribution curves for various temperatures $T_1 < T_2 < T_3$ are shown in Fig. 10.2. As is evident from these curves, the most probable velocity u_p (10.3.4) increases with temperature and the fraction of molecules travelling at this velocity decreases.

10.3.3 Another form of the Maxwellian distribution is

$$dn = n_0 \left(\frac{m}{2\pi kT} \right)^{3/2} e^{-\frac{mu^2}{2kT}} \, du_x \, du_y \, du_z,$$

where u_x, u_y and u_z are projections of the molecular velocity on the coordinate axes. Owing to the random nature of thermal motion of the molecules (8.1.1), distributions of the molecules

according to the projections u_i of their velocities on the coordinate axes $(i = x, y, z)$ are mutually independent. Hence

$$dn = n_0 f(u_x) f(u_y) f(u_z) du_x du_y du_z,$$

where

$$f(u_i) = \left(\frac{m}{2\pi kT}\right)^{3/2} e^{-\frac{mu_i^2}{2kT}} \quad (i = x, y, z)$$

is a distribution function of the projections of the molecular velocities.

The Maxwellian distribution is isotropic. This is evident from the fact that the function $f(u_x, u_y, u_z)$ depends only upon the speed (magnitude of the velocity), and because the function $f(u_i)$ is the same along all the axes.

10.3.4 The law of the distribution of molecular velocities (10.3.2) can be used to determine the *most probable velocity* u_p of the molecules, which corresponds to the maximum of the function

$$f(u) = \frac{dn}{du} = 4\pi n_0 \left(\frac{m}{4\pi kT}\right)^{1/2} e^{-\frac{mu^2}{2kT}}.$$

The condition $\left[\dfrac{d}{du}\left(e^{-\frac{mu^2}{2kT}} u^2\right)\right]_{u=u_p} = 0$ yields

$$u_p = \sqrt{\frac{2kT}{m}} = \sqrt{\frac{2RT}{\mu}} = v_{rms}\sqrt{\frac{2}{3}}.$$

10.3.5 A third form of the law of the distribution of molecular velocities is

$$dn = \frac{4n_0}{\sqrt{\pi}} e^{-u^2/u_p^2} \left(\frac{u}{u_p}\right)^2 \frac{du}{u_p}.$$

The fraction dn/n_0 of the molecules of a gas having velocities lying within the range from u to $u + du$ is numerically equal

to the hatched area dS under the curve in Fig. 10.3, which represents the dependence of $(u_p/n_0)\,(dn/du)$ on u/u_p:

$$dS = \left(\frac{u_p}{n_0} \frac{dn}{du} \right) \frac{du}{u_p} = \frac{dn}{n_0} \ .$$

The area under the curve in Fig. 10.3, i.e. bounded by the curve and the axis of abscissas, is equal to unity. This area
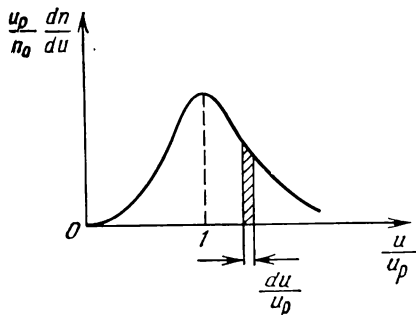

Fig: 10.3

represents the fraction of the molecules having all possible velocity values from 0 to ∞.

10.3.6 The *arithmetic mean velocity* $\langle u \rangle$ of translational motion of the molecules in an ideal gas, calculated by means of the distribution law (10.3.2), is

$$\langle u \rangle = v_{rms} \sqrt{\frac{8}{3\pi}} = \sqrt{\frac{8RT}{\pi\mu}} = \sqrt{\frac{8kT}{\pi m}} = 1.60 \sqrt{pv}.$$

10.3.7 The energy distribution of the molecules in an ideal gas determines the fraction dn_{Em}/n_0 of molecules, of their

total number n_0, having kinetic energies $E_k^m = mu^2/2$ lying within the range from E_k^m to $E_k^m + dE_k^m$:

$$dn_{E^m} = \frac{2n_0}{\sqrt{\pi}} (kT)^{-3/2} e^{-\frac{E_k^m}{kT}} \sqrt{E_k^m} \, dE_k^m.$$

Here $dn_{E^m}/n_0 = f(E_k^m) \, dE_k^m$, where $f(E_k^m)$ is the distribution function for the molecular energies in an ideal gas.

Example. The average kinetic energy $\langle E_k^m \rangle$ of a molecule in an ideal gas is

$$\langle E_k^m \rangle = \int_0^\infty E_k^m f(E_k^m) \, dE_k^m = \frac{2}{\sqrt{\pi}(kT)^{3/2}} \int_0^\infty E_k^m e^{-\frac{E_k^m}{kT}} \sqrt{E_k^m} \, dE_k^m$$

$$= \frac{3}{2} kT$$

(cf. Sect. 10.2.4).

10.3.8 The relative motion (7.1.2) of two particles with the masses m_1 and m_2 is equivalent to the motion of a single particle with the reduced mass $m_{red} = m_1 m_2/(m_1 + m_2)$. For a homogeneous gas $m_1 = m_2 = m$ and $m_{red} = m/2$. The distribution of molecules according to their relative velocities establishes the fraction $dn_{u_{rel}}/n_0$ of molecules, of their total number n_0, having relative velocities u_{rel} lying within the range from u_{rel} to $u_{rel} + du_{rel}$:

$$dn_{u_{rel}} = n_0 \left(\frac{m}{4\pi kT}\right)^{3/2} e^{-\frac{mu_{rel}^2}{4kT}} 4\pi u_{rel}^2 \, du_{rel}.$$

Here $dn_{u_{rel}}/n_0 = f(u_{rel}) \, du_{rel}$, where

$f(u_{rel}) = 4\pi \left(\frac{m}{4\pi kT}\right)^{3/2} e^{-\frac{mu_{rel}^2}{4kT}} u_{rel}^2$ is the relative velocity distribution of the molecules in an ideal gas.

Example. The *average relative velocity* $\langle u_{\text{rel}} \rangle$ of the molecules is

$$\langle u_{\text{rel}} \rangle = \int\limits_0^\infty u_{\text{rel}} f(u_{\text{rel}})\, du_{\text{rel}} = \sqrt{2}\, \sqrt{\overline{\frac{8kT}{\pi m}}} = \sqrt{2}\, \langle u \rangle,$$

where $\langle u \rangle$ is the arithmetic mean velocity of the molecules (10.3.6).

10.4 Particle Distribution in a Potential Force Field (Boltzmann Distribution)

10.4.1 Gases are ordinarily in the potential (conservative) gravitational fields of the earth (6.2.1). If there were no such gravitational field, the atmospheric air would be dispersed in the universe. If, on the other hand, there were no thermal motion, the molecules of atmospheric air would drop to the earth. Gravity and thermal motion lead to a steady state of gases, in which their concentration and pressure decrease with an increase of their altitude above the earth's surface.

10.4.2 If a gas (or other system of particles) is located in an external potential (conservative) force field (3.1.6), the distribution of the particles throughout the volume is described by the *Boltzmann distribution law*. This law establishes the number of particles $dn\,(x,\ y,\ z)$ that have coordinates within the ranges from x to $x + dx$, from y to $y + dy$, and from z to $z + dz$. In other words, $dn\,(x,\ y,\ z)$ is the number of molecules contained in the element of volume $dV = dx\,dy\,dz$. The Boltzmann distribution law is of the form

$$dn\,(x,\ y,\ z) = \text{const} \times e^{-\dfrac{E_{\text{p}}^{\text{m}}(x,\,y,\,z)}{kT}}\, dx\,dy\,dz$$

where $E_{\text{p}}^{\text{m}}\,(x,\ y,\ z)$ is the potential energy of a particle in the external force field, k is Boltzmann's constant, and T is the absolute temperature. The value of the constant in the preceding equation is determined from the normalization condition: $\int dn = n_0$, where n_0 is the total number of particles in unit volume.

Example. For particles of mass m, located in the earth's gravitational field, $E_{\mathrm{p}}^{\mathrm{m}} = mgh$ (3.3.3), where g is the acceleration due to gravity (7.3.3), and h is the altitude. At each altitude there is a Maxwellian molecular velocity distribution (10.3.1). The number of molecules contained in the volume dV is

$$dn\,(x,\ y,\ z) = \mathrm{const} \times e^{-\dfrac{mgh}{kT}}\ dV.$$

The gas density $\rho = m\,dn/dV$ decreases with an increase in the altitude according to the exponential law

$$\rho = \mathrm{const} \times e^{-\dfrac{mgh}{kT}}.$$

The value of the constant can be determined from the condition $\rho = \rho_0 = \mathrm{const}$ at $h = 0$. The density of the gas or its pressure varies according to the *barometric height formula*

$$\rho = \rho_0 e^{-\dfrac{mgh}{kT}} \quad \text{and} \quad p = p_0 e^{-\dfrac{mgh}{kT}}.$$

10.4.3 The *Maxwell-Boltzmann law*, or *distribution*, describes the distribution of the coordinates and velocities of gas molecules subject to the action of an arbitrary potential (conservative) force field:

$$dn = \mathrm{const} \times e^{-\dfrac{p_x^2 + p_y^2 + p_z^2}{2mhT}}\ dp_x\,dp_y\,dp_z e^{-\dfrac{F_{\mathrm{p}}^{\mathrm{m}}(x,\ y,\ z)}{kT}}\ dx\,dy\,dz,$$

where dn is the number of molecules contained in six-dimensional space within the element of volume $d\Gamma = dx\,dy\,dz\,dp_x\,dp_y\,dp_z$, and $E_{\mathrm{p}}^{\mathrm{m}}\,(x,\ y,\ z)$ is the potential energy of a molecule in the external force field at a point with the coordinates x, y and z, and the projections p_x, p_y and p_z of the momenta on the axes. The Maxwell-Boltzmann law is the product of two distribution functions. One of them describes the distribution with respect to coordinates, and the other with respect to momenta (or velocities).

10.5 Mean Free Path of Molecules

10.5.1 Gas molecules have finite dimensions (8.4.1) and are continuously colliding with one another in thermal motion. Between two successive collisions, molecules travel at uniform velocity in a straight line over certain distances called their free paths λ. The *mean free path* $\langle \lambda \rangle$ is the average distance that a molecule travels without a collision. The mean free path is a characteristic of the whole totality of gas molecules at given values of p and T.

10.5.2 In unit time each molecule undergoes a mean number of collisions $\langle z \rangle$ (*average collision frequency*) equal to

$$\langle z \rangle = \pi d^2 n_0 \langle u_{\text{rel}} \rangle = \sqrt{2}\, \pi d^2 n_0 \langle u \rangle.$$

Here d is the effective diameter of the molecule (8.4.1), n_0 is the number of molecules in unit volume of the gas, $\langle u_{\text{rel}} \rangle$ is their average relative velocity (10.3.8), and $\langle u \rangle$ is their arithmetic mean velocity (10.3.6).

10.5.3 The average distance travelled by a molecule in unit time is numerically equal to $\langle u \rangle$. Therefore $\langle u \rangle = \langle \lambda \rangle \langle z \rangle$. The mean free path $\langle \lambda \rangle$ is

$$\langle \lambda \rangle = \frac{\langle u \rangle}{\langle z \rangle} = \frac{1}{\sqrt{2}\, \pi d^2 n_0}.$$

At constant temperature, n_0 is proportional to the pressure p of the gas (8.4.5). Hence, for a given gas the mean free path is inversely proportional to the pressure:

$$p_1 \langle \lambda_1 \rangle = p_2 \langle \lambda_2 \rangle = \text{const,}$$

in which the subscripts 1 and 2 refer to two states of the gas.

10.5.4 If molecules are ejected from some source of particles and a beam is formed by means of a diaphragm, the free path distribution law is valid for the molecules of the beam:

$$N = N_0 e^{-x/\langle \lambda \rangle},$$

where N is the number of molecules in the beam that have travelled a distance x without collisions, and N_0 is the number of molecules in the beam at $x = 0$, i.e. at the output of the diaphragm.

10.6 Principle of the Equipartition of Energy.
Internal Energy of an Ideal Gas

10.6.1 The *number of degrees of freedom* of a body is the minimum number of coordinates (number of independent coordinates) that it is necessary to specify in order to completely determine the position of the body in space.

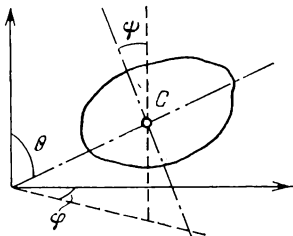A particle, for instance, travelling along one of the coordinate axes has a single degree of freedom. The same particle travelling in a plane has two degrees of freedom. The position of a particle travelling freely in space is determined by three degrees of freedom: the coordinates x, y and z. A rigid body (1.1.4) has six degrees of freedom: its position in space is determined by the three coordinates of its centre of mass (2.3.3), two angles, θ and φ, determining the direction of a certain axis, fixed in the body and passing through its centre of mass (Fig. 10.4), and, in addition, angle ψ, determining the direction of a second axis, fixed in the body and perpendicular to the first axis. Changes in the three coordinates of the centre of mass at given constant angles θ, φ and ψ correspond to translational motion of the rigid body. The coordinates of the centre of mass are the three degrees of freedom of translational motion. Changes in the angle θ, φ or ψ at a constant position of the centre of mass lead to rotation of the rigid body. Hence, the corresponding degrees of freedom are said to be rotary. Additional degrees of freedom of oscillatory, or vibratory, motion (10.6.5) are introduced to determine the position in space of an imperfectly rigid body, whose various parts can be displaced with respect to one another.

Fig. 10.4

10.6.2 In many problems a *molecule of a monatomic gas* can be regarded as a particle (cf Sect. 8.4.1). This is based on the fact that practically all the mass of such an atom is concentrated in its nucleus (42.1.1), which is of extremely small size. The

molecule of monatomic gas has three degrees of freedom of translational motion. Its average kinetic energy $\langle E_k^m \rangle$ is equal to the kinetic energy of molecules travelling at a velocity equal to the root-mean-square velocity v_{rms} (10.2.4):

$$\langle E_k^m \rangle = \frac{3}{2} \, kT.$$

Owing to the random nature of thermal motion of the molecules, this energy is uniformly distributed between the three degrees of freedom. Hence, on an average each degree of free-
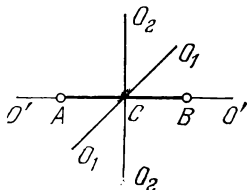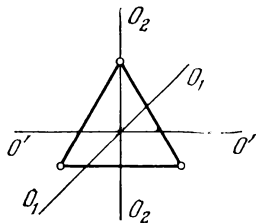

Fig. 10.5 .   Fig. 10.6

dom of translational motion of a monatomic molecule accounts for the same amount of kinetic energy $\langle E_{k0}^m \rangle = \langle E_k^m \rangle / 3 = kT/2$, where k is Boltzmann's constant and T is the absolute temperature.

10.6.3 As a first approximation, a *molecule of a diatomic gas* consists of two atoms, A and B, rigidly held together by an incompressible, undeformable bond (Fig. 10.5). In addition to the three degrees of freedom of translational motion, the centre of mass C (2.3.3) of such a molecule has two more degrees of freedom of rotary motion about axes O_1-O_1 and O_2-O_2. Rotation about the third axis O'-O' makes no contribution to the molecule's energy because the moment of inertia (4.2.1) of the atoms with respect to this axis is negligibly small. Thus, a diatomic molecule has five degrees of freedom. Molecules

11 *

consisting of three or more atoms have six degrees of freedom: three translational and three rotary degrees (Fig. 10.6).

10.6.4 The principle of equipartition of energy states that on an average each degree of freedom of a molecule accounts for the same amount of kinetic energy, equal to $kT/2$. If a molecule has i degrees of freedom its average kinetic energy $\langle E_k^m \rangle = ikT/2$.

The principle of equipartition of energy leads to the conclusion that all degrees of freedom of the molecule are equally justified: each makes the same contribution to its average energy. This conclusion actually has only a limited field of application, and has been revised for quantum statistics (41.1.1).

10.6.5 In real molecules, which are not absolutely rigid, vibratory degrees of freedom make a contribution to the energy of the molecule in accordance with the principle of equipartition of energy (10.6.4). Each vibratory degree of freedom accounts, not only for kinetic energy E_k^m, but for potential energy E_p^m as well. For harmonic vibrations (28.1.3), $\langle E_{k_0}^m \rangle = \langle E_{p_0}^m \rangle$. According to Sect. 10.6.4

$$\langle E_{k_0}^m \rangle = \langle E_{p_0}^m \rangle = \frac{1}{2} kT,$$

where $\langle E_{k_0}^m \rangle$ and $\langle E_{p_0}^m \rangle$ are the average kinetic and potential energies of a molecule per vibratory degree of freedom.

The average value $\langle E_0^m \rangle$ of the total energy of a molecule per vibratory degree of freedom is

$$\langle E_0^m \rangle = \langle E_{k_0}^m \rangle + \langle E_{p_0}^m \rangle = 2 \langle E_{k_0}^m \rangle = kT.$$

This equation is valid for harmonic vibrations of particles (atoms, molecules or ions) in the crystal lattice sites of solids (8.1.4).

10.6.6 The internal energy (9.1.2) of a polyatomic gas constitutes the kinetic energy of all kinds of motion of its particles. For one mole of such a gas

$$U = \langle E_k^m \rangle N_A = \frac{i}{2} kN_A T = i \frac{RT}{2} ,$$

where N_A is Avogadro's number (Appendix II), i is the number of degrees of freedom of the molecule (10.6.4), and R is the universal gas constant (8.4.4).

10.7 Heat Capacity of Monatomic, Diatomic and Polyatomic Gases

10.7.1 The molecules of a monatomic gas have three degrees of freedom (10.6.2) and, according to Sect. 10.6.6, the internal energy of one mole of such a gas is $U = 3kN_A T/2$. The molar heat capacity of such a gas at constant volume (9.5.4) is

$$C_{V\mu} = \frac{dU}{dT} = \frac{3}{2}\, N_A k = \frac{3}{2} R = 1246.5\, \frac{\text{J}}{\text{mol-K}} \left(\frac{\text{joules}}{\text{mole} \times \text{kelvin}} \right)$$

$$\approx 3\, \frac{\text{cal}}{\text{mol-K}} \quad \text{because} \quad R \approx 2\, \frac{\text{cal}}{\text{mol-K}}\ .$$

For a polyatomic gas whose molecules have i degrees of freedom

$$C_{V\mu} = \frac{iR}{2} = 4.16\, i\, \frac{\text{J}}{\text{mol-K}} \approx i\, \frac{\text{cal}}{\text{mol-K}}\ .$$

10.7.2 The molar heat capacity $C_{p\mu}$ at constant pressure (9.5.7 and 9.5.8) of a gas whose molecules have i degrees of freedom is

$$C_{p\mu} = \frac{(i+2)\,R}{2} = 4.16\,(i+2)\, \frac{\text{J}}{\text{mol-K}} \approx (i+2)\, \frac{\text{cal}}{\text{mol-K}}\ .$$

The principle of equipartition of energy (10.6.4) leads to the conclusion that the heat capacity of a gas depends upon the number of degrees of freedom of its molecules, but does not depend on the temperature. Experimental data disprove these conclusions of the classical theory of heat capacity. Investigations indicate that the heat capacity of gases increases and decreases with temperature.

10.7.3 A theoretical explanation of experimental data on the dependence of heat capacity on temperature in a wide temperature range is given in the quantum theory of heat capacity. This theory takes into account the following results obtained

in a quantum-mechanical description of the properties of diatomic and polyatomic molecules:

(a) Of the four parts that comprise the energy E^m of the molecule* (39.5.2), i.e.

$$E^m = E_t^m + E_e^m + E_v^m + E_r^m,$$

only E_t^m, the energy of translational motion of the molecule's centre of mass (2.3.3), varies continuously (not by discrete amounts). The values of all other kinds of internal energy of molecules are quantized and thereby assume only a discrete series of values (39.4.5). This refers to the following kinds of energy: E_e^m, the energy of motion of the electrons of the molecule; E_v^m, the energy of vibrational motion of the nuclei in the molecule; and E_r^m, the energy of rotary motion of the molecule.

(b) As a first approximation, the kinds of internal motion of the molecule can be assumed independent of one another. Thus, at small amplitudes of vibration of the nuclei, the changes in the moments of inertia (4.2.1) of the molecule due to vibration can be neglected, as can the effect of the vibrational motion in the molecule on its rotation.

(c) No changes in the energy of motion of the electrons occur when a substance is heated by one degree at ordinary temperatures. Adjacent energy levels of the electrons in the molecule are separated by energy of the order of several eV, corresponding to temperatures of several tens of thousands of degrees (10.7.6). Therefore, the energy of motion of electrons in the molecule can be neglected in dealing with the heat capacities of monatomic and polyatomic gases up to the very highest temperatures.

10.7.4 As a first approximation, the vibrational motion of the nuclei in the molecule is described as the vibration of a harmonic oscillator (38.5.4), whose mass m is equal to the reduced mass of a system of vibrating atoms, $m = m_1 m_2/(m_1 +$

---

* This ignores the nuclear energy of the molecule, which has no influence on the heat capacity of molecules.

$+ m_2$), where m_1 and m_2 are the masses of the atoms. Then the energy of vibrational motion of the molecule equals

$$E_v^m = h\nu_0 \left(n + \frac{1}{2} \right) ,$$

where n is the quantum number and assumes integral values: $n = 0, 1, 2, \ldots$, ν_0 is the natural frequency of vibrations (38.5.5), and h is Planck's constant (Appendix II). The energy of a harmonic oscillator, equal to $h\nu_0/2$ at $n = 0$, is conserved at any arbitrarily low temperatures, and even at $T \to 0$ (38.5.6). The difference ΔE_v^m between the energies of two adjacent vibrational energy levels, $\Delta E_v^m = h\nu_0$, is independent of the quantum number. Valid at room and lower temperatures is the condition: $\Delta E_v^m \gg kT$. No changes are observed in the energy of vibrational motion of molecules at such temperatures. In calculating the heat capacity of gases under such conditions, vibrational degrees of freedom of the molecule can be neglected. In the general case, vibrational motions make contributions to the internal energy U_v (9.1.2) and the heat capacity $C_{V\mu}$ (9.5.2). For one mole

$$U_v = \frac{N_A h\nu}{2} \frac{\left(e^{\frac{h\nu}{2kT}} + e^{-\frac{h\nu}{2kT}} \right)}{\left(e^{\frac{h\nu}{2kT}} - e^{-\frac{h\nu}{2kT}} \right)} ,$$

$$C_{V\mu_v} = \frac{N_A k}{4} \left(\frac{h\nu}{kT} \right)^2 \left(\frac{e^{\frac{h\nu}{2kT}} - e^{-\frac{h\nu}{2kT}}}{2} \right)^{-2} ,$$

where T is the absolute temperature, k is Boltzmann's constant (8.4.5) and N_A is Avogadro's number (Appendix II). At high temperatures $(T \gg h\nu/k)$

$$U_v \approx N_A kT = RT, \quad C_{V\mu_v} \approx N_A k = R.$$

These results coincide with those of classical theory which follow from the principle of the equipartition of energy (10.6.4). Under these conditions the energy of the oscillator can be

assumed to vary continuously because $\Delta E_{\mathrm{v}}^{\mathrm{m}} \ll kT$. At sufficiently low temperatures $(T \ll h\nu/k)$

$$U_{\mathrm{v}} \approx \frac{N_A h\nu}{2} + N_A h\nu e^{-\frac{h\nu}{kT}}, \quad \text{and}$$

$$C_{V\mu_{\mathrm{v}}} \approx N_A k \left(\frac{h\nu}{kT} \right)^2 e^{-\frac{h\nu}{kT}}.$$

As $T \to 0$, the energy $U_{\mathrm{v}} \to N_A h\nu/2$. This quantity is called the *zero-point vibrational energy* of the system (38.5.6). As $T \to 0$, the heat capacity $C_{V\mu_{\mathrm{v}}} \to$ $\to 0$ in accordance with the third law of thermodynamics (11.8.4). The dependence $C_{V\mu_{\mathrm{v}}} = f(T)$ for diatomic gases is shown in Fig. 10.7.
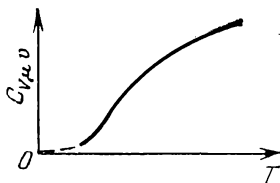
Fig. 10.7

10.7.5 As a first approximation, the rotary motion of diatomic molecules can be dealt with as the motion of a rigid dumbbell which rotates about its centre of mass with the moment of inertia $I = m_1 m_2 r_0^2/(m_1 + m_2)$, where $m = m_1 m_2/(m_1 + m_2)$ is the reduced mass of the molecule (m_1 and m_2 are the masses of the two atoms), and r_0 is the distance between the atoms in the molecule. The energy of such a system is

$$E_{\mathrm{r}}^{\mathrm{m}} = \frac{h^2}{8\pi I} J(J+1),$$

where J is a quantum number that assumes the values $J = = 0, 1, 2, \ldots$. The difference $\Delta E_{\mathrm{r}}^{\mathrm{m}}$ between the energies of two adjacent rotational energy levels is

$$\Delta E_{\mathrm{r}}^{\mathrm{m}} = \frac{2h^2}{8\pi^2 I} (J+1).$$

The value of $\Delta E_{\mathrm{r}}^{\mathrm{m}}$ is less than $\Delta E_{\mathrm{v}}^{\mathrm{m}}$ by a factor of 800 to 1000.

At ordinary temperatures $\Delta E_r^m \ll kT$ for diatomic and certain polyatomic gases (water vapour, methane, etc.), and the quantization of the energy of rotary motion of the molecules can be ignored. Under these conditions, the principle of equipartition of energy (10.6.4) can be applied in calculating the contribution of rotary motion to the heat capacity.

The contribution of rotary motion of diatomic molecules to the internal energy and heat capacity per mole equals:

(a) At high temperatures $(T \gg h^2/8\pi^2 kI)$

$$U_r \approx N_A kT \left(1 - \frac{h^2}{24\pi^2 IkT} \right), \quad \text{and} \quad C_{V\mu r} \approx N_A k = R.$$

The heat capacity $C_{V\mu r}$ at high temperatures has a value following from the principle of equipartition of energy (10.6.4).

(b) At low temperatures $(T \ll h^2/8\pi^2 kI)$

$$U_r \approx \frac{3h^2 N_A}{4\pi^2 I} e^{-\frac{h^2}{4\pi^2 IkT}}, \quad \text{and}$$

$$C_{V\mu r} \approx 3 \left(\frac{h^2}{4\pi^2 I} \right)^2 \frac{N_A}{kT^2} e^{-\frac{h^2}{4\pi^2 IkT}},$$

as $T \to 0$, the heat capacity $C_{V\mu r} \to 0$ (10.7.4). The general shape of the curve representing the dependence $C_{V\mu r} = f(T)$ is the same as that in Fig. 10.7.

As the gas cools, the reduction in the energy of its molecules leads to a reduction in the number of molecules transferring to higher rotational energy levels. At sufficiently low temperatures, the rotation of the molecules in a gas practically cannot be excited, and rotational degrees of freedom make no contribution to the heat capacity of the gas. The heat capacities of all gases at low temperatures become the same as for monatomic gases whose molecules have no rotational degrees of freedom.

10.7.6 At extremely high temperatures, of the order of tens of thousands of degrees, the heat capacity is increased. This is associated with the fact that the value of kT under such conditions becomes comparable, in order of magnitude, with ΔE_e^m, the change in the energy of the electrons in their transition

from lower to higher energy levels. Moreover, at high temperatures the processes of dissociation and ionization of the gases make contributions to their heat capacities.

10.8 Transport Phenomena in Gases

10.8.1 *Transport phenomena* unite a group of processes pertaining to nonuniformities in the density, temperature or velocity of ordered motion of various layers of a substance. The equalization of these nonuniformities leads to transport phenomena. Transport phenomena include diffusion, internal friction and heat conduction.

10.8.2 Transport phenomena in gases and liquids consist of ordered directional mass transport in these substances (diffusion), momentum transport (internal friction) and internal energy transport (heat conduction). In gases these phenomena violate the completely random motion of the molecules and Maxwell's molecular velocity distribution (10.3.2). Deviations from the Maxwell law can be explained by the directional transport of the physical characteristics of the substance in such phenomena. In the simplest cases of one-dimensional transport phenomena, the physical quantities that govern them depend on only a single Cartesian coordinate.

10.8.3 *Diffusion* is the spontaneous interpenetration and mixing of the particles of two adjoining gases, liquids or solids. In chemically pure gases at constant temperature, diffusion occurs owing to differing densities in various parts throughout the volume of the gas. In a mixture of gases, diffusion is caused by the difference in the densities of the various gases at various parts of the volume of the gas mixture.

In a chemically homogeneous gas the phenomenon of diffusion consists in the transport of a mass of the gas from places with higher density to places with lower density. This phenomenon obeys *Fick's law*:

$$m_{\text{sec}} = - D \; \frac{d\rho}{dx} \, .$$

Here m_{sec} is the *specific mass flow* and is numerically equal to the mass that passes by diffusion in unit time through a plane surface of unit area located perpendicular to the direction the

substance is being transported, ρ is the density of the gas, and D is the *diffusion coefficient*. The derivative $d\rho/dx$ is numerically equal to the density gradient, i.e. the change in density per unit length. The diffusion coefficient is numerically equal to the specific mass flow at unit density gradient. The minus sign in Fick's law indicates that mass transport takes place in the direction of decreasing density.

Another form of Fick's law is

$$j = - D \frac{dn_0}{dx} \, ,$$

where $j = m_{\text{sec}}/m$ is the density of the molecular flow in diffusion, i.e. the number of molecules diffusing in unit time through the surface indicated in the first form of Fick's law, n_0 is the molecular concentration, equal to the number of molecules per unit volume, and m is the mass of one molecule, so that $\rho = n_0 m$.

The expression obtained in the kinetic theory of gases for the diffusion coefficient is

$$D = \frac{1}{3} \langle u \rangle \langle \lambda \rangle,$$

where $\langle u \rangle$ is the arithmetic mean velocity of thermal motion of the molecules (10.3.6), and $\langle \lambda \rangle$ is the mean free path (10.5.1).

10.8.4 *Internal friction (viscosity)* is the occurrence of friction forces between layers of a gas or liquid that flow parallel to one another at different speeds. A layer flowing faster exerts an accelerating force on a slower layer. On the other hand, a slow layer retards a faster flowing layer of fluid. The forces of internal friction (viscosity) are directed tangent to the contacting surfaces of the layers. The cause of viscosity (as internal friction in gases and liquids is more commonly called) is the superposition of ordered motion of layers of fluid at various velocities **v** and the thermal random motion of the molecules at velocities depending upon the temperature. The random motion of the molecules transfers them from layer B flowing at the velocity $\mathbf{v}_2$ into layer A flowing at the velocity $\mathbf{v}_1$ (Fig. 10.8). This leads to transport of momenta $m\mathbf{v}$ of ordered motion of the molecules. When $\mathbf{v}_1 > \mathbf{v}_2$, the molecules that entered layer A from layer B have their velocities of ordered motion accelerated by collisions with the molecules from the

new layer. This slows down the velocities of ordered motion
of the molecules in layer A. On the other hand, molecules
passing from faster layer A to slower layer B have higher mo-
menta $m v_1$ so that the inter-
molecular collisions in layer B
accelerate the motion of its mol-
ecules.
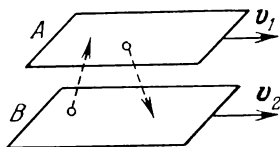
The phenomenon of viscosity is
described by the *Newtonian fric-
tion law*:

$$\tau = -\eta \, \frac{dv}{dn} \, ,$$

Fig. 10.8

where τ is the *friction stress*, i.e.
the physical quantity equal nu-
merically to the force of internal friction acting per unit
area of the surface of the layer, dv/dn is the velocity gradient,
i.e. the variation in the velocity of layer motion per unit length
in the direction of the inner normal n to the layer surface.
The internal frictional force is opposite in direction to the
derivative of the gas velocity vector v with respect to n.
The quantity η is called the *coefficient of dynamic viscosity*
(or just the *viscosity*). It is numerically equal to the friction
stress at a velocity gradient equal to unity.
The *coefficient of kinematic viscosity* is the quantity $v = \eta/\rho$,
where ρ is the density of the substance.
The coefficient of dynamic viscosity is calculated by the equa-
tion

$$\eta = \frac{1}{3} \, \langle u \rangle \, \langle \lambda \rangle \, \rho,$$

where $\langle u \rangle$ is the arithmetic mean velocity of thermal motion
of the molecules (10.3.6), and $\langle \lambda \rangle$ is their mean free path
(10.5.1). The coefficient of dynamic viscosity is independent
of the pressure (or density) of the gas since $\langle \lambda \rangle \propto 1/\rho$. This
can be explained by the fact that in an isothermal increase
in the density of a gas, for example, by two times, the number
of carriers of momenta is doubled, but each molecule (or atom)
travels only one half the distance without a collision and carries
only one half as much momentum. Hence, all in all, momenta
transport remains unchanged.

10.8.5 The third transport phenomenon, *heat conduction*, occurs under the condition that a difference in temperature exists in a body in a certain direction. For instance, the two opposing walls of a vessel holding a gas can have different temperatures that are maintained by external sources. Then the gas molecules at various locations throughout the volume have different average kinetic energies (10.2.4). Under such conditions, the random thermal motion of the molecules leads to directional energy transport (or transfer) in the form of heat (9.2.6). Molecules passing from the heated parts of the volume of gas to the colder parts yield some of their average kinetic energy to the surrounding molecules in the process of molecular collisions. Slow molecules, on the contrary, increase their average kinetic energy, in passing from the less heated to the more heated parts of the gas, owing to collisions with molecules having higher velocities.

In one-dimensional heat conduction, when the temperature of the gas depends only on a single coordinate $T = T(x)$, energy transfer in the form of heat proceeds along the OX axis according to the *Fourier law* of heat conduction (Fourier heat equation):

$$q_{\text{sec}} = -K \, \frac{dT}{dx} \, ,$$

where q_{sec} is the *specific rate of heat flow*, a physical quantity numerically equal to the energy transmitted in the form of heat in unit time through a plane surface of unit area located perpendicular to the direction of energy transport. The quantity K is called the *thermal conductivity*. It is numerically equal to the specific rate of heat flow at the temperature gradient dT/dx (the change in temperature over unit length) equal to unity. The minus sign in the Fourier law indicates that in heat conduction energy is transported in the direction of decreasing temperatures.

According to the kinetic theory of gases, the thermal conductivity is equal to

$$K = \frac{1}{3} \langle u \rangle \langle \lambda \rangle c_V \rho,$$

where c_V is the specific heat at constant volume (9.5.4). The other quantities are defined in Sect. 10.8.4. It follows

from the last equation that the thermal conductivity is inde-
pendent of the gas density. The reasons are similar to those
given in Sect. 10.8.4 to explain why η is independent of the
density.

10.8.6 The laws of transport phenomena, given in the subsec-
tions 10.8.3, 10.8.4 and 10.8.5, can be written in another form:

$$dM = -D\,\frac{d\rho}{dx}\,dS\,dt, \quad dF = -\eta\,\frac{dv}{dn}\,dS \quad \text{and}$$

$$dQ = -K\,\frac{dT}{dx}\,dS\,dt.$$

Here dM is the mass transported in diffusion during the time
dt through the element of area dS, located perpendicular to
the direction of diffusion; dQ is the amount of energy transported
in the form of heat during the time dt through the same element
of area dS, located perpendicular to the OX axis; and dF is
the force of internal friction (viscous force) exerted on an ele-
ment of surface of the layer having the area dS. The other
quantities are defined in the subsections mentioned above.

TABLE 10.1

Phenomenon	Transported physical quantity	Transport equation	Formula for the transport coefficient
Diffusion	Mass	$dM = -D\dfrac{d\rho}{dx}\,dS\,dt$	$D = \dfrac{1}{3}\,\langle u\rangle\,\langle\lambda\rangle$
Internal friction (viscosity)	Momentum	$dF = -\eta\dfrac{dv}{dn}\,dS$	$\eta = \dfrac{1}{3}\,\langle u\rangle\,\langle\lambda\rangle\,\rho$
Heat conduction	Energy in the form of heat	$dQ = -K\dfrac{dT}{dx}\,dS\,dt$	$K = \dfrac{1}{3}\,\langle u\rangle\,\langle\lambda\rangle\,c_V\rho$

10.8.7 Simple relations exist between the transport coefficients:

$$\eta = \rho D \quad \text{and} \quad \frac{K}{\eta c_V} = 1.$$

If any one of the transport coefficients is known the others can be found by these formulas (for known values of ρ and c_V). When the coefficients are known, such vital characteristics of a gas can be determined as the mean free path (10.5.1) of its molecules and the effective diameter of the molecules (8.4.1). Information on transport phenomena is listed in Table 10.1.

10.9 Properties of Rarified Gases

10.9.1 A gas in said to be *rarified* (a *dilute gas*) if its density is so low that the mean free path $\langle \lambda \rangle$ of the molecules is comparable to the linear dimensions d of the vessel confining the gas. Such a state of a gas is also called a *vacuum*. The degrees of vacuum of gases are classified as ultrahigh ($\langle \lambda \rangle \gg d$), high ($\langle \lambda \rangle > d$), medium ($\langle \lambda \rangle \leqslant d$) and low ($\langle \lambda \rangle \ll d$). The properties of rarified gases differ from those of nonrarified gases only for the first three of these degrees of vacuum. This is evident from Table 10.2 which lists certain characteristics for the various degrees.

10.9.2 *In the state of a high vacuum* (10.9.1) a reduction in the density of a rarified gas leads to a corresponding decrease in the number of particles without changing $\langle \lambda \rangle$. This, consequently, reduces the number of carriers of momentum or internal energy in viscosity and heat conduction. The transport coefficients in these phenomena are directly proportional to the gas density (cf. Sects. 10.8.4 and 10.8.5). Internal friction is essentially absent in highly rarified gases. Instead, only external friction exists between the moving gas and the walls of the vessel. This is due to the fact that the molecules change their momentum only in interacting with the walls. Under these conditions, as a first approximation, the friction stress is proportional to the density of the gas and velocity of its motion (cf. Sect. 10.8.4). The specific rate of heat flow in highly rarified gases is proportional to the temperature difference and to the gas density (cf. Sect. 10.8.5).

TABLE 10.2

Characteristic	Vacuum			
	Low	Medium	High	Ultrahigh
Pressure, mmHg	760 to 1	1 to 10^{-3}	10^{-3} to 10^{-7}	10^{-8} and less
Number of molecules in unit volume (per m^{-3})	10^{25} to 10^{22}	10^{22} to 10^{19}	10^{19} to 10^{13}	10^{13} and less
Dependence of coefficients K and η on the pressure	None	Pressure dependence is determined by the parameter $\langle\lambda\rangle/d$	Directly proportional to pressure	Heat conduction and viscosity are practically absent

10.9.3 A steady state of a rarified gas, contained in two vessels connected by a narrow tube, is possible under the condition of equality of the counterflow of particles passing between the vessels: $n_1 \langle u_1 \rangle = n_2 \langle u_2 \rangle$, where n_1 and n_2 are the numbers of molecules per cm^3 in the two vessels, and $\langle u_1 \rangle$ and $\langle u_2 \rangle$ are their arithmetic mean velocities (10.3.6).
If T_1 and T_2 are the gas temperatures in the two vessels, the preceding condition for a steady state can be rewritten in the form of an equation that expresses the *Knudsen effect*:

$$\frac{p_1}{p_2} = \sqrt{\frac{T_1}{T_2}} \, ,$$

where p_1 and p_2 are the pressures of the rarified gas in the two vessels.

CHAPTER 11   SECOND LAW OF THERMODYNAMICS

11.1 Cycles. The Carnot Cycle

11.1.1 A *cycle* is a succession of thermodynamic processes (8.3.7) as a result of which the system returns to its initial state. In pV, pT and other thermodynamic diagrams, equilibrium cycles (8.3.7) are represented in the form of closed curves, because two identical states, the beginning and end of the cycle, correspond to the same point of the diagram. A thermodynamic system that accomplishes a cycle and exchanges energy with other bodies is called the *working medium*, or *substance*. This medium is usually a gas.

Fig. 11.1

11.1.2 The arbitrary equilibrium cycle $C_1aC_2bC_1$ (Fig. 11.1), using an ideal gas, can be divided into the process of expansion of the gas from state C_1 to state C_2 (curve C_1aC_2) and the process of compressing the gas from state C_2 to state C_1 (curve C_2bC_1). In expansion the gas does positive work W_1, measured by the area under the curve C_1aC_2 (the area of the figure $V_1C_1aC_2V_2$; see Sect. 9.4.2). The gas is compressed by external forces that do positive work $W_2' = -W_2$, measured by the area under the curve C_1bC_2 (the area of the figure $V_1C_1bC_2V_2$). Since $W_1 > W_2'$ (Fig. 11.1), the gas does the *positive* work $W = W_1 + W_2 = W_1 - W_2'$ in the cycle. This work is measured by the hatched area, bounded by the curve $C_1aC_2bC_1$ of the process.
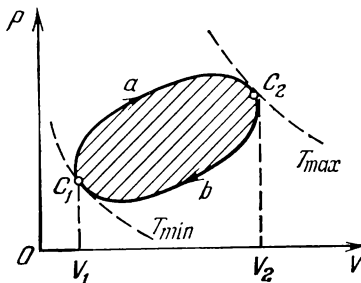
11.1.3 A *direct cycle* is one in which the system does positive work $W = \oint p\, dV > 0$. In a pV diagram a direct cycle is represented by a closed curve along which the working medium passes in the clockwise direction (Fig. 11.1). An example of a di-

rect cycle is that accomplished by the working medium in a
heat engine. In such an engine the working medium (11.1.1)
receives energy in the form of heat (9.2.1) from external sources
and gives a part of it up in the form of work (9.2.1) (see Sect.
11.1.5).

11.1.4 A *reverse cycle* is one in which the system does negative
work $W = \oint p \, dV < 0$. In a pV diagram a reverse cycle is
represented by a closed curve along which the working medium
passes in the counterclockwise direction. An example of a re-
verse cycle is that accomplished by the working medium in a
refrigerator. In such a device the working medium receives
energy in the form of work and
transmits energy in the form of
heat from a cold body to a hot-
ter one (see Sect. 11.1.5).
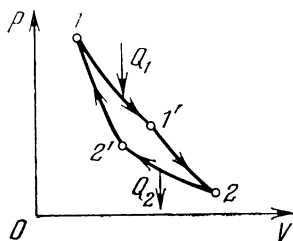
11.1.5 Since the total change
in the internal energy of a gas
in a cycle is equal to zero
(9.1.3), the first law of thermo-
dynamics (9.3.1) for such a
process is of the form

$$Q = \Delta U + W = W,$$

where Q is the total quantity of
heat delivered to the gas in the
cycle, and W is the work done
by the gas in this cycle.

Fig. 11.2

In a direct cycle (11.1.3) $Q > 0$ and $A > 0$, and the gas does
work owing to the heat delivered to it. In a reverse cycle (11.1.4)
the work done *on the gas* is $W' = -W$ (where $W > 0$) and a
quantity of heat equivalent to this work is removed from the gas.

11.1.6 The *Carnot cycle* is the one shown in Fig. 11.2. The
direct Carnot cycle consists of four consecutive processes: isother-
mal expansion 1-1' at the temperature T_1 (where $T_{1'} = T_1$),
adiabatic expansion 1'-2, isothermal compression 2-2' at
the temperature T_2 (where $T_{2'} = T_2$), and adiabatic compres-
sion 2'-1.

11.1.7 In practice a direct Carnot cycle could be carried out by
means of a gas enclosed in a vessel with a movable piston. In
the process 1-1' the gas is in thermal contact with a *heat*

source (*heater*) at the constant temperature T_1^*. From the heater
the gas receives a certain quantity of heat Q_1 (where $Q_1 > 0$).
It is assumed that this does not change the temperature of the
heater. This is possible, strictly speaking, only if the heater
has an infinite heat capacity. In the process 1'-2 the gas is
thermally isolated and its expansion is adiabatic** (9.5.10).
In the process 2-2' the gas is brought into thermal contact
with a *cooler* (*heat sink*) having the constant temperature T_2
(where $T_2 < T_1$). Here the gas is isothermically compressed
and gives off the quantity of heat $-Q_2$ to the cooler (assuming
that Q is the quantity of heat received by the gas from the cool-
er). The heat capacity of the cooler is assumed to be infinitely
high. In the state 2' the gas is again thermally insulated and
contracts adiabatically to its initial state 1.
11.1.8 The work done by the gas in an equilibrium (8.3.7)
direct Carnot cycle is

$$W = Q = Q_1 + Q_2 = Q_1 - |Q_2|.$$

It is evident from the equation that $W < Q_1$, i.e. the work done
by the working medium (11.1.1) in the Carnot cycle is less
than the energy received from the heater by the amount of
energy given up to the cooler in the form of heat. This is valid
for an arbitrary process: the work W done in the cycle is always
less than the sum Q_{del} of all the quantities of heat delivered to
the working medium from heaters.
The operation of heat engines is based on the direct Carnot
cycle.
11.1.9 The *thermal efficiency* η is the ratio of the work W done
by the working medium in a direct cycle to the sum Q_{del} of all
quantities of heat delivered in this cycle to the working medium
by heaters:

$$\eta = \frac{W}{Q_{del}} .$$

The quantity η characterizes the economy of operation of a heat
engine.

---

* For instance, a large tank filled with water.
** For instance, the cylinder with the gas is covered with
a thick layer of felt.

2*

11.1.10 The thermal efficiency of a direct equilibrium Carnot cycle carried out with an ideal gas is

$$\eta_C = \frac{Q_1 + Q_2}{Q_1} = \frac{T_1 - T_2}{T_1} = 1 - \frac{T_2}{T_1}.$$

The value of η_C depends only on the ratio of the temperatures of the cooler (T_2) and the heater (T_1).

11.1.11 The *reverse Carnot cycle* is shown in Fig. 11.3. In isothermal compression, carried out in the process 1'-1, the quantity of heat Q_1 is removed from the gas at the temperature T_1, which remains constant. In the process of isothermal expansion 2'-2 at the temperature $T_2 < T_1$, the quantity of heat Q_2 is delivered to the gas. In the reverse Carnot cycle $Q_1 < 0$, $Q_2 > 0$ and the work W done by the gas in one cycle is negative: $W = (Q_1 + Q_2) < 0$. This conclusion is valid for any reverse cycle. If the working medium carries out a reverse cycle, energy can be transferred in the form of heat (9.2.1) from a cold body to a hotter one at the expense of the corresponding work done by external forces. This principle is used in all refrigerating devices. The less the work $W' = -W$ done by external forces to extract the quantity of heat Q_2 from a cold body, the higher the economical efficiency (coefficient of performance) of a refrigerator:
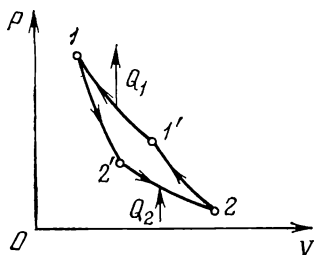
Fig. 11.3

$$Q_2 = W_1 - Q_1 = W - \frac{W}{\eta} = -\frac{1-\eta}{\eta} W,$$

where η is the thermal efficiency of the direct cycle between the same temperatures T_1 and T_2 (11.1.3 and 11.1.4).

11.2 Reversible and Irreversible Processes

11.2.1 A thermodynamic process is said to be reversible (*reversible process*) if the thermodynamic system (8.3.1), in carrying out the process first in the direct and then in the exactly reversed order, returns, together with all the external bodies with which it interacts, to its initial state. In other words, in a reversible process the thermodynamic system returns to its initial state without leaving any changes in the surrounding medium. A necessary and sufficient condition for the reversibility of a thermodynamic process is the equilibrium (8.3.3) of all the consecutive states in the process. Here all corresponding states of the thermodynamic system in the direct and reverse processes are absolutely identical.

Example 1. The mechanical motion of a body in a vacuum in the complete absence of friction is a reversible process. Assume, for instance, that a body under these conditions is thrown with a certain initial velocity into a gravitational field (6.2.1) at a definite angle to the horizon. After describing a parabolic path the body drops to the earth at a certain place. Now, if the body is thrown from this place at the same angle and the same initial velocity, but in the opposite direction, it describes the same path in the opposite direction and drops to the earth at the initial point. Any intermediate states of the moving body in the forward and reverse motions are entirely identical. The *reversibility of mechanical motions* indicates their symmetry with respect to the interchanging of the future by the past, i.e. with respect to the reversal of the sign of the time. The reversibility of mechanical motions follows from the differential equations of motion (2.4.4). When the sign of the time is changed, this changes the sign of the velocity as well, but the acceleration, in the motion equations, retains its sign.

Example 2. Another reversible process is the undamped vibrations of a body suspended in a vacuum on a perfectly elastic spring (28.2.3). The body-spring system is a conservative one (3.1.7). Its mechanical vibrations do not change the energy of thermal random motion of its particles. Only changes in the configuration and velocity of the system lead to a change of its state. But these changes are completely repeated after each period T (28.1.2) of vibration. Hence, the conditions for the reversibility (11.2.1) of the process are complied with.

11.2.2 Any process that does not comply with the conditions

for reversibility is said to be *irreversible (irreversible process).*
Example 1. The direct process of braking a moving body by
means of friction forces is an irreversible process. If these forces
are the only ones acting on the body, its velocity is reduced
and it finally stops. The energy of mechanical motion of the
body as a whole is reduced and is expended in increasing the
energy of random motion of particles of the body and the sur-
rounding medium. The internal energy (9.1.2) of the body and
the medium is increased; they are heated by the action of the
friction forces. The direct process dealt with here proceeds spon-
taneously; it is carried out without any processes taking place
in surrounding bodies. For the reverse process to occur, return-
ing the system to its initial state, it would be necessary for
the braked and stopped body to start moving again by cooling
it and the surrounding medium. As is evident from experiments,
the thermal random motion of particles of a body cannot spon-
taneously lead to the initiation of ordered motion of all the
particles of the body as a whole. To accomplish such motion it
is necessary to provide for an additional *compensating process* in
which the body and surrounding medium are cooled to the
initial temperature. In this process, the quantity of heat Q is
given up to the cooler and the work $A' = Q$ is performed on
the body.
Thus, the consecutive performance of such direct and reverse
processes returns the body-medium system to its initial state,
but the states of external bodies are changed. Hence, all pro-
cesses accompanied by friction are irreversible ones.
Example 2. The direct process of heat exchange (9.2.4) between
two bodies having different temperatures is a spontaneous one.
The reverse process, heating one body by cooling another that
initially has the same temperature as the first body, cannot
proceed spontaneously. To accomplish this process, a refrigerat-
ing device (11.1.11) is required. The process of heat exchange
at a finite temperature difference is an irreversible one.

11.3 Second Law of Thermodynamics

11.3.1 The reference in the preceding section to experiments
that prove the irreversibility of the processes of heat exchange
and motion with friction (11.2.2) was no chance one. The first
law of thermodynamics (9.3.1) cannot describe thermodynamic

processes comprehensively. An essential limitation of the first law is that it cannot be used to predict in which direction a thermodynamic process will proceed. Any process that does not violate the law of conservation of energy is feasible from the point of view of the first law of thermodynamics. Among others, the spontaneous transfer of energy in the form of heat from a cooler body to a hotter body is also possible. Also feasible is a process whose only result would be the obtaining of a certain quantity of heat from a body and the conversion of this heat into an equivalent amount of work. A periodic-action device, based on the first law of thermodynamics, which performs work by cooling a single source of heat (for instance, the internal energy of large bodies of water), is called a *perpetual motion machine of the second kind.*

11.3.2 The *second law of thermodynamics* is the experimentally established statement of the impossibility of building a perpetual motion machine of the second kind (11.3.1). The two most common statements of the second law, equivalent to each other, are:

(a) no process is possible whose sole result is the conversion of all the heat received from a certain body into an equivalent amount of work;

(b) no process is possible whose sole result is the transfer of energy in the form of heat from a cold body to a hotter one.

11.3.3 It follows from the second law of thermodynamics that work and heat are not equivalent forms of transferring energy. The conversion of orderly motion of a body as a whole into the random motion of its particles is an irreversible process, proceeding without compensating processes (11.2.2). The conversion of random motion of the particles of a body into ordered motion of the body as a whole requires the provision of some simultaneous compensating process*.

Example 1. In the isothermal expansion of an ideal gas work is done that is fully equivalent to the quantity of heat delivered to the gas (9.5.9). The heat received by the gas is completely converted into an equivalent amount of work. But the gas does not return to its initial state. It expands and its specific volume is increased. The conversion of heat into work* is not the sole result of the isothermal expansion of an ideal gas.

---

* Such a process is sometimes, not quite correctly, called the "conversion of heat into work".

Example 2. In a heat engine based on the direct Carnot cycle (11.1.8), work is done owing to the heat delivered from the heater. But a part of the quantity of heat obtained is given up to the cooler (11.1.7). Hence, the work done in a cycle is not equivalent to the whole quantity of heat delivered to the working medium.

Example 3. In refrigerating devices operating on the reverse Carnot cycle (11.1.11), a certain quantity of heat is transferred from a cold body to a hotter one. But external forces do work to accomplish this and, consequently, a compensating process is being carried out.

11.3.4 The *Carnot theorem* states that the thermal efficiency (11.1.9) of the reversible Carnot cycle does not depend upon the composition of the working medium and is always expressed by the formula (11.1.10):

$$\eta_C = \frac{Q_1 + Q_2}{Q_1} = \frac{T_1 - T_2}{T_1}.$$

The thermal efficiency $\eta_{C\,ir}$ of the irreversible Carnot cycle is always less than the thermal efficiency $\eta_{C\,rev}$ of the reversible Carnot cycle, carried out at the same temperatures T_1 and T_2: $\eta_{C\,ir} < \eta_{C\,rev}$.

The thermal efficiency of any reversible cycle cannot be higher than that η_C of a reversible Carnot cycle being accomplished by means of a heater and cooler having the temperatures T_1 and T_2:

$$\eta_{rev} \leqslant \eta_C = \frac{T_1 - T_2}{T_1}.$$

11.3.5 The Carnot theorem (11.3.4) is used to establish the *thermodynamic temperature scale.* It follows from the first equation in the preceding subsection that $T_2/T_1 = -Q_2/Q_1$ or, since $Q_2 < 0$,

$$\frac{T_2}{T_1} = \frac{|Q_2|}{Q_1}.$$

In order to compare the temperatures T_1 and T_2 of two bodies it is necessary to carry out a reversible Carnot cycle in which these bodies are the heater and cooler. Then, according to the ratio of the numerical values of the quantities of heat delivered

to and received from the bodies, the ratio of their temperatures is determined. The result of such a comparison of temperatures does not depend upon the chemical composition of the working medium (11.3.4). Therefore, the thermodynamic temperature scale is independent of the properties of the temperature-indicating substance (8.3.4) and, in this respect, has great generality. However, since all real thermodynamic processes are irreversible, the comparison of temperatures by the above-mentioned procedure is unfeasible; it has only a theoretical significance.

11.4 Entropy and Free Energy

11.4.1 The ratio of the quantity of heat Q, received by a body in an isothermal process, to the temperature T of the body giving up the heat is called the *reduced quantity of heat* Q^*, i.e. $Q^* = Q/T$. In heating a body ($Q > 0$), Q^* is positive; in cooling ($Q < 0$), Q^* is negative.

The reduced quantity of heat delivered to a body in an infinitely small portion of an arbitrary process, is equal to $\delta Q/T$, where T is the temperature of the corresponding heater*. The reduced quantity of heat Q^*_{1-2} for the arbitrary portion $1 \to 2$ of the process $C_1 C_2$ is

$$Q^*_{1-2} = \int_{C_1}^{C_2} \frac{\delta Q}{T} \cdot$$

11.4.2 The reduced quantity of heat Q^*_{rev} delivered to a body in any reversible cycle equals zero:

$$Q^*_{\text{rev}} = \oint_{\text{rev}} \frac{\delta Q}{T} = 0.$$

Here T is the temperature at which an element of the quantity of heat δQ is delivered to the body. It follows from the equation in the preceding subsection that the function $\delta Q/T$ is a com-

---

* In the case of a reversible process, T coincides with the temperature of the body that is carrying out the process.

plete differential of a certain function S [in contrast to δQ, which is not a complete differential (9.4.3)]:

$$dS = \left(\frac{\delta Q}{T} \right)_{\text{rev}}.$$

The single-valued state function (9.1.3) S, whose complete differential is determined by the preceding equation, is called the *entropy* of a body. It is evident from this equation that dS and δQ have the same sign. Consequently, from the way in which the entropy varies it is possible to determine in which direction heat exchange (9.2.4) takes place. In heating a body ($\delta Q > 0$) its entropy increases ($dS > 0$). If a body is cooled ($\delta Q < 0$), its entropy decreases ($dS < 0$).

Example. The complete differential of the entropy of an ideal gas is expressed by the equation

$$dS = \left(\frac{\delta Q}{T} \right)_{\text{rev}} = \frac{M}{\mu} \, C_{V\mu} \, \frac{dT}{T} + \frac{M}{\mu} \, R \, \frac{dV}{V} \, ,$$

where M is the mass of the gas, μ is its molar mass (8.4.3), $C_{V\mu}$ is its molar heat capacity at constant volume (9.5.4), R is the universal gas constant (8.4.4), T is its temperature, and V is its volume. This result is obtained on the basis of the first law of thermodynamics (9.3.1) for δQ, with the Mendeleev-Clapeyron equation (8.4.4) taken into account.

The change $\Delta S_{1 \to 2}$ in the entropy of an ideal gas going over from state 1 to state 2 does not depend upon the kind of process $1 \to 2$ used in this transition:

$$\Delta S_{1 \to 2} = S_2 - S_1 = \frac{M}{\mu} \left(C_{V\mu} \ln \frac{T_2}{T_1} + R \ln \frac{V_2}{V_1} \right).$$

11.4.3 Certain of the most important properties of the entropy of closed systems (2.2.4) are:

(a) The entropy of a closed system in a reversible Carnot cycle (11.1.6) does not change:

$$\Delta S_{\text{rev}} = 0 \text{ and } S = \text{const.}$$

(b) The entropy of a closed system increases in an irreversible Carnot cycle:

$$\Delta S_{\text{ir}} > 0.$$

(c) In any processes that occur in a closed system its entropy never decreases:

$$\Delta S \geqslant 0.$$

In an element of change in the state of a closed system, the entropy does not decrease: $dS \geqslant 0$.
The equality sign pertains to reversible processes and the greater-than sign to irreversible ones. Item (c) is one of the statements of the second law of thermodynamics.
11.4.4 The following relation is valid for an arbitrary process occurring in a thermodynamic system (8.3.1):

$$\delta Q \leqslant T \, dS,$$

where T is the temperature of the body that transfers the energy δQ to the thermodynamic system in the process of an infinitely small change in the state of the system. Making use of an equation of the first law of thermodynamics (9.3.1) for δQ, the preceding inequality can be rewritten in the form

$$T \, dS \geqslant dU + \delta W,$$

combining the first and second laws.

11.4.5 For a reversible process

$$\delta W = -(dU - T \, dS)$$

or

$$\delta W = -d \, (U - TS) - S \, dT = -dF - S \, dT,$$

where $F = U - TS$

is called the *free energy*. The free, or available, energy is the difference of two state functions (8.3.8) and is therefore also a state function of the thermodynamic system.
If the system carries out a reversible isothermal process, $dT = 0$ and $\delta W_{isot} = -dF$. When the system passes from state 1 to state 2 in a reversible isothermal process,

$$W_{isot} = F_1 - F_2.$$

The decrease in free energy is the measure of the work done by a system (or body) in a reversible isothermal process.

11.4.6 It follows from the equation $U = F + TS$ that the internal energy of a body (or system) is equal to the sum of the free energy F and the *bound energy* TS. The bound, or unavailable, energy is the part of the internal energy of a body (or system) that cannot be transferred in the form of work in an isothermal process. In this sense, this part of the internal energy is "depreciated". The more the entropy of the body (or system), the more the bound energy. Therefore, the entropy of a body (or system) serves as a measure of the "depreciation" of its energy.

11.5 Statistical Interpretation of the Second Law of Thermodynamics

11.5.1 The statement of the second law of thermodynamics on the impossibility of a decrease in the entropy in an isolated system (11.4.3) can be interpreted statistically, on the basis of the molecular-kinetic theory of the structure of matter, by means of Boltzmann's relation:

$$S = k \ln P + \text{const,}$$

where S is the entropy of the system, k is Boltzmann's constant (8.4.5), and P is the thermodynamic probability of the state. **11.5.2** The *thermodynamic probability* P of a state of a body (or system) is equal to the number of all possible kinds of coordinate and velocity distributions of its particles corresponding to the given thermodynamic state (8.3.3). By definition, P is a whole number not less than unity $(P \geqslant 1)$. The Boltzmann relation (11.5.1) leads to the following statistical interpretation of the second law of thermodynamics: the thermodynamic probability of a state of a closed system cannot decrease, regardless of the processes that occur in it. In any process that proceeds in a closed system and brings it from state 1 to state 2, the change ΔP in the thermodynamic probability is either positive or zero: $\Delta P = P_2 - P_1 \geqslant 0$.

In a reversible process $\Delta P = 0$, i.e. the thermodynamic probability P is constant. If the process is irreversible, $\Delta P > 0$ and P increases. This means that an irreversible process transfers the system from a less probable state to a more probable one and, in the limit, to an equilibrium state (8.3.3).

11.5.3 Being a statistical law, the second law of thermodynamics describes the laws of random motion of a great number of particles making up a closed system. In systems that consist of a relatively small number of particles, fluctuations (11.6.1) are observed. They are deviations from the second law of thermodynamics.

11.5.4 The second law of thermodynamics, established for closed systems on the earth, cannot be applied to the whole infinite universe. Such an application leads to the conclusion that the temperatures of all bodies in the universe equalize. This is incorrect from both the philosophical and physical points of view. According to this erroneous conclusion, all forms of motion, except random thermal motion, will cease and the universe will then reach a state of so-called "heat death". Actually, owing to the infinity of the universe, fluctuations (11.6.1) are inevitable in certain of its parts. These violate thermal equilibrium. The duration and magnitude of the fluctuations can be extremely great. It has been proven that there can be no equilibrium state, corresponding to "heat death", for the infinite universe.

11.6 Fluctuations

11.6.1 Substantial deviations of certain physical quantities, characterizing a system, from their average values may occur in systems consisting of a comparatively small number of particles. Such deviations are called *fluctuations* of physical quantities. For instance, in highly rarified gases the density at various places in the volume of the gas may differ from the average density corresponding to the equilibrium state for definite p and T values. Chance deviations of the temperature T, pressure p and other physical quantities may occur in exactly the same way.

11.6.2 If M is the true value of a physical quantity, and $\langle M \rangle$ is its average value, then the quantity ΔM, where $\Delta M = M - \langle M \rangle$, and its average value $\langle \Delta M \rangle = \langle M - \langle M \rangle \rangle$ cannot be measures of the fluctuation of the quantity M. The quantity ΔM is not constant with time, and the quantity

$$\langle \Delta M \rangle = \langle M \rangle - \langle M \rangle = 0*.$$

* Used here is the statement that the mean value $\langle M \rangle$ of a constant quantity coincides with the quantity.

This last equation follows from the fact that the deviation of the quantity M from $\langle M \rangle$ occurs in both directions, towards values larger than the average and towards values smaller than the average, with equal frequency.

11.6.3 A measure of the fluctuations of the physical quantity M is the mean square of the difference ΔM. It is called the *mean square fluctuation*:

$$\langle (\Delta M)^2 \rangle = \langle (M - \Delta M)^2 \rangle = \langle M^2 \rangle - (\langle M \rangle)^2.\text{*}$$

The mean square fluctuation is essentially positive or zero: $\langle (\Delta M)^2 \rangle \geqslant 0$.

The *absolute fluctuation* is the quantity $\sqrt{\langle (\Delta M)^2 \rangle}$, which also characterizes the deviation of M from $\langle M \rangle$. A small absolute fluctuation signifies that large deviations of M from $\langle M \rangle$ occur very rarely.

The *relative fluctuation* δ_M is the ratio of the absolute fluctuation to the average value $\langle M \rangle$ of a physical quantity:

$$\delta_M = \frac{\sqrt{\langle (\Delta M)^2 \rangle}}{\langle M \rangle}.$$

The relative fluctuations in the concentration of particles (or density) of a gas, and in its pressure and temperature, decrease with an increase in the number N of gas molecules contained in the vessel:

$$\delta_\rho = \frac{\sqrt{\langle (\Delta \rho)^2 \rangle}}{\langle \rho \rangle} \propto \frac{1}{\sqrt{N}}, \quad \delta_p = \frac{\sqrt{\langle (\Delta p)^2 \rangle}}{\langle p \rangle} \propto \frac{1}{\sqrt{N}},$$

$$\delta_T = \frac{\sqrt{\langle (\Delta T)^2 \rangle}}{\langle T \rangle} \propto \frac{1}{\sqrt{N}}.$$

When $N = N_A$ (Avogadro's number, see Appendix II), δ_ρ, δ_p and δ_T have values of the order of 10^{-14}.

---

* This last equation, following from the rules of algebraic operations with average values, underlines the fact that the average value $\langle M^2 \rangle$ of the square of a quantity should not be confused with the square of the average value $(\langle M \rangle)^2$.

If a system consists of N independent particles, the relative fluctuation of an additive state function (9.1.3) of the system is inversely proportional to the square root of N:

$$\delta_M \propto \frac{1}{\sqrt{N}} \, .$$

11.6.4 The following are examples of fluctuations of physical quantities.
Example 1. In measuring the temperature by means of a gas thermometer filled with an ideal gas (8.4.1), the thermometer readings are not constant owing to fluctuations of temperature. The variations in temperature measured by the thermometer cannot be less than the absolute fluctuation of the thermometer indication, equal to $\sqrt{\langle(\Delta T)^2\rangle}$, i.e. $\Delta t \geqslant \sqrt{\langle(\Delta T)^2\rangle}$.
The absolute fluctuation can be determined by the formula in 11.6.3:

$$\sqrt{\langle(\Delta T)^2\rangle} \propto \frac{\langle T \rangle}{\sqrt{N}} \, .$$

Thus

$$\Delta t \geqslant \sqrt{\langle(\Delta T)^2\rangle} \propto \frac{\langle T \rangle}{\sqrt{N}} \, .$$

If the gas thermometer contains 10^{-8} mole of gas, i.e. $N = 6.02 \times 10^{15}$, then the minimum changes in temperature Δt that can be detected by the instrument, are of the order of magnitude $\Delta t \approx 10^{-10} \langle T \rangle$. This value indicates the limit of sensitivity of the gas thermometer.
Real variations in temperature, commonly encountered in experiments, are incommensurably larger than Δt.
Example 2. Electrical fluctuations in circuits restrict the limits of sensitivity of radio receiving equipment. Among others, fluctuations in the number of electrons emitted by the incandescent cathode lead to fluctuations in the current supplied to the vacuum tube. This is called *shot*, or *Schottky*, *noise*. A measure of the shot noise is the mean square fluctuation of the current:

$$\langle(\Delta I)^2\rangle \approx \frac{e I_0}{t} \, ,$$

where e is the charge of the electron, and I_0 is the average current over the time t during which it is measured. Here $t \gg \tau$, where τ is the time of flight of the electron in the tube.

11.7 Brownian Movement

11.7.1 The *Brownian movement* is the continuous haphazard zigzag motion of small particles suspended in a gas or liquid and can be observed in a microscope. The Brownian movement is caused by fluctuations (11.6.1) in the pressure exerted by the molecules of gas or liquid on the surfaces of the minute suspended particles. As a result of pressure fluctuations the Brownian particles are subject on all sides to the action of unbalanced forces which cause their complex motions that are observed. **11.7.2** Under constant external conditions, no changes are observed in the motions of the Brownian particles and they continue indefinitely. This is testimony to the continuous nature of random thermal motion of the molecules to which the displacements of the Brownian particles are due. The velocity v and energy E of motion of these particles depend upon their size, but not on their chemical composition. The values of v and E increase with the temperature and with a reduction in the viscosity of the gas or liquid. **11.7.3** It has been experimentally established that Brownian particles can move upwards, rising in the fluid. This occurs when the molecules of the gas (or liquid) transmit an uncompensated momentum directed upwards to the particle. This increases the potential energy of the particle at the expense of the kinetic energy of the molecules surrounding it, thereby leading to localized cooling of the gas or liquid. This means that the mechanical energy of the Brownian particle has increased by cooling a single source of heat, the liquid or gas, thereby contradicting the second law of thermodynamics (11.3.2). In this manner, Brownian movement is proof of the limited application of the second law of thermodynamics and of its statistical nature (11.5.1). **11.7.4** The motions of Brownian particles are completely chaotic. Therefore, the average displacement $\langle x \rangle$ of a particle along an arbitrary direction equals zero. Then mean square displa-

cement $\langle x^2 \rangle$ is proportional to the time t of observation of the particle and is expressed by *Einstein's equation*:

$$\langle x^2 \rangle = 2Dt,$$

where D is the diffusion coefficient of the Brownian particles. For spherical particles of radius r

$$D = \frac{RT}{6\pi\eta r N_A} .$$

Here T is the absolute temperature, R is the universal gas constant (8.4.4), η is the coefficient of viscosity of the liquid or gas (10.8.4), and N_A is Avogadro's number (Appendix II).

11.8 Third Law of Thermodynamics

11.8.1 As a single-valued state function of a system entropy is introduced by means of a differential relation (11.4.2). Hence, the entropy can be determined up to an arbitrary constant, which cannot be found from the first and second laws of thermodynamics. It is consequently impossible to determine the absolute value of the entropy.

11.8.2 An experimental investigation of the properties of substances at ultralow temperatures led to the establishment of the *third law of thermodynamics*. This law, also called the *Nernst heat theorem*, states that the change in entropy equals zero in any isothermal process conducted at absolute zero:

$$\Delta S_{T \to 0} = 0 \text{ and } S = S_0 = \text{const}$$

regardless of the changes in any other state variables (for example, volume, pressure, and intensity of the external force field). The third law does not enable the absolute value of the entropy to be determined. But the fact that the entropy is constant as $T \to 0$ permits this constant value to be employed as the reference point for entropy values. This enables the change in entropy to be determined in processes being investigated.

11.8.3 Planck extended Nernst's heat theorem by stating that the entropy of a system equals zero at absolute zero temperature. If E_0, E_1, ..., E_n is a sequence of energy levels (38.2.5) of a system, then at absolute zero temperature the system in equilibrium is in a quite definite lowest state with energy exactly equal to E_0. Here the thermodynamic probability P of the

state (11.5.2) is equal to unity ($P = 1$) and, according to Boltz-mann's relation (11.5.1),

$$S_0 = k \ln P = 0.$$

11.8.4 The heat capacity $C_{V\mu}$ at constant volume (9.5.4) for all bodies vanishes at $T = 0$ K. As a matter of fact, if the temperature of a system is sufficiently low, so that the average kinetic energy kT of a particle (10.6.4) is substantially less than the difference ΔE between the lowest and first energy levels ($\Delta E \gg kT$), then the thermal excitation of the system is insufficient to transfer it from the state with the energy E_0 to one with the energy E_1. Therefore, at ultralow temperatures the system should be in a state having the least energy E_0. The internal energy U_0 of the system (9.1.2) is equal to E_0, i.e. $U_0 = E_0$. Hence the heat capacity of the system at constant volume

$$C_{V\mu} = \left(\frac{dU}{dT} \right)_V = \left(\frac{dE_0}{dT} \right)_V = 0 \quad \text{as} \quad T \to 0.$$

The coefficient of cubic expansion (40.2.3) also vanishes at absolute zero temperature.

It follows from the third law that a process in which a body could be cooled to absolute zero temperature is impossible (*principle of the unattainability of absolute zero temperature*, sometimes called the *Nernst principle*).

CHAPTER 12   REAL GASES AND VAPOURS

12.1 Forces of Intermolecular Interaction

12.1.1 The properties of gases that are not highly rarified differ from those of ideal gases, which comply with the Mendeleev-Clapeyron equation (8.4.4). Experiments indicate that the specific heat capacities (9.5.2), coefficients of viscosity (10.8.4) and other quantities have different values for real gases than for ideal gases.

12.1.2 A *real gas* is one in which forces of molecular interaction exist between the molecules.

A *vapour* is a real gas in states close to that in which it condenses.

12.1.3 The forces of intermolecular interaction decrease rapidly with an increase in the distances between the molecules (*short-range forces*). At intermolecular distances exceeding 10^{-9} m, these forces can be neglected. Forces of interaction between molecules can be subdivided into *forces of attraction* and *forces of repulsion*. Both types act simultaneously. Otherwise it would be impossible to determine the volume of liquids or solids: their constituent particles would either fly apart in all directions or would stick together, forming the minimum possible volume.

12.1.4 The forces of mutual attraction and repulsion depend differently on the distance r between the molecules. At distances commensurable with the linear dimensions of atoms and small inorganic molecules (10^{-10} m), repulsive forces F_1 predominate; at distances r of the order of 10^{-9} m forces of mutual attraction F_2 prevail. If **r** is the radius vector to the point where molecule A is located from a point where molecule B is located and acts on molecule A with the forces $\mathbf{F}_1$ and $\mathbf{F}_2$, then

Fig. 12.1

$$\mathbf{F}_1 = F_{1r}\,\frac{\mathbf{r}}{r} \quad \text{and} \quad \mathbf{F}_2 = F_{2r}\,\frac{\mathbf{r}}{r}.$$

Projections F_{1r} and F_{2r} of forces $\mathbf{F}_1$ and $\mathbf{F}_2$ on the direction of **r** depend upon r in the following way

$$F_{1r} = b/r^{13} \text{ and } F_{2r} = -a/r^7,$$

where a and b are factors that depend on the structure of the molecules and the type of intermolecular interaction. The curves in Fig. 12.1 show the dependence of F_{1r} and F_{2r} on r. Forces of repulsion are conventionally accepted as positive and attractive forces as negative (Fig. 12.1).
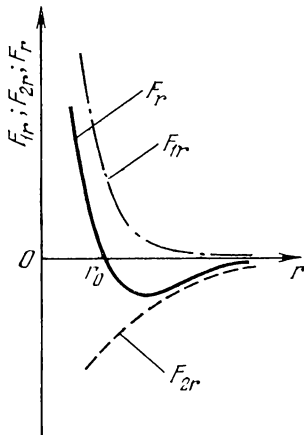
The resultant force is

$$\mathbf{F} = \mathbf{F}_1 + \mathbf{F}_2 = F_r \frac{\mathbf{r}}{r} \ ,$$

and $F_r = F_{1r} + F_{2r}$. The dependence of F_r on r is also shown in Fig. 12.1.

12.1.5 At $r = r_0$, forces $\mathbf{F}_1$ and $\mathbf{F}_2$ counterbalance each other and $\mathbf{F} = 0$. At $r > r_0$ force $\mathbf{F}_2 > \mathbf{F}_1$; at $r < r_0$ force $\mathbf{F}_2 < \mathbf{F}_1$. Thus r_0 is the equilibrium distance between the molecules and where they would be with respect to each other if there were no thermal motion.

12.1.6 The element of work δW done by the resultant force $\mathbf{F}$, when the distance between the molecules is increased by the amount $d\mathbf{r}$, is equal to the decrease in the *mutual potential energy* E_p of the two molecules (3.3.1):

$$\delta W = (\mathbf{F}, \ d\mathbf{r}) = F_r \ dr = -dE_p.$$

Integrating with respect to r from r to ∞:

$$\int_{E_p}^{E_{p\infty}} dE_p = - \int_r^\infty F_r \ dr \quad \text{and} \quad E_p - E_{p\infty} = \int_r^\infty F_r \ dr.$$

At $r = \infty$ the molecules do not interact and $E_{p\infty} = 0$. Hence,

$$E_p = \int_r^\infty F_r \ dr.$$

This integral can be calculated graphically from the known dependence of F_r on r (Fig. 12.1). It is proportional to the area bounded by the curve $F_r = F_r(r)$, the r axis and the value of r (where $r = \text{const}$) for which E_p is being calculated. At $r > r_0$, the energy $E_p < 0$ because $F_r < 0$; at $r = r_0$, the energy E_p reaches its minimum value: $E_p = E_{min}$. This follows from the equation

$$\left(\frac{dE_p}{dr} \right)_{r=r_0} = -F_r(r_0) = 0.$$

A system consisting of the interacting molecules, in a state of stable equilibrium $(r = r_0)$, possesses minimum potential energy. At $r < r_0$, the energy E_p begins to increase, becomes positive and then rises drastically due to the rapid increase in the repulsive forces as r drops (Fig. 12.2).

12.1.7 The value $E = E_{p\ min}$, the minimum potential energy of interaction of molecules, is the criterion for various states of aggregation of matter.

If $|E|_{p\ min} \ll kT$, the substance is in a gaseous state. When $|E|_{p\ min} \gg kT$, it is in the solid state. The condition
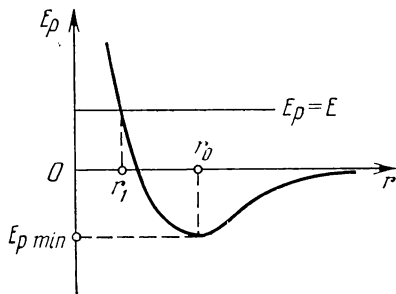
Fig. 12.2

$|E|_{p\ min} \approx kT$ corresponds to the liquid state. Here kT is twice the average energy per degree of freedom of thermal motion of the molecules (10.6.4).

12.1.8 *Van der Waals forces* are weak* forces of attraction that act between molecules at distances of the order of 10^{-9} m (12.1.4). They are the reason for the correction of the internal pressure in the Van der Waals equation of state for a real gas (12.2.4). Distinction is made between three types of Van der Waals forces of attraction.

(a) *Orientational forces* of attraction $\mathbf{F}_{or}$ are due to the fact that asymmetric polar molecules (18.1.4) have constant dipole

---

* Weak in the sense of their small magnitude compared to the forces of attraction that provide for the formation of stable molecules (12.1.9).

moments p_e (18.1.4). The result of interaction of these dipole moments is the attraction force F_{or}. This force is opposed by the thermal motion of the molecules. Calculations indicate that

$$F_{or} \propto -\frac{p_e^4}{kT}\frac{1}{r^7} ,$$

where r is the distance between the molecules, k is Boltzmann's constant (8.4.5), and T is the absolute temperature.

(b) *Induction forces* of attraction F_{in} are exerted when a neutral molecule of a real gas is in an electric field set up by another molecule, both molecules having high polarizability (18.1.3). If the molecules are sufficiently close together, the action of the electric field of the second molecule develops an induced dipole moment $p_e = \varepsilon_0 \beta E$ (18.1.4) in the first molecule, where β is the polarizability of the molecule (18.1.3), and E is the strength of the electric field set up by the second molecule. To constant factors

$$F_{in} \propto -\beta p_e^2 \frac{1}{r^7} .$$

(c) The vibration of electrons in one molecule (or atom) can excite vibration in another molecule (or atom). The vibrations of the electrons in the two molecules are in phase with each other, leading to attraction between the molecules (or atoms). The resulting resonance forces of attraction are called *dispersion forces*. They play the main role in the interaction of non-polar molecules (18.1.3). The force of dispersion attraction is

$$F_d \propto -\frac{e^4 h \nu_0}{k^2}\frac{1}{r^7} ,$$

where e is the charge of the electron, h is Planck's constant, $\nu_0 = (1/2\pi)\sqrt{k/m}$ is the frequency of vibration of the atoms (28.1.2), k is the coefficient of the quasi-elastic force (40.3.7), and m is the mass of the atom.

Dispersion forces owe their name to the analogy between the origin of these forces and dispersion phenomena occurring when electromagnetic waves propagate through matter (34.4.1).

12.1.9 The attraction force $\mathbf{F}_r$ between two molecules, shown in Fig. 12.1, is the resultant of all the kinds of attraction enumerated in Sect. 12.1.8. The potential energy of Van der Waals attraction amounts to $(0.4 \text{ to } 4) \times 10^3$ J/mol.

In addition to the electromagnetic interaction, a special kind of quantum interaction exists between molecules at distances $r \leqslant 10^{-10}$ m. This interaction leads either to the development of repulsive forces between the molecules (12.1.4), or to strong attraction of adjacent atoms (or groups of atoms) and the formation between them of ionic and covalent chemical bonds (39.4.4 and 39.4.5). The result of such bonds is the formation of stable molecules. The potential energy of chemical bonds exceeds the energy of Van der Waals attraction and ranges from $(0.4 \text{ to } 4) \times 10^4$ J/mol.

12.1.10 Being internal forces (2.2.4), the forces of interaction in a system of two molecules cannot change the total energy E of the system, which is the sum of the kinetic energy E_k of the molecules and their mutual potential energy E_p. Hence

$$dE = dE_k + dE_p = 0 \text{ or } dE_k = -dE_p = F_r \, dr.$$

Use is made here of the equation in Sect. 12.1.6.

When the molecules are brought closer together ($dr < 0$), to the distance r_0 (Fig. 12.1), the potential energy E_p decreases and the kinetic energy E_k accordingly increases. This occurs at the expense of the positive work done by the resultant force of mutual attraction between the molecules (12.1.4) ($F_r < 0$ when $r > r_0$). As the molecules are brought still closer, they do work in overcoming the resultant force of mutual repulsion ($F_r > 0$ when $r < r_0$). This reduces the kinetic energy of the molecules. At the instant they are closest together ($r = r_1$ in Fig. 12.2) all of the kinetic energy of the molecules has been completely expended in doing work to overcome the repulsive forces: $E_k = 0$ and the total energy E is equal to the potential energy E_p, i.e. $E = E_p$ (Fig. 12.2).

If all the state variables of real gases are maintained constant with the exception of the temperature, the distance r_1 decreases when the gas is heated. This decrease is very small, however, even at high temperatures. This is due to the extremely steep increase in the repulsive forces F_1 as r is decreased (Fig. 12.1). Thus the distance r_1 is the effective diameter d of the molecule

(8.4.1). The finite dimensions of molecules of real gases are explained by the action of repulsive forces between the molecules.

12.2 Van der Waals Equation of State

12.2.1 A *Van der Waals gas* is a model of a real gas in which the molecules are dealt with as perfectly rigid spheres of diameter d (12.1.10), with forces of mutual attraction acting between them. The finite size of the spheres signifies that the repulsive forces between the molecules of a real gas are also taken into account.

12.2.2 The molecules of a real gas, each having a volume of $\widetilde{v} = \pi d^3/6$, do not move around as freely in a vessel as the "point" molecules of an ideal gas. Therefore, in the Mendeleev-Clapeyron equation (8.4.4) $pV_\mu = RT$, the total volume V_μ of a vessel holding one mole of the gas should be replaced by the "free" volume

$$V'_\mu = V_\mu - b,$$

where b is Van der Waals's correction for the intrinsic volume of the molecules. Correction b is equal to four times the volume of all the molecules contained in one mole of the gas:

$$b = 4N_A\widetilde{v},$$

where N_A is Avogadro's number (Appendix II), and $\widetilde{v}$ is the volume of a single molecule.

12.2.3 The forces of mutual attraction between the molecules are taken into account for a Van der Waals gas by introducing a correction for the pressure in the Mendeleev-Clapeyron equation (8.4.4). Owing to the short-range nature of the attraction forces (12.1.3), each molecule interacts only with particles that are located at distances $r \leqslant R_m$, where R_m is the *range of molecular action*, with a magnitude of the order of 10^{-9} m. A sphere of radius R_m is called the *sphere of molecular action*.

The forces of attraction of a molecule within the bulk of the gas to other molecules are mutually counterbalanced and have no effect on the motion of the given molecule. If the molecule is within a layer of gas that borders on a wall of the vessel, it is subject to an uncompensated force of attraction directed in-

ward, into the bulk of the gas. Consequently, in its collision
with a wall, such a molecule imparts less momentum to the
wall (10.2.1), and the pressure p exerted by a real gas on the
walls is reduced in comparison to the pressure p_{id} of an ideal
gas having the same density at the same temperature,

$$p = p_{\mathrm{id}} - p^*,$$

or

$$p_{\mathrm{id}} = p + p^*,$$

where p^* is the Van der Waals correction due to the forces of
mutual attraction and is called the *internal pressure*. The inter-
nal pressure p^* is inversely proportional to the square of the
volume V_μ of the vessel occupied by one mole of the gas:

$$p^* = \frac{a}{V_\mu^2} \ ,$$

where a is the Van der Waals coefficient and depends upon the
chemical nature of the gas.
12.2.4 The Van der Waals equation describes the state of a real
gas and differs from the Mendeleev-Clapeyron equation (8.4.4)
in that the corrections b and p^* have been introduced (12.2.2
and 12.2.3). For a single mole of gas it is of the form

$$\left(p + \frac{a}{V_\mu^2} \right)(V_\mu - b) = RT.$$

12.2.5 The Van der Waals equation for an arbitrary mass of
real gas, having the molar mass μ (8.4.3), is

$$\left(p + \frac{M^2}{\mu^2}\, \frac{a}{V^2} \right)\left(V - \frac{M}{\mu}\, b \right) = \frac{M}{\mu}\, RT.$$

This equation is valid for gases that are not too highly com-
pressed. In highly rarefied gases $V_\mu \gg b$, $p^* \ll p$ and the Van der
Waals equation does not differ from the Mendeleev-Clapeyron
equation,

12.3 Isothermals of Real Gases. Phase Transitions

12.3.1 The curve representing the dependence of the molar volume of a gas on the pressure at constant temperature is called the *isothermal*, or *isotherm*, *of a real gas*. Isothermals of carbon dioxide gas are shown in Fig. 12.3. At temperatures T below $T_c = 340$ K, all the isothermals have horizontal steps along which, in addition to the constant temperature, the pressure is also constant while the molar volume varies. The difference $V_D - V_B$ of the molar volumes over the horizontal steps decreases as the temperature is raised (Fig. 12.3). At $T = T_c$ this difference vanishes. The temperature $T = T_c$, corresponding to the condition $V_D - V_B = 0$, is called the *critical temperature*. The isothermal of a real gas at $T = T_c$ is called the *critical* isothermal. On this isothermal points D and B merge into point C, which is called the *critical point*. The state variables (8.3.2) of the gas at the critical point are the critical variables $V_{\mu c}$, p_c and T_c.
Critical point C is an inflection point on the critical isothermal. Consequently, the tangent to the isothermal at this point is parallel to the OV_μ axis.
12.3.2 Any subcritical isothermal ($T < T_c$) is a curve representing the continuous transition of the substance from the gaseous to the liquid state. It consists of three portions: TD, DB and BA, each of which represents different states of the substance. Over the portion TD the substance is in the gaseous state, portion DB corresponds to the transition of the substance from the gaseous to the liquid state. Over the portion BA of the isothermal, the substance is a liquid. This portion is almost vertical owing to the low compressibility of liquids. Points D and B of the horizontal portion of the isothermal correspond to the beginning and end of condensation in isothermal compression of a real gas. On the contrary, in isothermal expansion of the liquid, points B and D are the beginning and end of the boiling process. Point B corresponds to the state of a *boiling liquid*; point D to *dry saturated vapour*. The mixture of the boiling liquid and dry saturated vapour that exists at any point M of portion BD is called wet vapour (Fig. 12.3).
12.3.3 In thermodynamics a *phase* is a totality of all the parts of a system, having the same chemical composition and being in the same state. Wet vapour, for instance, is a two-phase system consisting of boiling liquid and dry saturated vapour.

Shown in Fig. 12.4 are two boundary curves BC and DC which connect points B and D (Fig. 12.3) at various temperatures. Curves BC and DC converge at critical point C. Boiling curve BC separates the single-phase liquid region I from the two-phase wet vapour region II. The curve BC is one representing the beginning of the *first-kind phase transition* of the substance from the liquid to the gaseous state. The boundary curve DC
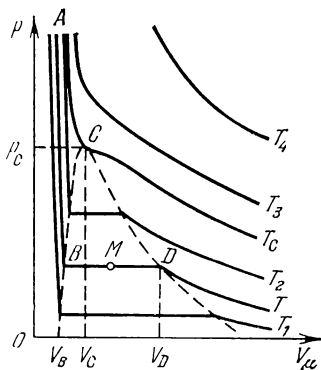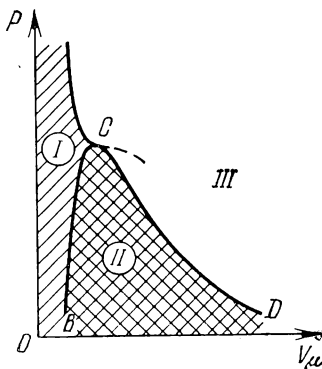
Fig. 12.3   Fig. 12. 4

separates the two-phase region II from the single-phase region III in which the substance is in the gaseous state.

12.3.4 The two-phase region II cannot exist at pressures above the critical one p_c (12.3.1) at which the substance can be in either of two states, gaseous or liquid. No pressure can convert gas at a temperature above the critical T_c into the liquid state by isothermal compression. The critical temperatures of many gases are very low: $T_c \approx 5$ K for helium and 33 K for hydrogen. This makes the liquefaction of such gases quite difficult.

12.3.5 Besides the difference in the molar volumes of boiling liquid and dry saturated vapour, the specific heat of vaporization (13.6.4) also vanishes when a substance is in the critical state, as does the surface tension (13.4.4). The difference be-

tween the liquid and gaseous states of a substance completely
disappears when it is in the critical state.

12.3.6 The Van der Waals equation (12.2.4) is a cubic equation
with respect to the molar volume V_μ, having factors that de-
pend upon the pressure, temperature and the chemical nature
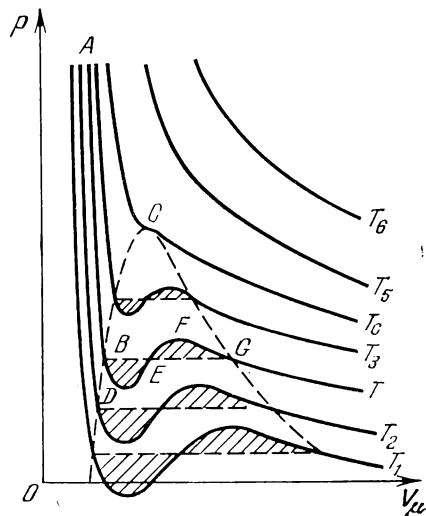
Fig. 12.5

of the gas. This equation has either one or three real roots, in
accordance with the numerical values of p and T. Shown in
Fig. 12.5 are the isothermals of a real gas that complies with
the Van der Waals equation. The isothermals correspond to the
different temperatures $T_1 < T_2 < T < T_3 < T_c < T_5 < T_6$.
All the subcritical isothermals have a hatched portion where
each pressure corresponds to three different states, which are
represented by the three points, B, E and G, of the isothermal.
The wave-shaped portion $BDEFG$ of the isothermal in Fig. 12.5
more accurately represents the transition of the substance from

the gaseous to the liquid state than the horizontal portions of experimentally plotted isothermals (dash straight lines BG). The portion BD of the isothermal corresponds to a *superheated liquid*, which can be obtained if the beginning of boiling is delayed at point B. The portion GF of the isothermal represents the state of *supersaturated vapour*, occurring upon slow isothermal compression in the absence of condensation centres. If such centres (dust specks or ions) are introduced into the supersaturated vapour, extremely rapid condensation of the vapour takes place. Over the portion DEF, the molar volume increases (decreases) as the pressure is increased (decreased). Such states of matter are impossible. The horizontal steps BG cut the portions $BDEFG$ of the isothermals in such manner that the hatched areas $BDEB$ and $EFGE$ in Fig. 12.5 are equal to each other (*Maxwell's equal area rule*).

12.3.7 The values of the critical variables p_c, $V_{\mu c}$ and T_c (12.3.1) can be expressed in terms of the coefficients a and b in the Van der Waals equation (12.2.4) and the universal gas constant R (8.4.4):

$$p_c = \frac{1}{27}\frac{a}{b^2}, \quad V_{\mu c} = 3b \quad \text{and} \quad T_c = \frac{8}{27}\frac{a}{bR}.$$

12.4 Superfluidity of Helium

12.4.1 Helium is a kind of substance which, after going over from the gaseous to the liquid state at standard atmospheric pressure and a temperature of 4.22 K, remains in the liquid state no matter how close to absolute zero (11.8.4) it is cooled.

12.4.2 At the temperature 2.19 K a second-order phase transition occurs: liquid helium I, existing at temperatures $T > 2.19$ K, undergoes the transition to liquid helium II, which exists at $T < 2.19$ K.

A *second-order phase transition* is a transformation of matter that is not associated with the evolution or absorption of heat as in a first-order transition. In second-order phase transitions the heat capacity, coefficient of thermal expansion and certain other characteristics of substances vary with discontinuous jumps. Some examples of second-order phase transitions are the transition of iron at the Curie point (26.5.2) from a ferromagnetic substance into a paramagnetic one (26.3.5) and the

transition of certain metals and alloys at extremely low temperatures to the superconductive state (41.6.1).

12.4.3 *Superfluidity,* a phenomenon observed *in liquid helium II*, is the practically complete absence of viscosity when such helium flows through narrow capillary tubes (13.5.5). The coefficient of viscosity (10.8.4) of helium II is less than 10^{-12} Pa-s, while that of helium I close to the temperature 4.22 K is of the order of 10^{-6} Pa-s.

12.4.4 According to the *two-fluid theory*, liquid helium II is a mixture of a superfluid component and a normal component. The superfluid component moves without friction and does not participate in the transfer of energy in the form of heat (9.2.1). The normal component moves with friction and does participate in energy transfer.

12.4.5 The theory of superfluidity is based on quantum mechanics (38.1.2). In the first place, quantum mechanics was required to explain why helium is the only nonfreezing liquid at ultralow temperatures and standard pressure. The zero-point vibrations (38.5.6) of the light-weight atoms of helium are sufficiently intensive and do not allow the weak forces of attraction between the helium atoms to form a crystalline structure at ordinary pressures.

The present-day theory of superfluidity is based on a study of the energy spectrum of helium at ultralow temperatures. Under these conditions the continuous energy spectrum can be regarded as the totality of quasi-particles with the energies $h\nu_i$, where h is Planck's constant and ν_i is the phonon frequency (41.7.5) corresponding to these quasi-particles. The quasi-particles cannot be identified with the real atoms of helium. They represent the motion of the whole system of helium atoms, and the energy values of the quasi-particle describe the energy spectrum of the whole quantum system, i.e. the liquid helium as a whole. At low temperatures the excited state of the helium constitutes sound waves that are quasi-particles in the normal part of helium II. Associated with the quasi-particles are the store of internal energy and the presence of friction in the liquid helium. But states are possible in helium II that correspond to its superfluid part, in which quasi-particles are not advantageous from the energy point of view and do not arise. As a result of the strong interaction between the particles of the superfluid part of helium II a bound body is formed in which no thermal excitations are generated: the superfluid part of helium II has

no store of internal energy and is not subject to viscosity. At
absolute zero temperature, there should actually not be any
normal part in helium II; all of the helium should be superfluid.
As the helium is heated, the number of phonons increases as
does the share of the normal part of helium II. But until the
temperature reaches 2.19 K, a superfluid part of helium II,
with its characteristic properties, still remains. At the temper-
ature 2.19 K, helium II is converted into helium I and all the
special properties of helium II disappear.

CHAPTER 13   LIQUIDS

13.1 Certain Properties of Liquids

13.1.1 *Liquids* are bodies that have a definite volume, but assume
the shape of the vessel that holds them. On the thermal mo-
tion in liquids, see Sect. 8.1.5. The nature of the thermal mo-
tion in liquids determines the similarity of their properties with
those of both solids and gases. Like solids, liquids are almost
incompressible. This property is associated with the strong
intermolecular interaction of the particles in liquids. When liq-
uids are compressed, the distances between the molecules are
reduced and the repulsive forces that impede compression in-
crease drastically (12.1.4). Liquids have relatively high densities
and, like solids, resist, not only compression, but tension as
well (40.3.6). This is manifested in the fact that the Van der
Waals isothermals enter the region of negative pressures (the
isothermal at $T = T_1$ in Fig. 12.5). The resemblance of the
properties of liquids and real gases at high temperatures and
low densities is observed, for instance, in that the surface ten-
sion of liquids (13.4.4) and the specific heat of vaporization
(13.6.4) are reduced as the temperature is raised. Besides, the
densities of dry saturated vapour (12.3.2) and of boiling liquid
(12.3.2) approach each other as the temperature is increased.
13.1.2 The similarity between liquids and solids is confirmed
by X-ray diffraction analysis (33.4.1). At temperatures close
to that of solidification, the arrangement of the particles in
liquids is similar to the ordered arrangement of particles that is

typical of solidified liquids. The mutual arrangement of neigh-
bouring particles in liquids resembles the ordered arrangement
of neighbouring particles in crystals. In liquids, however, this
ordered arrangement is observed only within small volumes. At
distances $r > (3 \text{ to } 4)$ d from a certain selected "central" mol-
ecule, order is no longer evident (here d is the effective diameter
of the molecule, see Sect. 8.4.1). Such order in the arrangement
of the particles is called *short-range order in liquids*.

13.1.3 X-ray patterns of liquids do not differ from those of
polycrystalline bodies (33.4.5) consisting of extremely fine
crystals (having linear dimensions of the order of 10^{-9} m) that
are randomly oriented with respect to one another in so-called
cybotaxic regions. Within the boundaries of these regions the
particle distribution is ordered, but the kind of order changes
from one cybotaxic region to another. In the course of time,
intensive thermal motion at temperatures that are not too low
rapidly alters the arrangement and structure of the cybotaxic
regions.

13.2 Frenkel's Hole Theory of the Liquid State

13.2.1 The most important parameter determining the structure
and physical properties of liquids is the specific volume. In
melting a crystalline body the specific volume increases only
slightly, by approximately 10 per cent. Such an increase in the
specific volume takes place in a solid subject to negative pres-
sure equal to the ultimate strength (40.3.7) of the solid. This
enables a liquid to be dealt with as a body whose integrity has
been violated at various places. In melting crystalline bodies,
their particles acquire higher mobility. This is due to the vital
property of *fluidity of liquids*, as well as the violation of long-
range order in crystals and the emergence of short-range order in
liquids (13.1.2). Moreover, owing to the high mobility of the
particles, microscopic gaps or microvoids, which shall be
called *holes*, appear in liquids. The thermal motion in liquids
causes the holes to disappear at random at certain places and to
appear simultaneously at others. This is equivalent to chaotic
movement of the holes.

13.2.2 The hole theory of the structure of liquids is inapplicable
to liquids subject to high external pressures of the order of
thousands of atmospheres, at which the compressibility of

liquids is comparable to that of solids. At high temperatures, close to the critical one*, the structure and properties of a liquid are close to those of a gas for which the concept of holes is meaningless. Hence, the hole theory is inapplicable for liquids under these conditions.

13.2.3 It follows from the nature of thermal motion in liquids (8.1.5) that the molecules vibrate about a certain equilibrium position for the length of time τ, after which this equilibrium position is displaced with a jump over a distance of an order of magnitude equal to the average distance $\langle d \rangle$ between adjacent molecules:

$$\langle d \rangle \approx \sqrt[3]{\frac{1}{n_0}} = \sqrt[3]{\frac{\mu}{N_A \rho}} \ ,$$

where n_0 is the number of molecules in unit volume, N_A is Avogadro's number (Appendix II), ρ is the density of the liquid, and μ is its molar mass. For water, for example, $\rho = 10^3$ kg/m³, $\mu = 0.018$ kg/mol and $\langle d \rangle \approx 3 \times 10^{-10}$ m.

13.2.4 The *relaxation time* is the average time $\langle \tau \rangle$ that a molecule of the liquid remains in a temporarily settled state close to a certain equilibrium position. As the temperature is raised $\langle \tau \rangle$ decreases rapidly (13.2.5). This explains the high mobility of molecules in a liquid at high temperatures and the low viscosity of liquids under these conditions.

13.2.5 The expenditure of a certain amount of *activation energy* E is required for a molecule to pass from one equilibrium position to another (13.2.4). Such a passage is regarded as overcoming a potential barrier of height E (38.7.1) because to accomplish the passage the potential energy of the molecule should increase by the amount E. Only after this can it pass over to a new equilibrium position. The whole process turns out to be possible because high energy is concentrated on some of the molecules as a result of collisions in thermal motion. It is transferred to them by other molecules. The dependence of the relaxation time (13.2.4) on E and the absolute temperature is of the form

$$\langle \tau \rangle = \tau_0 e^{E/kT},$$

---

* Substances with an exceptionally low critical temperature are not considered here.

14—0186

where k is Boltzmann's constant (8.4.5) and τ_0 is the average period of vibrations of the molecule about its equilibrium position.

13.2.6 When a liquid is subject to the action of an external force for the length of time $t \gg \langle \tau \rangle$ the particles of the liquid are displaced mainly in the direction of this force. This reveals the fluidity of the liquid. But if $t \ll \langle \tau \rangle$ a particle has no opportunity to change its equilibrium position during the time of force action. At this the liquid manifests elastic properties, resisting changes in both volume and shape.

During the time $\langle \tau \rangle$ a particle of liquid is displaced, on an average, over the distance $\langle d \rangle$, the average velocity $\langle v \rangle$ of displacement of the molecules is determined by the equation

$$\langle v \rangle = \frac{\langle d \rangle}{\langle \tau \rangle} \quad \text{or} \quad \langle v \rangle = \frac{\langle d \rangle}{\langle \tau_0 \rangle} \, e^{-E/kT} \, .$$

The average velocity of motion of molecules in liquids is quite considerable, as a rule, but an order of magnitude less than the average velocities of molecules of the vapour of the same substance at the same temperatures.

13.3 Diffusion and Viscosity Phenomena in Liquids

13.3.1 If conditions required for transport phenomena (10.8.1) arise, diffusion, heat conduction and viscosity are observed in liquids. The difference between transport phenomena in liquids and similar phenomena in gases affects the transport coefficients and their dependence on the characteristics of properties of the liquids.

13.3.2 For a chemically homogeneous liquid the diffusion coefficient D (10.8.3) is calculated by the equation

$$D = \frac{1}{6} \, \frac{\langle d \rangle^2}{\tau_0} \, e^{-E/kT} \, .$$

The quantities are defined in Sects. 13.2.3, 13.2.4 and 13.2.5. The diffusion coefficient grows rapidly with temperature. This is due, for the most part, to the sharp drop in the relaxation time $\langle \tau \rangle$ (13.2.4). Besides, the quantity $\langle d \rangle$ (13.2.3) increases somewhat with the temperature T.

13.3.3 When the temperature approaches the critical value (12.3.1), the average velocity $\langle v \rangle$ of the particles of the liquid approaches the average velocity of molecules of a real gas and the values of the diffusion coefficient D for liquids become close to those for gases.

At temperatures much lower than the critical one, the diffusion coefficients for liquids are extremely small compared to those for the corresponding vapours or gases at ordinary pressures. For example, $D \approx 1.5 \times 10^{-9}$ m²/s for water at $T = 300$ K, while for the vapour of water (steam) in air at the same temperature and at standard atmospheric pressure $D \approx 2 \times 10^{-5}$ m²/s.

13.3.4 At temperatures close to the critical one*, the thermal motion in liquids acquires features that differ from those described in Sect. 8.3.5 and approach those of thermal motion in gases. Under these conditions the internal friction (viscosity) in liquids is of the same nature as in gases (10.8.4).

At temperatures close to the melting point, the viscosity of liquids cannot be accounted for in the same way as for gases. The mechanism by means of which viscosity arises is a complex one. The coefficient of dynamic viscosity η (10.8.4) for liquids may be related to the *mobility* u_0 *of the molecule*. This is defined as the velocity u gained by a molecule subject to an external force F equal to unity: $u_0 = u/F$. The relation between η and u_0 is an inversely proportional one: $\eta \propto u_0^{-1}$. In its turn $u_0 \propto \propto D/kT$, where D is the diffusion coefficient, T is the absolute temperature, and k is Boltzmann's constant (8.4.5). Consequently, $\eta \propto T/D$ or $\eta \propto T$ exp (E/kT), where E is the activation energy (13.2.5). As the temperature is raised the viscosity of liquids drops rapidly, especially in the region of low temperatures. At high pressures the viscosity of liquids increases rapidly with pressure. This is due to the increase in the activation energy (13.2.5) and the corresponding increase in the relaxation time (13.2.4).

13.4 Surface Tension of Liquids

13.4.1 A molecule of a liquid in the surface layer is subject to uncompensated, inwardly directed forces of attraction exerted by the rest of the liquid. As a result, the surface layer exerts a

---

* See the footnote on page 209.

14*

high *internal pressure* on the whole liquid. It is of the order of tens of thousands of atmospheres.

13.4.2 Particles of the surface layer of a liquid have a higher potential energy than those inside the liquid. The reason for this is that in order to emerge from inside the liquid to its surface the molecules must do work to overcome the inwardly directed forces of internal pressure (13.4.1). This work increases the potential energy of the molecules that reach the surface.

The work W that must be performed to increase the area of the surface layer of a liquid isothermally is equal to

$$W = \langle (F_S - F_V) \rangle N,$$

where $\langle (F_S - F_V) \rangle$ is the average difference between the free energy (11.4.5) per molecule at the surface F_S and within the bulk F_V of the liquid, and N is the number of molecules in the surface layer of the liquid.

13.4.3 The condition of minimum potential energy of a liquid, required to maintain its stable equilibrium, is complied with when the liquid has its minimum possible free surface area. The state of stable equilibrium of a liquid incompressible body corresponds to the minimum ratio of its surface area to its volume. Therefore, small drops of liquid suspended in air are of spherical shape. A liquid tends to reduce the area of its free surface. Consequently, the surface layer resembles a stretched elastic film in which tensile forces act (see also Sect. 13.4.5).

13.4.4 The work done in the isothermal formation of unit surface area is called the *surface tension* σ of the given liquid at the boundary with another phase (12.3.3):

$$\sigma = \frac{W}{A} = \langle (F_S - F_V) \rangle \frac{N}{A} = \langle (F_S - F_V) \rangle n_1,$$

where $n_1 = N/A$ is the number of molecules per unit area of the surface layer. The surface tension is also calculated by the equation

$$\sigma = \frac{\Delta F}{\Delta A},$$

where ΔF is the change in the free energy of the surface layer, and ΔA is the change in the surface area.

The surface tension σ depends upon the chemical composition of the liquid and its temperature. It decreases when the tem-

perature is raised and vanishes at the critical temperature
(12.3.1). The surface tension can be reduced by adding *surface-
active substances* to the liquid. This is due to the fact that such
substances are adsorbed on the surface layer of the liquid and
reduce the free energy of this layer (11.4.5).
13.4.5 There is an essential difference between the surface
layer of a liquid and an elastic film (13.4.3). The surface ten-
sion of liquids does not depend upon the surface area and tends
to reduce this area to zero. The tension of an elastic film is pro-
portional to its deformation and equals zero at a definite finite
area of the film's surface.
This peculiarity in the properties of liquid films is associated
with the change in the number of molecules in the surface layer
when these films are stretched (or compressed). But the average
distances between the molecules and the forces of intermolecular
interaction determined by these distances do not change. This
is why the magnitude of the surface tension is independent of
the area of the free surface of the liquid.
13.4.6 If the surface of a liquid is bounded by a wetted per-
imeter (13.5.1), then the quantity σ is equal to the force acting
on unit length of the wetted perimeter, perpendicular to this
perimeter. This force lies in a plane tangent to the free surface
of the liquid.

13.5 Wetting and Capillary Phenomena

13.5.1 The free surface of a liquid, curved adjacent to the wall
of the vessel holding the liquid, is called the *meniscus*. The line
along which the meniscus contacts the wall is called the *wetted
perimeter*. The meniscus is characterized by the *contact angle* θ
between the wetted surface of the wall and the meniscus at their
point of contact. When $\theta < \pi/2$ (Fig. 13.1a), the liquid is said
to *wet* the wall (to be a *wetting* liquid); when $\theta > \pi/2$ (Fig. 13.1b)
the liquid *does not wet* the wall and is a *nonwetting* liquid.
Wetting (nonwetting) is said to be ideal when $\theta = 0$ ($\theta = \pi$).
Here the meniscus is of spherical shape and is concave or con-
vex. The absence of both wetting and nonwetting corresponds to
the condition $\theta = \pi/2$, in which case the liquid has a plane
free surface.
13.5.2 The formation of a meniscus is due to the interaction of
the molecules with one another and with the particles of the

rigid body (wall). Molecule A of the surface layer, located close to the wall of the vessel and having a sphere of molecular action (12.2.3) of radius R_m (range of molecular action) (Fig. 13.2), is subject to the resultant force of attraction exerted on it by all the other molecules of the liquid (F_1) and by all the particles of the wall (F_2). Force F_2 is perpendicular to the wall. This follows from considerations of symmetry. The direction of force F_1 depends upon the shape of the meniscus and the position of

(a)   (b)
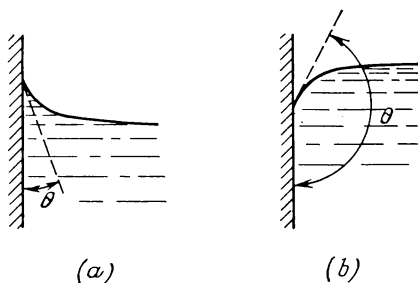
Fig. 13.1

molecule A with respect to the wall. For example, if the meniscus is flat (13.5.1) and molecule A is right at the wall, force F_1 is directed at an angle of 45° to the wall (Fig. 13.3). Molecule A can be in equilibrium only when the resultant force $F = F_1 + F_2$ is perpendicular to the surface of the liquid.* Otherwise, molecule A would move along this surface.

13.5.3 The shape of the meniscus is determined by the three possible directions of force F:

(a) force F is parallel to the surface of the wall, the liquid has a plane surface and $\theta = \pi/2$ (Fig. 13.4a);

(b) force F is directed towards the wall; the forces of attraction exerted by the wall on molecule A exceed the forces of attraction exerted on the molecule by all the other molecules of the

————————

* Here we ignore the gravity force acting on molecule A because it is negligibly small compared to forces F_1 and F_2.

liquid. The liquid has a concave meniscus with $\theta < \pi/2$, i.e. the liquid wets the wall (Fig. 13.4.b);
(c) force $\mathbf{F}$ is directed towards the liquid; the forces of attraction exerted by the molecules of the liquid on molecule A
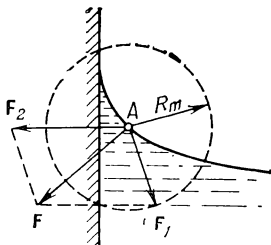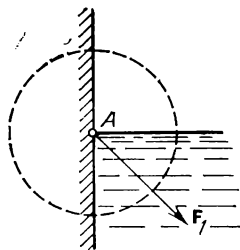
Fig. 13.2   Fig. 13.3

prevail over those exerted on it by the particles of the wall. The meniscus is convex with $\theta > \pi/2$, i.e. the liquid does not wet the wall (Fig. 13.4c).
13.5.4 The curved shape of the surface layer is the cause of the pressure Δp, additional to the external pressure, acting on the
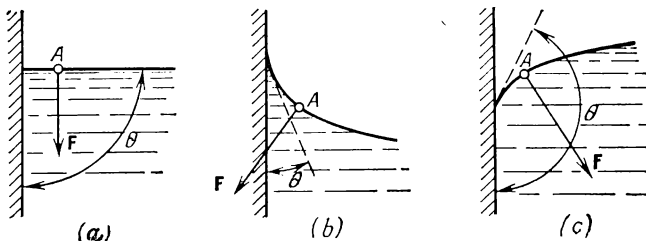
Fig. 13.4

liquid. This additional pressure is due to the surface tension. This is similar to the way in which a stretched elastic envelope (like a rubber balloon) exerts pressure on the gas it contains.

The additional pressure exerted on a liquid with a surface layer of arbitrary shape is calculated by the equation

$$\Delta p = \sigma \left(\frac{1}{R_1} + \frac{1}{R_2} \right) ,$$

where σ is the surface tension (13.4.4), and R_1 and R_2 are the radii of curvature of any two mutually perpendicular normal sections* of the surface of the liquid at the point where molecule A is located. The radius of curvature R_1 (or R_2) is taken positive if its centre in the corresponding section is inside the liquid. Otherwise, the radius of curvature is taken to be negative. Thus, $\Delta p > 0$ when the meniscus is convex and $\Delta p < 0$ when it is concave. In the case of a plane surface $R_1 = R_2 = \infty$, and there is no additional pressure ($\Delta p = 0$). The preceding equation is called the Laplace theorem. For a spherical surface $R_1 = R_2 = R$ and $\Delta p = 2\sigma/R$. Such a surplus pressure exists, for instance, within a bubble of gas of radius R, located inside a liquid near the surface.

The surplus pressure inside a soap bubble of radius R is due to the action of both superficial layers of the thin spherical film of soap (inside and outside the bubble). Thus $\Delta p = 4\sigma/R$.

13.5.5 The level of a liquid in narrow cylindrical vessels (*capillary tubes*) of radius r differs from the level of the liquid in a larger vessel into which the capillary tube is placed vertically. The level in the capillary tube is higher (lower) than in the larger vessel by the amount h if the liquid wets (does not wet) the walls of the vessel. Here

$$h = \frac{2\sigma \cos \theta}{rg\rho}$$

where θ is the contact angle (13.5.1), ρ is the density of the liquid, and g is the acceleration due to gravity (7.3.3). This is called *Jurin's rule*.

If the liquid is in a narrow slit of constant width δ between two parallel plates, the meniscus is of cylindrical shape having

---

* A normal section of the surface through point A is the curve obtained as the intersection of the surface and a plane passed through the normal to the surface at this point. Radii R_1 and R_2 are called the principal radii of curvature of the surface layer.

the radius $\delta/2$ and the level in the slit is above (for a wetting liquid) or below (for a nonwetting liquid) the level in the vessel in which the plates are immersed by the amount

$$h = \frac{2\sigma \cos \theta}{\delta g \rho} .$$

13.5.6 The pressure of saturated vapour (12.3.2) above the curved surface of a liquid depends upon the shape of the meniscus. If the meniscus is concave (convex), the pressure of the saturated vapour is less (greater) than that above a plane surface by the amount Δp_{ms}:

$$\Delta p_{\mathrm{ms}} = \frac{\rho}{\rho_1 - \rho} \Delta p,$$

where ρ is the density of the saturated vapour, ρ_1 is the density of the liquid, and Δp is the additional pressure due to the curvature of the surface (13.5.4).

13.6 Vaporization and Boiling of Liquids

13.6.1 *Vaporization*, or *evaporation*, is the process of vapour formation at the free surface of a liquid. It takes place at any temperature and increases as the temperature is raised. There are molecules in the surface layer of a liquid that have high velocities and kinetic energies of thermal motion. The escape of such molecules from the surface into the space above the liquid is the explanation for evaporation with the consequent reduction of the store of internal energy and cooling of the liquid. A measure of the process of vaporization is the *rate of evaporation u*, which is the amount of liquid converted into vapour in unit time. The rate u depends upon the external pressure and the motion of the gaseous phase above the free surface of the liquid. Thus

$$u = \frac{cA}{p_0} (p_{\mathrm{v}} - p),$$

where c is a constant, A is the area of the free surface of the liquid, p_{v} is the pressure of saturated vapour, p is the pressure of the vapour above the free surface of the liquid being vaporized, and p_0 is the external barometric pressure.

13.6.2 *Boiling* is the process of intense vaporization of a liquid, not only from its free surface, but throughout the whole volume of the liquid within the bubbles of vapour that are produced in boiling. The bubbles of vapour in a boiling liquid rapidly increase in size, rise to the surface and burst. This causes the turbulence in a boiling liquid. The pressure p within a bubble of vapour located within the liquid is the sum of the external pressure p_0, the hydrostatic pressure p_1 of the overlying layers of liquid, and the additional pressure Δp due to the surface tension (13.5.4). Thus

$$p = p_0 + p_1 + \Delta p,$$

and

$$p_1 = \rho gh \quad \text{and} \quad \Delta p = \frac{2\sigma}{r},$$

where r is the radius of the vapour bubble, h is the distance from its centre to the surface of the liquid, and ρ and σ are the density and surface tension of the liquid.
The boiling of a liquid begins at the temperature at which the pressure p_v of the saturated vapour inside a bubble is at least as high as the pressure p (13.6.2):

$$p_v \geqslant p_0 + \rho gh + \frac{2\sigma}{r}.$$

If this condition is not complied with, the bubble collapses and the vapour it contains is condensed.
13.6.3 With small radii r of the vapour bubbles, the pressure p_v should be quite high and it is necessary to heat the liquid to a high temperature before it starts boiling. If the liquid contains *centres of vaporization* (dust specks, bubbles of dissolved gases, etc.), boiling begins at a much lower temperature. This is due to the fact that the vapour bubbles that appear on the centres of vaporization are of such a size that the effect of the third member in the inequality of Sect. 13.6.2 can be neglected. Moreover, $\rho gh \ll p_0$ as a rule, so that the approximate condition for the beginning of boiling can be written in the form

$$p_v \approx p_0.$$

The temperature of a liquid at which the pressure of its saturat-ed vapour is equal to the external pressure is called the *boiling point* of the liquid.

13.6.4 When a liquid boils at constant pressure, its temper-ature remains constant. The quantity of heat transferred to a boiling liquid is completely expended in converting the mole-cules of the liquid into vapour. The heat r_b required to vapor-ize unit mass of a liquid heated to the boiling point is called the *heat of vaporization*. The value of r_b decreases as the boil-ing point is raised and vanishes at the critical temperature (12.3.1).

The change in the internal energy of a liquid (9.1.1) as a unit of its mass is converted into vapour at the boiling point is called the *internal heat of vaporization*.

13.6.5 The boiling of liquids and the condensation of vapour are examples of first-order phase transitions (cf. Sect. 12.4.2). Typ-ical of such phase transitions are the simultaneous constant values of pressure and temperature, but the changing ratio of the masses of the two phases (12.3.3). For a first-order phase transition to occur, the heat r_b (*specific latent heat of vaporiza-tion*) must be supplied to or taken away from the system. The heat r_b per unit mass is calculated by the *Clausius-Clapeyron equation*:

$$r_b = (v_2 - v_1)\, T\, \frac{dp}{dT}\,,$$

where v_1 and v_2 are the specific volumes of the substance in the initial and final phases, T and p are the temperature and pres-sure of the phase transition.

13.6.6 It follows from the Clausius-Clapeyron equation for the boiling of liquids that

$$\frac{dT}{dp} = \frac{(v_v - v_l)\, T}{r_b}\,,$$

where v_l and v_v are the specific volumes of the liquid and vapour at the boiling point T. Since $v_v > v_l$ and $r_b > 0$, $dT/dp > 0$, i.e. the boiling point increases with pressure. For instance, at the pressure $p = 1.25 \times 10^7$ Pa water can be heated without boiling to a temperature at which lead can be melted in the water. The melting point of lead is 900 K.

PART THREE
ELECTRODYNAMICS

CHAPTER 14   ELECTRIC CHARGES.
COULOMB'S LAW

14.1 Introduction

14.1.1 *Electrostatics* is the branch of the study of electricity
that deals with the interaction and properties of a system of
electric charges that are fixed with respect to the chosen inertial
frame of reference (2.1.2).
There are two kinds of electric charges, *positive* and *negative*.
Bodies having unlike charges attract one another; those having
like charges repulse one another.
14.1.2 The electric charge of any system of bodies consists of a
whole number of elementary charges, each equal to 1.6 $\times$
$\times$ 10^{-19} C (where C is the abbreviation of coulomb, the unit of
charge). The stable particle with the minimum rest mass (5.6.1)
and a negative elementary charge is the *electron*. The rest mass
of the electron is 9.1 $\times$ 10^{-31} kg. The stable particle with the
minimum rest mass and a positive elementary charge is the
positron (43.5.1).* Its rest mass is the same as that of the elec-
tron. Besides, there is a stable particle with a positive elemen-
tary charge called the *proton*. The rest mass of the proton is
1.67 $\times$ 10^{-27} kg. Electrons and protons are found in the atoms
of all the chemical elements.
14.1.3 The *law of conservation of electric charge* states: the
algebraic sum of the electric charges of bodies or particles that
constitute an electrically isolated system remains constant re-
gardless of any processes that occur in the system.
New electrically charged particles may be formed in the system
being considered. These may be electrons formed by the ion-
ization of atoms or molecules (22.4.1), ions formed by the phe-

---

* Here we do not consider the instability of the positron, com-
mon in our world, that is associated with the annihilation of
an electron-positron pair (43.5.1).

nomenon of electrolytic dissociation, etc. If the system is electrically isolated, however, then the algebraic sum of the charges of all the newly formed particles of the system always equals zero. The law of the conservation of electric charge is one of the fundamental conservation laws, like the laws of conservation of momentum (2.7.1) and of energy (3.4.2).

14.1.4 Upon contact of two electrically neutral bodies, charges go over from one body to the other as a result of friction. In each of such two bodies there is no longer an equal number of positive and negative charges and the bodies become oppositely charged.

When a body is electrized by induction, the charges in it are no longer uniformly distributed. The charges are redistributed so that one part of the body has an excess of positive charges and the other part, of negative charges. If these two parts of the body are separated in some way, they will be oppositely charged.

14.2 Coulomb's Law

14.2.1 Coulomb experimentally established that the force of interaction F_{12} between two small charged spheres, with charges equal to q_1 and q_2 (14.1.1), is proportional to the product $q_1 q_2$ and inversely proportional to the square of the distance r between the spheres:

$$F_{12} = k_1 \frac{q_1 q_2}{r^2},$$

where k_1 is the proportionality factor ($k_1 > 0$).

14.2.2 *Central* forces act on the charges along a straight line joining the centres of the charges. In the interaction of like charges the product $q_1 q_2 > 0$ and $F_{12} > 0$. This corresponds to mutual repulsion. For unlike charges $q_1 q_2 < 0$ and $F_{12} < 0$, which corresponds to mutual attraction. The vector form of the Coulomb law is

$$\mathbf{F}_{12} = k_1 \frac{q_1 q_2}{r^2} \frac{\mathbf{r}_{12}}{r},$$

where $\mathbf{F}_{12}$ is the force exerted on charge q_1 by charge q_2, $\mathbf{r}_{12}$ is the radius vector joining charge q_2 to q_1, and $r = |\mathbf{r}_{12}|$

(Fig. 14.1a). The force $\mathbf{F}_{21}$, exerted on charge q_2 by charge q_1, equals

$$\mathbf{F}_{21} = k_1 \frac{q_1 q_2}{r^2} \frac{\mathbf{r}_{21}}{r} \, ,$$

where $\mathbf{r}_{21} = -\mathbf{r}_{12}$ is the radius vector joining charges q_1 and q_2 (Fig. 14.1b). Forces $\mathbf{F}_{12}$ and $\mathbf{F}_{21}$ are said to be *Coulomb forces*.
14.2.3 Coulomb's law is valid for the interaction of *point electric charges* that are fixed in the given inertial frame of reference. Point charges are charged bodies whose linear dimensions d are very much smaller than the distance r between them ($d \ll r$).
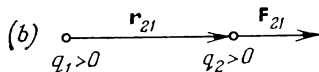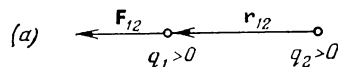
Fig. 14.1

Coulomb's law states: the force of electrostatic interaction between two point electric charges is proportional to the product of the magnitudes of the charges, inversely proportional to the square of the distance between them and acts along the straight line joining them.

In the form given in Sect. 14.2.2 Coulomb's law is also valid for the interaction of charged spheres of radii R_1 and R_2 if the charges q_1 and q_2 of the spheres are uniformly distributed over their surfaces or throughout their volumes. Radii R_1 and R_2 may be commensurable with the distance r between the centres of the spheres.
14.2.4 A Coulomb force (14.2.2) depends upon the properties of the medium in which the interacting charges are located. This dependence can be taken into consideration if we consider the factor k_1 in the formulas of Sects. 14.2.1 and 14.2.2 as the ratio of two factors:

$$k_1 = \frac{k}{\varepsilon_{\mathrm{r}}}$$

where k is a factor depending only on the choice of the system of units of measurement and ε_{r} is the dimensionless *relative permittivity*, or *dielectric constant*, of the medium and character-

izes its electric properties. It is assumed here that the medium
is boundless, homogeneous and isotropic, i.e. that its properties
are the same throughout its volume and are independent of the
direction. For a vacuum, ε_r is considered to equal unity.

14.2.5 For a medium with the relative dielectric constant ε_r,
Coulomb's law takes the form

$$F = k \; \frac{q_1 q_2}{\varepsilon_r r^2} \; .$$

For charges q_1 and q_2 located in vacuum ($\varepsilon_r = 1$),

$$F_0 = k \frac{q_1 q_2}{r^2} \; .$$

From this follows the meaning of the relative dielectric con-
stant

$$\varepsilon_r = F_0/F.$$

14.2.6 The concept of ε_r, introduced in Sects. 14.2.4 and 14.2.5,
has a meaning only under the conditions stipulated in Sect.
14.2.4. An ε_r-fold reduction of F is due to a phenomenon
known as *electrostriction*, i.e. the deformation of a dielectric
(18.1.1) from the effect of charged bodies. In their deformation,
liquid and gaseous dielectrics are directly adjacent to the charged
bodies and exert an additional mechanical action on them,
reducing the Coulomb force (14.2.2). In solid dielectrics, charges
q_1 and q_2 on the bodies may be located within certain cavities.
This makes the calculation of forces F_{12} and F_{21} (14.2.2) con-
siderably more difficult. These forces depend on the shapes of
the cavities in which the charges q_1 and q_2 are located. In this
case the interpretation of ε_r given in Sect. 14.2.5 cannot be
correct.

14.2.7 In SI units (Appendix I) the factor k in the formulas of
Sect. 14.2.5 are taken equal to

$$k = \frac{1}{4\pi\varepsilon_0}$$

and Coulomb's law is written in the form

$$F = \frac{1}{4\pi\varepsilon_0} \; \frac{q_1 q_2}{\varepsilon_r r^2} \; .$$

This form of Coulomb's law and of all laws following from it is said to be *rationalized*.

The quantity ε_0 is called the *permittivity of free space* or, according to the SI units, the *electric constant*. It is equal to 8.85 $\times$ $\times$ 10^{-12} C²/N-m² (Appendix II).

The product $\varepsilon_0\varepsilon_r$ is sometimes called the *absolute dielectric constant*, or *absolute permittivity* of the medium.

In the cgse system of units (Appendix I), Coulomb's law is written in the following irrational form:

$$F = \frac{q_1 q_2}{\varepsilon_r r^2} .$$

CHAPTER 15   ELECTRIC FIELD STRENGTH
AND DISPLACEMENT

15.1 Electric Field. Field Strength

15.1.1 A *force field* is a form of matter that is studied in physics along with its study of other forms of matter known commonly as substances. The most basic feature of force fields is that various kinds of interaction are accomplished by means of them. A gravitational field (6.2.1), for instance, accomplishes gravitational interaction between bodies of various masses located in the field. All force fields have vital properties that characterize their material nature. Primarily this concerns their energy. There are no impassible barriers between a field and material substances; one can be converted into the other (43.5.5).

A physical field cannot be defined as space in which certain forces act. Space, like time, is a form of the existence of matter. The fact that a field exists in space by no means infers that the field can be identified with space, because the form of existence of matter cannot be confused with the matter itself.

15.1.2 Coulomb interaction (14.2.2) between fixed electrically charged particles or bodies is accomplished through the *electrostatic field* set up by the charges. An electrostatic field is a *steady-state electric field*, i.e. one that does not change in the course of time, and is set up by fixed electric charges. This is

one of the forms of an *electromagnetic field* by means of which interaction is accomplished between electrically charged particles (or bodies) that travel, in the general case, in some arbitrary way with respect to the chosen frame of reference.

A typical property of an arbitrary electric field, distinguishing it from all other physical fields, is the action it exerts on both travelling and fixed electric charges (charged particles and bodies).

15.1.3 The quantitative force characteristic of the action exerted by an electric field on charged particles and bodies is the vector **E** of the *electric field strength* (formerly called the electric intensity). The electric field strength at a given point is the ratio of the force **F**, exerted by the field on a fixed point (14.2.3) *electric test charge* placed at the point of the field being considered, to the magnitude q_0 of the charge:

$$E = \frac{F}{q_0}.$$

It is assumed that the test charge q_0 is so small that its presence does not lead to a redistribution in space of the charges that set up the field being investigated. In other words, the test charge does not distort the field.

An electric field is said to be *uniform* (a *uniform electric field*) if the electric field vector **E** has a constant magnitude and direction at any point of the field.

15.1.4 The strength of the electrostatic field of point charge q at a point located at the distance r from the charge is

$$E = \frac{1}{4\pi\varepsilon_0} \frac{q}{\varepsilon_r r^3} \mathbf{r} \quad \text{(in SI units)}$$

$$E = \frac{q}{\varepsilon_r r^3} \mathbf{r} \quad \text{(in cgse units)}$$

where **r** is the radius vector joining charge q and the point at which the field strength is to be calculated. Vectors **E** are directed at all points of the field radially away from charge q when $q > 0$, and directed radially toward the charge when $q < 0$.

The projection E_r of the field strength on the direction of radius vector **r** is

$$E_r = \frac{1}{4\pi\varepsilon_0}\frac{q}{\varepsilon_r r^2} \text{ (in SI units)}$$

$$E_r = \frac{q}{\varepsilon_r r^2} \text{ (in cgse units)}.$$

These formulas are also used to calculate the strength of the field set up by a hollow sphere, having the charge q on its surface, at the distance r from the centre of the sphere whose radius equals R. Inside the hollow sphere $E_r = 0$.

15.1.5 The force F exerted by an electric field on an arbitrary charge q located at a given point in the field is

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

where **E** is the field strength at the point where charge q is located and is distorting the field, i.e. making it different from the field set up before introducing charge q.

15.1.6 Electrostatic fields are represented graphically by employing *lines of force*.

Lines of force are imaginary curves drawn in such a way that their direction at any point (i.e. the direction of the tangent to the line at the point) coincides with that of the electric field vector. Lines of force are assumed to have the same direction as the corresponding electric field vectors. Lines of force do not intersect because at each point of the field vector **E** has only a single direction.

Lines of force are not identical with the paths of motion of light charged particles in an electrostatic field. Tangent at each point of the path of a particle is its velocity. Tangent to a line of force is the force acting on a charged particle. This, consequently, corresponds to the acceleration of the particle.

15.2 Principle of Superposition of Electric Fields

15.2.1 The basic problem of electrostatics can be formulated as follows: find the magnitude and direction of the electric field vector **E** for each point of a field from the given space distribution and magnitudes of the electric charges that are the source of the field.

15.2.2 When the field is set up by a system of fixed charges $q_1, q_2, q_3, \ldots, q_n$, the resultant force **F** exerted on a test charge q_0 (15.1.3) at any point of the field being considered is equal to the vector sum of the forces $\mathbf{F}_i$ exerted on charge q_0 by the fields set up by each of the charges q_i:

$$\mathbf{F} = \sum_{i=1}^{i=n} \mathbf{F}_i.$$

According to Sect. 15.1.3, $\mathbf{F} = q_0\mathbf{E}$ and $\mathbf{F}_i = q_0\mathbf{E}_i$, where **E** is the strength of the resultant field, and $\mathbf{E}_i$ is the strength of the field set up by charge q_i. Following from the preceding formulas is the *principle of independent action of electric fields*, or the *principle of superposition of electric fields*:

$$\mathbf{E} = \sum_{i=1}^{n} \mathbf{E}_i.$$

The strength of an electric field of a system of charges is equal to the vector sum of the strengths of the fields set up by each charge separately. The strength of the resultant field is found by superposition of the strengths of the fields of the separate charges.

For charges distributed continuously in space (15.2.3), the principle of superposition of fields is of the form

$$\mathbf{E} = \int d\mathbf{E},$$

where integration extends over all the continuously distributed sources of the fields, setting up electric fields with the strength $d\mathbf{E}$.

Example. A system of fixed point charges $q_1, q_2, \ldots, q_n$ sets up an electrostatic field whose strength **E** equals

$$\mathbf{E} = \frac{1}{4\pi\varepsilon_0} \sum_{i=1}^{n} \frac{q_i}{\varepsilon_r r_i^3} \, \mathbf{r}_i \text{ (in SI units)}$$

$$\mathbf{E} = \sum_{i=1}^{n} \frac{q_i}{\varepsilon_r r_i^3} \, \mathbf{r}_i \text{ (in cgse units)}$$

15*

where $\mathbf{r}_i$ is the radius vector from point charge q_i to the point being investigated in the field. Any charged body on which the charge Q is discretely distributed (15.2.3) can be divided into extremely small parts, each having a point charge. Therefore, the preceding formula is of general significance for calculating electrostatic fields in a homogeneous, isotropic medium that fills the whole field.

15.2.3 Electric charges, the sources of electrostatic fields, can be distributed in space either *discretely* (*discrete charge distribution*) at various points in space, or *continuously* (*continuous charge distribution*). In the latter case, the charges are distributed either along a certain line, on the surface of some body or within a certain volume. The concept of charge density is introduced for continuous charge distribution. When the electric charge is distributed continuously along a line we have *linear charge density* τ:

$$\tau = \frac{dq}{dl} ,$$

where dq is the charge of an infinitesimal element of length dl. When the electric charge is distributed continuously over a certain surface, we speak of the *surface charge density* σ:

$$\sigma = \frac{dq}{dS} ,$$

where dq is the charge on an infinitesimal element of area dS. When the electric charge is distributed continuously throughout some volume, we speak of the *volume charge density* ρ:

$$\rho = \frac{dq}{dV} ,$$

where dq is the charge in an infinitesimal element of volume dV.

Example. The electrostatic field of an electric dipole. An *electric dipole* is a system of two charges of equal magnitude, but of opposite sign. These charges, q and $-q$ ($q > 0$) are separated by the distance l which is small in comparison with the distance to the points being considered in the field. The *arm of the dipole* ts the vector $\mathbf{l}$, directed along the dipole axis from the negative ioward the positive charge and equal to the distance between

them (Fig. 15.1). The product of the charge q of the dipole $(q > 0)$ by the arm l is called the *electric dipole moment* $\mathbf{p_e}$:

$$\mathbf{p_e} = q\mathbf{l}.$$

The strength E of the field of an electric dipole at an **arbitrary** point is:

$$\mathbf{E} = \mathbf{E_+} + \mathbf{E_-},$$

where $\mathbf{E_+}$ and $\mathbf{E_-}$ are the strengths of the fields set up by the charges q and $-q$ (Fig. 15.1).
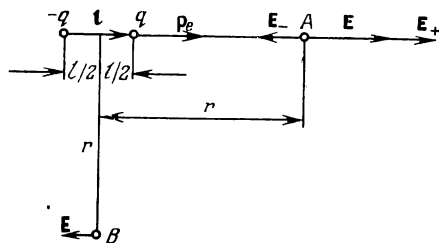
Fig. 15.1

At point A, located on the dipole axis at the distance r from its middle (Fig. 15.1) $(r \gg l)$, the strength of the field of a dipole equals:

$$E = \frac{1}{4\pi\varepsilon_0}\,\frac{2\mathbf{p_e}}{\varepsilon_r r^3}\ \text{(in SI units),}$$

$$E = \frac{2\mathbf{p_e}}{\varepsilon_r r^3}\ \text{(in cgse units).}$$

At point B, located on a perpendicular erected at the middle of the dipole axis and at the distance r from the middle of the axis $(r \gg l)$

$$E = -\frac{1}{4\pi\varepsilon_0}\frac{p_e}{\varepsilon_r r^3} \text{ (in SI units)},$$

$$E = -\frac{p_e}{\varepsilon_r r^3} \text{ (in cgse units)}.$$

At the arbitrary point O, sufficiently distant from the dipole
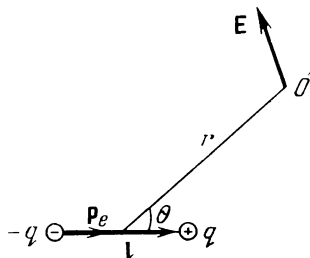
Fig. 15.2

$(r \gg l)$ (Fig. 15.2), the magnitude of its field strength is:

$$E = \frac{1}{4\pi\varepsilon_0}\frac{p_e}{\varepsilon_r r^3}\sqrt{3\cos^2\theta + 1} \text{ (in SI units)},$$

$$E = \frac{p_e}{\varepsilon_r r^3}\sqrt{3\cos^2\theta + 1} \text{ (in cgse units)}.$$

15.3 Electric Displacement. Ostrogradsky-Gauss Electric Flux Theorem

15.3.1 The electric field strength (15.1.3) depends upon the properties of the medium. In a homogeneous isotropic medium, the strength E is inversely proportional to ε_r (14.2.5). Another characteristic of an electric field, along with its strength E,

is the vector **D** of *electric displacement*, or *electric induction* (other names are dielectric displacement, dielectric flux density, displacement, electric displacement density and electric flux density). For a field in an electrically isotropic medium, the relation between **D** and **E** is of the form

$D = \varepsilon_0 \varepsilon_r E$ (in SI units),

$D = \varepsilon_r E$ (in cgse units).

The general relation for **D**, valid for anisotropic media is given in Sect. 18.3.4.

Example. For the field of point electric charge q (14.2.3)

$$D = \frac{1}{4\pi} \frac{q}{r^3} r \text{ (in SI units)},$$

$$D = \frac{q}{r^3} r \text{ (in cgse units)}.$$

The projection of **D** on the direction of radius vector **r**, from the point charge to the given point in the field, is

$$D_r = \frac{1}{4\pi} \frac{q}{r^2} \text{ (in SI units)},$$

$$D_r = \frac{q}{r^2} \text{ (in cgse units)}.$$

15.3.2 An *element* $d\Phi_e$ of *electric displacement flux* through an element of surface with the area dS is the scalar physical quantity defined by the equation

$$d\Phi_e = D \, dS = D \, dS \cos \alpha = D_n \, dS = D \, dS_\perp,$$

where **D** is the displacement vector (15.3.1), **n** is a unit vector normal to area dS, $d\mathbf{S} = dS \, \mathbf{n}$ is the vector of the element of surface dS, $D_n = D \cos \alpha$ is the projection of vector **D** on the direction of vector **n**, $dS_\perp = dS \cos \alpha$ is the area of the projection of the element of surface dS on a plane perpendicular to vector **D**, and α is the angle between vectors **D** and **n** (Fig. 15.3)

When the electric field is set up by point charge q, the element of displacement flux $d\Phi_e$ through the element of surface dS of the closed surface S, within which the point charge is located, **is**

$$d\Phi_e = \frac{q}{4\pi}\, d\omega,$$

where $d\omega$ is the solid angle subtended at point charge q by the element dS of surface S (Fig. 15.3).
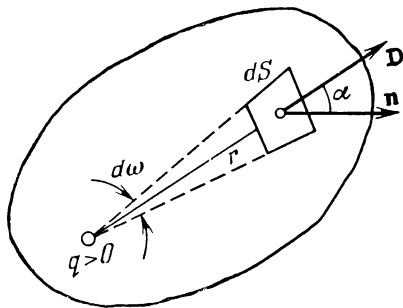
Fig. 15.3.

The *total displacement flux* Φ_e through the surface S is the summation or integration of all the elements of flux:

$$\Phi_e = \int_S \mathbf{D}\, d\mathbf{S} = \int_S D\, dS \cos\alpha = \int_S D_n\, dS = \int_S D\, dS_\perp,$$

where α is the angle between vectors $\mathbf{D}$ and $\mathbf{n}$. Here all the vectors and normals to the elements dS of surface are directed toward the same side of surface S. In the case of a closed surface S (Fig. 15.3), all the vectors $\mathbf{n}$ of the normals must be either outward or inward ones*.

---

* Only outward normals are used in the following.

15.3.3 The *Ostrogradsky-Gauss theorem* states that the displacement flux passing through an arbitrary closed surface is equal to or proportional to the algebraic sum of the electric charges q_1, q_2, q_3, . . ., q_h enclosed within the surface:

$$\oint_S \mathbf{D} \, d\mathbf{S} = \oint_S D_n \, dS = \sum_{i=1}^{h} q_i \text{ (in SI units)},$$

$$\oint_S \mathbf{D} \, d\mathbf{S} = \oint_S D_n \, dS = 4\pi \sum_{i=1}^{h} q_i \text{ (in cgse units).}$$

The displacement flux passing through an arbitrary closed surface that does not enclose any electric charges equals zero.
The differential form of the Ostrogradsky-Gauss theorem is one of Maxwell's equations for an electromagnetic field (27.4.2.)
15.3.4 The Ostrogradsky-Gauss theorem for the electrostatic field in vacuum states that vector flux of electrostatic field strength in vacuum passing through an arbitrary closed surface is proportional to the algebraic sum of the electric charges enclosed within the surface:

$$\oint_S \mathbf{E} \, d\mathbf{S} = \frac{1}{\varepsilon_0} \sum_{i=1}^{h} q_i \text{ [in SI units, where } \varepsilon_0 \text{ is the electric constant (14.2.7)]},$$

$$\oint_S \mathbf{E} \, d\mathbf{S} = 4\pi \sum_{i=1}^{h} q_i \text{ (in cgse units).}$$

The Ostrogradsky-Gauss theorem for the field in a dielectric is discussed in Sect. 18.3.3.
15.3.5 Together with the principle of superposition of fields (15.2.2), the Ostrogradsky-Gauss theorem is applied to calculate electric field vectors D. This is done by choosing a closed surface in such a way that in the expression for the displacement flux, the quantity D can be placed outside the sign of the surface integral. This can be done for fields set up by the simplest symmetrically arranged charges (charged lines, plane, sphere, etc.).

CHAPTER 16   ELECTRIC FIELD POTENTIAL

16.1 Work Done in Moving an Electric Charge in an Electrostatic Field

16.1.1 The work done in moving an electric charge q' in an electrostatic field of strength $\mathbf{E}$ (15.1.3) is independent of the shape of the path along which the charge is moved; it depends only upon the initial and final positions of the charge. In other words, electrostatic forces, like gravitational forces, are potential (conservative) forces (3.1.6). The work δW done by the force $\mathbf{F} = q'\mathbf{E}$ exerted on charge q' in moving it over the distance $d\mathbf{l}$, is equal to:

$$\delta W = F\,dl\cos\varphi = q'E\cos\alpha\,dl,$$

where α is the angle between the directions of vectors $\mathbf{E}$ and $d\mathbf{l}$, and φ is the angle between the directions of the vectors $\mathbf{F}$ and $d\mathbf{l}$. The work done in a finite movement of charge q' from point a to point b is equal to

$$W = q'\int\limits_a^b E\cos\alpha\,dl = q'\int\limits_a^b \mathbf{E}\,d\mathbf{l},$$

where $\mathbf{E}\,d\mathbf{l}$ is the scalar product of vectors $\mathbf{E}$ and $d\mathbf{l}$.

16.1.2 If the electric field has been set up by charge q, then

$$E = \frac{q}{4\pi\varepsilon_0\varepsilon_r r^2}, \quad dl\cos\alpha = dr,$$

and the work done in the movement of charge q' from point a to point b in this field is equal to

$$W = \frac{qq'}{4\pi\varepsilon_0\varepsilon_r}\int\limits_{r_1}^{r_2}\frac{dr}{r^2} = \frac{qq'}{4\pi\varepsilon_0\varepsilon_r}\left(\frac{1}{r_1} - \frac{1}{r_2}\right),$$

where r_1 and r_2 are the distances of points a and b from charge q (Fig. 16.1).

The work W done by the electric forces of repulsion of like charges q and q' is positive when the charges move away from

each other, and negative when they move toward each other. The work done by electric forces of attraction of unlike charges q and q' is positive when they move toward each other and negative when they move away from each other.

16.1.3 When an electric charge q' is moved in a field set up by a system of point charges q_1, q_2, ..., q_n, the charge q' is subject to the force

$$\mathbf{F} = \mathbf{F}_1 + \mathbf{F}_2 + \ldots + \mathbf{F}_n$$

and the work W done by the resultant force is equal to the algebraic sum of the work done by the component forces:

$$W = W_1 + \ldots + W_n = \sum_{i=1}^{n} \frac{q' q_i}{4\pi\varepsilon_0\varepsilon_r} \left(\frac{1}{r_{i1}} - \frac{1}{r_{i2}} \right),$$

where r_{i1} and r_{i2} are the distances of charge q_i to the initial and final positions of charge q'. Each of the amounts of work W_i and the total work W depend upon the initial and final positions of the charge q', but are independent of the shape of its path.

16.1.4 *Circulation of the field strength vector* along the closed path or loop L is the line integral

$$\oint_L E \, dl \cos \alpha = \oint_L \mathbf{E} \, d\mathbf{l}.$$
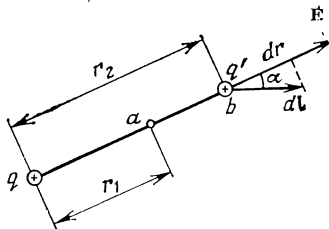
Fig. 16.1

This integral is numerically equal to the work done by electrostatic forces in moving a unit positive charge along a closed path. Since $r_{i1} = r_{i2}$ for a closed path, we have, according to Sect. 16.1.3, $\oint_L \mathbf{E} \, dl = 0$, i.e. the above-mentioned work equals zero.

A force field of strength E that complies with such a condition is said to be a *conservative one*. An electrostatic field is conservative.

The differential form of the condition indicating the conservative nature of an electrostatic field is one of Maxwell's equations for an electrostatic field (27.2.1).

16.2 Potential of an Electrostatic Field

16.2.1 The work δW (16.1.1) is equal to the decrease dE_p in potential energy (3.3.1) of the charge q' moving in an electrostatic field:

$$\delta W = -dE_p$$

and

$$W = \int_a^b q' E \, dl \cos \alpha = -\Delta E_p = E_{1p} - E_{2p},$$

where E_{1p} and E_{2p} are the values of the potential energy of charge q' at points a and b, and α is the angle between the directions of vectors $\mathbf{E}$ and $d\mathbf{l}$ (Fig. 16.1).

16.2.2 When point charge q' is moved by electrostatic forces in a field set up by charge q, the change dE_p in its potential energy in an infinitesimal displacement is

$$dE_p = -\delta W = -\frac{qq' \, dr}{4\pi\varepsilon_0\varepsilon_r r^2} \, .$$

Upon a finite displacement of charge q' from point a to point b (Fig. 16.1), the change ΔE_p in the potential energy of the charge is

$$\Delta E_p = E_{2p} - E_{1p} = \int_{r_1}^{r_2} dE_p = \frac{qq'}{4\pi\varepsilon_0\varepsilon_r r_2} - \frac{qq'}{4\pi\varepsilon_0\varepsilon_r r_1} \quad \begin{array}{l} \text{(in SI} \\ \text{units)} \end{array}$$

$$\Delta E_p = qq' \left(\frac{1}{\varepsilon_r r_2} - \frac{1}{\varepsilon_r r_1} \right) \text{ (in cgse units).}$$

16.2.3 When point charge q' is moved in a field set up by a system of point charges $(q_1, q_2, \ldots, q_n)$, the change in the

potential energy of charge q' is

$$\Delta E_p = q' \sum_{i=1}^{n} \left(\frac{q_i}{4\pi\varepsilon_0\varepsilon_r r_{i2}} - \frac{q_i}{4\pi\varepsilon_0\varepsilon_r r_{i1}} \right) \text{ (in SI units)},$$

$$\Delta E_p = q' \sum_{i=1}^{n} \left(\frac{q_i}{\varepsilon_r r_{i2}} - \frac{q_i}{\varepsilon_r r_{i1}} \right) \text{ (in cgse units)},$$

where r_{i1} and r_{i2} are the distances between charges q_i and q' before and after the displacement of the latter.
16.2.4 To find the absolute value of the potential energy of an electric charge at a given point of an electrostatic field, it is necessary to choose a reference point (3.3.4) for the potential energy. Integrating the equation in Sect. 16.2.2, we obtain for the general case:

$$E_p = \frac{qq'}{4\pi\varepsilon_0\varepsilon_r r} + C,$$

where C is an arbitrary constant. If we assume that $E_p = 0$ as $r \to \infty$, then $C = 0$ and the potential energy of charge q', located in the field of charge q at the distance r from this charge, equals

$$E_p = \frac{qq'}{4\pi\varepsilon_0\varepsilon_r r} \text{ (in SI units)},$$

$$E_p = \frac{qq'}{\varepsilon_r r} \text{ (in cgse units)}.$$

For like charges q and q', the potential energy of their interaction (repulsion) is positive and increases (decreases) as the charges move closer together (away from each other). In attraction of unlike charges $E_p < 0$ and increases to zero as one of the charges approaches infinity. The curves showing the dependence of the potential energy of two point charges on the distance between them are given in Fig. 16.2.

16.2.5 The potential energy E_p of charge q', located in a field set up by a system of point charges q_1, q_2, . . ., q_n is equal to the sum of the potential energies E_{ip} in the fields set up by each of the charges separately:

$$E_p = \sum_{i=1}^{n} E_{ip} = q' \sum_{i=1}^{n} \frac{q_i}{4\pi\varepsilon_0\varepsilon_r r_i} \quad \text{(in SI units)},$$

where r_i is the distance between charges q_i and q'.
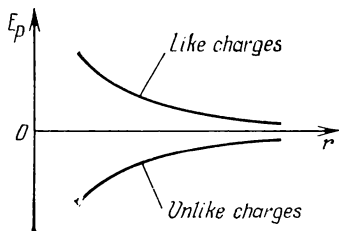
Fig. 16.2

16.2.6 *The potential of an electrostatic field* is an energy characteristic of the field, it is the ratio of the potential energy E_p of charge q' to the magnitude of the charge

$$V = \frac{E_p}{q'} = \sum_{i=1}^{n} \frac{q_i}{4\pi\varepsilon_0\varepsilon_r r_i} \quad \text{(in SI units)}.$$

This ratio is independent of the magnitude of the charge q' and is numerically equal to the potential energy of a unit test charge (15.1.3) placed at the point being considered in the field.

Examples:
(1) The potential of an electrostatic field set up by one point charge q in a homogeneous and isotropic dielectric is

$$V = \frac{q}{4\pi\varepsilon_0\varepsilon_r r} \text{ (in SI units)},$$

$$V = \frac{q}{\varepsilon_r r} \text{ (in cgse units)}.$$

(2) The potential of a field set up by a charged conductor (16.4.4) in a homogeneous and isotropic dielectric is

$$V = \frac{1}{4\pi\varepsilon_0\varepsilon_r} \oint_S \frac{\sigma\, dS}{r} \text{ (in SI units)},$$

$$V = \frac{1}{\varepsilon_r} \oint_S \frac{\sigma\, dS}{r} \text{ (in cgse units)},$$

where σ is the surface charge density on the conductor (15.2.3) and dS is an element of surface of the conductor.
(3) The potential of an isolated conducting sphere of radius R, having the charge q, is

$$V = \frac{q}{4\pi\varepsilon_0\varepsilon_r R} \text{ (in SI units)},$$

$$V = \frac{q}{\varepsilon_r R} \text{ (in cgse units)}.$$

16.2.7 The work W done by electric forces in moving charge q' from point a to point b of an electrostatic field is

$$W = E_{1p} - E_{2p} = q'\, (V_1 - V_2) = q'\, \Delta V,$$

where E_{1p} and E_{2p} are the potential energies of charge q' at points a and b, V_1 and V_2 are the potentials of the field at the same points, and $\Delta V = V_1 - V_2$ is the *potential difference*. If point b is at infinity, $E_{2p} = 0$ and $V_2 = 0$. Then the work W_∞ required to move charge q' from point a to infinity is equal to

$$W_\infty = E_{1p} = q'V_1.$$

Owing to the arbitrary nature of point a subindex 1 can be discarded. Then

$$V = \frac{W_\infty}{q'}.$$

Hence, the potential at a given point of an electrostatic field is numerically equal to the work done by electrostatic forces in moving a unit positive charge from the point in the field to infinity. This work is also numerically equal to that done by external forces (acting against the forces of the electrostatic field) in moving a unit positive charge from infinity to the given point in the field.

'16.2.8 In the study of electrostatic fields, all the problems require a knowledge of the potential difference between any points in a field rather than the absolute values of the potentials at these points. Therefore, the choice of a point of zero potential depends only on its convenience for solving the given problem. It frequently proves convenient to assume that the potential of the earth is equal to zero.

16.3 Relation Between the Potential and Strength of an Electrostatic Field

16.3.1 On the basis of the equations in Sects. 16.2.1 and 16.2.6, an element of work δW done in the infinitesimal displacement of charge q' in an electrostatic field is

$$\delta W = q'E \cos \alpha \, dl = -dE_p = -q' \, dV,$$

or

$$E \cos \alpha \, dl = -dV; \quad E \, dl = -dV.$$

But $dl \cos \alpha = dl_0$ where dl_0 is an element of length of the line of force (15.1.6) (Fig. 16.3). Hence, $E = -dV/dl_0$. The derivative dV/dl_0 is the rate of change of the potential along a line of force and is numerically equal to the change in potential per unit length of a line of force.

16.3.2 The projection E_l of vector **E** on the direction of motion $d\mathbf{l}$ is equal to: $E_l = E \cos \alpha$. Therefore

$$E_l = -\frac{dV}{dt} \bullet$$

It is obvious that $E_l \leqslant E$ and $|\, dV/dl\, | \leqslant |\, dV/dl_0\, |$. Hence, E_l and dV/dl reach their maximum values when $d\mathbf{l}$ is tangent to the line of force.

In the vicinity of a given point in an electrostatic field, the potential varies most rapidly in the direction of a line of force. The sense of vector **E** is toward the most rapid decrease in potential.

16.3.3 In a more general form the relation between the strength and potential of an electrostatic field is

E $= -\text{grad } V$,
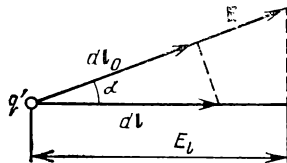
where grad V is the *potential gradient* vector, directed toward the most rapid increase in the potential and numerically equal to its rate of change per unit length in this direction. If the potential V is regarded as a function of the three Cartesian coordinates of the given point in the field, then

Fig. 16.3

$$\text{grad } V = \frac{\partial V}{\partial x}\, \mathbf{i} + \frac{\partial V}{\partial y}\, \mathbf{j} + \frac{\partial V}{\partial z}\, \mathbf{k},$$

where **i**, **j** and **k** are unit vectors along the OX, OY and OZ axes. The projections of the electrostatic field strength vector on the coordinate axes are related to the field potential by the equations:

$$E_x = -\frac{\partial V}{\partial x}\, ,\ E_y = -\frac{\partial V}{\partial y}\ \text{ and }\ E_z = -\frac{\partial V}{\partial z}$$

16.3.4 When an electric charge is moved in a direction $d\mathbf{l}$, perpendicular to a line of force, i.e. perpendicular to vector **E**, $E_l = 0$ and $dV/dl = 0$ (16.3.1), i.e. $V = \text{const}$. The potential is the same at all points of a curve orthogonal to a line of force.

The geometrical locus of all points having the same potential is called an *equipotential surface*. It follows from the aforesaid that equipotential surfaces are everywhere orthogonal to lines of force.

The work done in moving an electric charge along one and the same equipotential surface equals zero.

An infinite number of equipotential surfaces can be constructed about any system of electric charges. They are usually constructed so that the potential difference is the same between any two adjacent equipotential surfaces.

16.3.5 There are two methods for the graphical representation of electrostatic fields: by means of lines of force and by means of equipotential surfaces. If the arrangement of the lines of force of an electrostatic field is known, equipotential surfaces can be constructed. Conversely, knowing the arrangement of the equipotential surfaces, the magnitude and direction of the field strength can be found for each point of the field, i.e. the lines of force can be constructed.

16.4 Conductors in an Electrostatic Field

16.4.1 Solid metal conductors contain current carriers. These are conduction electrons (free electrons) that can be moved by the external electric field through the volume of the conductor. Conduction electrons appear when a conducting metal goes over from a less condensed to a more condensed state, i.e. from the gaseous to the liquid or solid state. At this valence electrons (39.3.9) are collectivized; they are separated from "their" atoms and form a special *electron gas*.

16.4.2 The electrical properties of conductors under electrostatic conditions are determined by the behaviour of the conduction electrons in the external electrostatic field. In the absence of an external electrostatic field, the electric fields of the conduction electrons and the positive ions of the metal (40.1.3), the atomic cores, mutually compensate each other. When a metal conductor is placed into an external electrostatic field, this field redistributes the conduction electrons in the conductor so that at any point within the conductor the conduction electrons and the positive ions compensate the external electrostatic field.

At any point inside a conductor located in an electrostatic field, the strength of the resultant steady-state electric field is equal to zero.

16.4.3 At the surface of the conductor the field strength vector **E** should be normal to the surface. If this were not so the tangential component $\mathbf{E}_\tau$ of vector **E** would cause the charges to move along the conductor's surface, and this would contradict the fixed distribution of the charges. This leads to the following consequences:

(a) $\mathbf{E} = 0$ at all points inside the conductor; at all points on its surface $\mathbf{E} = \mathbf{E}_n$ (and $\mathbf{E}_\tau = 0$), where $\mathbf{E}_n$ is the normal component of the field strength vector;

(b) the whole volume of a conductor in an electrostatic field is equipotential, because, for any point within the conductor,

$$\frac{dV}{dl} = -E \cos \alpha = 0 \quad \text{and} \quad V = \text{const};$$

(c) the surface of the conductor is also equipotential (16.3.4), because for any line on the surface

$$\frac{dV}{dl} = -E_\tau = 0 \quad \text{and} \quad V = \text{const};$$

(d) in a charged conductor the uncompensated charges are located only on its surface. This follows from the Ostrogradsky-Gauss electric flux theorem (15.3.3), according to which the total charge q inside a conductor and enclosed within a certain volume bounded by the arbitrary closed surface S is equal to

$$q = \Phi_e = \oint_S D \, dS \cos \alpha = 0,$$

because $D = 0$ at all points on the surface.

16.4.4 When the electrostatic field is set up by a charged conductor the displacement and strength of this field close to the surface of the conductor are found by the equations

$$D_n = \sigma \quad \text{and} \quad E_{\perp} = \frac{\sigma}{\varepsilon_0 \varepsilon_r} \quad \text{(in SI units)},$$

where n is the outward normal to the conductor's surface, σ is the surface charge density (15.2.3), ε_r is the relative permittivity (dielectric constant) of the medium (14.2.4) and ε_0 is the electric constant in SI units (14.2.7).

16*

CHAPTER 17   CAPACITANCE

17.1 Capacitance of an Isolated Conductor

17.1.1 Upon an increase in the amount of electricity transmitted to an isolated conductor, the charges are distributed over the conductor's surface with varying surface density. The way in which the charges are distributed depends only on the shape of the conductor, but not on the amount of electricity that was previously on the conductor. Each new portion of charges is distributed over the surface of the conductor similar to the preceding portion. It follows that, at any point on the conductor's surface, the surface charge density σ (15.2.3) increases in proportion to the charge q on the conductor: $\sigma = kq$, where $k = f(x, y, z)$ is a function of the coordinates of the point being considered on the surface.

The potential of the field set up by a charged conductor (16.2.6) is

$$V = \frac{1}{4\pi\varepsilon_0\varepsilon_r} \oint_S \frac{\sigma \, dS}{r} = \frac{q}{4\pi\varepsilon_0\varepsilon_r} \oint_S \frac{k \, dS}{r} \, .$$

For points on the surface S of the conductor this integral depends only on the size and shape of the conductor's surface. Hence the potential V of the conductor is proportional to its charge q.

17.1.2 The ratio of the charge q of an isolated conductor to its potential V is called the *capacitance* of the given conductor:

$$C = \frac{q}{V} \quad \text{or} \quad C = 4\pi\varepsilon_0\varepsilon_r \left(\oint_S \frac{k \, dS}{r} \right)^{-1} \text{(in SI units).}$$

The capacitance of an isolated conductor is numerically equal to the charge that changes the potential of the conductor by one unit.

The capacitance of an isolated conductor depends upon its shape and dimensions. For geometrically similar conductors, their capacitances are proportional to their linear dimensions.

The capacitance of an isolated conductor depends upon the dielectric properties of the surrounding medium. The capacitance of a conductor in a homogeneous isotropic medium is proportional to the relative permittivity (dielectric constant) (14.2.4) of the medium.

17.1.3 It follows from the equation for the potential of an isolated conducting sphere of radius R (16.2.6) that the *capacitance of* such a *sphere* is

$C = 4\pi\varepsilon_0\varepsilon_r R$ (in SI units),

$C = \varepsilon_r R$ (in cgse units).

17.2 Mutual Capacitance. Capacitors

17.2.1 When there are other conductors close to conductor A, the capacitance of the latter is greater than for an identical, but isolated, conductor. This is due to the fact that when the charge q is transferred to conductor A, charges are induced on the conductors surrounding it under the influence of the field set up by conductor A. The portions of the surrounding conductors nearest to the inducing charge q become oppositely charged (Fig. 17.1). These charges weaken the field set up by charge q, lower the potential of conductor A and increase its capacitance.
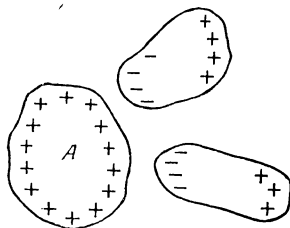
Fig. 17.1

17.2.2 In a system of two conductors arranged close to each other, having charges q equal in magnitude but opposite in sign, a potential difference $V_1 - V_2$ is produced that is proportional to q:

$$V_1 - V_2 = \frac{1}{C}\, q,$$

where $C = q/(V_1 - V_2)$ is the *mutual capacitance* of the two conductors.

The mutual capacitance of two conductors is numerically equal to the charge that it is necessary to transfer from one conductor to the other to change the potential difference between them by one unit.

17.2.3 The mutual capacitance C of two conductors depends upon their shapes, dimensions and relative positions. Moreover, C depends upon the dielectric properties of the medium surrounding the conductors. If the medium is homogeneous and isotropic, C is directly proportional to the relative permittivity (14.2.4) of the medium.

When one of the conductors is moved away to infinity, the potential difference $V_1 - V_2$ between them increases and their mutual capacitance is reduced, tending to the capacitance of the remaining isolated conductor.

17.2.4 A *capacitor* is a system of two charged conductors, insulated from each other and having charges of equal magnitude and opposite sign, provided that their shapes and relative position are such that the conductors set up a concentrated electrostatic field in a limited region of space. In the case of parallel-plate capacitors, the conductors are called *plates*. The capacitance of a capacitor is the mutual capacitance of its plates (or, in the general case, conductors).

17.2.5 The capacitance of a *parallel-plate capacitor*, consisting of two metal plates, is

$$C = \frac{\varepsilon_0 \varepsilon_\mathrm{r} S}{d} \text{ (in SI units)},$$

$$C = \frac{\varepsilon_\mathrm{r} S}{4\pi d} \text{ (in cgse units)},$$

where ε_r is the relative permittivity of the medium filling the space between the plates, S is the area of each plate, and d is the distance between them. For a multiple-plate capacitor consisting of n plates, the formula for the capacitance contains $S(n-1)$ instead of S. The capacitance formula is valid only for small d values, when the lack of uniformity of the electrostatic field at the edges of the capacitor plates can be ignored.

17.2.6 A *spherical capacitor* consists of two concentric hollow metal spheres A and B, of radii r_1 and r_2 (Fig. 17.2). The spheres are charged on their surfaces and set up fields only outside themselves (15.1.4). Hence, in the space between the spheres

the field is set up only by the charge of sphere A. Outside the capacitor, the fields of the oppositely charged spheres A and B mutually eliminate each other.

The capacitance of a spherical capacitor is calculated by the formula

$$C = \frac{4\pi\varepsilon_0\varepsilon_r r_1 r_2}{r_2 - r_1} \text{ (in SI units)},$$

$$C = \frac{\varepsilon_r r_1 r_2}{r_2 - r_1} \text{ (in cgse units)}.$$

As $r_2 \to \infty$ and $1/r_2 \to 0$, the inner sphere becomes an isolated one and $C = 4\pi\varepsilon_0\varepsilon_r r_1$ (cf. 17.1.3). For any finite values of r_1 and r_2

$$C = 4\pi\varepsilon_0\varepsilon_r r_1 \frac{r_2}{r_2 - r_1} > 4\pi\varepsilon_0\varepsilon_r r_1,$$

i.e. the capacitance of a spherical capacitor is greater than that of an isolated hollow sphere.
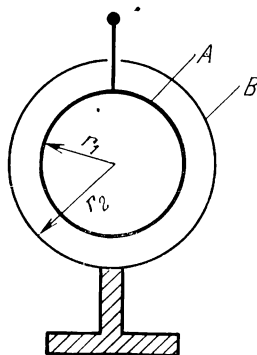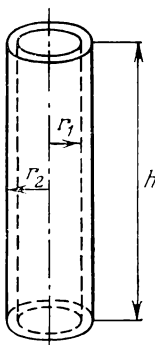
Fig. 17.2   Fig. 17.3

17.2.7 A *cylindrical capacitor* consists of two hollow coaxial metal cylinders of height h and radii r_1 and r_2 (Fig. 17.3). The

formula for the capacitance of a cylindrical capacitor (or of a coaxial cable) is of the form

$$C = \frac{2\pi\varepsilon_0\varepsilon_r h}{\ln\dfrac{r_2}{r_1}} \quad \text{(in SI units)},$$

$$C = \frac{\varepsilon_r h}{2\ln\dfrac{r_2}{r_1}} \quad \text{(in cgse units)}.$$

17.2.8 All types of capacitors are characterized by their *breakdown voltage*, which is a potential difference across the plates at which an electric discharge (22.5.1) passes through the layer of dielectric in the capacitor. The breakdown voltage depends upon the properties of the dielectric, its thickness and the shape of the plates.

17.2.9 To obtain a high capacitance *capacitors can be connected in parallel*. This is done by connecting the plates of like charge together. Here the total capacitance C is

$$C = \sum_{i=1}^{n} C_i,$$

where C_i is the capacitance of the ith capacitor.

17.2.10 When *capacitors are connected in series* their oppositely charged plates are connected together. The quantities added together in this case are reciprocals of the capacitance C_i of each capacitor:

$$\frac{1}{C} = \sum_{i=1}^{n} \frac{1}{C_i}.$$

The resultant capacitance C is always less than the minimum capacitance of any capacitor in the bank. Upon a connection in series, the possibility of a breakdown (17.2.8) of the capacitors is reduced because the potential difference across each capacitor is only a part of the total potential difference across the whole bank of capacitors.

CHAPTER 18   DIELECTRICS
IN AN ELECTRIC FIELD

18.1 Dipole Moments of Molecules of a Dielectric

18.1.1 *Dielectrics* are substances that do not conduct electric current. At not especially high temperatures and under conditions when dielectrics are not subject to very strong electric fields, these substances, in contrast to conductors, have no carriers of electric current.

18.1.2 The molecules of a dielectric are electrically neutral and contain an equal number of positive and negative charges. Nevertheless, the molecules do have electrical properties. As a first approximation, a molecule of a dielectric can be regarded as a dipole having the electric dipole moment $\mathbf{p}_e = ql$ (15.2.3), where q is the magnitude of the total positive (and also total negative) charges, located, respectively, at the centres of gravity of the charges, l is the distance between the centres of gravity of the positive and negative charges. As any dipole, a molecule of matter sets up an electric field in the surrounding space (15.2.3).

18.1.3 A dielectric is said to be *nonpolar* (a *nonpolar dielectric*) if the electrons of the atoms in its molecules are arranged symmetrically with respect to their nuclei (H_2, O_2, CCl_4, etc.). In such molecules the centres of gravity of the positive and negative charges coincide when there is no external electric field $[l = 0$ (18.1.2)] and the dipole moment $\mathbf{p}_e$ of the molecule is equal to zero. When a nonpolar dielectric is placed into an external electric field, the electron shells (39.3.6) in the atoms (or molecules) are deformed and the centres of gravity of the positive and negative charges are displaced with respect to each other ($l \neq 0$). At this an *induced electric dipole moment* appears in atoms (or molecules) of the dielectric. This moment is proportional to the strength E of the external electric field:

$$\mathbf{p}_e = \varepsilon_0 \alpha \mathbf{E} \quad \text{(in SI units)},$$

$$\mathbf{p}_e = \alpha \mathbf{E} \quad \text{(in cgse units)},$$

where α is the *polarizability* of the molecule (or **atom**) and ε_0 is the electric constant in SI units (14.2.7).

The polarizability of a molecule depends only on its volume. It is important that α does not depend upon temperature. The thermal motion of the molecules in nonpolar dielectrics has no effect on the induction of dipole moments in them. Molecules with such dipole moments are like *quasi-elastic (induced) dipoles*.

18.1.4 A *polar dielectric* is one in which the molecules (or atoms) have electrons arranged asymmetrically with respect to their nuclei (H_2O, HCl, NH_3, CH_3Cl, etc.). In such molecules the centres of gravity of the positive and negative charges do not coincide, but are at a practically constant distance l from each other. With respect to electrical properties, molecules of polar dielectrics are like *rigid dipoles* that have a *permanent dipole moment* $\mathbf{p}_e = $ const.

18.1.5 A *permanent dipole* placed in a homogeneous external electrostatic field is subject to the action of a force couple (4.1.6) with the moment of force (torque) $\mathbf{M}$ equal to

$$\mathbf{M} = [\mathbf{p}_e \mathbf{E}].$$

The moment $\mathbf{M}$ of the force couple is perpendicular to the plane passing through vectors $\mathbf{p}_e$ and $\mathbf{E}$. From the tip of vector $\mathbf{M}$ rotation through the smaller angle from vectors $\mathbf{p}_e$ to $\mathbf{E}$ is seen to be counterclockwise. In Fig. 18.1 the moment of force $\mathbf{M}$ is directed normal to the diagram into the page and tends to turn the dipole so that $\mathbf{p}_e$ and $\mathbf{E}$ are parallel to each other. In real molecules of polar dielectrics, the molecules are deformed in addition to the turning of the dipole axes into alignment with the field. This leads to the induction of a certain dipole moment (18.1.3).
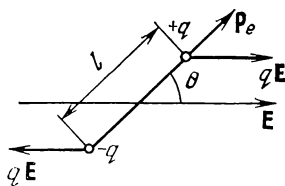
Fig. 18.1

18.1.6 When a permanent dipole is placed into a nonhomogeneous electric field, in which the field strength E varies along the length l of the dipole, then, in addition to the torque $\mathbf{M}$ (18.1.5), the dipole is subject to force F, which is equal to

$$\mathbf{F} = p_e \frac{\partial \mathbf{E}}{\partial l}$$

or, in the more general form:

$$\mathbf{F} = p_{ex} \frac{\partial \mathbf{E}}{\partial x} + p_{ey} \frac{\partial \mathbf{E}}{\partial y} + p_{ez} \frac{\partial \mathbf{E}}{\partial z},$$

where p_{ex}, p_{ey} and p_{ez} are the projections of vector $\mathbf{p_e}$ on the axes of a Cartesian coordinate system. In particular, if vector $\mathbf{E}$ is directed along the OX axis and depends only on the coordinate x [$E_x = E(x)$, $E_y = E_z = 0$ and $\mathbf{E} = E_x \mathbf{i}$], then

$$\mathbf{F} = p_{ex} \frac{\partial E}{\partial x} \mathbf{i}.$$

By the action of force $\mathbf{F}$ a free permanent dipole tends to move to a region of highest field strength.

18.1.7 The potential energy E_p of a permanent dipole in an external electric field is

$$E_p = -\mathbf{p_e}\mathbf{E} = -p_e E \cos\theta,$$

where $\mathbf{p_e}$ is the electric dipole moment, $\mathbf{E}$ is the field strength at the point where the dipole is located and θ is the angle between the dipole axis and vector $\mathbf{E}$. The minus sign indicates that the stable position of the dipole corresponds to the angle $\theta = 0$, at which the potential energy of the dipole has its minimum value and vector $\mathbf{p_e}$ is directed along vector $\mathbf{E}$.

18.2 Polarization of Dielectrics

18.2.1 When a polar dielectric (17.1.4) is not in an external electric field, the vectors of its dipole moments are chaotically oriented as a result of the random thermal motion of the molecules. Consequently, the sum of the dipole moments of all the molecules is equal to zero in any physically infinitesimal volume ΔV^*.

No dipole moments of the molecules whatsoever can be induced (18.1.3) in a nonpolar dielectric that is not in an external electric field.

---

* Here $\Delta V \gg v_0$, where v_0 is the volume of a single molecule, and an immense number of molecules are contained in volume ΔV.

18.2.2 When a *dielectric* is placed in an external electric field,
it *is polarized*. This means that in any element of volume ΔV,
the total dipole moment of its molecules is no longer equal to
zero. A dielectric in such a state is said to be polarized (*polar-
ized dielectric*). In accordance with the structure of the mole-
cules (or atoms), distinction is made between the following
three types of polarization:

(a) *Orientation polarization* of polar dielectrics (18.1.4). The
external electric field tends to align the dipole moments of the
permanent dipoles along the direction of the field (18.1.5).
This is hindered by the random thermal motion of the molecules
which tends to disarrange the dipoles arbitrarily. The combined
action of the field and thermal motion results in preferred orien-
tation of the electric dipole moments along the field. This
polarization increases with the strength of the electric field
and with a drop in temperature.

(b) *Electronic polarization* in nonpolar dielectrics (18.1.3) In
the molecules of dielectrics of this type dipole moments directed
along the field are induced (18.1.3) by the action of the external
electric field. The thermal motion of the molecules has no
effect on electronic polarization. In gaseous and liquid dielec-
trics, electronic polarization takes place practically simulta-
neously with orientational polarization.

(c) *Ionic polarization* in solid dielectrics having ionic crystal
lattices (40.1.3), such as NaCl and CsCl. The external electric
field causes displacement of all the positive ions in the direction
of the field strength E and all the negative ions in the opposite
direction.

18.2.3 A quantitative measure of the polarization of a dielec-
tric is the polarization vector $\mathbf{P}_e$. The *polarization vector* (or
simply the *polarization*) is the ratio of the electric dipole moment
of a small volume ΔV of the dielectric to this volume:

$$\mathbf{P}_e = \frac{1}{\Delta V} \sum_{i=1}^{n} \mathbf{p}_{ei},$$

where $\mathbf{p}_{ei}$ is the electric dipole moment of the ith molecule,
and n is the total number of molecules in the volume ΔV. This
volume should be so small that the electric field (15.1.2) within
it can be assumed uniform. At the same time, the number n

of molecules in it should be sufficiently large to allow statistical methods of investigation (8.2.2) to be applied.

18.2.4 For a homogeneous nonpolar dielectric (18.1.3) in a uniform electric field

$$\mathbf{P}_e = n_0 \mathbf{p}_e,$$

where n_0 is the number of molecules in unit volume, and $\mathbf{p}_e$ is the dipole moment of a single molecule. Making use of the formula for $\mathbf{p}_e$ (18.1.3) we obtain

$$\mathbf{P}_e = n_0 \varepsilon_0 \alpha \mathbf{E} = \varepsilon_0 \chi_e \mathbf{E} \quad \text{(in SI units)},$$
$$\mathbf{P}_e = n_0 \alpha \mathbf{E} = \chi_e \mathbf{E} \qquad \text{(in cgse units)},$$

where $\chi_e = n_0 \alpha$ is the *dielectric* (or *electric*) *susceptibility of the substance*.

18.2.5 For a homogeneous polar dielectric (18.1.4) in a uniform electric field, $\mathbf{P}_e = n_0 \langle \mathbf{p}_e \rangle$, where $\langle \mathbf{p}_e \rangle$ is the average value of the component along the field of the permanent dipole moment of the molecule. When a polar dielectric is in a weak external field, the dielectric susceptibility is calculated by the *Debye-Langevin formula*:

$$\chi_e = \frac{n_0 p_e^2}{3\varepsilon_0 kT} \quad \text{(in SI units)},$$

$$\chi_e = \frac{n_0 p_e^2}{3kT} \quad \text{(in cgse units)}.$$

Here k is Boltzmann's constant (8.4.5), and T is the absolute temperature. The remaining notation is given in Sect. 18.1.3. Shown in Fig. 18.2 is the dependence $\chi_e = \chi_e(1/T)$ for nonpolar (*a*) and for polar (*b*) molecules. The straight line (*b*) does not pass through the origin of coordinates because orientational and electronic polarization (18.2.2b) usually occur in polar molecules and the dielectric susceptibility consists of two parts: $\chi_e = \chi_e' + \chi_e''$, where χ_e' and χ_e'' are expressed by the equations in Sects. 18.2.4 and 18.2.5.

18.2.6 If a dielectric is in a uniform electric field, the electrical neutrality of any volume ΔV of the dielectric, containing a sufficiently large number of molecules, is provided for by the mutual

compensation of the dipole charges of opposite sign alongside one another (Fig. 18.3). In the thin layers at surfaces S_1 and S_2 of the dielectric, bounding its volume, *surface-bound polarization* charges appear as a result of the polarization of the dielectric. At the surface where the lines of force (15.1.6) of the external electric field enter the dielectric, negative charges of the ends
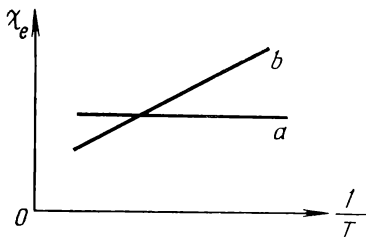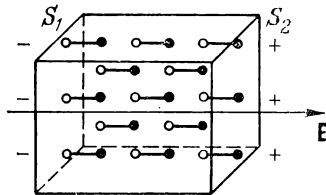
Fig. 18.2   Fig. 18.3

of the dipole molecules appear. Positive charges appear at the opposite surface. The surface density (15.2.3) σ_p of the polarization charges is calculated by the formula

$$\sigma_p = P_{en},$$

where P_{en} is the projection of vector $\mathbf{P}_e$ on the outward normal to the surface of the dielectric.

In a nonuniform electric field, the polarization of a dielectric is also nonuniform and the polarization vector $\mathbf{P}_e$ depends upon the coordinates. Hence, in addition to surface-bound charges, space polarization charges also appear. These are distributed with the volume charge density (15.2.3) ρ_p:

$$\rho_p = -\operatorname{div} \mathbf{P}_e,$$

where $\operatorname{div} \mathbf{P}_e = \dfrac{\partial P_{ex}}{\partial x} + \dfrac{\partial P_{ey}}{\partial y} + \dfrac{\partial P_{ez}}{\partial z}$ is the divergence of the polarization vector.

18.3 Relation Between Displacement, Field Strength and Polarization Vectors

18.3.1 Distinction is made in substances between two kinds of charges: free and bound.

Bound charges are those included in the composition of atoms and molecules, as well as the charges of ions in crystalline solids with an ionic lattice (40.1.3). Charges not bound to the afore-mentioned particles of matter are said to be free. *Free charges* include: (a) current-conducting charges in conducting media [conduction electrons in metals and semiconductors (16.4.1), holes in semiconductors (41.10.3), ions in electrolytes and gases], (b) surplus charges that are transmitted to a body by various means and violate its electrical neutrality. An example are charges imparted from outside onto the surface of a dielectric.

18.3.2 Generally, an electric field is set up in a dielectric by both free and bound charges. The field strength vector **E** characterizes the resultant field in the dielectric, set up by the two kinds of charges, and depends upon its electrical properties: the relative permittivity (dielectric constant) ε_r (14.2.4). But the primary source of an electric field in a dielectric is the free charges. As a matter of fact, a field of bound charges is set up in a dielectric as a result of its polarization when it is placed in an external electric field set up by free electric charges.

18.3.3 The information contained in Sects. 18.3.1 and 18.3.2 call for a certain revision of the Ostrogradsky-Gauss theorem (15.3.3). The algebraic sum of the charges $\sum\limits_{i=1}^{k} q_i$ on the right-hand side of the theorem is to be regarded as the algebraic sum of the free electric charges enclosed by surface S, i.e. $\sum\limits_{i=1}^{k} q_i = q_{\text{free}}$, and the theorem is then of the form

$$\oint\limits_S \mathbf{D}\, d\mathbf{S} = \oint\limits_S D_n\, dS = q_{\text{free}} \text{ (in SI units)},$$

$$\oint\limits_S \mathbf{D}\, d\mathbf{S} = \oint\limits_S D_n\, dS = 4\pi q_{\text{free}} \text{ (in cgse units)}.$$

In this form the Ostrogradsky-Gauss theorem is valid for an electric field either in homogeneous and isotropic or in nonhomogeneous and anisotropic media (cf. 15.3.3).

18.3.4 For an electric field in vacuum

$\mathbf{D} = \varepsilon_0 \mathbf{E}$ (in SI units),

$\mathbf{D} = \mathbf{E}$ (in cgse units).

The flux of the field strength vector $\mathbf{E}$ through an arbitrary closed surface S in vacuum is

$$\oint_S \varepsilon_0 E_n \, dS = q_{\text{free}} \text{ (in SI units)},$$

$$\oint_S E_n \, dS = 4\pi q_{\text{free}} \text{ (in cgse units)}.$$

In accordance with Sect. 18.3.2, the flux of vector $\mathbf{E}$ for a field in a substance is

$$\oint_S \varepsilon_0 E_n \, dS = q_{\text{free}} + q_{\text{bnd}} \text{ (in SI units)},$$

$$\oint_S E_n \, dS = 4\pi \, (q_{\text{free}} + q_{\text{bnd}}) \text{ (in cgse units)}.$$

The sum of bound charges q_{bnd} within closed surface S is calculated by the equation

$$q_{\text{bnd}} = -\oint_S P_{\text{en}} \, dS,$$

where P_{en} is the projection of the polarization vector on the outward normal to surface dS. Hence

$$\oint_S (\varepsilon_0 E_n + P_{\text{en}}) \, dS = q_{\text{free}} \text{ (in SI units)},$$

$$\oint_S (E_n + 4\pi P_{\text{en}}) \, dS = 4\pi q_{\text{free}} \text{ (in cgse units)}.$$

A correlation with the general formulation of the Ostrogradsky-Gauss theorem (18.3.3) enables a relation to be established between vectors $\mathbf{D}$, $\mathbf{E}$ and $\mathbf{P}_e$:

$$D_n = \varepsilon_0 E_n + P_{en} \quad \text{(in SI units)},$$

$$D_n = E_n + 4\pi P_{en} \quad \text{(in cgse units)},$$

or, by virtue of the arbitrary nature of outward normal $\mathbf{n}$,

$$\mathbf{D} = \varepsilon_0 \mathbf{E} + \mathbf{P}_e \quad \text{(in SI units)},$$

$$\mathbf{D} = \mathbf{E} + 4\pi \mathbf{P}_e \quad \text{(in cgse units)}.$$

These formulas are generalizations of the formulas in Sect. 15.3.1. Making use of the results obtained in Sect. 18.2.3, we have for an isotropic homogeneous medium

$$\mathbf{D} = \varepsilon_0 \mathbf{E} + \varepsilon_0 \chi_e \mathbf{E} = \varepsilon_0 (1 + \chi_e) \mathbf{E} \quad \text{(in SI units)},$$

$$\mathbf{D} = \mathbf{E} + 4\pi \chi_e \mathbf{E} = (1 + 4\pi \chi_e) \mathbf{E} \quad \text{(in cgse units)}.$$

Consequently

$$\mathbf{D} = \varepsilon_0 \varepsilon_r \mathbf{E}, \text{ where } \varepsilon_r = 1 + \chi_e \quad \text{(in SI units)},$$

$$\mathbf{D} = \varepsilon_r \mathbf{E}, \text{ where } \varepsilon_r = 1 + 4\pi \chi_e \quad \text{(in cgse units)}.$$

The quantity ε_r is the relative permittivity (dielectric constant) (cf. 14.2.4).
For vacuum $\varepsilon_r = 1$ and $\chi_e = 0$.
The concept of ε_r introduced in Sect. 14.2.4 has a meaning only for isotropic homogeneous media.

18.4 Ferroelectric Materials

18.4.1 *Ferroelectrics* belong to a group of crystalline dielectrics. The first substance of this type to be investigated was Seignette's salt (potassium sodium tartrate) $NaKC_4H_4O_6 \cdot 4H_2O$ and, consequently, they are sometimes called seignette-electrics. Another example of a ferroelectric is barium titanate $BaTiO_3$. Typical of ferroelectrics is a drastic increase in the relative permittivity (dielectric constant) (14.2.4) in a definite temperature range (Fig. 18.4).
18.4.2 The relative permittivity ε_r and the dielectric susceptibility χ_e (18.2.4) of ferroelectrics are functions of the field strength

E in the substance (Fig. 18.5). As a result, no linear relation between vectors P_e and E is observed in these substances. The dependence of the electric displacement D on the field strength
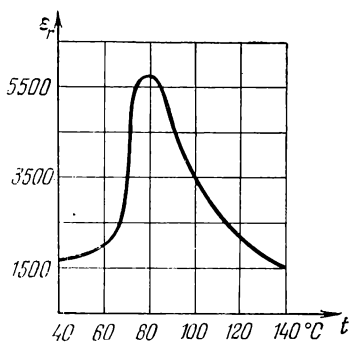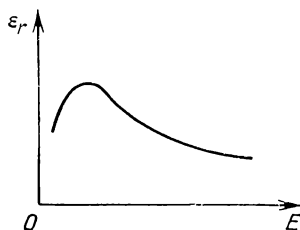
Fig. 18.4   Fig. 18.5
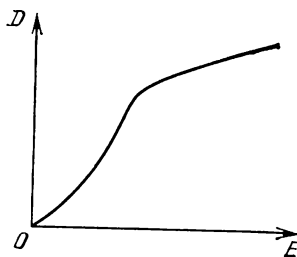
Fig. 18.6

is of a complex nature and a linear relation between D and E exists only at very high E values (Fig. 18.6).

18.4.3 A monocrystal of a ferroelectric is divided into spontaneously polarized regions, called *domains* (cf. 26.5.4). This *spontaneous polarization* of the domains is a result of the orientation of the dipole moments of the molecules in the domain in a defi-

nite direction. In the absence of an external electric field, the polarization vectors of the various domains are directed at random and for a large monocrystal or a polycrystal the total polarization is equal, on an average, to zero. When a ferroelectric is placed in an electric field, the electric moments of the domains are reoriented and the crystal acquires a total polarization not equal to zero.

18.4.4 Domains are formed in ferroelectrics within a definite temperature range, between the upper and lower *Curie points* θ_C^{up} and θ_C^{low}(cf. 26.5.2). For Seignette's salt (also called Rochelle salt) (18.4.1), $\theta_C^{up} = 298$ K and $\theta_C^{low} = 258$ K. At $T \geqslant \theta_C^{up}$ the forces of interaction between the dipoles cannot resist the thermal motion, the spontaneous polarization of the domains is violated and the ferroelectric is converted into an ordinary polar dielectric. The drastic increase in heat capacity of the substance indicates that a second-order phase transition (12.4.2) occurs at the Curie point. A disordered phase exists above the Curie point θ_C^{up} and the dielectric is not polarized in the absence of an external field. At $T < \theta_C^{up}$ there is an ordered phase indicated by the spontaneous polarization in the domains*.

18.4.5 The phenomenon of *dielectric*, or *ferromagnetic*, *hysteresis* (lag) is observed in ferroelectrics. As is evident from Fig. 18.7, with an increase in the strength of the external field, the magnitude of vector $\mathbf{P_e}$ increases and reaches saturation at point *a*. Then, as E is reduced to zero, a ferroelectric has *remanent polarization*, characterized by the value of the polarization vector $\mathbf{P_{e0}}$. Polarization completely disappears only after applying a depolarization electric field of the strength $-E_c$ in the opposite
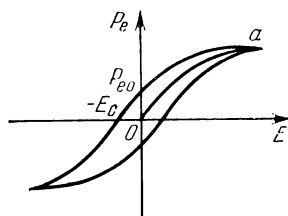
Fig. 18.7

---

* A discussion of the phenomena that occur in ferroelectrics at $T = \theta_C^{low}$ and lead to the destruction of the domains is beyond the scope of the present handbook.

direction. This is called the *coercive force*. Periodic reversal of polarization of a ferromagnetic material is accompanied by the consumption of electric energy, spent in heating the substance. The area of the hysteresis loop is proportional to the electric energy that is converted into internal energy per unit volume of the ferroelectric per cycle.

CHAPTER 19   ENERGY OF AN ELECTRIC FIELD

19.1 Energy of a Charged Conductor and an Electric Field *

19.1.1 The charging of an electric conductor requires the performance of work that must be done in overcoming the Coulomb forces of repulsion between like charges. This work increases the electrical energy of a charged conductor in the same sense as the potential energy in mechanics (3.3.1). The work δW done in bringing a charge dq from infinity to the conductor is

$$\delta W = V\,dq = CV\,dV,$$

where C and V are the capacitance and potential of the conductor.

The work required to charge a conductor from zero potential to the potential V is

$$W = \int_0^V CV\,dV = \frac{CV^2}{2}.$$

Correspondingly, the energy of a *charged isolated conductor* (*intrinsic energy of a charged conductor*) is

$$E_e = \frac{CV^2}{2} = \frac{q^2}{2C} = \frac{1}{2}\,Vq.$$

---

* It is assumed in this chapter that the electric charges and charged bodies are in an isotropic, homogeneous, nonferromagnetic (18.4.1) medium.

The *energy of a charged capacitor* is

$$E_e = \frac{C\,(\Delta V)^2}{2} = \frac{q^2}{2C} = \frac{1}{2}\,\Delta Vq,$$

where C and q are the capacitance and charge of the capacitor, and ΔV is the potential difference between oppositely charged plates of the capacitor.

19.1.2 The energy of a system of charges, distributed continuously throughout the volume of a dielectric or vacuum, and along the surfaces of charged conductors and an electrified dielectric, is

$$E = \frac{1}{2} \int\limits_{S_{\text{char}}} V\sigma\,dS + \frac{1}{2} \int\limits_{v_{\text{char}}} V\rho\,dv$$

where σ and ρ are the surface and volume density of free charges, V is the potential of the resultant field of all the surface and volume charges at the points of the elements dS or dv of the charged surface or volume (to avoid confusion, volume is denoted here by the lower-case v). Integration is carried out over all charged surfaces S_{char} and throughout all charged volumes v_{char}. The effect of the dielectric is that with a constant distribution of the free charges the values of V differ for the different dielectrics. Thus, in a homogeneous isotropic dielectric that fills the whole field, V is ε_r-fold less than in vacuum.

19.1.3 The intrinsic energy of a charged conductor is, at the same time, the *energy* of its *electrostatic field*. Hence, for the uniform electrostatic field (15.1.3) of a parallel-plate capacitor (17.2.5) we have

$$E_e = \frac{1}{2}\,(\Delta V)^2\,C = \frac{\varepsilon_0\varepsilon_r E^2}{2}\,Sd = \frac{\varepsilon_0\varepsilon_r E^2}{2}\,v \text{ (in SI units),}$$

$$E_e = \frac{\varepsilon_r E^2}{8\pi}\,v \text{ (in cgse units),}$$

where $v = Sd$ is the volume of the electrostatic field between the capacitor plates. The energy of the field is proportional to its volume, and the energy per unit volume of the electro-

static field, called the *volume energy density* E_c^d, is the same at all points of a uniform field:

$$E_e^d = \frac{E_e}{v} = \frac{\varepsilon_0 \varepsilon_r E^2}{2} = \frac{ED}{2} \quad \text{(in SI units)},$$

$$E_c^d = \frac{E_e}{v} = \frac{\varepsilon_r E}{8\pi} = \frac{ED}{8\pi} \quad \text{(in cgse units).}$$

19.1.4 In the case of nonuniform electrostatic fields, set up by arbitrary charged bodies, the volume energy density at each point of an isotropic medium is expressed by the formulas in the preceding subsection. But when the medium is electrically anisotropic, the volume energy density of the electric field is

$$E_c^d = \frac{1}{2} \, \mathbf{DE} \quad \text{(in SI units),}$$

$$E_c^d = \frac{1}{8\pi} \, \mathbf{DE} \quad \text{(in cgse units).}$$

19.1.5 The energy dE_e of an infinitesimal volume of an arbitrary electrostatic field in an isotropic medium is

$$dE_e = E_c^d \, dv = \frac{\varepsilon_0 \varepsilon_r E^2}{2} \, dv \quad \text{(in SI units).}$$

The total energy E_e of an electrostatic field is

$$E_e = \int\limits_{v_{\text{field}}} \frac{\varepsilon_0 \varepsilon_r E^2}{2} \, dv$$

where integration is carried out throughout the whole volume v_{field} of the field.

19.1.6 The total energy of an electrostatic field set up by an arbitrary charged body is equal to the intrinsic energy of the body (19.1.1):

$$E_e = \frac{CV^2}{2} = \int\limits_{v_{\text{field}}} \frac{\varepsilon_0 \varepsilon_r E^2}{2} \, dv.$$

This equation can be generalized for the case of an electrostatic field set up by an arbitrary system of charges. The total energy of such a system (19.1.2) coincides with the total energy of the electrostatic field set up by this system of charges:

$$\int_{field} E_e^d \, dv = \frac{1}{2} \int_{S_{char}} V\sigma \, dS + \frac{1}{2} \int_{v_{char}} V\rho \, dv.$$

19.2 Energy of a Polarized Dielectric

19.2.1 The process of polarizing a homogeneous isotropic dielectric, placed in an external electric field, is accompanied by work done in deforming the electron shells in the atoms and molecules or in turning the dipole axes of the molecules into alignment with the field strength. Therefore, a polarized dielectric has a margin of electric energy.

19.2.2 The volume energy density of a polarized dielectric at a point with the field strength **E** equals:

$$E_{e\,(diel)}^d = \frac{\varepsilon_0 \, (\varepsilon_r - 1)}{2} \, E^2 \text{ (in SI units).}$$

The volume energy density of a field with the same strength **E** in vacuum is

$$E_{e\,(vac)}^d = \frac{\varepsilon_0 E^2}{2} .$$

The volume energy density of a field in a dielectric is $E_e^d = \varepsilon_0 \varepsilon_r E^2/2$, and according to the energy conservation law it equals

$$E_e^d = E_{e(vac)}^d + E_{e(diel)}^d.$$

CHAPTER 20   DIRECT ELECTRIC CURRENT

20.1 Concept of an Electric Current

20.1.1 *Electrodynamics* is the main branch of the science of electricity that deals with the phenomena and processes involved in the motion of electric charges or macroscopic charged bodies. The most vital concept in electrodynamics is that of electric current.

20.1.2 An *electric current* is any ordered motion of electric charges. An electric current produced in conductive media as the result of the ordered motion of free charges by the action of an electric field is called *conduction current*. Examples of conduction currents are the currents in metals and semiconductors, associated with the ordered motion of "free" electrons, and currents in electrolytes, which are the ordered displacement of ions of opposite signs.

A *convection current* is the ordered motion in space of charged macroscopic bodies. An example of such a current is that associated with the motion of the earth, having a surplus negative charge, along its orbit.

20.1.3 Upon ordered motion of electric charges in a conductor, the equilibrium distribution of charges is violated and the surface of the conductor is no longer an equipotential surface (16.4.3). On the surface of the conductor there is a tangential component of the field strength ($E_\tau \neq 0$) and there should be an electric field inside the conductor (cf. Sect. 16.4.3). The electric current continues until all the points of the conductor become equipotential.

20.1.4 The conditions required for the initiation and existence of an electric current in a conducting medium are as follows:

(a) The availability of free *current carriers*—charged particles— in the medium, capable of ordered motion in it. In metals and semiconductors suitable particles are the conduction electrons (16.4.1); in liquid conductors (electrolytes) they are the positive and negative ions; in gases they are the oppositely charged ions and electrons.

(b) The existence in the given medium of an external electric field whose energy is spent in providing ordered motion of the electric charges. To maintain an electric current, the energy of the electric field must be continuously replenished, i.e.

a *source of electric energy* is required. This is a device that converts some kind of energy into the energy of an electric field.
20.1.5 The direction of ordered motion of the positive electric charges is conventionally taken as the direction of the electric current. As a matter of fact, in metallic conductors the current consists of the ordered motion of electrons, which move in the direction opposite to that of the current.

20.2 Current and Current Density

20.2.1 The *current strength* (or simply the *current*) I is a scalar quantity, the rate of flow of electricity. It is equal to the ratio of the charge dq, conveyed through the surface* being considered during a small interval of time, to the amount of this time dt:

$$I = \frac{dq}{dt}.$$

An electric current is said to be *direct* (*direct electric current*) if its magnitude and direction remain constant with time. For direct current

$$I = \frac{q}{t},$$

where q is the electric charge conveyed through the surface* being considered during the finite time interval from 0 to t.
20.2.2 If the current is a direct one, the charges cannot accumulate or be depleted at any part of the conductor. A direct current circuit should be closed and comply with the condition $Q_{S_1} = Q_{S_2}$, where Q_{S_1} is the total electric charge passing in unit time through surface S_1 into the volume of the conductor enclosed between the cross sections S_1 and S_2, and Q_{S_2} is the total electric charge emerging from this volume in unit time through surface S_2.
20.2.3 The current density determines the current direction at various points of the surface being considered and the distribution of the current over this surface. The *current density*

---

* In the case of conduction current, through the cross section of the conductor.

vector **j** is opposite in direction to the motion of the electrons that carry the current in metals* and is numerically equal to the ratio of the current dI through an element of surface normal to the motion of the charged particles to the quantity dS', which is the area of this element of surface:

$$j = \frac{dI}{dS'}.$$

A more general relation between the current density **j** and an element of current dI is

$$dI = j\,d\mathbf{S},$$

where $d\mathbf{S} = \mathbf{n}\,dS$ is the vector of an element of surface, and **n** is the unit vector of the normal to surface dS and makes angle α with vector **j**.

20.2.4 The current through an arbitrary surface S is

$$I = \int_S \mathbf{j}\,d\mathbf{S} = \int_S j_\mathrm{n}\,dS,$$

where $j_\mathrm{n} = j\cos\alpha$ is the projection of vector **j** on the direction of normal **n** (20.2.3), and integration is carried out over the whole area of surface S. If the conduction current is being calculated on the basis of the cross section of the conductor for which $j_\mathrm{n} = j$, then

$$I = \int_S j\,dS.$$

20.2.5 The density of direct current is the same over the whole cross section S of a homogeneous conductor. Here

$$I = jS.$$

In a direct current circuit, the current density of two cross sections S_1 and S_2 is inversely proportional to the cross-sectional areas:

$$\frac{j_1}{j_2} = \frac{S_2}{S_1}.$$

* In other conducting media vector **j** coincides in direction with that of the positively charged current carriers,

20.3 Fundamentals of the Classical Electron Theory of Electrical Conduction in Metals

20.3.1 The high electrical conduction of metals is due to the fact that metals contain an immense number of current carriers, the *conduction electrons*, which are formed of the valence electrons (39.3.9) of the atoms of metal. These electrons do not belong to any definite atom, but are *collective electrons.* According to the classical Drude-Lorentz electron theory, these electrons are regarded as an electron gas (16.4.1) that possesses all the properties of a monatomic ideal gas (8.4.1).

The number of conduction electrons in unit volume of a monovalent metal is

$$n_0 = \frac{N_A}{m_a} \, \rho,$$

where N_A is Avogadro's number (Appendix II), m_a is the atomic mass of the metal and ρ is its density. In order of magnitude $n_0 \approx (10^{28}$ to $10^{29})$ m^{-3}.

In the absence of an electric field inside the metal the conduction electrons move at random and collide with the ions of the crystal lattice of the metal (40.1.1). It is assumed that the mean free path $\langle \lambda \rangle$ of the electron (10.5.1) should be of the order of magnitude of crystal lattice constants of the metal, i.e. $\langle \lambda \rangle \approx 10^{-10}$ m.

The average kinetic energy of thermal motion of the electrons (10.2.4) is

$$\frac{m v_{\mathrm{rms}}^2}{2} = \frac{3}{2} \, kT,$$

where m is the mass and v_{rms} is the root-mean-square velocity of the electrons (10.2.3). At the temperature $T = 273$ K, the velocity $v_{\mathrm{rms}} \approx 10^5$ m/s.

The arithmetic mean velocity $\langle u \rangle$ of thermal motion of the electrons (10.3.6) is of the same order of magnitude.

20.3.2 Electric current in a metal is initiated by the action of the external electric field (20.1.4), which leads to ordered motion of the electrons. The current density $\mathbf{j}$ is equal to the charge

of all the electrons that pass in unit time through unit area of the conductor's cross section,

$$\mathbf{j} = n_0 e \langle \mathbf{v} \rangle,$$

where n_0 is the number of conduction electrons per unit volume, e is the magnitude of the charge of the electron, and $\langle \mathbf{v} \rangle$ is the average velocity of ordered motion of the electrons due to the action of the external field. At the highest current densities $\langle \mathbf{v} \rangle$ is 10^{-4} m/s and is negligibly small compared to the thermal velocities of the electrons (20.3.1).

20.3.3 A steady electric current is established in a circuit in the time $t = L/c$, where L is the length of the circuit and c is the velocity of light in vacuum. The time t coincides with the time required to establish a steady-state electric field along the circuit and the initiation of ordered motion of the electrons along the whole circuit. Therefore, electric current is started practically simultaneously with the closing of the circuit.

20.3.4 *Ohm's law for current density (Ohm's law in differential form)* is

$$\mathbf{j} = \varkappa \mathbf{E} = \frac{1}{\rho}\,\mathbf{E}.$$

The current density in a conductor is the product of the *electrical conductivity* $\varkappa$ by the electric field strength $\mathbf{E}$. The quantity $\rho = 1/\varkappa$ is called the *electrical resistivity*.

According to the classical electron theory, electrical conductivity and resistivity are calculated by the formulas

$$\varkappa = \frac{n_0 e^2 \langle \lambda \rangle}{2m \langle u \rangle} \quad \text{and} \quad \rho = \frac{2m \langle u \rangle}{n_0 e^2 \langle \lambda \rangle}\,,$$

where n_0 is the number of electrons in unit volume of the metal, $\langle \lambda \rangle$ is the mean free path of the electron (10.5.1), $\langle u \rangle$ is the arithmetic mean velocity of thermal motion of the electrons (10.3.6), and e and m are the magnitude of the charge of the electron and its mass.

20.3.5 By the action of the field the electron acquires a velocity equal to $v_{\max}$ at the end of its mean free path. Upon collision with an ion at this point, the electron loses its velocity and the energy of ordered motion of the electron is converted into internal energy of the conductor, which is heated as the current passes through.

The *volume density* E_h^d *of the heating capacity of a current* is the amount of energy generated in unit volume of a conductor in unit time. The *Joule-Lenz law* for the volume density of the heating capacity of a current is

$$E_h^d = \mathbf{jE} = \varkappa E^2 = \frac{1}{\rho}\, E^2.$$

The *Joule-Lenz law in its differential form* states: the volume density of the heating capacity of a current is equal to the scalar product of the current density vector by the electric field strength vector.

The volume density of the heating capacity of a current is independent of the kind of collisions of the electrons with the crystal lattice points [elastic or inelastic collision (3.5.3)]. It follows from the laws of conservation of energy and momentum that the energy, ΔE transmitted to the ion upon a collision of the electron with an ion, is only a small part of the electron's energy E_{el}. In an inelastic collision $\Delta E/E_{el} = m/(m + M)$, whereas in an elastic collision $\Delta E/E_{el} = 4mM/(m + M)^2$, where m is the mass of the electron and M is the mass of the ion. Practically, in either case, $\Delta E/E_{el} \approx m/M \approx 10^{-4}$.

20.3.6 The *Wiedemann-Franz law* states that for all metals the ratio of the thermal conductivity K (10.8.5) to the electrical conductivity $\varkappa$ is directly proportional to the absolute temperature T:

$$\frac{K}{\varkappa} = 3\left(\frac{k}{e}\right)^2 T,$$

where k is Boltzmann's constant (8.4.5) and e is the charge of the electron.

20.3.7 The classical electron theory of electrical conduction of metals has the following shortcomings:

(a) it cannot explain the experimentally observed linear relation, in a wide temperature range, between the resistivity ρ and the absolute temperature: $\rho \propto T$;

(b) it yields an incorrect value for the molar heat capacity of metals. This should be equal, according to this theory, to 9 cal/mol-K and be the sum of the heat capacity of the ionic crystal lattice (6 cal/mol-K) and the heat capacity of the monatomic electron gas (3 cal/mol-K). It is known, however, from

the experimentally established Dulong and Petit law (41.7.2)
that the molar heat capacity of metals differs only slightly from
that of other solids and is approximately 6 cal/mol-K. It is
impossible on the basis of classical theory to explain the absence
of an electron component of heat capacity of metals;
(c) the experimental value of the resistivity ρ and the theoretical
value of the arithmetic mean velocity $\langle u \rangle$ of the electrons lead,
in formulas of Sect. 20.3.4, to a value of the mean free path
$\langle \lambda \rangle$ of the electron that exceeds the crystal lattice constant of
the metal by two orders of magnitude. This contradicts the
assumptions of the classical electron theory of the electrical
conduction of metals.

CHAPTER 21   DIRECT CURRENT LAWS

21.1 Extraneous Forces

21.1.1 A metallic conductor has an electrostatic field that is
set up by the electrons and positive ions of the crystal lattice.
This is a field of Coulomb forces (14.2.2). Coulomb interaction
between the charges in a metal leads to such an equilibrium
distribution of charges that the electric field inside the conductor
equals zero and the whole conductor is equipotential (16.4.3).
Consequently, an electrostatic Coulomb field cannot be the
cause of the steady-state process of ordered motion of electrons,
i.e. it cannot generate and maintain a direct (constant) current.
21.1.2 *Extraneous forces* are nonelectrostatic forces whose
action exerted on the conduction electrons in the conductor
initiates their ordered motion and maintains a direct (constant)
electric current in the circuit. In contrast to Coulomb forces,
extraneous forces do not join unlike charges; they lead to their
separation and maintain a potential difference at the ends of
the conductor.
Extraneous forces are the cause of the nonelectrostatic electric
field, existing in the conductor and providing for ordered motion
of the charges from points with a higher potential to points
with a lower potential. A steady-state electric field of extraneous

forces is set up by *sources of electrical energy* (galvanic cells, electric generators, etc.).

21.1.3 In a closed hydraulic system that provides for constant circulation of a liquid (Fig. 21.1), the liquid travels from point A to point B, in the direction opposite to the pull of gravity, by the action of certain "extraneous forces" exerted on the liquid by pump P. The pump produces a constant difference in hydrostatic pressure between points B and A, and the liquid flows by gravity from B to A. A source of electrical energy plays a similar role in a direct current circuit. Owing to the electric field set up by extraneous forces and existing inside the source, the electric charges move within the source against the forces of the electrostatic field. This maintains a potential difference at the ends of the external circuit that is required for a direct current to flow. At the expense of the energy consumed in the source, work is done to obtain ordered motion of the electric charges.
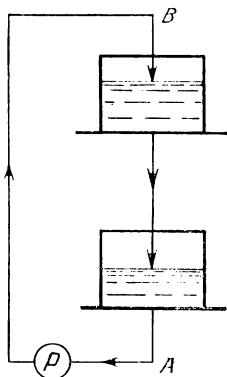
Fig. 21.1

In a direct current generator, for example, the work of the extraneous forces (21.1.2) is done at the expense of the mechanical energy required to drive the rotor of the generator.

21.2 Ohm's Law and the Joule-Lenz Law

21.2.1 At any point inside a portion of a conductor that contains a source of electrical energy there is an electrostatic field of Coulomb forces with the strength $\mathbf{E}_{coul}$ and an electric field of extraneous forces with the strength $\mathbf{E}_{extr}$.

According to the principle of superposition of fields (15.2.2) the strength of the resultant field is

$$\mathbf{E} = \mathbf{E}_{coul} + \mathbf{E}_{extr}.$$

Ohm's law for current density (20.3.4),

$$\mathbf{j} = \frac{1}{\rho} \left(\mathbf{E}_{\mathrm{coul}} + \mathbf{E}_{\mathrm{extr}} \right),$$

enables the following relation to be obtained for portion 1-2 of a homogeneous conductor of cross-sectional area S:

$$I \int_1^2 \rho \frac{dl}{S} = \int_1^2 \mathbf{E}_{\mathrm{coul}} \, d\mathbf{l} + \int_1^2 \mathbf{E}_{\mathrm{extr}} \, d\mathbf{l},$$

where I is the current in the conductor, $d\mathbf{l}$ is a vector with the magnitude dl, equal to an element of conductor length, and is directed tangent to the conductor toward the current density vector, and S is the cross-sectional area of the conductor.

21.2.2 The integral $\int_1^2 \mathbf{E}_{\mathrm{coul}} \, d\mathbf{l}$ is numerically equal to the work done by the Coulomb forces in moving a unit positive charge from point 1 to point 2. According to Sect. 16.3.1

$$\int_1^2 \mathbf{E}_{\mathrm{coul}} \, d\mathbf{l} = V_1 - V_2,$$

where V_1 and V_2 are the potentials at points 1 and 2 of the conductor.

The *electromotive force* (emf) $\mathscr{E}_{12}$, acting over the portion 1-2 of the circuit, is the name given to the linear integral

$$\mathscr{E}_{12} = \int_1^2 \mathbf{E}_{\mathrm{extr}} \, d\mathbf{l}.$$

The electromotive force $\mathscr{E}_{12}$ is numerically equal to the work done by the extraneous forces in moving a unit positive charge along the conductor from point 1 to point 2. This work is done at the expense of the energy consumed in the source.

The *voltage* (or *voltage drop*) U_{12} across the portion 1-2 is the physical quantity numerically equal to the work done by the

resultant field of Coulomb and extraneous forces in moving a unit positive charge along the circuit from point 1 to point 2. Thus

$$U_{12} = \int_{1}^{2} (\mathbf{E}_{coul} + \mathbf{E}_{extr}) \, d\mathbf{l} = \int_{1}^{2} \mathbf{E} \, d\mathbf{l},$$

or

$$U_{12} = (V_1 - V_2) + \mathscr{E}_{12}.$$

The voltage across the ends of the circuit coincides with the potential difference only when no emf is applied to this portion.
The integral

$$R_{12} = \int_{1}^{2} \rho \frac{d\mathbf{l}}{S}$$

is the *resistance* R_{12} in the portion of the circuit between sections 1 and 2. For a homogeneous cylindrical conductor (wire), $\rho = \text{const}$, $S = \text{const}$

$$R_{12} = \rho \frac{l_{12}}{S},$$

where l_{12} is the length of the conductor between cross sections 1 and 2.

21.2.3 The *generalized Ohm's law* for an arbitrary portion of a circuit is

$$IR_{12} = U_{12} = (V_1 - V_2) + \mathscr{E}_{12}.$$

The product of the current by the resistance of a portion of the circuit is equal to the potential difference across this portion plus the emf of all sources of electrical energy connected in the given portion of the circuit. In this form Ohm's law is applicable both for *passive portions of the circuit*, not containing sources of electrical energy, as well as for *active portions* that contain such sources.

21.2.4 The *sign rule* for the emf of sources of electrical energy included in portion 1-2 states that if the current inside the

source is from the cathode to the anode, i.e. the strength of the
field of extraneous forces coincides in direction with the current
in this portion of the circuit, then, in calculations, the emf $\mathscr{E}_{12}$
of this source is considered positive. If the current inside the
source is from the anode to the cathode, the emf $\mathscr{E}_{12}$ of this
source is considered negative (Fig. 21.2).

21.2.5 In a closed series electric circuit, the current is the same
at all cross sections, and such a circuit is a portion with coin-
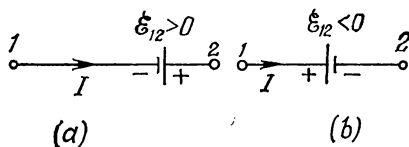
Fig. 21.2

ciding ends (points 1 and 2 coincide). In this circuit $V_1 = V_2$
and $R_{12} = R$, the total resistance of the whole circuit.

Ohm's law for a closed electric circuit is

$$IR = \mathscr{E},$$

where $\mathscr{E}$ is the algebraic sum of all the emf's applied in the
circuit.

If a closed circuit consists of a source of electrical energy with
an emf $\mathscr{E}$ and internal resistance r, and the resistance of the
external part of the circuit equals R, Ohm's law is of the form

$$I = \frac{\mathscr{E}}{R+r} \, .$$

The potential difference across the terminals of the source is
equal to the voltage across the external part of the circuit:

$$V_1 - V_2 = RI = \mathscr{E} - Ir.$$

If the circuit is open and has no current ($I = 0$), the potential
difference across the terminals of the source is equal to its emf:

$$V_1 - V_2 = \mathscr{E}.$$

A voltmeter connected in parallel with portion 1-2 of a direct current (d.c.) electric circuit measures the potential difference across the ends of this portion, and not the voltage:

$$R_V I_V = V_1 - V_2,$$

where R_V and I_V are the resistance of the voltmeter and its current (Fig. 21.3). This follows from the generalized Ohm's law (21.2.3) as written for portion 1-2 of the voltmeter circuit, in which there is no emf (see also Sect. 21.3.5).

21.2.6 In a d.c. electric circuit with fixed conductors, the work done by the extraneous forces is completely consumed in heating the conductors (20.3.5). The energy E, evolved in a circuit during the time t throughout the whole volume of the conductor, is equal to

$$E = IUt,$$

where I is the current and U is the voltage. The quantity of heat Q (9.2.1) in calories, corresponding to this energy and evolved in the conductor, is

$$Q = 0.24\ IUt.$$

Fig. 21.3

All the other quantities are expressed in SI units (Appendix I). The *Joule-Lenz law* states that the quantity of heat generated by current in a conductor is proportional to the current, voltage and time that the current flows. Other expressions of the Joule-Lenz law are:

$$Q = 0.24 I^2 R t = 0.24 \frac{U^2}{R}\ t.$$
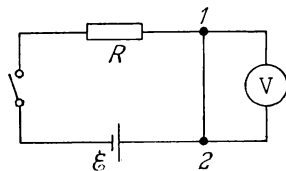
21.3 Kirchhoff's Laws

21.3.1 The analysis of *complex* (*branched*) *circuits* consists in finding the currents in the various portions of the circuit in accordance with their given resistances and applied emf's.

18*

21.3.2 A *branch point*, or *junction*, in a branched circuit, or network, is a point where there are more than two possible directions of the current (Fig. 21.4).

Kirchhoff's first law (*point rule*) states that the algebraic sum of the currents that meet at any junction of a network equals zero:

$$\sum_{i=1}^{i=n} I_i = 0,$$

where n is the number of conductors joined at the branch point and I_i is the current at the point. Currents flowing toward the branch point are considered positive; those flowing away are considered negative.
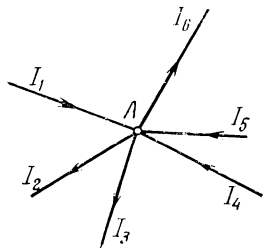
Kirchhoff's second law (*mesh rule*) states that in any closed loop (mesh of a network), arbitrarily chosen in a branched circuit (network), the algebraic sum of the products of the currents I_i by the resistances R_i of the corresponding portions of the loop is equal to the algebraic sum of the emf's in the loop. Thus

Fig. 21.4

$$\sum_{i=1}^{i=n_1} I_i R_i = \sum_{i=1}^{i=n_1} \mathscr{E}_i,$$

where n_1 is the number of portions into which the loop is divided by branch points (junctions). In applying the mesh rule, a definite direction around the loop is chosen (either clockwise or counterclockwise). Currents coinciding in direction with the chosen one are considered positive. The emf's of the sources of electrical energy are considered positive if they produce currents whose direction coincides with the chosen one.

21.3.3 The following procedure is employed to analyse a complex (branched) d.c. circuit:

(a) The directions of the currents are arbitrarily assigned in all the portions of the circuit.

(b) For m branch points (junctions), $m - 1$ independent equations are written on the basis of Kirchhoff's first law.

(c) Arbitrary closed loops are separated out and, after choosing the direction for going around each loop, a system of equations is written on the basis of Kirchhoff's second law. In a branched circuit consisting of p portions (branches) between adjacent branch points and m branch points (junctions), the number of independent equations based on the mesh rule equals $p - m + 1$. In selecting the loop, it is necessary for each new loop to contain at least one portion (branch) not included in the previously considered loops.

21.3.4 To *shunt an ammeter* means to connect an additional resistor R_{sh} in parallel to the ammeter. This is done so that
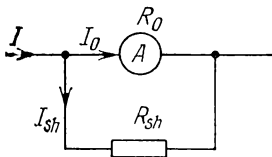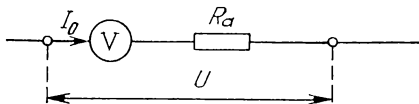
Fig. 21.5   Fig. 21.6

the ammeter, having the resistance R_0 and designed to measure a maximum current of I_0, can be applied to measure currents I that exceed current I_0 (Fig. 21.5). The required resistance of the shunt is found according to Kirchhoff's laws: $I = I_0 + I_{sh}$ and $I_0 R_0 = I_{sh} R_{sh}$, from which I_{sh} is eliminated:

$$R_{sh} = \frac{I_0 R_0}{I - I_0}.$$

21.3.5 A potential difference $V_1 - V_2 = U$ across a portion of a circuit is to be measured by a *voltmeter* designed for a maximum measurement of U_0 volts (21.2.5) at a maximum current of I_0 (with $U_0 = I_0 R_0$) and this potential difference exceeds U_0 (i.e. $U > U_0$). This can be done by connecting an additional *resistor* R_a in series with the voltmeter (Fig. 21.6). The required additional resistance is determined from the equation $U = (R_0 + R_a) I_0$, from which

$$R_a = \frac{U}{I_0} - R_0.$$

CHAPTER 22   ELECTRIC CURRENT IN LIQUIDS AND GASES

22.1 Faraday's Laws of Electrolysis. Electrolytic Dissociation

22.1.1 Liquids are conductors of electric current (*electrolytes*, *ion conductors*) provided that ordered motion of ions can be accomplished in them by the action of an external electric fields.

The ordered motion of ions in conducting liquids takes place in an electric field set up by *electrodes*, which are conductors connected to the terminals of a source of electrical energy. The positive electrode is called the *anode*, the negative electrode, the *cathode*. Positive ions, called *cations*, are the ions of metals and hydrogen, and move toward the cathode. Negative ions, called *anions*, are the ions of acid radicals and of the hydroxyl group, and move toward the anode. Electric current in electrolytes is accompanied by the phenomenon of *electrolysis*. This is the deposition on the electrodes, or the liberation at them, of the components of the substances contained in the electrolyte or of other substances produced by secondary reactions.

22.1.2 *Faraday's first law of electrolysis* states that the mass M of a substance deposited on or liberated at an electrode is directly proportional to the electric charge Q passing through the electrolyte. Thus

$$M = kQ = kIt$$

if direct current I is passed during the time t through the electrolyte.

The proportionality factor k is called the *electrochemical equivalent* of the substance. It is numerically equal to the mass of the substance liberated when a unit electric charge passes through the electrolyte and depends upon the chemical nature of the substance.

22.1.3 *Faraday's second law of electrolysis* states that the electrochemical equivalents of the elements are directly proportional to their chemical equivalents:

$$k = Ck_{ch},$$

where C is a certain constant universal for all the elements, and k_{ch} is the chemical equivalent and is equal to

$$k_{ch} = 10^3 \frac{m_a}{z} .$$

Here m_a is the atomic mass of the element (in kg/mol), and z is its valency.

Hence, $k = \dfrac{1}{F} \dfrac{m_a}{z}$, where $F = 10^{-3}/C$ is *Faraday's constant* (Appendix II).

22.1.4 *Faraday's united law of electrolysis*,

$$M = \frac{1}{F} \frac{m_a}{z} Q,$$

enables the physical meaning of F to be cleared up: at $M = m_a/z$, Faraday's constant $F = Q$. The amount of substance equal to $1/z$ moles is called its *gram equivalent*. At $z = 1$ the gram equivalent of a substance is equal to one mole. Faraday's constant is numerically equal to the electric charge that must be passed through an electrolyte to deposit one gram equivalent of any substance on the electrode.

22.1.5 The splitting of neutral molecules into oppositely charged ions as a result of interaction between the dissolved substance (solute) and the solvent is called *electrolytic dissociation*. It is caused by the thermal motion of the polar molecules (18.1.4) of the solute, consisting of interrelated, oppositely charged ions (39.4.3), and by the interaction of these molecules with the polar molecules of the solvent. Both causes lead to weakening of the heteropolar bonds in ionic molecules (39.4.3) and to the conversion of such molecules into two oppositely charged ions. The *degree of dissociation* α is the ratio of the number n' of molecules dissociated into ions in a certain volume to the total number n_0 of molecules of the solute in the same volume: $\alpha = n'/n_0$.

22.1.6 The process opposite to electrolytic dissociation (22.1.5) is called *deionization*. It consists in the recombination of ions of opposite signs into neutral molecules. Under conditions of

dynamic mobile equilibrium between the dissociation and
deionization processes, α is determined by the equation

$$\frac{1-\alpha}{\alpha^2} = \text{const} \times n_0.$$

As $n_0 \to 0$, we have $\alpha \to 1$, i.e. in weak solutions almost all
the molecules dissociate. The degree of dissociation α decreases
with an increase in the concentration of the solution. In highly
concentrated solutions

$$\alpha \approx \frac{\text{const}}{\sqrt{n_0}}.$$

22.2 Atomicity of Electric Charges

22.2.1 It follows from Faraday's laws of electrolysis that all
electric charges consist of a whole number of elementary indi-
visible charges.
22.2.2 The charge Q of any ion is equal to

$$Q = \pm \frac{zF}{N_A},$$

where z is the valency of the ion, F is Faraday's constant
(22.1.4), and N_A is Avogadro's number (Appendix II). The
charge of a monovalent ion is equal to the charge e of the elec-
tron or proton:

$$Q_1 = e = 1.602 \times 10^{-19}\text{ C} = 4.803 \times 10^{-10}\text{ electrostatic units.}$$

Any electric charge consists of a whole number of elementary
charges e (14.1.2).

22.3 Electrolytic Conduction of Liquids

22.3.1 The current density $\mathbf{j}$ (20.2.3) in an arbitrary section SS,
perpendicular to the direction of ion motion (Fig. 22.1), is
equal to the sum of the current densities of the positive and
negative ions:

$$\mathbf{j} = \mathbf{j}_+ + \mathbf{j}_-.$$

Here $j_+ = q_+ n_{0+} \langle v_+ \rangle$ and $j_- = q_- n_{0-} \langle v_- \rangle$, where q_+ and q_-, n_{0+} and n_{0-}, and $\langle v_+ \rangle$ and $\langle v_- \rangle$ are the charges, concentrations and average velocities of ordered motion (*drift* due to the action of the electric field) of the positive and negative ions.

22.3.2 The average velocities of ion drift are proportional to the electric field strength E:

$$\langle v_+ \rangle = u_+ E \quad \text{and}$$

$$\langle v_- \rangle = -u_- E,$$

where u_+ and u_- are the *mobilities* of the positive and negative *ions*. The mobility of an ion is equal to the ratio of the magnitude of the average drift velocity vector to that of the field strength vector. The mobility is independent of the

Fig. 22.1

field strength vector E. Since there are no bulk charges in electrolytes, $q_+ n_{0+} + q_- n_{0-} = 0$.

Moreover, $q_+ = e z_+ = \dfrac{F}{N_\Lambda} z_+$ (22.2.2).

22.3.3 *Ohm's law of current density in electrolytes* (cf. Sect. 20.3.4) is:

$$j = \frac{F}{N_\Lambda} z_+ n_{0+} (u_+ + u_-) E.$$

The resistivity ρ of an electrolyte (20.3.4) is equal to

$$\rho = \frac{N_A}{F z_+ n_{0+} (u_+ + u_-)} .$$

If a molecule of the solute dissociates into k_+ positive and k_- negative ions, then

$$k_+ z_+ = k_- z_-; \quad n_{0+} = k_+ \alpha n_0 \quad \text{and} \quad n_{0-} = k_- \alpha n_0,$$
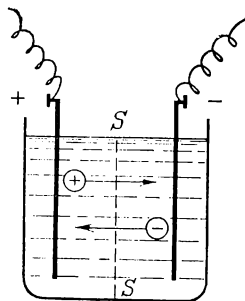
where α is the degree of dissociation, n_0 is the concentration of the solute (22.1.5) and

$$\rho = \frac{N_A}{F z_+ k_+ \alpha n_0 (u_+ + u_-)} \; .$$

The ratio N_A/z_+ is the number of positive ions in one gram equivalent (22.1.4). We introduce the quantity

$$C = \frac{k_+ n_0 z_+}{N_A} = \frac{k_- n_0 z_-}{N_A} \; ,$$

which is called the *equivalent concentration of the solution.* It is the number of gram equivalents of the ions of one sign contained in unit volume of the electrolyte (either in the free state or bound in the molecules). Then

$$\rho = \frac{1}{F C \alpha (u_+ + u_-)} \; .$$

22.4 Electrical Conduction in Gases

22.4.1 Gases, consisting of neutral atoms and molecules, are insulators and do not conduct an electric current. Gases are capable of conduction after they are ionized.

The *ionization of a molecule* (or *atom*) consists in the detachment of one or several electrons, after which the molecule (or atom) is converted into a positive ion. If a molecule (or atom) gains electrons, it becomes a negative ion.

The reverse process, in which electrons join a positive ion to form a neutral molecule (or atom), is called *recombination.*

22.4.2 To ionize a molecule (or atom) it is necessary to perform the *work of ionization* W_1 to overcome the forces of attraction between the electron being detached and the other particles of the molecule (or atom). The work W_1 depends upon the energy state of the electron being detached (39.1.9) in the atom or molecule of the given gas. The ionization energy increases with the electron detachment factor, i.e. with the number of electrons detached from each atom.

22.4.3 The *ionization potential* V_1 is the potential difference that the charged particle must traverse in an accelerating electric field in order to accumulate energy equal to the work of

ionization W_1. Thus $V_1 = W_1/e$, where e is the magnitude of the charge of the particle.

22.4.4 A gas is ionized by external effects: raising the temperature sufficiently, the action of various kinds of radiation, cosmic radiation, bombardment of the molecules (or atoms) of the gas by high-speed electrons or ions, etc. The *intensity*, or *rate*, *of ionization* is measured by the number of pairs of particles of opposite sign formed in unit time in unit volume of the gas.

22.4.5 *Collision ionization* is the process in which a gas is ionized by the action of fast-moving electrons or by ions. The minimum kinetic energy (3.2.1) that an ionizing particle must have is assessed by applying the laws of conservation of momentum and energy. Thus

$$\frac{mv^2}{2} = W_1 \left(1 + \frac{m}{M} \right) ,$$

where W_1 is the work of ionization, m is the mass of the electron and M is the mass of the atom.

The smaller the ratio m/M, the closer this energy is to W_1. An electron and a single-charge ion accumulate the same energy $E = e \, \Delta V$ if they traverse the same potential difference ΔV.

It follows from the preceding formula that for collision ionization by electrons and ions, the latter of a mass 10^4 times that of the electron, ions must travel across a greater potential difference in an accelerating field than electrons.

22.5 Various Types of Gas Discharges

22.5.1 A *gas discharge* is the process of current conduction through a gas.

A *semi-self-maintained gas discharge* is one caused by external ionizing agents (22.4.4) and ceasing when the ionizing action stops. The curve in Fig. 22.2 shows the dependence of the current I in a semi-self-maintained discharge on the voltage U across the electrodes. In the 1st region of the curve, at low voltages, Ohm's law, similar to the one for electrolytes (22.3.3), is valid. If electrons and monovalent ions are formed in the gas, then

$$\mathbf{j} = en_0 \, (u_+ + u_-) \, \mathbf{E},$$

where n_0 is the number of pairs of oppositely charged particles in unit volume, u_+ and u_- are the mobilities of the positive and negative ions, and e is the magnitude of the charge of the electron. In a wide pressure range, from 10 to 10^7 Pa, the mobility of gaseous ions is inversely proportional to the pressure.

Fig. 22.2

22.5.2 The relation between the current I and the voltage U is no longer linear in the 2nd region of the curve in Fig. 22.2 because the ion concentration decreases in the gas. In this region the current increases slower and slower with an increase in U.
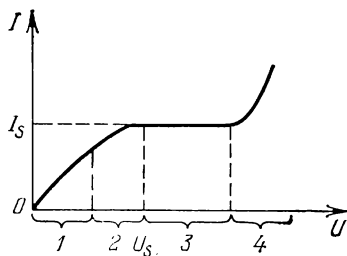
Beginning with a certain voltage U_S, the current remains constant in the 3rd region of the curve in Fig. 22.2 when the voltage is increased. This occurs because, at a constant rate of ionization (22.4.4) in strong electric fields, all the ions formed in unit time in the gas reach the electrodes. No further increase in the current takes place at a fixed rate of ionization. Saturation results because all the charged particles produced in the gas travel in a strong electric field and reach the electrodes before an appreciable part of them have enough time to recombine with particles of opposite charge.

The saturation current I_S is the maximum current that can be obtained at a given rate (intensity) of ionization:

$$I_S = eN_0,$$

where N_0 is the maximum number of pairs of monovalent ions formed in the gas in unit time at a given rate of ionization. The linear relation between I_S and N_0 confirms the ionic nature of electrical conduction in gases.

22.5.3 A *self-maintained gas discharge* is one that continues after the action of the external ionizing agent ceases. To obtain such a discharge it is necessary for new pairs of oppositely charged particles to be formed continuously in the gas. The main

source of these particles is collision ionization (22.4.5). At a certain sufficiently high voltage across the electrodes, the electrons are so highly accelerated by the electric field that their energy proves adequate to ionize the gas molecules (22.4.5). This is called *volume ionization*. Secondary electrons, accelerated in the electric field, also ionize the gas molecules. As a result, the number of current carriers in the gas and its electrical conductance are greatly increased (4th region in Fig. 22.2). But, by itself, ionization by the action of electrons is insufficient to produce a self-maintained discharge. Electrons travelling in the direction from the cathode toward the anode ionize the molecules of gas located closer to the anode than the place where the electrons are emitted. If the energy of the positive ions is insufficient for collision ionization of the gas molecules or for knocking electrons out of the metallic cathode (*surface ionization*), electrons can be produced near the cathode of a gas-discharge tube only by external ionizing agents.
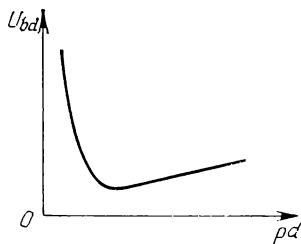
Fig. 22.3

When the action of these agents is stopped, the region of collision ionization by electrons contracts toward the anode as the electrons move in this direction. Under these conditions the gas discharge is extinguished as soon as collision ionization ceases.

22.5.4 Surface ionization (22.5.3), initiated at high voltage U, produces secondary electrons and sets up a two-way avalanche of electrons and positive ions. After this the action of the external ionizing agent is no longer essential for continuing the gas discharge. When the voltage U across the electrodes of a gas-discharge tube is raised sufficiently, an *electric breakdown of the gas* occurs. This is the transition from a semi-self-maintained to a self-maintained gas discharge. The voltage $U = U_{bd}$, corresponding to electric breakdown, is called the *ignition potential*, or *breakdown voltage*. The breakdown voltage for a gas discharge in a tube having flat parallel electrodes at the distance d from each other depends upon the product pd, where p is the gas pressure (Fig. 22.3). In addition, U_{bd} depends

on the chemical nature of the gas and the material of the cathode.

Self-maintained gas discharges are classified into glow, corona, spark and arc discharges. They are not discussed in the present handbook.

22.6 Certain Information on Plasma

22.6.1 *Plasma* is a special state of matter, whose most important property is a predominant, or even complete, ionization of the particles of matter. The *degree of ionization* α is the ratio of the number of ionized particles to their initial number. Plasma is classified as *weakly ionized* (when α is a fraction of one per cent), *moderately ionized* (when α equals several per cent) and *fully ionized* (when α is close to 100 per cent).

Weakly ionized plasma is found in the *ionosphere*, the conducting layer of the atmosphere. The ionosphere extends from 60 to 2×10^4 km above the earth's surface.

Fully ionized plasma, formed at ultra-high temperatures (high-temperature plasma), exists on the sun and hot stars. Under laboratory conditions, plasma is produced in gas discharges (22.5.1) and gaseous-discharge lamps. Accelerated plasma is employed as the working medium (11.1.1) in jet engines. Plasma can also be used for the direct conversion of internal energy into electrical energy (magnitohydrodynamic generators and plasma sources of electrical energy).

The great number of charged particles in a plasma provides for its high electrical conductance; in this aspect its properties approach those of conductors of electric current.

22.6.2 A condition for the existence of a plasma is a certain minimum density ρ_{min} of charged particles. Beginning with this density, we can speak of a plasma rather than a simple accumulation of separate charged particles. The density ρ_{min} is determined by the inequality $L \gg D$, where L is a linear dimension of the system of charged particles and D is a typical "plasma" parameter, the length called the *Debye-Hückel screening radius* (it is also called the *Debye shielding length* or *Debye length*):

$$D = \left(\sum_i \frac{4\pi q_i^2 n_i}{kT_i} \right)^{-1/2} \text{ (in cgse units),}$$

$$D = \left(\sum_i \frac{q_i^2 n_i}{\varepsilon_0 k T_i} \right)^{-1/2} \text{ (in SI units)},$$

where q_i is the charge of the ith species of particles, n_i is their concentration, T_i is their temperature, k is Boltzmann's constant (8.4.5) and ε_0 is the electric constant. Summation is to be carried out over all the species of particles. The quantity D is the distance over which the Coulomb field of any charge of the plasma is shielded. The shielding occurs because any charge is predominantly surrounded by oppositely charged particles. According to a more exact definition, a plasma is a quasi-neutral assembly of a large number of charged particles that occupy a region of space with the linear dimensions $L \gg D$. The violation of quasi-neutrality in the plasma is eliminated by the strong electric fields set up in it. The Debye-Hückel screening radius characterizes the interaction of the particles in the plasma. It was found that $E_p/E_k \propto N^{2/3}$, where $N = 4\pi n D^3/3$. Here E_p is the potential energy of interaction between two particles located at an average distance of $n^{-1/3}$ from each other (n being the concentration of particles), and E_k is the kinetic energy of these particles. The quantity N is the total number of charged particles inside a sphere of radius D, and is called the *Debye number*. If N is large, the plasma is said to be *gaseous* and can be dealt with thermodynamically as an ideal gas with the state equation $p = nkT$ (8.4.5).

22.6.3 Coulomb long-range interaction of charged particles in a plasma leads to its qualitative uniqueness that enables it to be regarded as a special, *fourth state of aggregation of matter. The most important properties of plasma* are:

(a) its strong interaction, due to its high electrical conduction, with external magnetic and electric fields;

(b) specific collective interaction of particles in the plasma (by means of a special field whose nature and origin is beyond the scope of this handbook);

(c) the fact that due to long-range interaction it is a special kind of elastic medium in which various types of vibrations and waves can be readily excited and propagated.

22.6.4 The motion of plasma in a magnetic field is employed to directly convert the internal energy of an ionized gas into electrical energy. This method is made use of in a *magneto-hydrodynamic (MHD) generator*. This generator operates as

follows: gas produced in the combustion of fuel and constituting a plasma passes through a strong traverse magnetic field. In a conducting plasma, as in any conductor moving in such a field, electromagnetic induction (24.1.1) occurs. The induced emf can be taken off by means of electrodes in the external circuit. A schematic diagram of an MHD generator is shown in
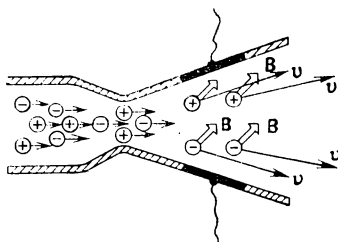
Fig. 22.4

Fig. 22.4. The ionized gas produced in fuel combustion passes through the nozzle and its internal energy is converted into kinetic energy. As this gas passes through the transverse magnetic field **B**, the positive ions move toward the upper electrode by the action of the induced electric field (24.1.1) and the free electrons move toward the lower electrode. When we connect the electrodes to an external load, a current is obtained in the circuit.

22.6.5 A state of thermodynamic equilibrium is possible at a definite temperature in the plasma when the loss of charged particles due to recombination (22.4.1) is recompensed by new acts of ionization. The average kinetic energies of the various particles making up such a plasma are the same. The processes of energy exchange between the particles of this plasma, as well as the energy exchange of the plasma with blackbody radiation (36.1.8) are equilibrium processes (8.3.7). A *plasma* with the aforesaid properties is said to be *isothermal*. It exists in the atmospheres of high-temperature stars.

A condition required for a high degree of ionization of a thermodynamically equilibrium plasma, consisting of two kinds of charged particles, equal in magnitude and opposite in sign, is the maximum reduction in the recombination (22.4.1) of the particles. A completely ionized plasma can be obtained at $kT \gg eV_1$, where V_1 is the ionization potential (22.4.3) of the gas atoms and kT is the average energy of thermal motion of the plasma particles. For hydrogen and deuterium this corresponds to $T \approx 160\ 000$ K. Essential under these conditions are the radiation of the plasma and the difficulty of isolating the plasma from the walls of its container (22.6.7).

22.6.6 In the plasma of a gas discharge (22.5.1) (*gas-discharge plasma*) thermodynamical equilibrium (22.6.5) is absent. The charged particles of such a plasma are in an accelerating electric field.

The average kinetic energy of the electrons in a gas-discharge plasma is characterized by a certain *electron temperature* T_e corresponding to the Maxwellian energy distribution of the electrons (10.3.7). Owing to the lack of thermodynamical equilibrium in such a plasma, T_e is of a conventional nature. The average kinetic energy of the neutral particles is substantially less than that of the electrons. Besides the electron temperature T_e, the parameters of a gas-discharge plasma are: the electron concentration n_e, the number of ionizations per electron per second, the density of the ion or electron current, and the longitudinal electric field strength E_z established along the symmetry axis of the plasma.

22.6.7 The existence of a thermodynamically nonequilibrium gas-discharge plasma is made possible by the energy of the discharge current passing through it. In the absence of an external electric field, the gas-discharge plasma disappears.

The disappearance of an unmaintained gas-discharge plasma that has been left to itself is called *deionization of the gas*.

In addition to the ionization and recombination processes (22.4.1), a large share of the energy balance of a plasma that exists in a limited volume consists in the interaction between the plasma and the walls confining its volume, as well as the radiation of the plasma and the transport of radiation in it. The diffusion of the charged particles to the walls and their recombination on the walls, and the transfer of energy to the walls by heat conduction (10.8.2) contaminate the plasma with impurities and reduce its energy. To avoid these phenomena, measures are taken to confine the plasma by a magnetic field to prevent contact with the walls.

Radiation of the plasma in the optical range and in the far ultraviolet consists of: braking radiation (bremsstrahlung) of the electrons that occurs when they are retarded by the ions, ordinary radiation produced by excited particles, and recombination radiation that occurs in the recombination process (22.4.1). In a magnetic field, a plasma also has a special beta-tron (synchrotron) radiation, which is beyond the scope of this handbook.

CHAPTER 23   MAGNETIC FIELD
OF DIRECT CURRENT

23.1 Magnetic Field. Ampere's Law

23.1.1 A *magnetic field* is one of the forms of an electromagnetic field (15.1.2). A magnetic field is set up by moving charged particles, as well as moving bodies carrying electric charges. Magnetic fields act only on moving electric charges and moving charged bodies.

Variable electric fields (displacement currents) (27.3.1) are also sources of magnetic fields.

23.1.2 The main force characteristic of a magnetic field is *vector of magnetic induction* B (also called *magnetic flux density, magnetic displacement*, or, simply, *magnetic vector*). Vector B is determined by one of three methods:

(a) from Ampere's law (23.1.4);

(b) from the effect of the magnetic field on a current-carrying loop (23.4.2);

(c) from the expression for the Lorentz force (24.1.3).

23.1.3 Magnetic fields are represented graphically by lines of magnetic induction. *Lines of magnetic induction (lines of magnetic force)* are curves whose tangent at each point coincides with vector B at this point of the field. The directions of the magnetic vector and the lines of magnetic induction are determined by the *corkscrew (Maxwell) rule* (also called the *right-hand screw rule*): if a corkscrew with right-hand thread is screwed in the direction of the current density vector in the conductor (20.2.3), the direction that the corkscrew handle turns indicates the direction of the lines of magnetic induction and the magnetic vector.

Lines of magnetic induction are not interrupted at any point of the magnetic field, i.e. they neither begin nor end. These lines are either closed curves or they extend from infinity to infinity or they are wound infinitely on some surface, covering it densely all over, but never returning again to any point of the surface. A similar situation is observed, for example, in a field set up by a system of a circular current and an infinite rectilinear current passing through the centre of the circular current and perpendicular to its plane.

A magnetic field is said to be *uniform*, or *homogeneous*, if vector **B** is constant at any point in the field. Otherwise, it is said to be *nonuniform* (*nonhomogeneous*).

23.1.4 The force exerted by a magnetic field on a current-carrying conductor placed in the field is called *Ampere's force*.

Ampere's law states that an element of force $d\mathbf{F}$ exerted on a small element of length dl of a conductor, carrying a current and located in a magnetic field, is proportional to the vector product of the element $d\mathbf{l}$ of conductor length by the magnetic induction **B**:

$$d\mathbf{F} = I\,[d\mathbf{l}\,\mathbf{B}] \text{ (in SI units),}$$

$$d\mathbf{F} = \frac{I}{c}\,[d\mathbf{l}\,\mathbf{B}] \text{ (in Gaussian units, see Appendix I).}$$

Here $d\mathbf{l}$ is a vector with the magnitude dl and has the same direction as the density vector **j** of the current in the conductor (20.2.3).

Ampere's force **F**, acting in a magnetic field on a current-carrying conductor of finite length, is

$$\mathbf{F} = \int I\,[d\mathbf{l}\,\mathbf{B}],$$

in which integration is carried out over the whole length of the conductor. In the case of a uniform magnetic field (23.1.3)

$$F = IBl \sin \alpha \text{ (in SI units),}$$

$$F = \frac{I}{c}\,Bl \sin \alpha \text{ (in Gaussian units),}$$

where α is the angle between the density vector of the current in the conductor and vector **B**. The relative positions of vectors $d\mathbf{F}$, **B** and $d\mathbf{l}$ are shown in Fig. 23.1. If $d\mathbf{l}$ is perpendicular to **B**, the direction of force $d\mathbf{F}$ is found by the *left-hand rule*: if the left hand, opened flat with the fingers parallel to one another and the thumb perpendicular to the fingers, is held so that the magnetic induction vector enters the palm and the fingers indicate the direction of the electric current, the thumb indicates

the direction of the force exerted by the magnetic field on the current-carrying conductor. It is evident from Fig. 23.1 that vector $d\mathbf{F}$ is perpendicular to the plane passing through vectors $d\mathbf{l}$ and $\mathbf{B}$, and that its direction is such that from the head of vector $d\mathbf{F}$ the direction of shortest rotation from vector $d\mathbf{l}$ to vector $\mathbf{B}$ is counterclockwise. In other words, vector $d\mathbf{F}$ coincides in direction with the vector product $[d\mathbf{l}\ \mathbf{B}]$. **23.1.5** It follows from Ampere's law that the magnetic induction vector in SI units is numerically equal to the limit of the following ratio: the force exerted by a magnetic field on an element of a current-carrying conductor to the product of the current by the length of an element of the conductor, the limit being taken as the length of the element approaches zero and is positioned in the field so that the limit reaches its maximum value. Thus
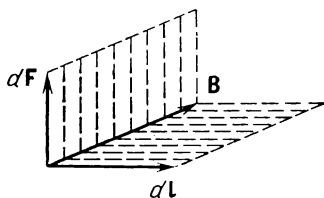
Fig. 23.1

$$B = \frac{1}{I}\left(\frac{dF}{dl}\right)_{\max}.$$

In Gaussian units (Appendix I),

$$B = \frac{c}{I}\left(\frac{dF}{dl}\right)_{\max},$$

where c is the electrodynamic constant (Appendix I).
23.1.6 In contrast to electrostatic forces (14.2.2), Ampere's force is not a central one (3.3.4). Ampere's force is perpendicular to the lines of magnetic induction and to the current-carrying conductors.

23.2 The Biot-Savart-Laplace Law

23.2.1 The *Biot-Savart-Laplace law* determines the magnetic induction at any point in a magnetic field set up by a direct electric current flowing along a conductor of any shape. The

vector $d\mathbf{B}$ of magnetic induction at any point C of a magnetic field, set up by an element of a conductor carrying the current I and of the length dl is determined by the formula

$$dB = k\mu_r \frac{I}{r^3} [dl\ \mathbf{r}],$$

where $d\mathbf{l}$ is the vector of an element of length of the conductor (23.1.4), $\mathbf{r}$ is the radius vector from the element $d\mathbf{l}$ of the conductor to point C (Fig. 23.2), r is the magnitude of the radius vector $\mathbf{r}$, k is a factor depending only upon the chosen system of units, and μ_r is a dimensionless quantity that characterizes the magnetic properties of the medium and called the *relative magnetic permeability* (or simply *relative permeability*) of the medium. The relative permeability does not depend upon the choice of units and equals unity for vacuum. For all substances, except ferromagnetic ones, μ_r differs only slightly from unity (26.5.1). See Sect. 26.4.5 for a general definition of μ_r.
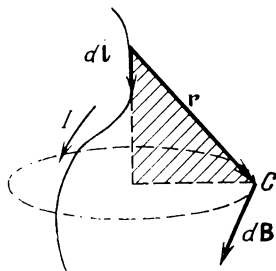
Fig. 23.2

23.2.2 In SI units $k = \mu_0/4\pi$, where $\mu_0 = 4\pi \times 10^{-7}$ henry/m is the *permeability* of free space, or the *magnetic constant* (Appendix II). In Gaussian units, $k = 1/c$, where $c = 3 \times 10^{10}$ cm/s is the electrodynamic constant (Appendix I).
The Biot-Savart-Laplace law is of the form

$$dB = \frac{\mu_0\mu_r}{4\pi} \frac{I}{r^3} [dl\ \mathbf{r}].$$

This form of the Biot-Savart-Laplace law and of all the electromagnetic field equations is said to be *rationalized*.
The product $\mu_0\mu_r$ is sometimes called the *absolute magnetic permeability* (or simply *permeability*) of the medium.

In Gaussian units $d\mathbf{B} = \dfrac{1}{c}\dfrac{I}{r^3}[d\mathbf{l}\ \mathbf{r}]$. The magnitude of vector $d\mathbf{B}$ is equal to

$$dB = \frac{\mu_0\mu_r}{4\pi}\frac{I\,dl\sin\alpha}{r^2}\ \text{(in SI units)},$$

$$dB = \frac{\mu_r}{c}\frac{I\,dl\sin\alpha}{r^2}\ \text{(in Gaussian units)},$$

where α is the angle between vectors $d\mathbf{l}$ and $\mathbf{r}$.

23.2.3 The *magnetic field strength* **H** is the vector characteristic of a magnetic field that, for a homogeneous isotropic medium, is related to **B** as follows:

$$\mathbf{H} = \frac{\mathbf{B}}{\mu_0\mu_r}\ \text{(in SI units)},$$

$$\mathbf{H} = \frac{\mathbf{B}}{\mu_r}\ \text{(in Gaussian units)}.$$

The universal relation between vectors **B** and **H** for a magnetic field in an arbitrary medium and the more general definition of the magnetic field strength vector **H** are discussed in Sect. 26.4.4.

The strength of a magnetic field set up by a direct current in a homogeneous isotropic medium is independent of the magnetic properties of the medium. Thus

$$d\mathbf{H} = \frac{I}{4\pi r^3}[d\mathbf{l}\ \mathbf{r}], \quad dH = \frac{I\,dl\sin\alpha}{4\pi r^2}\ \text{(in SI units)},$$

$$d\mathbf{H} = \frac{I}{cr^3}[d\mathbf{l}\ \mathbf{r}], \quad dH = \frac{I\,dl\sin\alpha}{cr^2}\ \text{(in Gaussian units)},$$

where α is the angle between vectors $d\mathbf{l}$ and $\mathbf{r}$.

23.2.4 It follows from a comparison of the characteristics of electric (**E** and **D**) and magnetic (**B** and **H**) fields that the electric field strength vector **E** is analogous to the magnetic vector **B**. Both vectors specify the force effects of the fields and depend upon the properties of the medium in which the field is set up.

The analogue of vector **D** of electric displacement (15.3.1) is the magnetic field strength vector **H**.

23.2.5 An electric charge q, travelling in an unbounded homogeneous and isotropic medium at the constant velocity $\mathbf{v}$, sets up a magnetic field whose magnetic induction $\mathbf{B}_q$ can be calculated by the formula[*]

$$\mathbf{B}_q = \frac{\mu_0 \mu_r}{4\pi} \frac{q}{r^3} [\mathbf{v} \mathbf{r}], \quad B_q = \frac{\mu_0 \mu_r}{4\pi} \frac{qv \sin \beta}{r^2} \text{ (in SI units),}$$

$$\mathbf{B}_q = \frac{\mu_r}{c} \frac{q}{r^3} [\mathbf{v} \mathbf{r}], \quad B_q = \frac{\mu_r}{c} \frac{qv \sin \beta}{r^2} \text{ (in Gaussian units),}$$

where β is the angle between vectors $\mathbf{v}$ and $\mathbf{r}$, and $\mathbf{r}$ is the radius vector from the travelling charge to the point A being considered in the field. Vectors $\mathbf{B}_q$ and $\mathbf{H}_q$ are perpendicular to the plane passing through vectors $\mathbf{v}$ and $\mathbf{r}$.
When $q > 0$, the shortest rotation from vector $\mathbf{v}$ to vector $\mathbf{r}$ is seen from the head of vector $\mathbf{B}_q$ (and $\mathbf{H}_q$) to be counter-
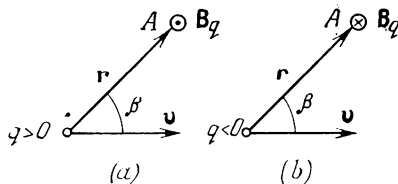
Fig. 23.3

clockwise (Fig. 23.3a). When $q < 0$, then $\mathbf{B}_q$ (and $\mathbf{H}_q$) have the opposite direction (Fig. 23.3b). The magnetic field of a moving charge is variable, because the radius vector $\mathbf{r}$ varies in magnitude and direction as charge q travels even if $\mathbf{v} = \text{const.}$ The magnetic field of a moving charge, depending upon angle β between vectors $\mathbf{v}$ and $\mathbf{r}$, is not spherically symmetrical, as in the case of the electrostatic field of a point charge (14.2.3). The magnetic field being considered has mirror symmetry with respect to the direction of $\mathbf{v}$. The field strength is maximum at points in a plane passed through the charge perpendicularly to vector $\mathbf{v}$ (under the condition that $v \ll c$). At all points on a straight line coinciding with vector $\mathbf{v}$ there is no magnetic field induction,

23.3 Simplest Cases of Magnetic Fields Set Up by Direct Currents

23.3.1 Making use of the Biot-Savart-Laplace law we can determine the characteristics (B and H) of a magnetic field set up by an electric current flowing in a conductor of finite dimensions and arbitrary shape. According to the principle of superposition of fields (15.2.2) the magnetic induction **B** at an arbitrary point of a magnetic field of a conductor carrying the current I is equal to

$$\mathbf{B} = \int_L d\mathbf{B}$$

where $d\mathbf{B}$ is the magnetic induction of the field set up by an element of the conductor with the length dl. Integration is carried out over the whole length L of the conductor.
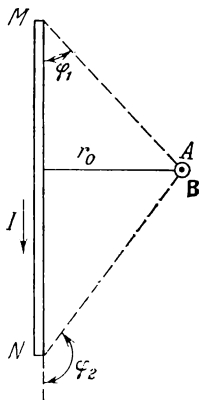
Fig. 23.4

23.3.2 *A straight conductor MN carrying the current I sets up a magnetic field with induction* **B** and strength **H** at an arbitrary point A equal to*

$$B = \frac{\mu_0 \mu_r}{4\pi} \frac{I}{r_0} (\cos \varphi_1 - \cos \varphi_2), \quad H = \frac{B}{\mu_0 \mu_r} \text{ (in SI units)},$$

$$B = \frac{1}{c} \mu_r \frac{I}{r_0} (\cos \varphi_1 - \cos \varphi_2), \quad H = \frac{B}{\mu_r} \text{ (in Gaussian units)},$$

where r_0 is the distance from point A to the conductor, φ_1 and φ_2 are the angles formed by the radius vectors from point A to the beginning and end of the conductor (Fig. 23.4), μ_r is the

---

* It is assumed in all the examples of Sect. 23.3 that the medium is homogeneous, isotropic and completely fills all the space in which the magnetic field exists.

relative magnetic permeability of the medium, and μ_0 is the magnetic constant in SI units (Appendix II).
For an infinitely long conductor ($\varphi_1 = 0$ and $\varphi_2 = \pi$)

$$B = \frac{\mu_0 \mu_r}{4\pi} \frac{2I}{r_0} , \quad H = \frac{B}{\mu_0 \mu_r} \text{ (in SI units),}$$

$$B = \frac{1}{c} \mu_r \frac{2I}{r_0} , \quad H = \frac{B}{\mu_r} \text{ (in Gaussian units).}$$

23.3.3 For the magnetic field at the centre of a *rectangular turn (loop) carrying the current I*

$$B = \frac{\mu_0 \mu_r}{4\pi} \frac{8I \sqrt{a^2 + b^2}}{ab} , \quad H = \frac{B}{\mu_0 \mu_r} \text{ (in SI units),}$$

$$B = \frac{1}{c} \mu_r \frac{8I \sqrt{a^2 + b^2}}{ab} , \quad H = \frac{B}{\mu_r} \text{ (in Gaussian units),}$$

where a and b are the sides of the rectangle.
23.3.4 The *magnetic moment* $\mathbf{p_m}$ *of a closed loop* of arbitrary shape, carrying the current I, is equal to

$$\mathbf{p_m} = I \int_S \mathbf{n} \, dS \text{ (in SI units),}$$

$$\mathbf{p_m} = \frac{1}{c} I \int_S \mathbf{n} \, dS \text{ (in Gaussian units),}$$

where $\mathbf{n}$ is a unit vector of the outward normal to element dS of the surface S, bounded by the current-carrying loop. In the case of a flat loop, surface S is plane and all the normals have the same direction. Hence

$$\mathbf{p_m} = IS\mathbf{n}, \quad p_m = IS \text{ (in SI units),}$$

$$\mathbf{p_m} = \frac{1}{c} IS\mathbf{n}, \quad p_m = \frac{1}{c} IS \text{ (in Gaussian units).}$$

Vector $\mathbf{p}_m$ is directed so that from its head the current in the circuit is seen to be flowing counterclockwise (Fig. 23.5).
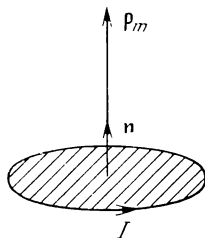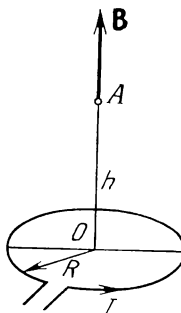
Fig. 23.5   Fig. 23.6

23.3.5 The magnetic field set up by a circular turn, or loop, carrying the current I, at an arbitrary point A on the axis of the turn (loop) (Fig. 23.6), is specified by

$$\mathbf{B} = \frac{\mu_0 \mu_r}{4\pi} \frac{2\mathbf{p}_m}{(R^2 + h^2)^{3/2}} \,, \quad \mathbf{H} = \frac{\mathbf{B}}{\mu_0 \mu_r} \text{ (in SI units),}$$

$$\mathbf{B} = \mu_r \frac{2\mathbf{p}_m}{(R^2 + h^2)^{3/2}} \,, \quad \mathbf{H} = \frac{\mathbf{B}}{\mu_r} \text{ (in Gaussian units).}$$

Here $\mathbf{p}_m$ is the magnetic moment of a current-carrying turn, or loop, (23.3.4).
The magnitudes of vectors $\mathbf{B}$ and $\mathbf{H}$ equal

$$B = \frac{\mu_0 \mu_r}{2} \frac{I R^2}{(R^2 + h^2)^{3/2}} = \frac{\mu_0 \mu_r I S}{2\pi (R^2 + h^2)^{3/2}} \,,$$

$$H = \frac{B}{\mu_0 \mu_r} \text{ (in SI units),}$$

$$B = \frac{1}{c} \mu_r \frac{2\pi I R^2}{(R^2 + h^2)^{3/2}} = \frac{1}{c} \mu_r \frac{2IS}{(R^2 + h^2)^{3/2}}$$

$$H = \frac{B}{\mu_r} \text{ (in Gaussian units),}$$

where h is the distance from point A to the centre of the turn (loop), R is the radius and S is the area of the turn.

23.3.6 The magnetic field at the centre of a circular turn, or loop, has the characteristics (23.3.5):

$$B = \frac{\mu_0 \mu_r}{4\pi} \frac{2p_m}{R^3}, \quad H = \frac{B}{\mu_0 \mu_r} \text{ (in SI units),}$$

$$B = \mu_r \frac{2p_m}{R^3}, \quad H = \frac{B}{\mu_r} \text{ (in Gaussian units).}$$

The magnitudes of vectors **B** and **H** equal

$$B = \mu_0 \mu_r \frac{I}{2R}, \quad H = \frac{I}{2R} \text{ (in SI units),}$$

$$B = \frac{2\pi}{c} \mu_r \frac{I}{R}, \quad H = \frac{1}{c} \frac{2\pi I}{R} \text{ (in Gaussian units).}$$

The direction of the magnetic field is along the axis of the loop and perpendicular to its plane (Fig. 23.6).
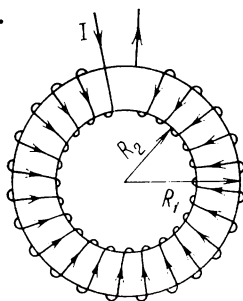
Fig. 23.7

23.3.7 A *toroid* is an annular coil wound on a core having the shape of a torus (Fig. 23.7). The magnetic field of a toroid is completely confined within its volume,

The characteristics of this magnetic field are calculated by the formulas

$$B = \mu_0 \mu_r \frac{NI}{2\pi r}, \quad H = \frac{NI}{2\pi r} \text{ (in SI units)},$$

$$B = \frac{1}{c} \mu_r \frac{2NI}{i}, \quad H = \frac{1}{c} \frac{2NI}{r} \text{ (in Gaussian units)}.$$

The magnetic induction B and the strength H of the magnetic field on the circular centre line of the toroid are equal to

$$B_{c1} = \mu_0 \mu_r \frac{NI}{2\pi R_{c1}} = \mu_0 \mu_r nI, \quad H = nI \text{ (in SI units)}.$$

Here N is the number of turns of the toroid, I is the current in the turns, r is the radius of a circle in the toroid, $R_{c1} = (R_1 + R_2)/2$, R_1 and R_2 are the external and internal radii of the torus, and n is the number of turns per unit length of the toroid's centre line.

23.3.8 A *solenoid* is a cylindrical coil consisting of a large number of turns of wire in the form of a helix. If the turns are wound tight up against or suficiently close to one another, the solenoid is dealt with as a system of circular closed current-carrying loops connected in series and having the same radius and a common axis.

The magnetic moment p_m (23.3.4) of the solenoid is equal to the vector sum of the magnetic moments of all N of its turns. Thus

$$p_m = NIS n,$$

where I is the current in the solenoid's turns, S is the area of a turn, and n is a unit vector of the normal to the plane of a turn. Vector p_m is along the axis of the solenoid and coincides in direction with the magnetic field as determined by the corkscrew rule (23.1.3).

The magnetic induction B and strength H of the solenoid's magnetic field at an arbitrary point A, lying on its axis, are

numerically equal to

$$H = \frac{\mu_0\mu_r}{2} nI (\cos\alpha_2 - \cos\alpha_1), \quad H = \frac{B}{\mu_0\mu_r} \text{ (in SI units),}$$

$$B = \frac{1}{c} 2\pi\mu_r nI (\cos\alpha_2 - \cos\alpha_1), \quad H = \frac{B}{\mu_r} \text{ (in Gaussian units),}$$

where $n = N/L$ is the number of turns per unit length of the solenoid, α_2 and α_1 are angles at which the ends of the solenoid are seen from point A $(\alpha_2 < \alpha_1)$,

$$\cos\alpha_1 = -\frac{l_1}{\sqrt{R^2 + l_1^2}}, \quad \cos\alpha_2 = \frac{L - l_1}{\sqrt{R^2 + (L - l_1)^2}},$$

L is the length of the solenoid (Fig. 23.8), and R is the radius of the cylindrical coil.
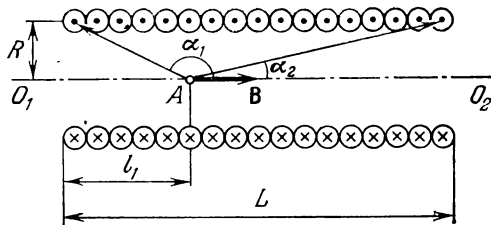
Fig. 23.8

23.3.9 Under the condition that $L \gg R$, the magnetic field inside the solenoid at points on its axis sufficiently distant from its ends is characterized by the values

$$B = \mu_0\mu_r nI, \quad H = nI \text{ (in SI units),}$$

$$B = \frac{1}{c} 4\pi\mu_r nI, \quad H = \frac{1}{c} 4\pi nI \text{ (in Gaussian units).}$$

The magnetic induction B and strength H of the magnetic field of a solenoid of sufficient length, at points of its axis coinciding with its ends, are equal to

$$B = \frac{\mu_0 \mu_r}{2} \, nI, \quad H = \frac{1}{2} \, nI \text{ (in SI units)},$$

$$B = \frac{1}{c} \, 2\pi \mu_r nI, \quad H = \frac{1}{c} \, 2\pi nI \text{ (in Gaussian units)}.$$

23.4 Interaction of Conductors. Effect of a Magnetic Field on Current-Carrying Conductors

23.4.1 Ampere's force (23.1.4), which acts on an element dl of length of a straight conductor carrying the current I_1, and is exerted by a long straight conductor carrying the current I_2, located parallel to the first conductor at a distance of a, is numerically equal to

$$dF = \frac{\mu_0 \mu_r}{4\pi} \, \frac{2I_1 I_2}{a} \, dl \text{ (in SI units)},$$

$$dF = \frac{1}{c^2} \, \mu_r \, \frac{2I_1 I_2}{a} \, dl \text{ (in Gaussian units)},$$

where μ_0 is the magnetic constant in SI units (23.2.2) and c is the electrodynamic constant (Appendix I). The force dF is a force of magnetic interaction (24.1.4).

It is assumed that the lengths of the conductors are many times greater than the distance a between them, and that element dl is located far from the ends of the first conductor. It is moreover assumed that the conductors are in a homogeneous isotropic medium having the relative magnetic permeability μ_r.

The force F exerted on a current-carrying conductor of length l is equal to

$$F = \frac{\mu_0 \mu_r}{4\pi} \, \frac{2I_1 I_2}{a} \, l \text{ (in SI units)},$$

$$F = \frac{1}{c^2} \, \mu_r \, \frac{2I_1 I_2}{a} \, l \text{ (in Gaussian units)}.$$

Conductors carrying currents I_1 and I_2 that flow in the same direction attract each other. If the currents flow in opposite directions, the conductors repulse each other (see also Sect. 24.1.4).

23.4.2 A closed flat current-carrying loop, of arbitrary geometric shape, placed in a uniform magnetic field (23.1.3), is subject to a moment of force (t orque) **M** equal to

$$\mathbf{M} = [\mathbf{p_m B}],$$

where $\mathbf{p_m}$ is the magnetic moment vector of a current-carrying loop (23.3.4), and **B** is the vector of magnetic induction (23.1.2) of the field. Torque **M** is perpendicular to the plane passing through vectors $\mathbf{p_m}$ and **B** in such a manner that the shortest rotation from $\mathbf{p_m}$ to **B** is seen from the head of vector **M** to be counterclockwise. The torque tends to turn the loop to a position of stable equilibrium, in which vectors $\mathbf{p_m}$ and **B** are parallel to each other.

The preceding formula provides for a definition of the magnetic induction B (23.1.2): the magnitude of the magnetic induction vector at a given point of a uniform magnetic field is equal to the maximum value of the torque M_{max} acting in the vicinity of the point on a small flat closed current-carrying loop, having a magnetic moment p_m of unit magnitude:

$$B = \frac{M_{max}}{p_m} = \frac{M_{max}}{.IS} \text{ (in SI units),}$$

$$B = \frac{M_{max}}{p_m} = \frac{M_{max}}{\frac{1}{c}IS} \text{ (in Gaussian units).}$$

23.4.3 When a closed current-carrying loop is placed into a nonuniform magnetic field (23.1.3) in which the magnetic induction **B** of the field varies over distances comparable to the linear dimensions of the loop, the loop is subject to the force **F** equal to

$$\mathbf{F} = p_{mx} \frac{\partial \mathbf{B}}{\partial x} + p_{my} \frac{\partial \mathbf{B}}{\partial y} + p_{mz} \frac{\partial \mathbf{B}}{\partial z} ,$$

where p_{mx}, p_{my}, and p_{mz} are the projections of vector $\mathbf{p_m}$ on the axes of Cartesian coordinates. In particular, when vector **B**

is along the OX axis and depends only on coordinate x [i.e. $B_x = B\,(x)$, $B_y = B_z = 0$ and $\mathbf{B} = B_x\mathbf{i}$] then

$$\mathbf{F} = p_{mx}\,\frac{dB}{dx}\,\mathbf{i}.$$

By the action of force $\mathbf{F}$ an unfastened current-carrying loop is attracted into a region with a stronger magnetic field.

23.5 Total Current Law. Magnetic Circuits

23.5.1 The *circulation of* the magnetic induction *vector* $\mathbf{B}$ along a closed circuit L is an integral of the form:

$$\oint_L \mathbf{B}\,dl = \oint_L B\,dl \cos\alpha,$$

where L is the closed loop of arbitrary shape, α is the angle between vectors $\mathbf{B}$ and dl, and dl is the vector of an element of length of the loop in the direction it is circuited.

23.5.2 The *total current law for a magnetic field in vacuum* states: the circulation of the magnetic induction vector of a magnetic field along a closed loop in vacuum is proportional to the algebraic sum of the currents bounded by the loop. Thus

$$\oint_L \mathbf{B}\,dl = \mu_0 \sum_{k=1}^{n} I_k \quad \text{(in SI units)},$$

$$\oint_L \mathbf{B}\,dl = \frac{4\pi}{c} \sum_{k=1}^{n} I_k \quad \text{(in Gaussian units)},$$

where μ_0 is the magnetic constant in SI units (23.2.2), c is the electrodynamic constant (23.2.2) and n is the number of current-carrying conductors bounded by loop L of arbitrary shape. This law is valid for current-carrying conductors of any shape or size.

In calculating the algebraic sum of the currents, a current is considered positive if from the head of the current density

vector (20.2.3) the circuiting of loop L is seen to be counter-clockwise. Otherwise, the current is considered to be negative. For a generalization of the total current law to cover magnetic fields in an arbitrary media, see Sect. 26.4.2.

23.5.3 In contrast to an electrostatic conservative field, in which the circulation of the strength E along any closed path equals zero (16.1.4), a magnetic field has a *curl* (it is called a *rotational, vortical* or *circuital field*). In such a field the circulation of the magnetic induction vector B of the field along a closed path is not equal to zero. If no currents are bound by loop L, the circulation of vector B along this loop equals zero. But this does not alter the rotational nature of the magnetic field.

23.5.4 The total current law can be written for either vacuum or for an arbitrary medium in the form of the circulation of field strength vector H (23.2.3) along an arbitrary closed loop L that bounds the currents:

$$\oint_L H \, dl = \sum_{k=1}^{n} I_k \text{ (in SI units)},$$

$$\oint_L H \, dl = \frac{4\pi}{c} \sum_{k=1}^{n} I_k \text{ (in Gaussian units)}.$$

If no currents are bounded by loop L, the circulation of vector H along this loop equals zero.

23.5.5 The *flux of the magnetic induction vector B (magnetic flux)* through a surface element of area dS is the scalar physical quantity equal to

$$d\Phi_m = B \, dS = B_n \, dS = B \, dS \cos \gamma,$$

where $dS = n \, dS$, n is the unit vector of the normal to the surface of area dS, B_n is the projection of vector B on the direction of the normal (Fig. 23.9), and γ is the angle between vectors B and n. The magnetic flux Φ_m through an arbitrary surface S is

$$\Phi_m = \int_S B \, dS = \int_S B_n \, dS.$$

20—0186

In calculating this integral, the vectors n of the normals to the elements of surface dS should be all in the same direction with

respect to surface S. Thus, if S is a closed surface, vectors n should be directed either all outward or all inward.
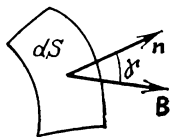
When the field is uniform and S is a plane surface perpendicular to vector B, then $B_n = B = \text{const}$ and $\Phi_m = BS$.

Fig. 23.9

23.5.6 The *Ostrogradsky-Gauss theorem for a magnetic field* states that the magnetic flux through an arbitrary closed surface equals zero. Thus

$$\oint_S \mathbf{B}\,d\mathbf{S} = \oint_S B_n\,dS = 0.$$

This theorem is a mathematical expression of the fact that in nature there are no magnetic charges at which lines of magnetic induction begin or end (23.1.3).

The differential form of the Ostrogradsky-Gauss theorem for a magnetic field is one of Maxwell's equations for an electromagnetic field (27.4.2).

23.5.7 A *magnetic circuit* is a set of bodies or regions of space to which a magnetic field is confined. For example, the internal region of a toroid (23.3.7) and that of an infinitely long solenoid (23.3.9) are magnetic circuits. To strengthen their magnetic fields, magnetic circuits are made of materials having a high relative magnetic permeability μ_r (23.2.1), such as iron. The design of magnetic circuits, which are the essential components of electrical machinery and such electric devices as transformers and electromagnets, is based on the laws of magnetic circuits.

23.5.8 *Ohm's law for a closed magnetic circuit* (the *Hopkinson formula*) is

$$\Phi_m = \frac{F_m}{R_m},$$

where Φ_m is the magnetic flux, which is constant along each portion of the magnetic circuit, $F_m = IN$ is the *magnetomotive force (mmf)*, or *magnetizing force* (in SI units), N is the number of turns carrying the electric current I for magnetizing the

circuit, and R_m is the *total magnetic resistance (reluctance) of the circuit*. The magnetic resistance of a magnetic circuit of length l_i is

$$R_{mi} = \int_0^{l_i} \frac{dl}{\mu_0 \mu_r S} \text{ (in SI units),}$$

where μ_r is the relative magnetic permeability of the given circuit, μ_0 is the magnetic constant in SI units (23.2.2), and S is the cross-sectional area of the circuit. When $S = \text{const}$, $R_{mi} = l_i/\mu_0 \mu_r S$ (in SI units).

23.5.9 The total magnetic resistance R_m of the parts of a magnetic circuit connected in series is

$$R_m = \sum_{i=1}^n R_{mi},$$

where R_{mi} is the magnetic resistance of the ith part of the magnetic circuit, and n is the number of parts that make up the circuit.

If n magnetic resistances are connected in parallel, the total magnetic resistance R_m of the circuit is

$$R_m = \frac{1}{\sum\limits_{i=1}^n \dfrac{1}{R_{mi}}}.$$

23.5.10 A *branch point* of a magnetic circuit is a region of space or bodies where there are more than two possible directions of the lines of magnetic induction (23.1.3).

Kirchhoff's first law for branched magnetic circuits states that the algebraic sum of the magnetic fluxes in the branches that join in a branch point is equal to zero. Thus

$$\sum_{i=1}^n \Phi_{mi} = 0,$$

where n is the number of branches joined at the branch point (cf. Sect. 23.3.1).

20*

A magnetic flux is considered positive if the lines of induction enter the branch point, and negative if the lines emerge from it. **23.5.11** *Kirchhoff's second law for magnetic circuits* states that in any closed loop (mesh of a network), arbitrarily chosen in a branched magnetic circuit (network), the algebraic sum of the products of the magnetic fluxes by the magnetic resistances of the corresponding portions of the circuit (23.5.8) equals the algebraic sum of the magnetomotive forces in this loop (23.5.8):

$$\sum_{i=1}^{k} \Phi_{\mathrm{m}i} R_{\mathrm{m}i} = \sum_{i=1}^{k} F_{\mathrm{m}i},$$

where k is the number of portions that make up the closed loop (cf. 21.3.2). The fluxes $\Phi_{\mathrm{m}i}$ and mmf's $F_{\mathrm{m}i}$ are considered positive if the directions of the corresponding lines of magnetic induction (23.1.3) coincide with the arbitrarily chosen direction for circuiting the loop.

23.6 Work Done in Moving a Current-Carrying Conductor in a Magnetic Field

23.6.1 An unfixed current-carrying conductor is moved by Ampere's force (23.1.4) when it is placed in a magnetic field. The element of work δW^* done by the Ampere force in moving an element, of length dl, of a conductor carrying the current I is

$$\delta W^* = I\, d\Phi_{\mathrm{m}}^* \quad \text{(in SI units)},$$

$$\delta W^* = \frac{1}{c}\, I\, d\Phi_{\mathrm{m}}^* \quad \text{(in Gaussian units)},$$

where $d\Phi_{\mathrm{m}}^*$ is the element of magnetic flux (23.5.5) through the element of surface dS generated by the element of the conductor, of length dl, in its small displacement, and c is the electrodynamic constant (23.2.2).
23.6.2 In a small displacement of a current-carrying conductor of finite length, the work δW done by the Ampere forces is the sum of the elements of work for all the elements of length of the

conductor. It is equal to the integral of δW^* calculated along the whole length l of the conductor. Thus

$$\delta W = \int_l I\, d\Phi_m^* = I \int_l d\Phi_m^* = I\, d\Phi_m \text{ (in SI units)},$$

$$\delta W = \frac{1}{c}\, I\, d\Phi_m \text{ (in Gaussian units)},$$

where $d\Phi_m$ is the magnetic flux through the surface generated by the conductor of length l in a small displacement.
23.6.3 If a conductor of finite length l carries a constan tcurrent I and is moved a finite distance, the required work of the Ampere forces for this motion is

$$W = I\Phi_m,$$

where Φ_m is the magnetic flux through the surface generated by the conductor in its motion.
23.6.4 The work done by Ampere forces in the displacement of a closed flat loop carrying the constant current I in a magnetic field from initial position 1 to final position 2 is equal to

$$W_{12} = I\,(\Phi_{m2} - \Phi_{m1}),$$

where Φ_{m1} and Φ_{m2} are magnetic fluxes "linked to the loop" in positions 1 and 2, i.e. the magnetic fluxes through a surface bounded by the loop. In calculating these magnetic fluxes, the direction of normal n (23.5.5) is coordinated with the direction of the current in the loop by the corkscrew rule: from the head of the vector of the normal, the current is seen flowing counter-clockwise in the loop.
23.6.5 The formulas given above are valid for a current-carrying coil of N turns moving in a magnetic field by Ampere forces, or, in general, for a loop of arbitrary shape. For example, the work done in a small displacement of a coil carrying the current I is equal to

$$\delta W = I\sum_{i=1}^{N} d\Phi_{mi} = I d\Big(\sum_{i=1}^{N} \Phi_{mi}\Big) = I\, d\Psi,$$

where $\Psi = \sum\limits_{i=1}^{N} \Phi_{mi}$ is the total magnetic flux through all N
turns of the coil; it is called the *flux linkage of the circuit.*
23.6.6 The work required to move a conductor or closed loop
carrying a constant current in a magnetic field is done at the
expense of the energy consumed in the current source (21.1.3).

CHAPTER 24   MOTION OF CHARGED PARTICLES
IN ELECTRIC AND MAGNETIC FIELDS

24.1 Lorentz Force

24.1.1 Not only current-carrying conductors (23.1.4) are sub-
ject to the action of a magnetic field. This is also true of separate
charged particles moving in the field. The force $\mathbf{F_L}$, acting on
an electric charge q moving in a magnetic field at the velocity
$\mathbf{v}$, is called the *Lorentz force* and is equal to (see also Sect. 24.1.5)

$\mathbf{F_L} = q\,[\mathbf{vB}]$ (in SI units),

$\mathbf{F_L} = \dfrac{q}{c}\,[\mathbf{vB}]$ (in Gaussian units),

where q is the algebraic value of the moving charge, $\mathbf{B}$ is the
induction of the magnetic field in which the charge is moving
(23.1.2), and c is the electrodynamic constant (23.2.2). The
relative arrangement of vectors $\mathbf{v}$, $\mathbf{B}$ and $\mathbf{F_L}$ is shown in
Fig. 24.1 for positive $(q > 0)$ and negative $(q < 0)$ charges.
The magnitude of the Lorentz force is

$F_L = qvB \sin \alpha,$

where α is the angle between vectors $\mathbf{v}$ and $\mathbf{B}$.
24.1.2 The Lorentz force is always perpendicular to the velocity
of the charged particle and imparts normal acceleration (1.4.6)
to it. Since the Lorentz force does not change the magnitude of
the velocity, but only its direction, this force does no work and
the kinetic energy of the charged particle remains constant
when it moves in a magnetic field.

24.1.3 Making use of the Lorentz force, magnetic induction B (23.1.2) can be defined as follows: the magnitude of the magnetic induction vector at a given point of the magnetic field is equal to the maximum Lorentz force $F_{L\ max}$ acting on a unit
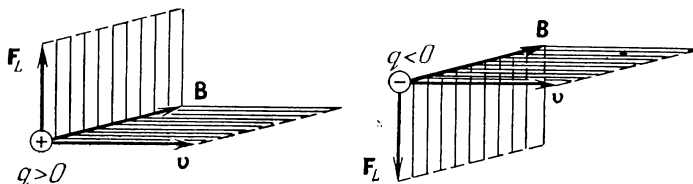
Fig. 24.1

positive charge, which, at the given point, travels at unit velocity. Thus

$$B = \frac{F_{L\ max}}{qv} .$$

Here $F_L = F_{L\ max}$ under the condition that $\alpha = \pi/2$ (24.1.2). See also Sects. 23.1.2 and 23.4.2.

24.1.4 The force exerted on moving charge q_2 by moving charge q_1 is called the *force of their magnetic interaction (magnetic force)*. For the special case of two positive charges q_1 and q_2, travelling in vacuum at the same velocity $v_1 = v_2 = v \ll c$ along the OX axis, the force F_{int} of magnetic interaction is a force of attraction and is numerically equal to

$$F_{int} = \frac{\mu_0}{4\pi} \frac{q_1 q_2}{r^2} v^2 \ \text{(in SI units)},$$

where r is the distance between the charges, and μ_0 is the magnetic constant (23.2.2).

The force F_{int} of magnetic interaction can be presented in the form

$$F_{int} = q_2 v \frac{\mu_0 q_1 v}{4\pi r^2} = q_2 v B,$$

where

$$B = \frac{\mu_0 q_1 v}{4\pi r^2} \, .$$

In this form the force F_{int} coincides with the Lorentz force at $\sin \alpha = 1$ (24.1.1) if we assume that B is the induction of the magnetic field (23.1.2) set up by moving charge q_1. In its turn, this field acts on moving charge q_2. Magnetic forces and a magnetic field are the result of the way in which forces are transformed in going over from a fixed to a moving frame of reference, and follows from the formulas of the special theory of relativity (5.1.1).

Comparing force F_{int} with the force F_{e} of electrostatic repulsion between mutually fixed charges q_1 and q_2 in vacuum $(\varepsilon_r = 1)$ located at the same distance r from each other (14.2.6)

$$F_{\text{e}} = \frac{1}{4\pi\varepsilon_0} \frac{q_1 q_2}{r^2} \text{ (in SI units)},$$

(where ε_0 is the electric constant in SI units), we readily find that

$$\frac{F_{\text{int}}}{F_{\text{e}}} = \frac{v^2}{c^2} \, ,$$

where c is the electrodynamic constant (23.2.2) and is related to ε_0 and μ_0 by the equation $\varepsilon_0 \mu_0 = 1/c^2$ (Appendix I). When the velocity of the charges is small compared to the velocity of light in free space $(v \ll c)$, the magnetic interaction between the moving charges is considerably less than their electrostatic interaction. But in the case when the charges are moving in an electrically neutral conductor, the electric forces are compensated (16.4.2), leaving only magnetic interaction. This explains the magnetic interaction of current-carrying conductors (23.4.1). Though the force of magnetic interaction between each pair of electrons is small, the number of pairs is so immense that the resultant force of magnetic interaction of parallel current-carrying conductors is an appreciable amount (23.4.1).

24.1.5 If, in addition to a magnetic field of induction **B**, a moving electric charge is also subject to the action of an electric field of strength **E** (15.1.2), the resultant force **F** applied to the

charge is equal to the vector sum of the force $F_e = qE$, exerted on the charge by the electric field and the Lorentz force (24.1.1),

$F = qE + q\,[vB]$ (in SI units),

$F = qE + \dfrac{q}{c}\,[vB]$ (in Gaussian units).

The last expression is also called the Lorentz force and, sometimes, the *generalized Lorentz force* or the *Lorentz formula*.
24.1.6 By the action of the Lorentz force (24.1.1), a charged particle travels in a circular path of constant radius r in a uniform magnetic field perpendicular to the particle's velocity. This circular path is in a plane perpendicular to vector B, and the Lorentz force is a centripetal one (2.4.3). The radius of the circular path is

$r = \dfrac{m}{|q|}\dfrac{v}{B}$ (in SI units),

$r = \dfrac{cmv}{|q|B}$ (in Gaussian units),

where $|q|$ is the absolute value of the particle's charge, m is the mass of the particle, v is its velocity, B is the magnetic field induction and c is the electrodynamic constant (23.2.2). If the particle travels in the plane of the drawing (Fig. 24.2), its deflection in a field perpendicular to the velocity and directed from behind the drawing depends upon the sign of the charge. This is the basis for determining the sign of the charge of a particle travelling in a magnetic field.
The period T of revolution (1.5.5) of a charged particle in a uniform magnetic field (23.1.3) is independent of its velocity (when $v \ll c$):

$T = \dfrac{2\pi}{B}\dfrac{m}{|q|}$ (in SI units),

$T = \dfrac{2\pi}{B}\dfrac{mc}{|q|}$ (in Gaussian units).

The principle of cyclic orbit accelerators of charged particles (24.4.6) is based on the aforesaid.

24.1.7. If the velocity vector **v** of a charged particle makes the angle α with the direction of vector **B** of a uniform magnetic
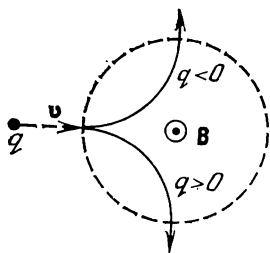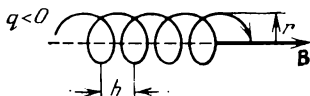
Fig. 24.2   Fig. 24.3

field, the particle travels along a helix (Fig. 24.3) of radius r and pitch h equal to

$$r = \frac{m}{|q|} \frac{v \sin \alpha}{B} \ , \quad h = \frac{2\pi}{B} \frac{m}{|q|} \ v \cos \alpha \ \text{(in SI units),}$$

$$r = \frac{mc}{|q|} \frac{v \sin \alpha}{B} \ , \quad h = \frac{2\pi}{B} \frac{mc}{|q|} \ v \cos \alpha \ \text{(in Gaussian units).}$$

If this motion occurs in a nonuniform magnetic field (23.1.3) whose magnetic induction increases in the direction of motion of the particle, then r and h decrease with an increase in B. This phenomenon is used for focusing charged particles in a magnetic field.

24.2 Hall Effect

24.2.1 The *Hall effect* refers to the setting up of a transverse electric field and potential difference in a current-carrying metal or semiconductor placed in a magnetic field which is perpendicular to the direction of the current density vector (20.2.3). The Hall effect results from the deflection of electrons moving in a magnetic field by the action of the Lorentz force (24.1.1).

Shown in Fig. 24.4a are the directions of the magnetic induction
B, current density **j**, electron velocity **v**, Lorentz force F_L and
the signs of the charges accumulated on the opposite, upper and
lower, sides in the case of a metal or electron (n-type) semicon-
ductor (41.10.2). In a hole (p-type) semiconductor (41.10.3)
the signs of the charges on the surfaces are the opposite of those
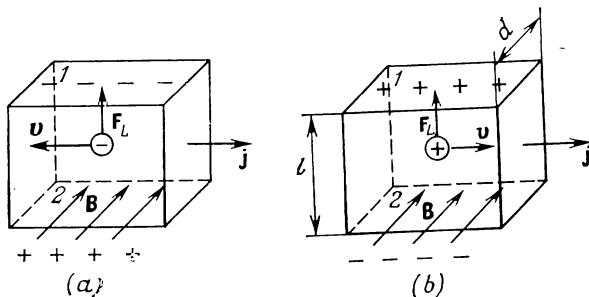
Fig. 24.4

in the preceding case (Fig. 24.4b). The charges are deflected by
the magnetic field until the action of the force in the transverse
electric field counterbalances the Lorentz force.
24.2.2 The equilibrium potential difference in the Hall effect is

$$\Delta V = V_1 - V_2 = R_H \frac{IB}{d} ,$$

where I is the current, B is the magnetic field induction (23.1.2),
d is the linear dimension of the metal or semiconductor in the
direction of vector **B**, and R_H is the Hall coefficient.
The strength E_H of the transverse electric field in the Hall
effect is

$$E_H = R_H [Bj],$$

where **j** is the current density vector.

24.2.3 For metals and extrinsic semiconductors (41.10.5), having a single kind of conduction, the Hall coefficient equals

$$R_{\mathrm{H}} = \frac{A}{n_0 q} \quad \text{(in SI units)},$$

$$R_{\mathrm{H}} = \frac{A}{cn_0 q} \quad \text{(in Gaussian units)},$$

where c is the electrodynamic constant (23.2.2), q and n_0 are the charge and concentration of the current carriers, and $A \approx 1$ is a dimensionless factor depending upon how the current-carrier velocities are statistically distributed. The sign of the Hall coefficient coincides with that of the current-carrier charge q. A measurement of the Hall coefficient of a semiconductor indicates its type of electrical conduction. For an n-type semiconductor (electron conduction) (41.10.2), $q = -e$ and $R_{\mathrm{H}} < 0$; for a p-type semiconductor (hole conduction) (41.10.3), $q = e$ and $R_{\mathrm{H}} > 0$.
If both types of electrical conduction are observed in a semiconductor, the predominant conduction can be determined from the sign of the Hall coefficient. The formula for R_{H} given in this subsection is inapplicable in such cases.
24.2.4 Provided the type of electrical conduction is known, a measurement of the Hall coefficient enables the concentration n_0 of the current carriers to be determined. For example, the concentration of conduction electrons equals the concentration of atoms in monovalent metals. This means that there is one free electron per atom in the electron gas (20.3.1) of the metal. A known concentration of current carriers enables the mean free path $\langle \lambda \rangle$ of the electron in the metal to be assessed. From the formula of (20.3.4),

$$\langle \lambda \rangle = \frac{2\varkappa m \langle u \rangle}{n_0 e^2},$$

we find that $\langle \lambda \rangle \approx 10^{-8}$ m, which exceeds the lattice constants of the metal by two orders of magnitude (see also Sects. 20.3.7 and 41.5.5).

24.3 Charge-to-Mass Ratio of Particles. Mass Spectroscopy

24.3.1. The *charge-to-mass ratio* of a particle, q/m, is a characteristic of charged particles. Experimental determination of this charge-to-mass ratio is based on measuring the deflection of the particles in electric and magnetic fields acting jointly on them. The mass of the particles is determined from the charge-to-mass ratio q/m and the known **charge q.**

24.3.2 The *mass spectrum* of particles is the set of their mass values. Applying mass spectrometers and mass spectrographs, special instruments used in *mass spectrometry*, the relative concentrations of isotopes of chemical elements (42.1.3) and their masses can be measured with exceptionally high accuracy.

24.3.3 In the *Aston mass spectrograph* (Fig. 24.5), which was employed to discover the isotopes of various chemical elements, the par-
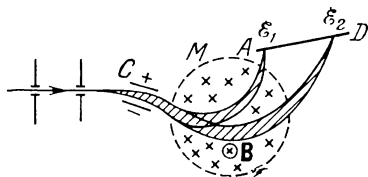
Fig. 24.5

ticles are deflected in the uniform electric field of capacitor C and the magnetic field of coil M, the fields being perpendicular to each other. The lower the velocity of the particles and the greater their charge-to-mass ratio, the more strongly they are deflected toward the capacitor plates. In a uniform magnetic field with the induction **B** directed into the drawing, the particles travel along circular arcs (24.1.6). The higher the velocity of the particles and the lower their charge-to-mass ratio, the greater the radii of these arcs. In a magnetic field a beam of particles is split into several beams, each corresponding to a definite charge-to-mass ratio. The magnetic field focuses particles having different velocities, but the same charge-to-mass ratio, on a single point. The magnetic induction **B** is chosen so that the charged particles are focused on a photographic plate AD, which is perpendicular to the plane of the drawing. A series of narrow parallel lines obtained on the plate corresponds to the various values of the particle charge-to-mass ratio. In Fig. 24.5, line $\mathscr{E}_1$ corresponds to particles with a higher charge-to-mass ratio, and line $\mathscr{E}_2$ to particles with a lower one. If

the charge-to-mass ratio of the particles of line $\mathscr{E}_1$ is known, as well as the distance between lines $\mathscr{E}_1$ and $\mathscr{E}_2$ and the parameters of the instrument, the charge-to-mass ratio of the particles corresponding to line $\mathscr{E}_2$ can be determined.

If the parallel-plate capacitors are replaced by tubular ones and special electromagnets are used that produce a plane-parallel beam of particles entering the magnetic field, we can achieve *double focusing* of the particles with respect to both energy and direction. With sufficient intensity of the lines on the photographic plate, this technique ensures high precision in particle mass measurements. For instance, the masses of the ions of light-weight elements can be measured to an accuracy within 10^{-4} per cent.

24.3.4 The relative concentrations of the isotopes of chemical elements in their natural mixtures are measured by instruments called *mass spectrometers* that electrically record the ion currents. Used in mass spectrometers are ion beams in which the kinetic energies of the ions are close in value. These are called monochromatic (monoenergetic) beams, and are produced by ion sources of special design. Even when they are strongly divergent and contain a great number of ions, such beams focus well in a transverse magnetic field. This raises the precision with which the concentrations of various isotopes can be measured.

24.4 Charged Particle Accelerators

24.4.1 Special installations for obtaining directed beams of charged particles (electrons, protons, atomic nuclei and ions of chemical elements) having an extremely high kinetic energy, under laboratory conditions are called *accelerators*. According to the shape of the path of the particles and the mechanism for accelerating them, distinction is made between *linear*, *cyclic* and *induction accelerators*. In linear accelerators (also called *linacs*) the paths of motion of the particles are close to straight lines; in the cyclic and induction accelerators the paths of the particles are circles or unwinding spirals.

24.4.2 The energy of the accelerated particles is increased by the electric field set up in the accelerator. Depending on the type of accelerator, this field can be electrostatic (15.1.2), induced (25.1.2) or high-frequency alternating. The principle of the betatron, based on the action of an induced electric field,

is discussed in Sect. 27.2.3. In a *linear electrostatic accelerator*, a charged particle passes only once through an accelerating electric field with a potential difference $(V_1 - V_2)$. If q is the charge of the particle, the energy acquired by the particle in the accelerator (15.2.7) is equal to

$$E = q (V_1 - V_2).$$

In such an accelerator the electric field is set up, for example, by a *Van de Graaff* high-voltage *electrostatic generator*, operating by the principle of multiple transfer of charges to a hollow conductor (16.4.3). This raises the potential of the conductor to values limited only by the leakage of the charge from the conductor.

Fig. 24.6

24.4.3 In *linear resonant accelerators*, the energy of the charged particles is increased by an alternating electric field of super-high frequency, which varies in synchronism with the motion of the particles. Using this type of accelerator, electrons can be accelerated to an energy of the order of dozens of GeV over a path length of several kilometres.

24.4.4 Protons, deuterons and other, heavier, charged particles are accelerated by *resonant cyclic accelerators*, in which a particle repeatedly passes through an electric field synchronized with its motion, increasing its energy each time. A strong transverse magnetic field controls the motion of the particles and returns them periodically to the region of the accelerating electric field. The particles pass definite points of the alternating electric field approximately when the field is in the same phase ("in resonance").

24.4.5 In the *cyclotron*, the simplest resonant cyclic accelerator, an accelerating alternating electric field is set up in the gap between the two halves (*dees*) N and M (Fig. 24.6) of a cylindrical box. The dees are placed in a flat closed chamber located be-

tween the poles of a strong electromagnet, whose magnetic field is perpendicular to the plane of the drawing. The alternating electric field in the gap between the dees is set up by an electric generator whose terminals are connected to electrodes m and n.
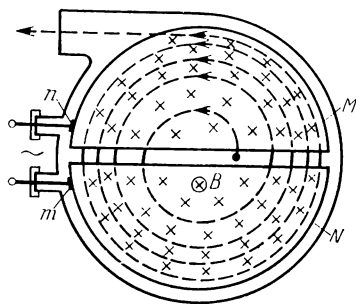
24.4.6 A particle is accelerated in the gap between the dees M and N each time it crosses this gap after describing semicircles of continuously increasing radii due to the action of the magnetic field. The same time is required for each semicircle (24.1.6). To continuously accelerate the particles it is necessary to comply with the condition of synchronism (condition of "resonance"): $T = T_0$, where T is the period of revolution of the particle in the magnetic field (24.1.6), and T_0 is the period of oscillation of the magnetic field (28.1.2). This condition is violated at relativistic velocities v of the particle, commensurate with the velocity of light c in free space. At such high velocities, the mass m of the particle increases with the velocity (5.6.1) as does the period T (24.1.6).

24.4.7 The feasibility of accelerating charged particles travelling at relativistic velocities in cyclic accelerators follows from the *principle of phase stability (autophasing)*, which states that each deviation of the period T from the resonance value T_0 (24.4.6) leads to such a change in the energy E of the particle at each acceleration that T varies in the neighborhood of T_0 and, on an average, remains equal to T_0. Thus

$$T_0 \approx T = \frac{2\pi}{B} \frac{m}{|q|} = \frac{2\pi}{B} \frac{E}{|q| \, c^2},$$

where $E = mc^2$ (5.7.2), $m = m_0/\sqrt{1 - v^2/c^2}$ (5.6.1), m_0 is the rest mass (5.6.1) of the particle, and c is the velocity of light in free space; the remaining notation is the same as in Sect. 24.1.6.

If, for example, as a result of the increase in mass m and the period T, the particle enters the gap between the dees (Fig. 24.6) by the action of the decelerating instead of accelerating electric field, the reduction in particle velocity leads to the reduction in T and the equality $T = T_0$ is reached again.

24.4.8 It follows from the principle of phase stability (24.4.7) that at a sufficiently slow increase in the period T_0 of the electric field, the period T of revolution of the particle in the magnetic field of the accelerator increases correspondingly. This

leads to an increase in the average energy $\langle E \rangle$ of all the particles because at constant induction of the magnetic field any increase in T is possible only owing to the increase in mass as the velocity of the particles increases.

This principle has been applied in an accelerator called the *synchrocyclotron* (also *frequency-modulated cyclotron* and *phasotron*). This device has a constant magnetic field, whereas the frequency $\nu_0 = 1/T_0$ (28.1.2) of the alternating electric field varies slowly with the period $\tau \gg T_0$. The radius of the orbit of the particles in the synchrocyclotron increases with their velocity (24.1.6). This requires a corresponding increase in the size of the synchrocyclotron to reach the maximum energy of the charged particles. The synchrocyclotron built and operated in the USSR, accelerating protons to an energy of 680 MeV, has a magnet with poles 6 m in diameter and a mass of 7×10^3 metric tons.

24.4.9 The *synchrotron* is an accelerator in which the frequency of the accelerating electric field is constant, and the induction B of the magnetic field varies with time. The period of revolution of a particle in the magnetic field of a synchrotron (24.4.6) is

$$T = \frac{2\pi}{|e|c^2} \frac{E}{B} ,$$

where e is the charge of the electron, and E is its energy. The condition of synchronism (24.4.6) is complied with in a synchrotron at $T_0 = \text{const}$, provided that the induction of the magnetic field increases proportionally to the particle's energy. Thus

$$B = \frac{2\pi}{|e|c^2} \frac{E}{T_0} ,$$

where T_0 is the period of a high-frequency accelerating electric field.

The following condition is complied with in a synchrotron:

$$\frac{m}{B} = \frac{eT_0}{2\pi} = \text{const.}$$

The particles travel along orbits close to circular ones (24.1.6) and, consequently, a synchrotron is equipped with annular

magnets that set up a magnetic field in a comparatively narrow region of the circular orbit.

24.4.10 A *proton synchrotron*, the most powerful accelerator of protons, combines the principles applied in the synchrocyclotron (24.4.8) and in the synchrotron (24.4.9). In this accelerator, the frequency ν_0 of the accelerating electric field is reduced and the induction B of the magnetic field is increased simultaneously and in coordination. The protons being accelerated travel along a circular orbit of constant radius. Therefore, the magnetic field is set up by an annular electromagnet as in the synchrotron.

24.4.11 The condition for simultaneous maintenance of vertical (axial) and radial stability of the design circular orbit in the synchrotron and proton synchrotron is the variation of magnetic induction B near the design orbit according to the law

$$B = \frac{\text{const}}{r^n},$$

where r is the distance from the centre of the orbit, and n varies within the limits $0 < n < 1$ (*condition of weak focusing* in the accelerator). With an increase in the maximum energy $E_{\max}$ acquired by the particles in an accelerator having weak focusing, the required mass of the electromagnet increases approximately in proportion to $E_{\max}^3$.

24.4.12 Accelerators having *strong focusing* are resorted to for increasing the maximum energy $E_{\max}$ of particles accelerated in synchrotrons and proton synchrotrons. In these accelerators, two kinds of sector-shaped magnets are used alternately along the almost circular orbit of particle travel. In one kind of sector the magnetic field varies according to the law of Sect. 24.4.11, in which n is much less than zero (for example, $n = -100$), and in the other kind $n \gg 1$. Sectors of the first type provide for radial focusing of the beam of particles being accelerated; sectors of the second kind, for vertical focusing. The application of strong focusing enables substantial reductions to be obtained in the overall size of the accelerator, the mass of the electromagnet and the total cost of the installation. This method is also called alternating-gradient focusing.

24.4.13 To increase the share of energy utilized by accelerated particles for various nuclear reactions (42.9.1), the bombardment of stationary targets by high-energy particles is replaced

by the *colliding beam technique*. It follows from the laws of conservation of energy (3.4.3) and momentum (2.7.1) that in bombarding a stationary target, the share of the kinetic energy E_k of the incident particle, utilized in the nuclear reaction, decreases as E_k is increased. When the technique of colliding beams is used, the total momentum of the particles after collision is reduced and the share of their useful energy is increased. Assume, for example, that in an accelerator with colliding beams of protons the energy of the protons in each beam is equal to 26 GeV. The total momentum of two protons colliding at equal and opposing velocities equals zero. But the collision energy of such two particles reaches 50 GeV. To obtain the same collision energy in bombarding a stationary hydrogen target, a beam of protons with energy of the order of 1400 GeV is required.

CHAPTER 25   ELECTROMAGNETIC INDUCTION *

25.1 Basic Law of Electromagnetic Induction

25.1.1 The *phenomenon of electromagnetic induction* is the setting up of an *induced electric field* in a conducting loop, or circuit, located in a varying magnetic field. The energy measure of this induced field is the *electromotive force* $\mathscr{E}_1$ *of electromagnetic induction*. If the loop is closed, the induced electric field causes ordered motion of electrons in the loop, i.e. a current, called the *induced current*, is produced.

25.1.2 *Faraday's law of electromagnetic induction* states that the emf $\mathscr{E}_1$ of electromagnetic induction in a loop, or circuit, is proportional to the rate of change of the magnetic flux Φ_m (23.5.5) through the surface area bounded by this loop. Thus

$$\mathscr{E}_1 = -\frac{d\Phi_m}{dt} \text{ (in SI units)},$$

$$\mathscr{E}_1 = -\frac{1}{c}\frac{d\Phi_m}{dt} \text{ (in Gaussian units)}.$$

_____

* In all the formulas written in Gaussian units, c is the electrodynamic constant (23.2.2).

The cause of the change in the magnetic flux is unessential. It may be due to distortion of the loop or its motion in the external magnetic field or to any other reasons for the change in the magnetic field with time. When a closed conducting circuit (or loop) moves in a magnetic field, emf's $\mathscr{E}_1$ are induced in all of the parts of the loop that cut the lines of magnetic induction. The algebraic sum of these emf's (25.1.3) is equal to the total emf of the loop. The work that must be done to move a closed loop, or circuit, in a magnetic field is equal to the work of the current induced in the loop.

In calculating Φ_m and $\mathscr{E}_1$ the directions around the loop (21.3.2) and of the outward normal **n** (23.5.5 and 23.5.8) are to be coordinated so that from the head of vector **n** the direction around the circuit (or loop) is seen to be counterclockwise. If a closed circuit consists of N turns connected in series (for example, the solenoid in Sect. 23.3.7), then the magnetic flux Φ_m in Faraday's law is to be replaced by the flux, or magnetic, linkage Ψ of the circuit (23.6.5):

$$\mathscr{E}_1 = -\frac{d\Psi}{dt} \text{ (in SI units),}$$

$$\mathscr{E}_1 = -\frac{1}{c}\frac{d\Psi}{dt} \text{ (in Gaussian units).}$$

25.1.3 The emf of electromagnetic induction in a loop (or circuit) is said to be positive if the magnetic moment vector p_m of the corresponding induction current (23.3.4) makes an acute angle with the lines of magnetic induction of the field that induces this current. Otherwise, $\mathscr{E}_1$ is said to be negative. In Fig. 25.1a, the emf $\mathscr{E}_1 < 0$, whereas in Fig. 25.1b, $\mathscr{E}_1 > 0$.

25.1.4 The minus sign in the formula for $\mathscr{E}_1$ complies with the *Lenz law*, which states that upon any change in the magnetic flux through a surface bounded by a closed loop, the current induced in the loop is always of a direction in which the magnetic field set up by the induced current opposes the change in the magnetic flux that induced the current.

25.1.5 An emf of electromagnetic induction is developed in a part of a conductor which, in its motion, cuts lines of induction of a magnetic field (23.1.3). In the case illustrated in Fig. 25.2, the conduction electrons of the metal (20.3.1) are subject to the Lorentz force (24.1.1). Thus $F_L = -e\,[(v + v')\,B]$, where **v**

is the velocity of the part AC of the conductor in the magnetic field, whose induction vector **B** is perpendicular to the plane passing through the part of the conductor and the velocity

(a) $\dfrac{d\Phi_m}{dt} > 0$ (b) $\dfrac{d\Phi_m}{dt} < 0$

P_m

I_i

I_i

$\mathcal{E}_i < 0$

P_m $\mathcal{E}_i > 0$
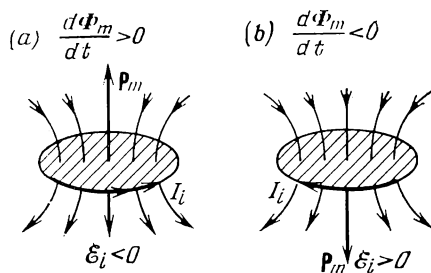
Fig. 25.1

vector of its motion. The electrons travel orderly along the conductor at the velocity v' by the action of the component of the Lorentz force tangent to the conductor in the direction from

B

C

l

O

v X

A

Z
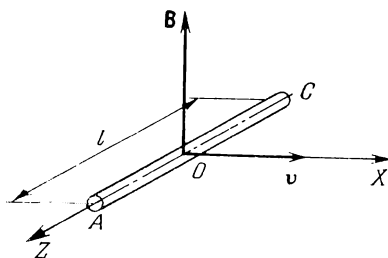
Fig. 25.2

A to C. The motion of the electrons ceases when the electrostatic field set up in the conductor and acting on the electrons with the force eE counterbalances the Lorentz force,

According to Ohm's law for an open circuit (at $I = 0$) (21.2.5), the equilibrium potential difference $\Delta V = V_A - V_C$ between points A and C that is established at $\mathbf{v}' = 0$ is equal to

$$\Delta V = V_A - V_C = -\mathscr{E}_1,$$

where $\mathscr{E}_1$ is the electromotive force of induction, because there are no sources of electrical energy in part AC of the conductor. The emf of electromagnetic induction in a piece of conductor of length l, moving at the velocity v, is

$$\mathscr{E}_1 = -Bvl,$$

or

$$\mathscr{E}_1 = -\frac{d\Phi_m}{dt} \text{ (in SI units)},$$

where $d\Phi_m/dt$ is the ratio of the magnetic flux through the surface generated by the conductor in its motion during an infinitesimal length of time to the value dt of this length of time or, in other words, the velocity at which the conductor cuts the lines of induction of the magnetic field (cf. Sect. 25.1.2). 25.1.6 The phenomenon of electromagnetic induction in stationary closed conducting circuits, located in an external varying magnetic field, cannot be explained by means of the Lorentz force, which has no effect on fixed charges (24.1.1). The phenomenon of electromagnetic induction in stationary conductors is due to the fact that the varying magnetic field excites a rotational electric field. The circulation of the strength vector of this field along the closed path L of the conductor (16.1.4) is the emf of electromagnetic induction. Thus

$$\mathscr{E}_1 = \oint\limits_L \mathbf{E} \, d\mathbf{l} = -\frac{\partial \Phi_m}{\partial t} \text{ (in SI units)},$$

$$\mathscr{E}_1 = \oint\limits_L \mathbf{E} \, d\mathbf{l} = -\frac{1}{c} \frac{\partial \Phi_m}{\partial t} \text{ (in Gaussian units)},$$

where the partial derivative $\partial \Phi_m/\partial t$ takes into account the dependence of the magnetic induction flux only on time. The choice of the direction of normal $\mathbf{n}$ in calculating the magnetic flux is discussed in Sect. 25.1.2.

25.1.7 The magnitude q of the electric charge passing through the cross section of the wire of the turn in which the current is induced is

$$q = -\frac{\Phi'_m - \Phi''_m}{R} \quad \text{(in SI units)},$$

$$q = \frac{1}{c}\frac{\Phi'_m - \Phi''_m}{R} \quad \text{(in Gaussian units)},$$

where Φ'_m and Φ''_m are the values of the magnetic flux through the area of the turn in its initial and final positions, and R is the electrical resistance of the turn.

25.2 Phenomenon of Self-Induction

25.2.1 *Self-induction* is the production of an induced electric field in a circuit as a result of the variations of current in the same circuit. The energy characteristic of the induced field is the emf of self-induction $\mathscr{E}_s$.

25.2.2 The intrinsic magnetic field of the current in the circuit produces the magnetic flux Φ_{ms} through the surface S bounded by the current-carrying loop. Thus

$$\Phi_{ms} = \int_S B_n\, dS,.$$

where B_n is the projection of the induction vector **B** of the magnetic field (23.1.2) on the normal **n** to the element of surface dS. The quantity Φ_{ms} is called the *magnetic flux of self-induction of the loop*. Making use of the Biot-Savart-Laplace law (23.2.2), Φ_{ms} can be calculated for the case in which the loop is in a nonferromagnetic medium (26.5.2):

$$\Phi_{ms} = I\,\frac{\mu_0}{4\pi}\int_S dS \oint_L \frac{\mu_r}{r^3}\,[d\mathbf{l}\,\mathbf{r}]_n \quad \text{(in SI units)},$$

$$\Phi_{ms} = \frac{1}{c}\,I\int_S dS \oint_L \frac{\mu_r}{r^3}\,[d\mathbf{l}\,\mathbf{r}]_n \quad \text{(in Gaussian units)},$$

or

$\Phi_{ms} = LI$ (in SI units),

$\Phi_{ms} = \dfrac{1}{c} LI$ (in Gaussian units),

where L is called the *self-inductance* of the loop and is determined by the formulas

$$L = \frac{\mu_0}{4\pi} \int\limits_{S} dS \oint\limits_{L} \frac{\mu_r}{r^3} [d\mathbf{l}\, \mathbf{r}]_n \text{ (in SI units)},$$

$$L = \int\limits_{S} dS \oint\limits_{L} \frac{\mu_r}{r^3} [d\mathbf{l}\, \mathbf{r}]_n \text{ (in Gaussian units)}.$$

Here μ_0 is the magnetic constant in SI units (23.2.2), μ_r is the relative magnetic permeability of the medium, $\mathbf{r}$ is the radius vector from element dl of the loop to element dS of surface S, bounded by the loop, and subscript n indicates the projection on the normal to the element dS.

25.2.3 The self-inductance L of the loop is numerically equal to the self-induction flux at a current of $I = 1$ A (in SI units) or $I = c$ (in Gaussian units). Self-inductance L depends only upon the geometric shape and dimensions of the loop, as well as on the magnetic properties of the medium it is located in. The self-inductance of a sufficiently long solenoid (23.4.8) is

$$L = \frac{\mu_0 \mu_r N^2 S}{l} = \mu_0 \mu_r n^2 V \text{ (in SI units)},$$

$$L = \frac{4\pi \mu_r N^2 S}{l} = 4\pi \mu_r n^2 V \text{ (in Gaussian units)}.$$

where l and S are the length and cross-sectional area of the solenoid, N is the total number of turns, $n = N/l$ is the number of turns per unit length, and $V = Sl$ is the volume of the solenoid.

25.2.4 The expression for the emf of self-induction follows from Faraday's law (25.1.2):

$$\mathscr{E}_s = -\frac{d\Phi_{ms}}{dt} = -\frac{d}{dt}(LI) \text{ (in SI units),}$$

$$\mathscr{E}_s = -\frac{1}{c}\frac{d\Phi_{ms}}{dt} = -\frac{1}{c^2}\frac{d}{dt}(LI) \text{ (in Gaussian units).}$$

For a rigid (undeformable) loop in a nonferromagnetic medium (26.5.2) $L = $ const. Hence

$$\mathscr{E}_s = -L\frac{dI}{dt} \text{ (in SI units),}$$

$$\mathscr{E}_s = -\frac{L}{c^2}\frac{dI}{dt} \text{ (in Gaussian units).}$$

25.2.5 The electromotive force of self-induction causes a *self-induction current* to flow in the loop. According to the Lenz
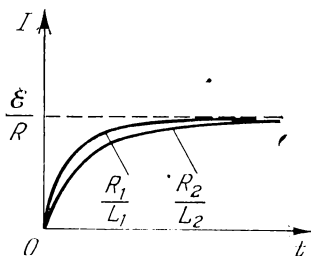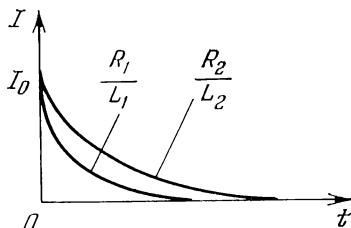
Fig. 25.3   Fig. 25.4

law (25.1.4), this current counteracts the change in the main current in the circuit, impeding its increase or decrease. The self-inductance of the loop (25.2.3) is a measure of its inertia to changes of the current.
25.2.6 The law for the change of the current in closing and opening a circuit, i.e. with a nonsteady state in the circuit, is

$$I = I_0 e^{-Rt/L} + \frac{\mathscr{E}}{R}(1 - e^{-Rt/L}),$$

where I_0 is the current at the initial instant of time ($t = 0$), R is the electrical resistance of the circuit, L is its self-inductance, and $\mathscr{E}$ is the algebraic sum of the emf's of the sources of electrical energy connected into the circuit (21.2.2).
When the source of emf is switched in there is no initial current in the circuit ($I_0 = 0$):

$$I = \frac{\mathscr{E}}{R}\left(1 - e^{-Rt/L}\right).$$

The current in the circuit increases from zero to the value $\mathscr{E}/R$, corresponding to that of the constant current. The higher the
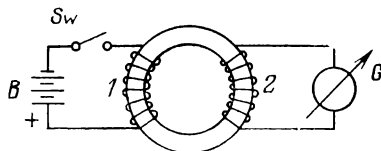
Fig. 25.5

ratio R/L (Fig. 25.3), the more rapidly the current increases. When the source of emf is switched off, $\mathscr{E} = 0$:

$$I = I_0 e^{-Rt/L}.$$

The current decreases exponentially from the initial value I_0. The higher the ratio R/L, the more rapidly the current decreases (Fig. 25.4).

25.3 Mutual Induction

25.3.1 An emf is induced in all conductors in the vicinity of a circuit in which the current varies. This phenomenon is called *mutual induction*. Upon a change in the current I_1 in circuit 1 (Fig. 25.5), for example, an emf $\mathscr{E}_{21}$ of mutual induction is induced in circuit 2. It equals

$$\mathscr{E}_{21} = -\frac{d\Phi_{m21}}{dt} \quad \text{(in SI units)},$$

$$\mathscr{E}_{21} = -\frac{1}{c}\frac{d\Phi_{m21}}{dt} \quad \text{(in Gaussian units)},$$

where Φ_{m21} is the magnetic flux through the surface bounded by the second circuit. This flux is due to the magnetic field set up by the current in the first circuit (*magnetic flux of mutual induction*).

25.3.2 By analogy with Sect. 25.2.2, the magnetic flux Φ_{m21} is proportional to the current I_1. Thus

$$\Phi_{m21} = M_{21}I_1 \text{ (in SI units)},$$

$$\Phi_{m21} = \frac{1}{c} M_{21}I_1 \text{ (in Gaussian units)},$$

where M_{21} is called the *mutual inductance* of the second and first circuits. The quantity M_{21} depends upon the shape, dimensions and relative arrangement of the circuits, as well as the relative magnetic permeability of the medium surrounding them. The magnetic flux Φ_{m12} through the surface bounded by the first circuit is due to the magnetic field set up by current I_2 in the second circuit. Thus

$$\Phi_{m12} = M_{12}I_2 \text{ (in SI units)},$$

$$\Phi_{m12} = \frac{1}{c} M_{12}I_2 \text{ (in Gaussian units)},$$

where M_{12} is the mutual inductance of the first and second circuits.

If the circuits are in a nonferromagnetic medium (26.5.2), then $M_{21} = M_{12}$. For ferromagnetic media M_{21} and M_{12} are not equal and depend, in addition to the above-mentioned factors, on the currents in the two circuits and the way in which they vary.

25.3.3 The expressions for the emf of mutual induction, under the condition that M_{21} and M_{12} are both constant values, are

$$\mathscr{E}_{21} = - M_{21} \frac{dI_1}{dt} \text{ (in SI units)},$$

$$\mathscr{E}_{21} = - \frac{1}{c^2} M_{21} \frac{dI_1}{dt} \text{ (in Gaussian units)},$$

$$\mathscr{E}_{12} = - M_{12} \frac{dI_2}{dt} \text{ (in SI units)},$$

$$\mathscr{E}_{12} = - \frac{1}{c^2} M_{12} \frac{dI_2}{dt} \text{ (in Gaussian units)}.$$

An example of the application of mutual induction is the prin-
ciple of the transformer which serves to raise or lower the volt-
age of alternating current.

25.4 Energy of a Magnetic Field Set up by an Electric Current *

25.4.1 To produce an electric current in a closed circuit (or loop)
and to increase this current from zero to I it is necessary to do
work W to overcome the emf of self-induction that impedes
the increase in current (25.2.6). Thus

$$W = \frac{\Phi_{ms}I}{2} = \frac{LI^2}{2} \quad \text{(in SI units)},$$

$$W = \frac{1}{c} \frac{\Phi_{ms}I}{2} = \frac{1}{c^2} \frac{LI^2}{2} \quad \text{(in Gaussian units)},$$

where Φ_{ms} is the magnetic flux of self-induction of the loop
(25.2.2) and L is the self-inductance of the loop (25.2.2).
According to the energy conservation law, W determines the
intrinsic energy E_m of current I in a circuit. Thus

$$E_m = \frac{LI^2}{2} \quad \text{(in SI units)},$$

$$E_m = \frac{1}{c^2} \frac{LI^2}{2} \quad \text{(in Gaussian units)}.$$

25.4.2 Increasing together with the current in a circuit is the
magnetic field set up by this current. Hence, the intrinsic energy
of a current is dealt with as *energy of* this *magnetic field*. In
a long solenoid (23.3.8), for instance, having a uniform magnet-
ic field (23.1.2),

$$E_m = \frac{1}{2} \mu_0 \mu_r n^2 I^2 V \quad \text{(in SI units)},$$

$$E_m = \frac{\mu_r}{c^2} 2\pi n^2 I^2 V \quad \text{(in Gaussian units)},$$

---

* It is assumed in this section that the conductors carrying
currents are in a nonferromagnetic, homogeneous and iso-
tropic medium.

where V is the volume of the solenoid, n is the number of turns per unit length, μ_0 is the magnetic constant (23.2.2), and μ_r is the relative permeability of the medium.

25.4.3 The *volume energy density* E_m^d of a magnetic field is the energy per unit volume of the field. Thus

$$E_m^d = \frac{dE_m}{dV}.$$

For a uniform magnetic field (23.1.3), $E_m^d = E_m/V$.
The volume energy density of any, including a nonuniform magnetic field, is expressed by the formula

$$E_m^d = \frac{1}{2}\,BH = \frac{1}{2}\,\mu_0\mu_r H^2 = \frac{1}{2}\,\frac{B^2}{\mu_0\mu_r} \quad \text{(in SI units),}$$

$$E_m^d = \frac{BH}{8\pi} = \frac{\mu_r H^2}{8\pi} = \frac{1}{8\pi}\,\frac{B^2}{\mu_r} \quad \text{(in Gaussian units).}$$

Here B and H are the magnitudes of the vectors of magnetic induction (23.1.2) and strength (23.2.3) at the point being considered in the magnetic field. The remaining notation is defined in Sect. 25.4.2.

25.4.4 The energy E_m of a magnetic field, localized in the volume V is

$$E_m = \int_V \frac{BH}{2}\,dV \quad \text{(in SI units),}$$

$$E_m = \int_V \frac{BH}{8\pi}\,dV \quad \text{(in Gaussian units).}$$

25.4.5 If the magnetic field is set up by an arbitrary system of n circuits, or loops, with the currents $I_1, I_2, I_3, \ldots, I_n$, the energy of the field can be expressed by the formula of Sect. 25.4.4, in which B and H are understood to be the magnitudes of vectors **B** and **H** of the resultant magnetic field, in accordance

with the principle of superposition of fields (15.2.2). Moreover, the energy of the magnetic field in this case is

$$E_m = \sum_{k=1}^{n} \frac{\Phi_{mk} I_k}{2},$$

where Φ_{mk} is the total magnetic flux linked to the kth loop (23.6.5). In calculating this flux, the normal n_k to the surface bounded by the loop is such that from the head of vector n_k the current is seen to be counterclockwise. The magnetic flux Φ_{mk} equals

$$\Phi_{mk} = \Phi_{msk} + \Phi_{mmk},$$

where Φ_{msk} is the magnetic flux of self-induction of the kth loop (25.2.2) and Φ_{mmk} is the magnetic flux of its mutual induction, produced by all the other loops carrying current (25.3.1). Hence the energy E_m of the magnetic field is equal to

$$E_m = \sum_{k=1}^{n} \frac{L_k I_k^2}{2} + \frac{1}{2} \sum_{k=1}^{n} \sum_{l=1(l \neq k)}^{n} M_{kl} I_k I_l \text{ (in SI units)},$$

$$E_m = \frac{1}{c^2} \sum_{k=1}^{n} \frac{L_k I_k^2}{2} + \frac{1}{2c^2} \sum_{k=1}^{n} \sum_{l=1(l \neq k)}^{n} M_{kl} I_k I_l \text{ (in Gaussian units)}.$$

The first term is the sum of the intrinsic energies of all the currents (25.4.1). The second term is called the *mutual energy of the currents*. Here M_{kl} is the mutual inductance of the kth and lth loops, or circuits, (25.3.2) with the currents I_k and I_l.

CHAPTER 26   MAGNETIC MATERIALS
IN A MAGNETIC FIELD

26.1 Magnetic Moments of Electrons and Atoms

26.1.1 *Magnetic materials* are all substances capable of acquiring magnetic properties, i.e. of being magnetized, in an external magnetic field. This means that they can set up their own magnetic field. The magnetic properties of substances are determined by the magnetic properties of their electrons and atoms.* Magnetic materials are divided according to their magnetic properties into three main groups: diamagnetic (26.3.2), paramagnetic (26.3.5) and ferromagnetic (26.5.1) materials.

26.1.2 The motion of an electron along its orbit in an atom is equivalent to a certain closed loop carrying a current (*orbital current*). According to Sect. 23.3.4, the *orbital magnetic moment* p_m of an electron equals

$$p_m = IS\mathbf{n} \quad \text{(in SI units)},$$

$$p_m = \frac{1}{c} IS\mathbf{n} \quad \text{(in Gaussian units)},$$

where $I = e\nu$ is the current, e is the magnitude of the charge of the electron, ν is the number of revolutions of the electron along its orbit in unit time, S is the area of the electron orbit, $\mathbf{n}$ is a unit vector of the normal to area S, and c is the electrodynamic constant (23.2.2).

An electron moving along its orbit has an orbital angular momentum $\mathbf{L_e}$ (4.1.4). The orbital magnetic moment is proportional to the orbital angular momentum:

$$p_m = \gamma \mathbf{L_e},$$

where $\gamma = -\dfrac{e}{2m}$ (in SI units),

$$\gamma = -\frac{e}{2mc} \quad \text{(in Gaussian units)}.$$

---

* The magnetic properties of atomic nuclei are dealt with in Sect. 42.1.6.

The quantity γ is called the *gyromagnetic ratio of orbital moments*.

In the last formulas m is the mass of the electron, e is the magnitude of its charge, and c is the electrodynamic constant (23.2.2).
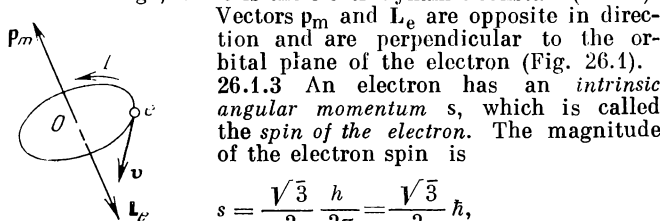
Vectors p_m and L_e are opposite in direction and are perpendicular to the orbital plane of the electron (Fig. 26.1).

26.1.3 An electron has an *intrinsic angular momentum* s, which is called the *spin of the electron*. The magnitude of the electron spin is

$$s = \frac{\sqrt{3}}{2}\,\frac{h}{2\pi} = \frac{\sqrt{3}}{2}\,\hbar,$$

Fig. 26.1

where h is Planck's constant (Appendix II), and $\hbar = h/2\pi$. The most important feature of electron spin is that it has only two projections on the direction of vector **B** of magnetic induction.* Thus

$$s_B = \pm\,\frac{\hbar}{2}\,.$$

26.1.4 Electron spin s corresponds to the *spin magnetic moment* p_{ms} which is proportional to the spin and opposite in direction. Thus

$p_{ms} = \gamma_s s.$

The quantity γ_s is called the *gyromagnetic ratio of spin moments*:

$$\gamma_s = -\,\frac{e}{m}\ \text{(in SI units),}$$

$$\gamma_s = -\,\frac{e}{mc}\ \text{(in Gaussian units).}$$

where the notation is the same as in Sect. 26.1.2.

---

* Regardless of whether it is an external magnetic field set up, for instance, by current-carrying conductors, or the internal magnetic field of the substance itself (26.4.2).

The projection of the electron spin magnetic moment on the direction of the field (26.1.3) equals

$$p_{ms_B} = \pm \frac{e\hbar}{2m} = \pm \mu_B \text{ (in SI units)},$$

$$p_{ms_B} = \pm \frac{e\hbar}{2mc} = \pm \mu_B \text{ (in Gaussian units)},$$

where μ_B is called the *Bohr magneton*, and is a unit for measuring magnetic moments (Appendix II).
26.1.5 The information presented in Sects. 26.1.1 through 26.1.4 is valid for each of the Z electrons in an atom. The number Z coincides with the atomic number in Mendeleev's periodic table (39.3.5).
The *orbital magnetic moment* $\mathbf{P}_m$ of an atom is the vector sum of the orbital magnetic moments $\mathbf{p}_m$ of all its electrons. Thus

$$\mathbf{P}_m = \sum_{i=1}^{Z} \mathbf{p}_{mi}.$$

The *orbital angular momentum* $\mathbf{L}$ of the atom is the vector sum of the orbital angular momenta $\mathbf{L}_e$ of all Z electrons. Thus

$$\mathbf{L} = \sum_{i=1}^{Z} \mathbf{L}_{ei}.$$

The atomic orbital magnetic moment $\mathbf{P}_m$ and angular momentum $\mathbf{L}$ have the following relationship:

$$\mathbf{P}_m = \gamma \mathbf{L},$$

where γ is the gyromagnetic ratio (26.1.2).

26.2 An Atom in a Magnetic Field

26.2.1 If a substance is in an external magnetic field, this field can be assumed uniform (23.1.3) within the limits of an atom. This follows from the smallness of atomic dimensions. Let us assume that an electron in the atom travels along a circular

orbit whose plane is perpendicular to the induction vector **B** of the magnetic field. The action of the Lorentz force F_L (24.1.1) reduces the force of attraction of the electron to the nucleus. The centripetal force (2.4.3) is found to equal the difference $F_e - F_L$, where F_e is the Coulomb force (14.2.2) of attraction of an electron to the nucleus (Fig. 26.2). This results in a change in the angular velocity ω (1.5.3) of the electron along its circular orbit.
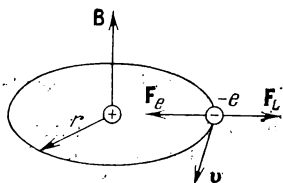
Fig. 26.2

26.2.2 The change in the angular velocity of the electron occurs in the process of growth of the magnetic field into which the atom is placed. This growth takes place during a finite length of time. At this an induced rotational electric field (25.1.1) is developed that acts on the electron in the atom. The strength **E** of this field is directed along a tangent to the electron's orbit and the force acting on the electron is $F = eE$ (15.1.2).

26.2.3 Upon an arbitrary position of the electron's orbit with respect to vector **B**, the orbital magnetic moment p_m of the electron (26.1.2) makes an angle α with the direction of the magnetic field (Fig. 26.3). In this case there will be a precessional motion of the orbit about the direction of vector **B** (4.3.2). This means that vector p_m, perpendicular to the orbit and maintaining its angle α with the field constant, rotates about the direction of vector **B** with the angular velocity ω_L. Thus

$$\omega_L = \frac{eB}{2m} \quad \text{(in SI units)},$$

$$\omega_L = \frac{eH}{2mc} \quad \text{(in Gaussian units)}.$$

Here e is the magnitude of the charge of the electron, m is its mass, H is the magnetic field strength, c is the electrodynamic constant (23.2.2), and ω_L is called the *angular velocity of Larmor precession*.

Larmor's theorem states that the only result of the effect of a magnetic field on an electron's orbit in an atom is the preces-

sion of the orbit and vector p_m with the angular velocity ω_L about an axis through the nucleus of the atom and parallel to the induction vector $\mathbf{B}$ of the magnetic field.

26.2.4 The precessional motion of the electron's orbit results in an additional orbital current ΔI_{orb} (Fig. 26.3) and the cor-
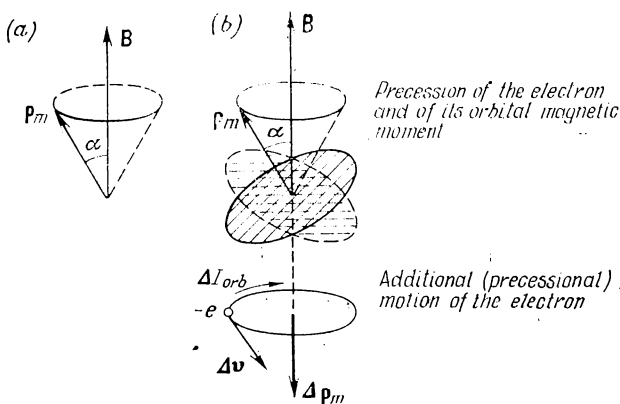
Fig. 26.3

responding induced orbital magnetic moment $\Delta \mathbf{p}_m$ (26.1.2) with the magnitude

$$\Delta p_m = \Delta I_{\text{orb}} S_\perp = \frac{e^2 S_\perp}{4\pi m} B \text{ (in SI units)},$$

$$\Delta p_m = \frac{1}{c} I_{\text{orb}} S_\perp = \frac{e^2 S_\perp}{4\pi m c^2} B \text{ (in Gaussian units)},$$

where $S_\perp$ is the area of the projection of the orbit on a plane perpendicular to the direction of vector $\mathbf{B}$. Vector $\Delta \mathbf{p}_m$ has

22*

the direction opposite to the magnetic induction vector **B**. Thus

$$\Delta \mathbf{p}_m = -\frac{e^2 S_\perp}{4\pi m}\,\mathbf{B}\quad\text{(in SI units)},$$

$$\Delta \mathbf{p}_m = -\frac{e^2 S_\perp}{4\pi m c^2}\,\mathbf{B}\quad\text{(in Gaussian units)}.$$

The notation is the same as in Sect. 26.2.3.
The total induced orbital moment $\Delta \mathbf{P}_m$ of the atom (26.1.5) is

$$\Delta \mathbf{P}_m = -\frac{e^2 Z \langle S_\perp \rangle}{4\pi m}\,\mathbf{B}\quad\text{(in SI units)},$$

$$\Delta \mathbf{P}_m = -\frac{e^2 Z \langle S_\perp \rangle}{4\pi m c^2}\,\mathbf{B}\quad\text{(in Gaussian units)},$$

where Z is the number of electrons in the atom, and $\langle S_\perp \rangle = \left(\sum\limits_{i=1}^{z} S_{\perp i}\right)/Z$ is the average area of the projections of the electron orbits on a plane perpendicular to the direction of vector **B**.

26.3 Diamagnetic and Paramagnetic Materials in a Uniform Magnetic Field

26.3.1 The magnetization of matter is characterized by the *magnetization vector (magnetization intensity)* **M**, which is the ratio of the magnetic moment of a small volume ΔV of the substance to the magnitude of this volume. Thus

$$\mathbf{M} = \frac{1}{\Delta V}\sum_{i=1}^{N}\mathbf{P}_{mi},$$

where $\mathbf{P}_{mi}$ is the magnetic moment of the ith atom (or molecule), and N is the total number of atoms (or molecules) in the small volume ΔV. Within the limits of the volume ΔV the magnetic field is assumed uniform. This volume should simultaneously

contain a sufficiently large number of particles N (i.e. $N \gg 1$) to enable the physical quantities characterizing the system of particles to be expediently averaged.

26.3.2 *Diamagnetic materials* are substances in which the magnetic moments of the atoms (or molecules) equal zero in the absence of an external magnetic field. This means that in diamagnetic materials the vector sum of the orbital magnetic moments of all the electrons in an atom equals zero (26.1.5) and that induced magnetic moments (26.2.4) exist only in a magnetic field.

26.3.3 In the volume ΔV of an isotropic diamagnetic material, the induced magnetic moments $\Delta \mathbf{P}_m$ of all the atoms (or molecules) are the same and are opposite in direction to vector $\mathbf{B}$ (26.2.4).

The magnetization vector $\mathbf{M}$ is equal to

$$\mathbf{M} = \frac{N \Delta \mathbf{P}_m}{\Delta V} = n_0 \Delta \mathbf{P}_m,$$

and, taking Sect. 26.2.4 into account:

$$\mathbf{M} = -\frac{n_0 e^2 Z \langle S_\perp \rangle}{4\pi m} \mathbf{B} \text{ (in SI units)},$$

$$\mathbf{M} = -\frac{n_0 e^2 Z \langle S_\perp \rangle}{4\pi m c^2} \mathbf{B} \text{ (in Gaussian units)},$$

where n_0 is the number of atoms (or molecules) in unit volume. The other notation is the same as in Sect. 26.2.4. Introducing the notation:

$$\chi'_m = -\frac{n_0 e^2 Z \langle S_\perp \rangle \mu_0}{4\pi m} \text{ (in SI units)},$$

$$\chi'_m = -\frac{n_0 e^2 Z \langle S_\perp \rangle}{4\pi m c^2} \text{ (in Gaussian units)},$$

we have

$$\mathbf{M} = \chi'_m \frac{\mathbf{B}}{\mu_0} \text{ (in SI units)},$$

$$\mathbf{M} = \chi'_m \mathbf{B} \text{ (in Gaussian units)},$$

where χ'_m is a dimensionless quantity that characterizes the magnetic properties of magnetic materials. For all diamagnetic materials $\chi'_m < 0$.

26.3.4 The *magnetic susceptibility* χ'_m is a quantity that is related to χ'_m as follows:

$$1 + \chi_m = \frac{1}{1 - \chi'_m} \quad \text{(in SI units),}$$

$$1 + 4\pi\chi_m = \frac{1}{1 - 4\pi\chi'_m} \quad \text{(in Gaussian units),}$$

from which

$$\chi_m = \frac{\chi'_m}{1 - \chi'_m} \quad \text{(in SI units),}$$

$$\chi_m = \frac{\chi'_m}{1 - 4\pi\chi'_m} \quad \text{(in Gaussian units).}$$

Practically, $\chi_m = \chi'_m$ for diamagnetic materials, because in absolute value χ'_m is very small: $|\chi'_m| \approx 10^{-6}$.

26.3.5 *Paramagnetic materials* are substances in which the atoms (or molecules) have a certain constant magnetic moment $\mathbf{P_m}$ in the absence of an external magnetic field. This means that the vector sum of the orbital magnetic moments of all the electrons of an atom (or molecule) is not equal to zero (26.1.5).

26.3.6 When a paramagnetic material is placed into a uniform magnetic field (23.1.3), the constant magnetic moments of the atoms (or molecules) precess about the direction of the induction vector **B** of the magnetic field at Larmor's angular velocity ω_L (26.2.3).

Thermal motion and collisions of the atoms (or molecules) of a paramagnetic material lead to a gradual damping of the precession of the magnetic moments and to a reduction in the angles between the directions of the magnetic moment vectors and vector **B**. The combined actions of the interatomic collisions and the magnetic field lead to a predominant orientation of the magnetic moments of the atoms in alignment with the external

field. Though the constant magnetic moment $\mathbf{P}_m$ of an atom (or molecule) is of the order of 10^{-23} joule/tesla (10^{-20} erg/gauss), the magnetic moments of all the particles in unit volume produce a magnetization that considerably exceeds diamagnetic phenomena (26.3.3). A paramagnetic material in an external magnetic field has an intrinsic magnetic field aligned with the external magnetic field.

26.3.7 In the classical theory of paramagnetism the magnitude of the magnetization vector (26.3.1) is expressed by the formula

$$M = n_0 P_m L\,(a),$$

where n_0 is the number of atoms (or molecules) in unit volume, and $L\,(a)$ is the classical Langevin function:

$$L\,(a) = \frac{e^a + e^{-a}}{e^a - e^{-a}} - \frac{1}{a}\,.$$

The parameter a is of the form $a = P_m B/kT$, where B is the magnetic induction of the field, k is Boltzmann's constant (8.4.5), and T is the absolute temperature. At room temperature and not especially strong external fields, $a \ll 1$ and the function $L\,(a)$ is simplified, after expansion into a series, and we obtain $L\,(a) \approx a/3$. Hence, the magnetization vector is equal to

$$\mathbf{M} = \chi_m' \frac{\mathbf{B}}{\mu_0} \quad \text{(in SI units)},$$

$$\mathbf{M} = \chi_m' \mathbf{B} \quad \text{(in Gaussian units)},$$

where χ_m' is determined by the formula

$$\chi_m' = \frac{n_0 P_m^2 \mu_0}{3kT} \quad \text{(in SI units)},$$

$$\chi_m' = \frac{n_0 P_m^2}{3kT} \quad \text{(in Gaussian units)}.$$

The quantity χ_m' is related to the magnetic susceptibility χ_m of paramagnetic materials by the formulas given in Sect. 26.3.4. The value of the quantity χ_m' is positive for a paramagnetic material and ranges from 10^{-5} to 10^{-3}; therefore, $\chi_m' = \chi_m$ with a high degree of accuracy.

The *Curie law* states that the paramagnetic susceptibility of a substance is inversely proportional to the absolute temperature.

26.3.8 In very strong external magnetic fields, *saturation magnetization* is reached; here $a \gg 1$ and the Langevin function $L(a) \to 1$. This means that the magnetic moments of all the atoms (or molecules) are aligned with the external magnetic field and $M = n_0 P_m$.

26.4 Magnetic Field in Magnetic Materials

26.4.1 Two types of currents that set up magnetic fields, macrocurrents and microcurrents, can be distinguished in matter. What we have called *macrocurrents* here are conduction currents (20.1.2) and convection currents (20.1.2). *Microcurrents (molecular currents)* are those due to the motion of electrons in atoms, molecules and ions.

The magnetic field in a substance is the vector sum of two fields (15.2.2): the *external* magnetic field, set up by the macrocurrents, and the *internal*, or *intrinsic*, *magnetic field*, set up by microcurrents. The magnetic induction vector **B** (23.1.2) in the magnetic field in a substance characterizes the resultant magnetic field and is equal to the vector sum of the magnetic inductions of the external $\mathbf{B_0}$ and intrinsic $\mathbf{B_{intr}}$ magnetic fields. Thus

$$\mathbf{B} = \mathbf{B_0} + \mathbf{B_{intr}}.$$

The primary source of magnetic fields in magnetic materials are the macrocurrents. Their magnetic fields are what leads to the magnetization of substances placed in an external magnetic field.

26.4.2 The *total current law for a magnetic field in a substance* is a generalization of the law formulated in Sect. 23.5.2:

$$\oint_L \mathbf{B} \, dl = \mu_0 \left(I_{mac} + I_{mic} \right) \quad \text{(in SI units),}$$

$$\oint_L \mathbf{B} \, dl = \frac{4\pi}{c} \left(I_{mac} + I_{mic} \right) \quad \text{(in Gaussian units)},$$

where I_{mac} and I_{mic} are the algebraic sums of the macro- and microcurrents passing through a surface bounded by the loop L.
26.4.3 The algebraic sum of the microcurrents is related to the magnetization vector $\mathbf{M}$ as follows:

$$I_{mic} = \oint_L \mathbf{M} \, d\mathbf{l} \quad \text{(in SI units)},$$

$$I_{mic} = c \oint_L \mathbf{M} \, d\mathbf{l} \quad \text{(in Gaussian units)},$$

where $\oint \mathbf{M} d\mathbf{l}$ is the circulation of the magnetization vector $\mathbf{M}$ (26.3.1) along the closed loop L through which the microcurrents are threaded.

The final form of the total current law (26.4.2) is

$$\oint_L \left(\frac{\mathbf{B}}{\mu_0} - \mathbf{M} \right) d\mathbf{l} = I_{mac} \quad \text{(in SI units)},$$

$$\oint_L (\mathbf{B} - 4\pi\mathbf{M}) \, d\mathbf{l} = \frac{4\pi}{c} I_{mac} \quad \text{(in Gaussian units)}.$$

26.4.4. The vector

$$\mathbf{H} = \frac{\mathbf{B}}{\mu_0} - \mathbf{M} \quad \text{(in SI units)},$$

$$\mathbf{H} = \mathbf{B} - 4\pi\mathbf{M} \quad \text{(in Gaussian units)}$$

is called the *strength of the magnetic field* existing in an arbitrary medium (cf. Sect. 23.2.3). The total current law for a magnetic field in an arbitrary medium is written in a form identical to that in Sect. 23.5.4:

$$\oint_L \mathbf{H} \, d\mathbf{l} = I_{mac} \quad \text{(in SI units)},$$

$$\oint_L \mathbf{H} \, d\mathbf{l} = \frac{4\pi}{c} I_{mac} \quad \text{(in Gaussian units)}.$$

The circulation of the magnetic field strength vector along an arbitrary closed loop is equal (or proportional) to the algebraic sum of the macrocurrents through the surface bounded by this loop.

26.4.5 For an isotropic medium, the relation between the magnetic induction vector **B** and the magnetization vector **M** (26.3.1) leads to the following result for vector **H** (26.4.4):

$$\mathbf{H} = (1 - \chi'_m) \frac{\mathbf{B}}{\mu_0} \quad \text{(in SI units),}$$

$$\mathbf{H} = (1 - 4\pi\chi'_m) \mathbf{B} \quad \text{(in Gaussian units).}$$

Substituting for $(1 - \chi'_m)$ or $(1 - 4\pi\chi'_m)$ on the basis of Sect. 26.3.4, we have

$$\mathbf{H} = \frac{\mathbf{B}}{\mu_0\mu_r} \quad \text{(in SI units),}$$

$$\mathbf{H} = \frac{\mathbf{B}}{\mu_r} \quad \text{(in Gaussian units),}$$

where

$$\mu_r = 1 + \chi_m \quad \text{(in SI units),}$$

$$\mu_r = 1 + 4\pi\chi_m \quad \text{(in Gaussian units).}$$

The quantity μ_r thus introduced is called the *relative magnetic permeability of the substance* (see also Sect. 23.2.1). In these formulas χ_m is the magnetic susceptibility (26.3.4).

26.5 Ferromagnetic Materials

26.5.1 *Ferromagnetic materials* are substances in which the intrinsic (internal) magnetic field (26.4.1) is hundreds or even thousands of times stronger than the external magnetic field applied to the substance.

Ferromagnetism is observed in the crystals of the transition metals (39.3.8), iron, cobalt and nickel and of certain alloys, provided that $d/a \geqslant 1.5$, where d is the diameter of the atom

and a is the diameter of the unfilled electronic shell of the atom (39.3.6).

26.5.2 The principal properties of ferromagnetic materials, distinguishing them from other magnetic materials are:

(a) The dependence of the magnetization (26.3.1) on the strength H of the external magnetic field is characterized by *magnetic saturation* M_S, which begins at $H \geqslant H_S$ (Fig. 26.4).

(b) The dependence of the magnetic induction **B** on **H** is characterized by the increase according to a linear law at $H \geqslant H_S$ (Fig. 26.5).
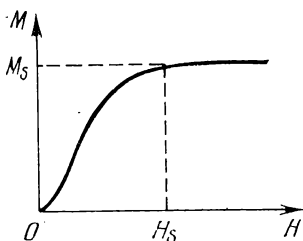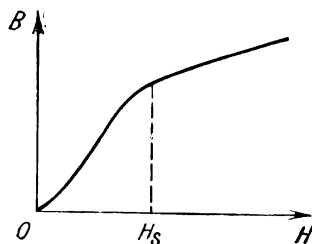
Fig. 26.4   Fig. 26.5

(c) The dependence of the relative magnetic permeability μ_r on the field strength H is of a complex nature (Fig. 26.6).

(d) *Magnetic hysteresis* exists in ferromagnetic materials. This refers to the lag in the variation of magnetization from the variation in the strength of the external magnetizing field, which is variable in magnitude and direction. This lag is due to the dependence of M on its previous values (past history) in magnetizing the substance.

(e) The properties of ferromagnetic materials enumerated above* are observed at temperatures below the *Curie point* θ_C (18.4.4). At temperatures $T \geqslant \theta_C$, thermal motion breaks up the regions of spontaneous magnetization (26.5.4) and the

---

* As well as many others that are beyond the scope of this handbook.

ferromagnetic material, losing its special features, is converted into a paramagnetic substance (26.3.5). The Curie points for iron, nickel, cobalt and Permalloy are 1043, 631, 1403 and 823 K.

26.5.3 The *hysteresis loop* is a curve showing the change in the magnetization of a ferromagnetic material located in an external magnetic field as the strength of this field is varied from $+H_s$ to $-H_s$ and back again, where H_s is the field strength corresponding to magnetic saturation (Fig. 26.7). The magnetization $\pm M_s$ at $H = \pm H_s$ is called *saturation magnetization*. The magnetization $\pm M_r$ that remains in a ferromagnetic material
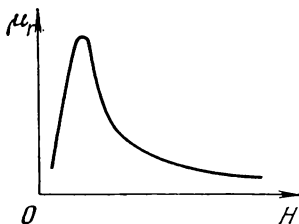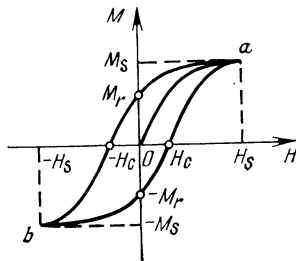
Fig. 26.6   Fig. 26.7

in the absence of an external field ($H = 0$) is called the *residual*, or *remanent*, *magnetization*. This factor is the basis for making permanent magnets. The strength $\pm H_c$ of the external field that completely demagnetizes a substance is called the *coercive force*. The coercive force specifies the property of a ferromagnetic material to retain its residual magnetization. *Hard*, or *retentive*, *magnetic materials* have a high coercive force. They have a wide hysteresis loop and are used to make permanent magnets. *Soft*, or *non-retentive*, *magnetic materials* have a low coercive force. They have a narrow hysteresis loop and are used to make transformer cores.

The reversal of magnetization is associated with a change in the orientation of the regions of spontaneous magnetization (26.5.4) and requires that work be done at the expense of the

energy of the external magnetic field. The amount of heat evolved
in reversing magnetization is proportional to the area of the
hysteresis loop.

26.5.4 At temperatures below the Curie point (26.5.2) a ferro-
magnetic material is made up of *domains*, which are small *re-
gions of* uniform *spontaneous magnetization*. The linear dimen-
sions of these domains range from 10^{-5} to 10^{-4} m. Within each
domain the substance is magnetized to saturation (26.5.2).
In the absence of an external magnetic field, the magnetic mo-
ments of the various domains are oriented in space so that the
resultant magnetic moment of the whole ferromagnetic body
equals zero.

Owing to the effect of the external magnetic field, the magnetic
moments of whole domains are oriented in a ferromagnetic
material, rather than those of the separate particles as in the
case of paramagnetic materials (26.3.6). As a result, the sub-
stance is magnetized.

Ferromagnetic properties can only exist in the crystalline state
of substances in which the interaction between the neighbouring
atoms of the crystal lattice leads to the total energy of the sys-
tem of electrons that provides for compliance with the condi-
tions for the existence of ferromagnetism (26.5.1).

26.5.5 Measurement of the gyromagnetic ratio (26.1.4) for
ferromagnetic bodies indicates that the elementary carriers
of magnetism in ferromagnetic materials are the spin magnetic
moments of the electrons (26.1.4). The modern quantum mechan-
ical theory of ferromagnetism explains the nature of the spon-
taneous magnetization of ferromagnetic materials (26.5.4) and
the reasons for the development of a strong intrinsic magnetic
field (26.4.1).

Ferromagnetic properties are possessed by crystals of substances
whose atoms have inner shells and subshells (39.3.6) unfilled
by electrons and having a nonzero value of the projections of
the resultant spin moment on the direction of the magnetic
field (26.1.3). Developed between the spins of such electrons
is a specific kind of quantum mechanical interaction that is
not of a magnetic origin. It is called exchange interaction
(39.4.5). As a result, the state of the system of electrons in fer-
romagnetic materials is found to be stable upon parallel align-
ment of the spins, and spontaneous magnetization (26.5.4)
is developed along with a strong intrinsic magnetic field.

CHAPTER 27 FUNDAMENTALS OF MAXWELL'S THEORY

27.1 General Features of Maxwell's Theory

27.1.1 *Maxwell's theory* is a systematic theory of a unified electromagnetic field (15.1.2) set up by an arbitrary system of charges and currents. Maxwell's theory enables the *principal problem of electrodynamics* to be solved: determining the characteristics of electric and magnetic fields on the basis of a given distribution of charges and currents that set up these fields. Maxwell's theory is a generalization of fundamental laws that describe electrical and electromagnetic phenomena. These include the Ostrogradsky-Gauss theorem (18.3.3), the total current law (26.4.2), and the law of electromagnetic induction (25.1.2).

27.1.2 Maxwell's theory is of a phenomenological nature. This is manifested in the fact that it does not deal with the internal mechanism of the phenomena that occur in a medium and lead to the production of electric and magnetic fields. The medium is described in this theory by three quantities that specify its electrical and magnetic properties. These are the relative permittivity (dielectric constant) ε_r (18.3.4), relative magnetic permeability μ_r (26.4.5) and the electrical conductivity $\varkappa$ (20.3.4).

27.1.3. Maxwell's theory deals with macroscopic fields set up by macroscopic charges and currents that are concentrated in volumes incommensurably greater than those of atoms (or molecules). It is assumed that the distances from the sources of the fields to the points in space being considered substantially exceed the linear dimensions of atoms (or molecules). Therefore, macroscopic fields change appreciably only at distances that are huge compared to the linear dimensions of atoms (or molecules). Moreover, the periods of time required for changes in variable electric and magnetic fields are considered to be considerably longer than those required for intramolecular processes.

27.1.4 The macroscopic charges and currents are the sets of microscopic charges and currents that set up their microfields (electric and magnetic), varying continuously with time at each point in space.

The macroscopic fields dealt with in Maxwell's theory are *averaged microfields.* Averaging is carried out over time intervals considerably longer than those typical of interatomic processes, and over volumes of fields many times greater than the volumes of atoms and molecules (27.4.5).

27.1.5 Maxwell's theory is one of *short-range action,* according to which electrical and magnetic interactions take place in electric and magnetic fields and propagate at a finite velocity equal to that of light in the given medium. This important consequence is taken into account in the electromagnetic theory of light developed by Maxwell.

27.2 Maxwell's First Equation

27.2.1 *Maxwell's first equation in integral form* is a generalization of Faraday's law of electromagnetic induction in the form (25.1.2):

$$\oint_L \mathbf{E}\,d\mathbf{l} = -\frac{\partial \Phi_m}{\partial t} \quad \text{(in SI units),}$$

$$\oint_L \mathbf{E}\,d\mathbf{l} = -\frac{1}{c}\frac{\partial \Phi_m}{\partial t} \quad \text{(in Gaussian units).}$$

According to Maxwell, this law is valid for any closed circuit, or loop, and not only current-conducting ones, conceivably selected in a variable magnetic field. This means that a variable magnetic field sets up a rotational electric field at any point in space, regardless of whether or not there is a conductor at that point.

27.2.2 Using the following expression for the magnetic flux (23.5.5):

$$\Phi_m = \oint_S \mathbf{B}\,d\mathbf{S} = \oint_S B_n\,dS$$

and the Stokes theorem from vector analysis:

$$\oint_S \mathbf{E}\,d\mathbf{l} = \oint_S \text{curl}\,\mathbf{E}\,d\mathbf{S},$$

where $d\mathbf{S} = \mathbf{n}\, dS$ ($\mathbf{n}$ being the unit vector of the normal to the element dS of surface), we can convert Maxwell's first equation as given in Sect. 27.2.1 to *Maxwell's first equation in differential form*:

$$\text{curl } \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \quad \text{(in SI units)},$$

$$\text{curl } \mathbf{E} = -\frac{1}{c}\frac{\partial \mathbf{B}}{\partial t} \quad \text{(in Gaussian units)}.$$

Here curl $\mathbf{E}$ is expressed in Cartesian coordinates by the determinant

$$\text{curl } \mathbf{E} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ E_x & E_y & E_z \end{vmatrix}.$$
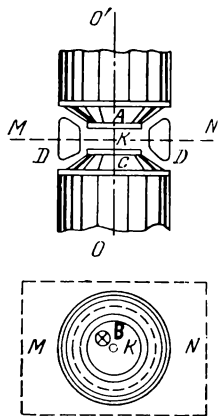
27.2.3 An inductive type of electron accelerator, the *betatron* (24.4.2), is based on setting up a rotational electric field in space by the action of an alternating magnetic field.

In the alternating magnetic field produced by an electromagnet with conical pole pieces A and C (Fig. 27.1), a rotational electric field is set up in accelerating chamber D, having the form of a closed ring. The lines of force (15.1.6) of the rotational electric field lie in plane MN, which is perpendicular to the axis OO'

Fig. 27.1

of symmetry of the pole pieces. These lines are circles with their centre at point K. At all points of any circle the field strength vector $\mathbf{E}$ has a constant numerical value and is tangent to the circle. Electrons in the accelerating chamber travel along circular paths and after repeated travel along the stable circular orbit the electron is accelerated to considerable energy.

27.2.4 The strength **E** of the rotational electric field in a betatron is equal to

$$E = - \frac{1}{2} r \frac{d \langle B \rangle}{dt} \, ,$$

where $\langle B \rangle$ is the average value of magnetic induction at the instant of time t within the limits of the circular electron orbit of radius r.
A condition for a stable orbit of the electron in the betatron is: $B = \langle B \rangle / 2$, where B is the instantaneous value of the magnetic induction on the orbit.
The orbit of an electron in the betatron is stable if:
(a) The whole orbit lies in a single plane. This condition, called *axial focusing*, is achieved due to the special shape of the pole pieces of the electromagnet. This shape provides for gradual weakening of the magnetic field induction in the direction from the centre of the orbit to its periphery.
(b) Provision is made to return electrons that have accidentally left the stable orbit back to it again. This is called *radial focusing* and is achieved by a space distribution of the magnetic field in which the magnetic induction decreases more slowly from the axis to the periphery than $1/r$, where r is the distance from a given point in the field to the axis of symmetry OO'.

27.3 Displacement Current. Maxwell's Second Equation

27.3.1 Maxwell generalized the total current law, (26.4.2) and (26.4.4), by assuming that a variable electric field, in the same way as an electric current, is a source of a magnetic field. The quantitative measure of the magnetic effect of a variable electric field is the displacement current.
27.3.2 The *displacement current density* (20.2.3) is

$$j_{dis} = \frac{\partial \mathbf{D}}{\partial t} \quad \text{(in SI units)},$$

$$j_{dis} = \frac{1}{4\pi} \frac{\partial \mathbf{D}}{\partial t} \quad \text{(in Gaussian units)}.$$

23—0186

The displacement current through the arbitrary surface S is a physical quantity numerically equal to the flux of the displacement current density vector through this surface. Thus

$$I_{\text{dis}} = \int_S \mathbf{j}_{\text{dis}} \, d\mathbf{S} = \int_S \frac{\partial \mathbf{D}}{\partial t} \, d\mathbf{S} = \frac{\partial \Phi_e}{\partial t} \quad \text{(in SI units)},$$

$$I_{\text{dis}} = \int_S \mathbf{j}_{\text{dis}} \, d\mathbf{S} = \int_S \frac{1}{4\pi} \frac{\partial \mathbf{D}}{\partial t} \, d\mathbf{S} = \frac{1}{4\pi} \frac{\partial \Phi_e}{\partial t} \quad \begin{array}{l}\text{(in Gaussian} \\ \text{units)},\end{array}$$

where $\Phi_e = \int_S \mathbf{D} \, d\mathbf{S}$ is the flux of the electric displacement vector through the surface S (15.3.2).

Taking the displacement currents into account, any variable current has a closed circuit. Displacement currents "flow" through a part of a circuit where there is no conductor, for example, between the plates of a capacitor being charged or discharged. Shown in Fig. 27.2 are the vectors $\mathbf{j}_{\text{dis}}$ and the lines of induction of the magnetic fields of displacement currents in charging (Fig. 27.2a) and in discharging (Fig. 27.2b) a capacitor.

27.3.3 According to Sect. 18.3.4, the displacement vector in any dielectric is equal to

$$\mathbf{D} = \varepsilon_0 \mathbf{E} + \mathbf{P}_e \quad \text{(in SI units)},$$

$$\mathbf{D} = \mathbf{E} + 4\pi \mathbf{P}_e \quad \text{(in cgse units)},$$

where $\mathbf{P}_e$ is the polarization vector (18.2.3).

The displacement current density in a dielectric is

$$\mathbf{j}_{\text{dis}} = \varepsilon_0 \frac{\partial \mathbf{E}}{\partial t} + \frac{\partial \mathbf{P}_e}{\partial t} \quad \text{(in SI units)},$$

$$\mathbf{j}_{\text{dis}} = \frac{1}{4\pi} \frac{\partial \mathbf{E}}{\partial t} + \frac{\partial \mathbf{P}_e}{\partial t} \quad \text{(in cgse units)}.$$

In these last formulas, the first term $\varepsilon_0 \dfrac{\partial \mathbf{E}}{\partial t}$ (or, correspondingly, $\dfrac{1}{4\pi} \dfrac{\partial \mathbf{E}}{\partial t}$) is called the *displacement current density in vacuum*, whereas the second term $\partial \mathbf{P}_e / \partial t$ is called the *polarization current density*. The second term is the density of the current due to
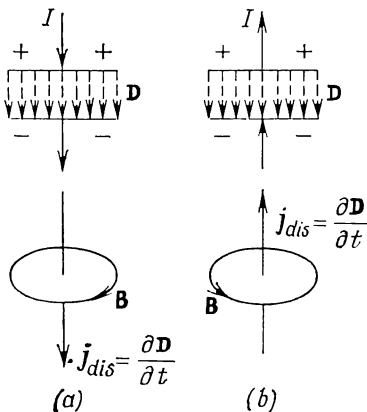
Fig. 27.2

the ordered motion of charges in the dielectric: either the displacement of charges in a molecule of a nonpolar dielectric (18.1.3) or the turning of the dipoles in polar dielectrics (18.1.5). The displacement current does not evolve Joule heat (21.2.6) in vacuum or in metals. In this it differs from conduction currents.

27.3.4 Maxwell added the displacement current to the right-hand side of the total current law in the form given in Sect. 26.4.4, writing this law in the form

$$\oint_L \mathbf{H}\, d\mathbf{l} = I_{\mathrm{mac}} + I_{\mathrm{dis}} \quad \text{(in SI units)},$$

$$\oint_L \mathbf{H}\, dl = \frac{4\pi}{c}\, (I_{\mathrm{mac}} + I_{\mathrm{dis}}) \quad \text{(in Gaussian units)}.$$

This equation is called *Maxwell's second equation in integral form*. It shows that the circulation of the magnetic field strength vector along an arbitrary closed circuit, or loop, L is equal to the algebraic sum of the macrocurrents and the displacement current through a surface bounded by this loop.

27.3.5 Making use of the Stokes theorem from vector analysis

$$\oint_L \mathbf{H}\, dl = \oint_S \operatorname{curl}\ \mathbf{H}\, d\mathbf{S}, \text{ where } d\mathbf{S} = \mathbf{n}\, dS, \text{ and } \mathbf{n} \text{ is a unit}$$

vector of the normal to the element dS of surface, and of the expression for the total current

$$I = I_{\mathrm{mac}} + I_{\mathrm{dis}} = \int_S (\mathbf{j} + \mathbf{j}_{\mathrm{dis}})\, dS,$$

Maxwell's second equation can be written *in differential form*:

$$\operatorname{curl} \mathbf{H} = \mathbf{j} + \frac{\partial \mathbf{D}}{\partial t} \quad \text{(in SI units)},$$

$$\operatorname{curl}\ \mathbf{H} = \frac{4\pi}{c}\ \mathbf{j} + \frac{1}{c}\ \frac{\partial \mathbf{D}}{\partial t} \quad \text{(in Gaussian units)},$$

In these equations, curl H has the same meaning as curl E in Sect. 27.2.2.

27.3.6 In the absence of conduction currents ($\mathbf{j} = 0$), Maxwell's first and second equations are of symmetrical form except for sign in their right-hand sides:

$$\operatorname{curl} \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}\ ,\ \operatorname{curl} \mathbf{H} = \frac{\partial \mathbf{D}}{\partial t} \quad \text{(in SI units)},$$

$$\operatorname{curl} \mathbf{E} = -\frac{1}{c}\ \frac{\partial \mathbf{B}}{\partial t}\ ,\ \operatorname{curl}\ \mathbf{H} = \frac{1}{c}\ \frac{\partial \mathbf{D}}{\partial t} \quad \text{(in Gaussian units)}.$$

Comparing Maxwell's equations*, we reach the following conclusions:

---

* The numeration of Maxwell's equations is a conditional one and is frequently the opposite of that accepted in this handbook.

(a) Electric and magnetic fields are interrelated: a change of an electric field with time sets up a magnetic field*. In its turn, a variable magnetic field is the source of a rotational electric field.

(b) The different signs at the right-hand sides of Maxwell's equations comply with the law of conservation of energy and with the Lenz law (25.1.4). If the signs of $\partial B/\partial t$ and $\partial D/\partial t$ were the same, an infinitesimal increase in one of the fields would lead to an increase without limit of both fields, whereas an infinitesimal decrease in one of the fields would lead to the disappearance of both fields. The above-mentioned difference in the signs of the right-hand sides of Maxwell's equations is a necessary condition for the existence of a stable electromagnetic field.
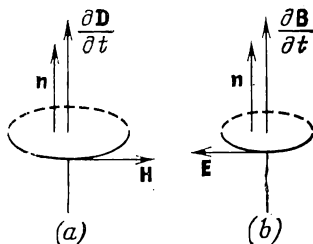
Fig. 27.3

The different signs in the right-hand sides of Maxwell's equations correspond to the fact that the directions of $\partial D/\partial t$ and H form a "right-hand screw" system (Fig. 27.3a), whereas the directions of $\partial B/\partial t$ and E form a "left-hand screw" system (Fig. 27.3b).

27.4 Complete Set of Maxwell's Equations for an Electromagnetic Field

27.4.1 A complete set of Maxwell's equations for an electromagnetic field includes, in addition to the equations dealt with in Sects. 27.2.1, 27.2.2, 27.3.4 and 27.3.5, the Ostrogradsky-Gauss theorem for an electric field (18.3.3):

$$\oint_S D\, dS = q_{\text{free}} \quad \text{(in SI units)},$$

* The magnetic field is always a rotational one (23.5.3).

$$\oint_S \mathbf{D}\, d\mathbf{S} = 4\pi q_{\text{free}} \quad \text{(in cgse units)}$$

and the same theorem for a magnetic field (23.5.6):

$$\oint_S \mathbf{B}\, d\mathbf{S} = 0.$$

Maxwell proposed that the theorem for the flux of an electric field displacement vector is valid, not only for a steady-state electrostatic field, but for a variable electric field as well.

27.4.2 Making use of the Gauss theorem from vector analysis:

$$\oint_S \mathbf{A}\, d\mathbf{S} = \int_V \operatorname{div} \mathbf{A}\, dV, \text{ and introducing the volume density } \rho$$

of free charges: $\rho = dq_{\text{free}}/dV$ (where dV is an element of volume), it is possible to obtain *Maxwell's third equation in differential form*:

div $\mathbf{D} = \rho$ (in SI units),

div $\mathbf{D} = 4\pi\rho$ (in Gaussian units).

In these equations, div $\mathbf{A}$ (where $\mathbf{A}$ is an arbitrary vector) is determined in Cartesian coordinates as follows:

$$\operatorname{div} \mathbf{A} = \frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z},$$

where $\mathbf{A} = A_x\mathbf{i} + A_y\mathbf{j} + A_z\mathbf{k}$, and $\mathbf{i}$, $\mathbf{j}$, and $\mathbf{k}$ are unit vectors along the coordinate axes.

27.4.3 The *complete set of Maxwell's equations* includes four equations:

curl $\mathbf{E} = -\dfrac{\partial \mathbf{B}}{\partial t}$,   div $\mathbf{D} = \rho$

                       (in SI units),

curl $\mathbf{H} = \mathbf{j} + \dfrac{\partial \mathbf{D}}{\partial t}$,   div $\mathbf{B} = 0$

$$\text{curl } \mathbf{E} = -\frac{1}{c}\frac{\partial \mathbf{B}}{\partial t}, \quad \text{div } \mathbf{D} = 4\pi\rho$$

(in Gaussian units).

$$\text{curl } \mathbf{H} = \frac{4\pi}{c}\mathbf{j} + \frac{1}{c}\frac{\partial \mathbf{D}}{\partial t}, \quad \text{div } \mathbf{B} = 0.$$

27.4.4 The system of Maxwell's equations is supplemented by equations that specify the electric and magnetic properties of the medium. For macrocurrents in an isotropic medium, which comply with Ohm's law (20.3.4), these equations are of the form

$$\mathbf{D} = \varepsilon_0\varepsilon_r\mathbf{E}, \quad \mathbf{B} = \mu_0\mu_r\mathbf{H}, \quad \mathbf{j}_{mac} = \varkappa\mathbf{E} \quad \text{(in SI units)},$$

$$\mathbf{D} = \varepsilon_r\mathbf{E}, \quad \mathbf{B} = \mu_r\mathbf{H}, \quad \mathbf{j}_{mac} = \varkappa\mathbf{E} \quad \text{(in Gaussian units)}.$$

Here ε_0 and μ_0 are the electric constant (14.2.7) and the magnetic constant (23.2.2) in SI units, ε_r and μ_r are the relative dielectric constant and relative magnetic permeability, and $\varkappa$ is the electrical conductivity.

To solve a system of Maxwell's equations, it is also necessary to specify the boundary conditions for the vectors characterizing the electromagnetic field. Thus

$$D_{n_1} - D_{n_2} = \sigma, \qquad E_{t_1} = E_{t_2},$$

(in SI units),

$$B_{n_1} = B_{n_2}, \qquad H_{t_1} - H_{t_2} = j_{sur}$$

$$D_{n_1} - D_{n_2} = 4\pi\sigma, \qquad E_{t_1} = E_{t_2},$$

(in Gaussian units),

$$B_{n_1} = B_{n_2}, \qquad H_{t_1} - H_{t_2} = \frac{4\pi}{c}j_{sur}$$

where σ is the surface density of free charges, $\mathbf{n}$ is a unit vector of the normal to the interface (boundary) between media and extends from medium 2 to medium 1, $\mathbf{t}$ is a unit vector of the tangent to the interface, and j_{sur} is the projection of the density vector of surface conduction currents on the direction $[\mathbf{tn}]$. For given boundary and initial conditions, i.e. known values of vectors $\mathbf{E}$ and $\mathbf{H}$ at the initial instant of time $(t = 0)$, the system of Maxwell's equations has a unique solution.

27.4.5 Maxwell's equations are invariant with respect to the Lorentz transformations (5.3.2).

In the special theory of relativity (5.1.1) it is shown that a common electromagnetic field can be different in different inertial frames of reference (2.1.2). In particular, one of the fields, either electric or magnetic, may be absent in one of the coordinate systems and be present in the other. The following formulas are used for the Lorentz transformations of the projections on the coordinate axes of the vectors **E**, **H**, **D** and **B** of the electric and magnetic fields, in going over from fixed inertial frame of reference K to frame K', which travels with respect to frame K at the constant velocity V in a straight line along the OX axis:

$$E_x = E'_x, \quad E_y = \frac{E'_y + V B'_z}{\sqrt{1 - V^2/c^2}}, \quad E_z = \frac{E'_z - V B'_y}{\sqrt{1 - V^2/c^2}},$$

(in SI units),

$$H_x = H'_x, \quad H_y = \frac{H'_y - V D'_z}{\sqrt{1 - V^2/c^2}}, \quad H_z = \frac{H'_z + V D'_y}{\sqrt{1 - V^2/c^2}},$$

$$D_x = D'_x, \quad D_y = \frac{D'_y + \dfrac{V}{c^2} H'_z}{\sqrt{1 - V^2/c^2}}, \quad D_z = \frac{D'_z - \dfrac{V}{c^2} H'_y}{\sqrt{1 - V^2/c^2}},$$

$$B_x = B'_x, \quad B_y = \frac{B'_y - \dfrac{V}{c^2} E'_z}{\sqrt{1 - V^2/c^2}}, \quad B_z = \frac{B'_z + \dfrac{V}{c^2} E'_y}{\sqrt{1 - V^2/c^2}}.$$

(in Gaussian units).

$$E_x = E'_x, \quad E_y = \frac{E'_y + \dfrac{V}{c} H'_z}{\sqrt{1 - V^2/c^2}}, \quad E_z = \frac{E'_z - \dfrac{V}{c} H'_y}{\sqrt{1 - V^2/c^2}},$$

$$H_x = H'_x, \quad H_y = \frac{H'_y - \dfrac{V}{c} E'_z}{\sqrt{1 - V^2/c^2}}, \quad H_z = \frac{H'_z + \dfrac{V}{c} E'_y}{\sqrt{1 - V^2/c^2}}.$$

27.4.6 The classical Lorentz theory of the electron is a further development of Maxwell's theory of the electromagnetic field.

The Lorentz theory is based on a definite model that implies certain conceptions of the structure of matter. It is assumed that atoms consist of negatively and positively charged particles, and that the great variety of electrical and magnetic phenomena can be explained by a definite spatial arrangement, motion and interaction of charges and microcurrents. At each point in space certain electric and magnetic microfields exist with the strengths e and h. These fields result from the combined action of all the charges and microcurrents. The microfields comply with a system of equations similar to Maxwell's equations (27.4.3). Maxwell's equations for macroscopic fields E and B (27.1.3) can be obtained by averaging the equations of the electron theory (27.1.4), and they are found to be the averaged microfields e and h. Thus

$$E = \langle e \rangle, \ \ B = \mu_0 \langle h \rangle \qquad \text{(in SI units)}.$$

Vectors D and H are found to be related to $\langle e \rangle$ and $\langle h \rangle$ by the polarization vector P_e (18.2.3) and the magnetization vector M (26.3.1) as indicated in Sects. 18.3.4 and 26.4.4. Thus

$$D = \varepsilon_0 E + P_e, \ \ \ H = \frac{B}{\mu_0} - M \ \ \ \text{(in SI units)},$$

$$D = E + 4\pi P_e, \ \ \ H = B - 4\pi M \ \ \ \text{(in Gaussian units)}.$$

PART FOUR OSCILLATIONS AND WAVES

CHAPTER 28 FREE HARMONIC OSCILLATIONS

28.1 Harmonic Oscillations

28.1.1 *Oscillations*, including vibrations and variations, are processes (motions or changes of state) that are repeated more or less regularly with time. According to the physical nature of the oscillatory process and the "mechanism" by which it is initiated, distinction is made between *mechanical oscillations*, commonly called vibrations (swinging of pendulums; vibration of strings; reciprocation and oscillation of machine components; vibration of buildings, bridges and other structures; compression and rarefaction of air as sound is propagated through it; rolling, pitching and heaving of ships; sea waves; etc.), *electromagnetic oscillations* (oscillations of alternating electric current in a circuit, variations in the field strength and magnetic induction vectors, **E** and **B**, of an alternating electromagnetic field, etc.), and *electromechanical vibrations* (vibration of the diaphragm of a telephone receiver, of the diaphragm in an electromagnetic loudspeaker, etc.) and others.

The system executing the oscillation (or vibration) is called the *oscillatory (vibratory) system*. *Free (natural) oscillations* are ones that occur in the absence of externally applied variable action on the system. They are due to some initial deviation of the system from a state of stable equilibrium. *Forced oscillations (vibrations)* are ones set up in a system as a result of variable external effects (for instance, current oscillation in an electric circuit caused by a varying emf, swinging of a pendulum due to varying external forces, etc.).

28.1.2 Oscillations (vibrations) are said to be *periodic* if all the physical quantities that characterize the oscillatory (vibratory) system and vary during its oscillation are repeated in successive time intervals. The minimum time interval T com-

plying with this condition is called the *period of oscillation* (or vibration). This is the time required for one *complete oscillation* (vibration) of the system. The *frequency* of periodic *oscillations* (vibrations) is the quantity $v = 1/T$, equal to the number of complete oscillations (vibrations) executed by the system in unit time. The angular, or circular, frequency is the quantity $\omega = 2\pi v = 2\pi/T$, which is equal to the number of complete oscillations accomplished in 2π units of time.

28.1.3 In periodic oscillations the dependence of a varying quantity s on the time t complies with the condition $s\,(t + T) = = s\,(t)$.

The periodic oscillation of the quantity $s\,(t)$ is said to be *harmonic* if

$$s\,(t) = A \sin (\omega t + \varphi_0) \quad \text{or} \quad s\,(t) = A \cos (\omega t + \varphi_1),$$

where $\omega = 2\pi v = (2\pi/T) = $ const is the *angular*, or *circular*, *frequency of harmonic vibration*, $A = s_{\max} = $ const > 0 is the maximum value of the varying quantity s and is called the *amplitude of oscillation*, and φ_0 and $\varphi_1 = \varphi_0 - \pi/2$ are constant quantities. The value of s at some arbitrary instant of time t is determined by the *phase of the oscillation* $\Phi\,(t) = = \omega t + \varphi_0$ [and, respectively, $\Phi_1\,(t) = \omega t + \varphi_1$]. The quantities φ_0 and φ_1 are the *initial phases of the oscillation*, i.e. the values of $\Phi\,(t)$ and $\Phi_1\,(t)$ at the instant $(t = 0)$ of time reference: $\varphi_0 = \Phi\,(0)$ and $\varphi_1 = \Phi_1\,(0)$.

28.1.4 The first and second time derivatives of a harmonically varying quantity $s\,(t)$ also vary harmonically at the same angular frequency:

$$\frac{ds}{dt} = A\omega \cos (\omega t + \varphi_0) = A\omega \sin \left(\omega t + \varphi_0 + \frac{\pi}{2} \right),$$

$$\frac{d^2s}{dt^2} = - A\omega^2 \sin (\omega t + \varphi_0) = A\omega^2 \sin (\omega t + \varphi_0 + \pi),$$

where the amplitudes of ds/dt and d^2s/dt^2 are equal, respectively, to $A\omega$ and $A\omega^2$. The initial phase of ds/dt is equal to $(\varphi_0 + \pi/2)$, i.e. the phase difference of the variations of ds/dt and s is constant and equal to $\pi/2$ (the quantity ds/dt has a phase lead of $\pi/2$ with respect to s). The initial phase of d^2s/dt^2 is equal to $(\varphi_0 + \pi)$, i.e. the phase difference of the variations of d^2s/dt^2

and s is constant and equal to π (the quantity d^2s/dt^2 has a phase lead of π with respect to s). Curves of the dependence on the time t of the quantities s, ds/dt and d^2s/dt^2 for harmonic variations and with $\varphi_0 = 0$ are shown in Fig. 28.1.
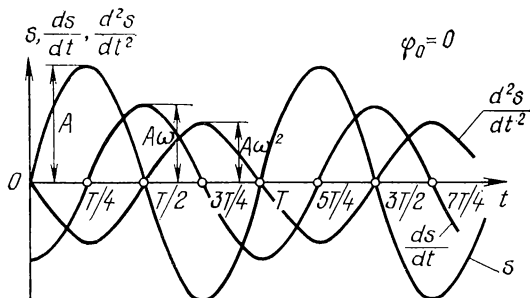
Fig. 28.1

28.1.5 It is evident from the second equation in Sect. 28.1.4 that a harmonically varying quantity s satisfies the differential equation

$$\frac{d^2s}{dt^2} + \omega^2 s = 0.$$

The general solution of this equation is of the form

$$s = A_1 \sin \omega t + A_2 \cos \omega t,$$

where A_1 and A_2 are arbitrary constants of integration. The values of A_1 and A_2 can be found from the initial conditions, i.e. when the values of s and ds/dt are known for the initial instant of time $(t = 0)$. Thus

$$A_1 = \frac{1}{\omega} \left(\frac{ds}{dt} \right)_{t=0} \text{ and } A_2 = s(0).$$

The general solution can be transformed to the standard form for harmonic oscillations (vibrations) given in Sect. 28.1.3:

$$s = A \sin (\omega t + \varphi_0),$$

where $A = \sqrt{A_1^2 + A_2^2}$ and $\varphi_0 = \text{arc tan} (A_2/A_1)$.
Thus, the quantity s has harmonic variations if and only if it satisfies the differential equation given at the beginning of this subsection. This equation is called the *differential equation of harmonic variations* (or vibrations or oscillations).

28.1.6 Harmonic oscillations (vibrations) can be represented graphically by means of a vector in a plane. For this purpose vector **A** is drawn from the origin of coordinates O in the plane XY (Fig. 28.2). The magnitude of vector **A** is equal to the amplitude A of the oscillation being considered and the vector makes the angle $\varphi = \omega t + \varphi_0$, equal to the phase of the oscilla-

Fig. 28.2

tion at the given instant of time t, with the coordinate axis OX. In the course of time angle φ increases so that vector **A** rotates at a constant angular velocity about point O equal to the angular frequency of oscillation (vibration) ω. The corresponding projection of vector **A** on the vertical axis OY varies harmonically according to the relation

$$A_y = s = A \sin (\omega t + \varphi_0).$$

The graphical representation of harmonic vibrations (or oscillations or variations) by means of a rotating amplitude vector is called the *vector diagram method*. It is extensively applied, for instance, in adding together harmonic oscillations having the same direction (28.4.2).

28.1.7 According to the Euler formula for complex numbers,

$$e^{i\varphi} = \cos \varphi + i \sin \varphi,$$

where $i = \sqrt{-1}$. Hence the harmonic oscillation (vibration) $s = A \sin (\omega t + \varphi_0) = A \cos (\omega t + \varphi_1)$, where $\varphi_1 = \varphi_0 - \pi/2$,
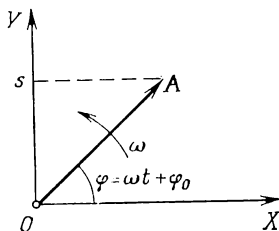
can be written in the exponential form

$$\widetilde{s} = \widetilde{A}e^{i\omega t} = Ae^{i(\omega t + \varphi_1)},$$

where $\widetilde{A} = Ae^{i\varphi_1}$ is the complex amplitude. Only the real part of the complex function $\widetilde{s}$, denoted by Re $\{\widetilde{s}\}$, has a physical meaning:

$$\text{Re } \{\widetilde{s}\} = s = A \cos (\omega t + \varphi_1) = A \sin (\omega t + \varphi_0),$$

where $\varphi_0 = \varphi_1 + \pi/2$.

28.2 Mechanical Harmonic Vibrations

28.2.1 If a particle vibrates harmonically in a straight line along the coordinate axis OX about the equilibrium position, taken as the origin of coordinates, the dependence of the coordinate x of the particle on the time t is of the form given in Sect. 28.1.3 (with $s = x$). Thus

$$x = A \sin (\omega t + \varphi_0).$$

The projections of the velocity **v** and acceleration **a** of the particle on the OX axis are:

$$v_x = v_0 \cos (\omega t + \varphi_0) \quad \text{and} \quad a_x = - a_0 \sin (\omega t + \varphi_0),$$

where $v_0 = A\omega$ is the amplitude of the velocity, and $a_0 = A\omega^2 = v_0\omega$ is the amplitude of the acceleration.
The force **F** acting on the particle equals

$$\mathbf{F} = m\mathbf{a} \quad \text{and} \quad F_x = - m\omega^2 x,$$

where m is the mass of the particle. Consequently, force **F** is proportional to the displacement of the particle from the equilibrium position and opposite in direction:

$$\mathbf{F} = - m\omega^2 x\mathbf{i},$$

where **i** is a unit vector of the OX axis.
Such a dependence of the force on the displacement is typical for an elastic force (3.3.6). Hence, forces of other physical origin, satisfying the same kind of relationship, are said to be *quasi-elastic forces*.

28.2.2 The *kinetic energy* of a particle vibrating harmonically in a straight line is equal to

$$E_k = \frac{mv^2}{2} = \frac{mv_0^2}{2}\cos^2(\omega t + \varphi_0) = \frac{m\omega^2 A^2}{2}\cos^2(\omega t + \varphi_0),$$

or

$$E_k = \frac{m\omega^2 A^2}{4}[1 + \cos(2\omega t + 2\varphi_0)].$$

The kinetic energy of the particle varies periodically from 0 to $m\omega^2 A^2/2$ as it vibrates harmonically with the angular frequency 2ω and the amplitude $m\omega^2 A^2/4$ about its average value of $m\omega^2 A^2/4$.

The *potential energy* of a particle vibrating harmonically in a straight line due to the effect of a quasi-elastic force is equal to

$$E_p = -\int_0^x F_x \, dx = \frac{m\omega^2 x^2}{2} = \frac{m\omega^2 A^2}{2}\sin^2(\omega t + \varphi_0),$$

or

$$E_p = \frac{m\omega^2 A^2}{4}[1 - \cos(2\omega t + 2\varphi_0)]$$

$$= \frac{m\omega^2 A^2}{4}[1 + \cos(2\omega t + 2\varphi_0 + \pi)].$$

The potential energy of the particle varies periodically from 0 to $m\omega^2 A^2/2$ as it vibrates harmonically with the angular frequency 2ω and the amplitude $m\omega^2 A^2/4$ about its average value of $m\omega^2 A^2/4$. The variations of the potential and kinetic energies have a phase difference of π, so that the total mechanical energy of the particle remains constant in vibration:

$$E = E_k + E_p = \frac{m\omega^2 A^2}{2} = \text{const}.$$

Curves of the dependence of E_k, E_p and E on the time t for the case when $\varphi_0 = 0$ are shown in Fig. 28.3.

28.2.3 Example 1. A *linear harmonic oscillator* is a particle of mass m having harmonic vibration along a straight line by the action of elastic force $\mathbf{F}_{el} = -kx\mathbf{i}$ (3.3.6). Serving as an example of such a system is the *spring pendulum*, where a weight of mass m is suspended by an absolutely elastic spring (k is a
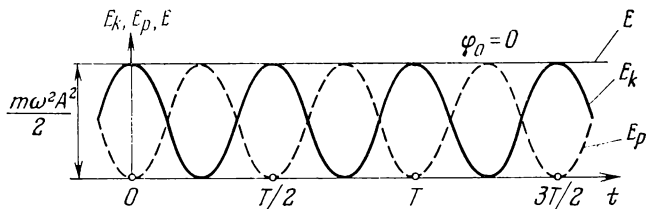
Fig. 28.3

factor specifying the elastic properties of the spring). The motion equation is

$$m\frac{d^2x}{dt^2} = -kx \quad \text{or} \quad \frac{d^2x}{dt^2} + \frac{k}{m}\,x = 0.$$

It follows from Sect. 28.1.5 that an oscillator (spring pendulum) has harmonic vibration according to the equation $x = = A\sin(\omega t + \varphi_0)$ with an angular frequency ω and period T equal to

$$\omega = \sqrt{\frac{k}{m}} \quad \text{and} \quad T = 2\pi\sqrt{\frac{m}{k}}.$$

The potential energy of a linear harmonic oscillator is $E_p = \frac{kx^2}{2}$.

28.2.4 Example 2. A *compound*, or *physical*, *pendulum* is a rigid body capable of oscillating under the action of the gravity force mg about a horizontal axis O, which does not pass through the centre of gravity of the body, and is called the *axis*

of oscillation, or *swing* (Fig. 28.4). The centre of gravity of the pendulum coincides with its centre of mass C (7.3.2). Point O, the intersection of the axis of oscillation with a vertical plane passing through the centre of gravity of the pendulum and perpendicular to its axis of oscillation, is called its *point of suspension*.

In the absence of friction forces in the suspension, the motion equation of the pendulum is of the form (4.3.4)

$$I \frac{d^2\alpha}{dt^2} = - mgd \sin \alpha,$$

where α is the angle of deviation of the pendulum about its axis of oscillation from the equilibrium position, $d = \overline{OC}$ is the distance from the centre of mass of the pendulum to its axis of

Fig. 28.4

oscillation, I is the moment of inertia (4.2.1) about the same axis, m is the mass of the pendulum, and g is the free fall acceleration. Upon small oscillations of the pendulum $\sin \alpha \approx \alpha$ and the motion equation of the pendulum is of the form

$$\frac{d^2\alpha}{dt^2} + \frac{mgd}{I} \alpha = 0,$$

i.e. angle α satisfies the differential equation of harmonic vibrations (28.1.5). Hence, in the absence of friction, small oscillations of a compound pendulum are harmonic:

$$\alpha = \alpha_0 \sin (\omega t + \varphi_0),$$

where α_0 is the amplitude of variation of angle α, and

$$\omega = \sqrt{\frac{mgd}{I}} \quad \text{and} \quad T = 2\pi \sqrt{\frac{I}{mgd}}$$

are the angular frequency and period of small oscillations of a compound pendulum.

24—0186
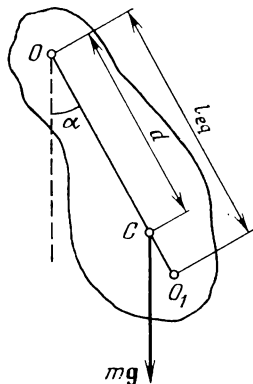
28.2.5 Example 3. A *simple pendulum* is a particle suspended by a weightless, inextensible thread and oscillating in a vertical plane by the action of the gravity force. The simple pendulum is the limiting case of the compound (physical) pendulum, whose whole mass is concentrated at the centre of mass, so that $d = l$, the length of the simple pendulum. The moment of inertia of a simple pendulum with respect to its axis of oscillation is $J = ml^2$. Correspondingly, the angular frequency and period of small oscillations of this pendulum are

$$\omega = \sqrt{\frac{g}{l}} \quad \text{and} \quad T = 2\pi \sqrt{\frac{l}{g}}.$$

Small oscillations of compound and simple pendulums are examples of *isochronous vibrations*, i.e. vibrations whose frequency and period are independent of their amplitude.

In the general case, the period of a compound pendulum depends upon its amplitude α_0. Thus

$$T = 2\pi \sqrt{\frac{I}{mgd}} \left[1 + \left(\frac{1}{2} \right)^2 \sin^2 \frac{\alpha_0}{2} \right.$$
$$\left. + \left(\frac{1}{2} \cdot \frac{3}{4} \right)^2 \sin^4 \frac{\alpha_0}{2} + \dots \right].$$

The variation in the value of T when α_0 is increased to 15^c does not exceed 0.5 per cent.

28.2.6 The *equivalent length* l_{eq} *of a compound pendulum* is the length of a simple pendulum that has the same period of oscillation. Thus

$$l_{eq} = \frac{I}{md} = d + \frac{I_C}{md} > d,$$

where I_C is the moment of inertia of the compound pendulum with respect to an axis passing through the centre of mass C of the pendulum and parallel to its axis of oscillation. Point O_1, lying on the straight line OC at the distance of l_{eq} from the point of suspension O (Fig. 28.4), is called the *centre of oscillation of the compound pendulum*. The centre of oscillation O_1 and the point of suspension O are interchangeable: if the pendulum is suspended so that its axis of oscillation passes through

point O_1, then point O will coincide with the new position of the centre of oscillation, i.e. the equivalent length and period of oscillation of the pendulum remain unchanged.

28.2.7 Example 4. *Small free vibrations of electrons in a plasma* (22.6.1) are called *Langmuir oscillations of the plasma.* They are caused by the forces of an electric field, which is set up in an electrically neutral plasma as the result of some random deviation of the spatial distribution of the electrons from their equilibrium distribution. Assume that in a flat layer of plasma of thickness l (Fig. 28.5) the electrons are displaced over the small distance s along the positive direction of axis OX. This leads to the formation of a surplus positive charge at the left-hand part of the layer and a surplus negative charge at the

Fig. 28.5

right-hand part. Consequently, an electric field is set up, whose strength **E** is along the OX axis and its projection on this axis is $E_x = n_0 es/\varepsilon_0$, where n_0 is the electron concentration in the plasma, e is the magnitude of the charge of the electron, and ε_0 is the electric constant. According to Newton's second law, the motion equation for the electrons of the plasma in this electric field is of the form:
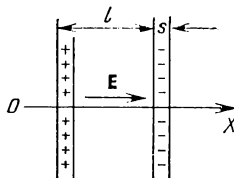
$$m \frac{d^2 s}{dt^2} = -eE_x = -\frac{e^2 n_0}{\varepsilon_0} s \quad \text{or} \quad \frac{d^2 s}{dt^2} + \frac{e^2 n_0}{m\varepsilon_0} s = 0,$$

where m is the mass of the electron. Thus, the electrons of the plasma have free harmonic vibrations at the angular frequency

$$\omega = e \sqrt{\frac{n_0}{m\varepsilon_0}},$$

which is called the *Langmuir plasma frequency*, or simply, the *plasma frequency.*

24*

28.3 Free Harmonic Oscillations in an Oscillatory Electric Circuit

28.3.1 An example of an electric circuit in which free electric oscillations may occur is the simple *oscillatory circuit* (Fig. 28.6). It consists of a capacitor of capacitance C connected in series to an induction coil with the self-inductance L. When the previously charged capacitor is connected to the coil, free oscillations of the capacitor charge and the current in the coil are produced. This sets up a variable electromagnetic field that propagates through space at a velocity equal to that of light. Hence, if the linear dimensions l of the circuit are not too large ($l \ll c/\nu$, where $c = 3 \times \times 10^8$ m/s is the velocity of light in vacuum, and ν is the frequency of oscillation in the circuit), it can be assumed that at each instant of time t the current I is the same at all parts of the circuit. Such a variable current is said to be *quasi-stationary*. According to Ohm's law (21.2.3) for the portion *1-L-2* of the circuit (Fig. 28.6)
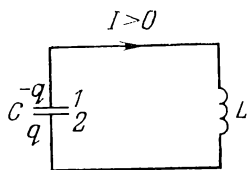
Fig. 28.6

$$ IR = V_1 - V_2 + \mathcal{E}_s \quad \text{or} \quad IR = -\frac{q}{C} - L\frac{dI}{dt} \, . $$

Here q and $V_1 - V_2 = -q/C$ are the charge of the capacitor and the potential difference of its plates at the arbitrary instant t of time being considered, R is the electric resistance of the oscillatory circuit, i.e. of the part *1-L-2* of the circuit, and $\mathcal{E}_s = -L \ (dI/dt)$ is emf of self-induction in the coil (25.2.4). If follows from the law of conservation of electric charge (14.1.3) that the quasi-stationary current in the circuit is $I = dq/dt$. Therefore, the differential equation for the variation of charge q is of the form

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

28.3.2 Free electric oscillations in an oscillatory circuit are *harmonic* if its electrical resistance $R = 0$. Thus

$$\frac{d^2q}{dt^2} + \frac{1}{LC}\, q = 0.$$

The angular frequency ω and the period T of these oscillations satisfy the *Thomson formula*:

$$\omega = \frac{1}{\sqrt{LC}} \quad \text{and} \quad T = 2\pi \sqrt{LC}\,.$$

The charge q of the capacitor and current I in the circuit vary according to the equations

$$q = q_0 \sin(\omega t + \varphi_0) \text{ and}$$

$$I = I_0 \cos(\omega t + \varphi_0) = I_0 \sin\left(\omega t + \varphi_0 + \frac{\pi}{2}\right),$$

where q_0 is the amplitude of the capacitor charge, $I_0 = \omega q_0 = q_0/\sqrt{LC}$ is the amplitude of the current, and φ_0 is the initial phase of capacitor charge oscillation. The current in the circuit has a phase lead of $\pi/2$ over the capacitor charge. The potential difference $u = V_2 - V_1$ of the capacitor plates also varies according to a harmonic law and is in phase with charge q. Thus

$$u = \frac{q}{C} = U_0 \sin(\omega t + \varphi_0),$$

where $U_0 = q_0/C$ is the amplitude of the potential difference. The amplitude of the current is

$$I_0 = U_0 \sqrt{\frac{C}{L}}\,.$$

The quantity $\sqrt{L/C}$ is called the *characteristic*, or *wave*, *impedance of the oscillatory circuit*.
28.3.3 With free harmonic oscillations in an oscillatory circuit, there is a periodic transfer of the energy E_e of the electric field

of the capacitor into energy E_m of the magnetic field of the inductance coil and back again. Thus

$$E_e = \frac{q^2}{2C} = \frac{q_0^2}{2C} \sin^2 (\omega t + \varphi_0) = \frac{q_0^2}{4C} [1 - \cos (2\omega t + 2\varphi_0)],$$

$$E_m = \frac{LI^2}{2} = \frac{LI_0^2}{2} \cos^2 (\omega t + \varphi_0) = \frac{LI_0^2}{4} [1 + \cos (2\omega t + 2\varphi_0)].$$

For this reason, the oscillations in an electrical oscillatory circuit are frequently said to be *electromagnetic oscillations in the circuit*.
The values of E_e and E_m vary in electromagnetic oscillations within the limits from 0 to their maximum values, equal, respectively, to $q_0^2/2C$ and $LI_0^2/2$, in which $q_0^2/2C = LI_0^2/2$. The variations of E_e and E_m are displaced in phase. At the instants of time when $E_e = 0$, $E_m = E_{m\,max} = LI_0^2/2$. On the other hand, when $E_m = 0$, $E_e = E_{e\,max} = q_0^2/2C$. The total energy of electromagnetic oscillations is constant with time. Thus

$$E = E_e + E_m = \frac{q_0^2}{2C} = \frac{LI_0^2}{2} = \text{const.}$$

28.4 Adding Harmonic Oscillations

28.4.1 The *addition of oscillations* (vibrations or variations) consists in finding the equation for the resulting oscillation of a system when this system simultaneously participates in several oscillatory processes. Distinction is made between two limiting cases: the combining of oscillations having the same direction and the combining of mutually perpendicular oscillations. The first case corresponds, for example, to the vibration of weight *1* (Fig. 28.7) which vibrates with respect to weight *2* on spring *a*, and together with weight *2* on spring *b*. This same situation occurs in superposing the variations of scalar physical characteristics of an oscillatory or variable system (pressure, temperature, density, electric charge, current, etc.).
28.4.2 *Two harmonic oscillations having the same direction*, such as $s_1 = A_1 \sin (\omega_1 t + \varphi_1)$ and $s_2 = A_2 \sin (\omega_2 t + \varphi_2)$ *can be added together* by using the vector diagram method (28.1.6).

Shown in Fig. 28.8 are the vectors $A_1(t)$ and $A_2(t)$ of the first and second oscillations at the arbitrary instant t of time when the phases of these oscillations equal: $\Phi_1(t) = \omega_1 t + \varphi_1$ and $\Phi_2(t) = \omega_2 t + \varphi_2$. The resultant oscillation $s = s_1 + s_2$ cor-
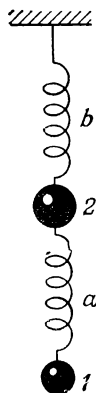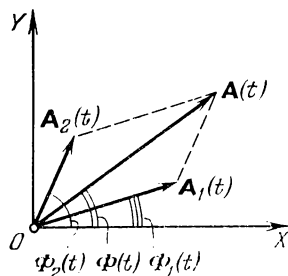
Fig. 28.7   Fig. 28.8

responds to the vector $A(t) = A_1(t) + A_2(t)$ whose projection on the vertical axis OY is equal to s. Thus

$$s = A(t) \sin \Phi(t).$$

According to the cosine law

$$[A(t)]^2 = A_1^2 + A_2^2 + 2A_1 A_2 \cos[\Phi_2(t) - \Phi_1(t)],$$

and

$$\tan \Phi(t) = \frac{A_1 \sin \Phi_1(t) + A_2 \sin \Phi_2(t)}{A_1 \cos \Phi_1(t) + A_2 \cos \Phi_2(t)}.$$

28.4.3 Two harmonic oscillations s_1 and s_2 are said to be *coherent* if their phase difference is independent of time. Thus

$$\frac{d}{dt}[\Phi_2(t) - \Phi_1(t)] \equiv 0 \quad \text{and} \quad \Phi_2(t) - \Phi_1(t) = \text{const}.$$

Since $\Phi_2(t) - \Phi_1(t) = (\omega_2 - \omega_1)\,t + (\varphi_2 - \varphi_1)$, the angular frequency of coherent oscillations should be the same, i.e. $\omega_2 = \omega_1 = \omega$. At any instant of time the phase difference of coherent oscillations is equal to the difference of their initial phases: $\Phi_2(t) - \Phi_1(t) = \varphi_2 - \varphi_1$. Consequently, the resultant oscillation is harmonic and has the same angular frequency ω, i.e.

$$s = s_1 + s_2 = A \sin(\omega t + \varphi_0),$$

where

$$A^2 = A_1^2 + A_2^2 + 2A_1 A_2 \cos(\varphi_2 - \varphi_1)$$

and

$$\tan \varphi_0 = \frac{A_1 \sin \varphi_1 + A_2 \sin \varphi_2}{A_1 \cos \varphi_1 + A_2 \cos \varphi_2}.$$

In accordance with the values of the initial phase differences of the oscillations being added together, the amplitude A of the resultant oscillation varies within the limits

from $A = |\,A_1 - A_2\,|$ at $\varphi_2 - \varphi_1 = \pm(2m + 1)\,\pi$

to $A = A_1 + A_2$ at $\varphi_2 - \varphi_1 = \pm 2m\pi$

where $m = 0, 1, 2, \ldots$ is any integral nonnegative number. When $\varphi_2 - \varphi_1 = \pm 2m\pi$, the oscillations being added are said to be *cophased* (or simply, *in phase*), whereas at $\varphi_2 - \varphi_1 = \pm (2m + 1)\,\pi$ they are said to be *antiphased* (or *in opposition*).
28.4.4 Harmonic oscillations of different frequency ($\omega_2 \neq \omega_1$) are *incoherent* because their phase difference, equal to $(\omega_2 - \omega_1)\,t + (\varphi_2 - \varphi_1)$, varies continuously with time. If such oscillations are superposed, an anharmonic resultant oscillation is obtained. The amplitude vectors $\mathbf{A}_1$ and $\mathbf{A}_2$ of the added oscillations (Fig. 28.8) rotate at different velocities, so that the parallelogram constructed on the vectors is continuously deformed and its diagonal, vector $\mathbf{A}$ of the resultant oscillation, varies in length and rotates at variable angular velocity.
Two harmonic oscillations having different angular frequencies ω_1 and ω_2 can be approximately considered to be coherent only during the time interval Δt in which the phase difference of these oscillations varies only slightly: $|\,\omega_2 - \omega_1\,|\,\Delta t \ll 2\pi$ or $\Delta t \ll$

$\ll \tau_{coh}$, where $\tau_{coh} = 2\pi/|\ \omega_2 - \omega_1\ |$ is the *coherence time* of the oscillations being considered.

28.4.5 Anharmonic oscillation obtained as a result of super-posing two harmonic oscillations having the same direction and slightly differing frequencies ($|\ \omega_2 - \omega_1\ | \ll \omega_1$) produces what are called *beats*. It proves convenient to take as the time reference t the instant when the phases of the two oscillations being super-posed, s_1 and s_2, coincide and are equal to φ_0. Then $s_1 = A_1 \sin(\omega_1 t + \varphi_0)$ and $s_2 = A_2 \sin(\omega_2 t + \varphi_0) = A_2 \sin[\omega_1 t + \varphi_0 + \varphi(t)]$, where $\varphi(t) = (\omega_2 - \omega_1)t$. The resultant oscillation satisfies the relation

$$s = A(t) \sin[\omega_1 t + \varphi_0 + \psi(t)],$$

where

$$[A(t)]^2 = A_1^2 + A_2^2 + 2A_1 A_2 \cos\varphi(t)$$

and

$$\tan\psi(t) = \frac{A_2 \sin\varphi(t)}{A_1 + A_2 \cos\varphi(t)}.$$

In particular, when $A_1 = A_2 = A_0$,

$$A(t) = 2A_0 \cos\frac{\omega_2 - \omega_1}{2} t \quad \text{and} \quad \psi(t) = \frac{\omega_2 - \omega_1}{2} t$$

so that

$$s = 2A_0 \cos\frac{\omega_2 - \omega_1}{2} t \sin\left(\frac{\omega_2 - \omega_1}{2} t + \varphi_0\right).$$

The quantity $|\ A(t)\ |$, characterizing the swing of the oscilla-tions in beating, varies within the limits from $|\ A_1 - A_2\ |$ to $A_1 + A_2$ with the angular frequency $\Omega = |\ \omega_2 - \omega_1\ |$, called the *angular beat frequency*. Since the beat frequency is much lower than the frequency of oscillation ($\Omega \ll \omega_1$), the variable quantity $|\ A(t)\ |$ is conditionally called the *beat amplitude*. The *beat period* T_b and the *beat frequency* ν_b are equal to

$$T_b = \frac{2\pi}{\Omega} = \frac{2\pi}{|\ \omega_2 - \omega_1\ |} = \frac{1}{\left|\dfrac{1}{T_2} - \dfrac{1}{T_1}\right|}$$

and

$$\nu_{\text{b}} = \frac{1}{T_{\text{b}}} = |\, \nu_2 - \nu_1\,|,$$

where T_1, ν_1, T_2 and ν_2 are the periods and frequencies of the superposed oscillations. Curves of the dependence of s on t in beating are shown in Fig. 28.9 for the case when $A_1 = A_2 = A_0$.
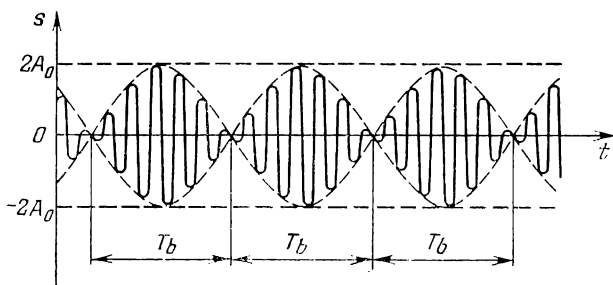
Fig. 28.9

28.4.6 As a result of the superposition of harmonic oscillations, coinciding in direction and having angular frequencies that are integral *multiples* of their lowest frequency, i.e. ω, 2ω, 3ω, etc., a *periodic* nonharmonic oscillation with the period $T = 2\pi/\omega$ is obtained. On the other hand, any complex *periodic oscillation* $s = f(t)$ can be represented as the sum of simple harmonic oscillations having angular frequencies, which are integral multiples of the *basic angular frequency* $\omega = 2\pi/T$, where T is the period of the oscillation. Thus

$$s = f(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos n\omega t + b_n \sin n\omega t)$$

$$= \frac{a_0}{2} + \sum_{n=1}^{\infty} A_n \sin(n\omega t + \varphi_n),$$

where

$$a_n = \frac{2}{T} \int_{-T/2}^{T/2} f(t) \cos n\omega t \, dt \quad (n = 0, 1, 2, \ldots),$$

$$b_n = \frac{2}{T} \int_{-T/2}^{T/2} f(t) \sin n\omega t \, dt \quad (n = 1, 2, \ldots).$$

Such a representation of the periodic function $f(t)$ is called the *expansion of this function into a Fourier series*, or the *harmonic analysis of a complex periodic oscillation*. The terms of the Fourier series corresponding to the harmonic oscillations with the angular frequencies ω, 2ω, 3ω, etc. are called the *first*, or *fundamental*, *second*, *third*, etc. *harmonics* of the complex periodic oscillation $s = f(t)$. The whole set of these harmonics forms the *spectrum of the oscillations* $s = f(t)$. The composition of the spectrum depends on the form of the periodic function $s = f(t)$. In the simplest cases, the spectrum can consist of a small number of harmonics.

The oscillation spectrum is often understood to be the *frequency spectrum*, i.e. the set of frequencies of simple harmonic oscillations that can be superimposed to obtain the complex oscillation being considered. Periodic oscillations have *discrete (line) frequency spectra*.

28.4.7 *Nonperiodic oscillations* have, as a rule, *continuous frequency spectra*, i.e. they can be represented as the result of the superposition of a great many harmonic oscillations, whose frequencies have all possible values in a certain range (from 0 to ∞ in the general case). The harmonic analysis of such oscillations consists in presenting them in the form of a *Fourier integral*:

$$s = f(t) = \int_{0}^{\infty} [a(\omega) \cos \omega t + b(\omega) \sin \omega t] \, d\omega,$$

where

$$a(\omega) = \frac{1}{\pi} \int_{-\infty}^{\infty} f(\xi) \cos \omega\xi \, d\xi, \qquad b(\omega) = \frac{1}{\pi} \int_{-\infty}^{\infty} f(\xi) \sin \omega\xi \, d\xi.$$

Certain nonperiodic oscillations, called *almost periodic (quasi-periodic)* oscillations have a line frequency spectrum. But the included angular frequencies are incommensurable with one another, because their ratios are expressed by irrational numbers.
28.4.8 *Modulation* is the change of any parameter of a periodic oscillation (for instance, the amplitude or frequency) according to a definite law during a length of time considerably longer than the period of the oscillation.

In *amplitude modulation*, for example, *of the harmonic oscillation* $s = A_0 \sin (\omega_0 t + \varphi_0)$, the modulated oscillation is of the form

$$s = A_0 [1 + b(t)] \sin (\omega_0 t + \varphi_0),$$

where $| b(t) | < 1$.
If amplitude modulation is carried out according to the harmonic law $b(t) = b_0 \cos \Omega t$, where $b_0 = $ const and $\Omega \ll \omega_0$, then

$$s = A_0 (1 + b_0 \cos \Omega t) \sin (\omega_0 t + \varphi_0).$$

This modulated oscillation has a line frequency spectrum because it can be represented as the sum of three harmonic oscillations with the angular frequencies ω_0, $\omega_0 - \Omega$ and $\omega_0 + \Omega$, and amplitudes equal, respectively, to A_0, $A_0 b_0/2$ and $A_0 b_0/2$. Thus

$$A_0 (1 + b_0 \cos \Omega t) \sin (\omega_0 t + \varphi_0)$$

$$= A_0 \sin (\omega_0 t + \varphi_0) + \frac{A_0 b_0}{2} \{\sin [(\omega_0 + \Omega) t + \varphi_0]$$

$$+ \sin [(\omega_0 - \Omega) t + \varphi_0]\}.$$

In *frequency modulation of the harmonic oscillation* $s = A_0 \times \sin (\omega_0 t + \varphi_0)$, carried out according to a harmonic law, the modulated oscillation is of the form

$$s = A_0 \sin [\omega_0 (1 + b_0 \cos \Omega t) t + \varphi_0],$$

where $b_0 < 1$ and $\Omega \ll \omega_0$.
Accordingly, in *phase modulation*, the initial phase of the oscillation is changed. Thus

$$s = A_0 \sin (\omega_0 t + \Delta \varphi \cos \Omega t), \quad \text{where } \Omega \ll \omega_0.$$

In the general case oscillations can be modulated simultaneously in amplitude and in phase (or frequency). An example of such modulated oscillations is beating (28.4.4).

28.4.9 *Addition of mutually perpendicular harmonic oscillations of equal frequency.* Assume that point M oscillates simultaneously along the coordinate axes OX and OY according to the law: $x = A_1 \sin(\omega t + \varphi_1)$ and $y = A_2 \sin(\omega t + \varphi_2)$, where x and y are the Cartesian coordinates of point M. The path equation for the resultant motion of point M in plane XOY can be found by eliminating parameter t from the expression for x and y. Thus

$$\frac{x^2}{A_1^2} + \frac{y^2}{A_2^2} - \frac{2xy}{A_1 A_2} \cos(\varphi_2 - \varphi_1) = \sin^2(\varphi_2 - \varphi_1).$$

This path is an ellipse (Fig. 28.10) and point M describes this ellipse during a time equal to the period of the oscillations being
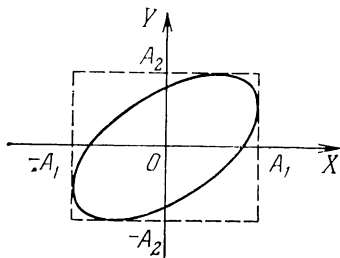
Fig. 28.10

added together ($T = 2\pi/\omega$). Hence, the resultant motion of point M is called *elliptically polarized oscillation*.

The orientation of the axes of the ellipse in plane XOY, as well as its dimensions depend upon the amplitudes A_1 and A_2 of the oscillations being combined and on their initial phase difference $\varphi_2 - \varphi_1$. If $\varphi_2 - \varphi_1 = (2m + 1)\,\pi/2$, where $m = 0, \pm 1, \pm 2, \ldots$, the axes of the ellipse coincide with the coordinate axes OX and OY, and the lengths of the semiaxes are equal to the amplitudes A_1 and A_2. Thus

$$\frac{x^2}{A_1^2} + \frac{y^2}{A_2^2} = 1.$$

If, in addition, $A_1 = A_2$, the path of point M is a circle. Such resultant motion of point M is called *circularly polarized oscillation*.

When $\varphi_2 - \varphi_1 = m\pi$, where $m = 0, \pm 1, \pm 2, \ldots$, the ellipse degenerates into a straight line. Thus

$$y = \pm \frac{A_2}{A_1}\, x.$$

The plus sign corresponds to even values of m, i.e. to the addition of cophased oscillations (Fig. 28.11a); the minus sign, to odd values of m, i.e. to the addition of oscillations that are in
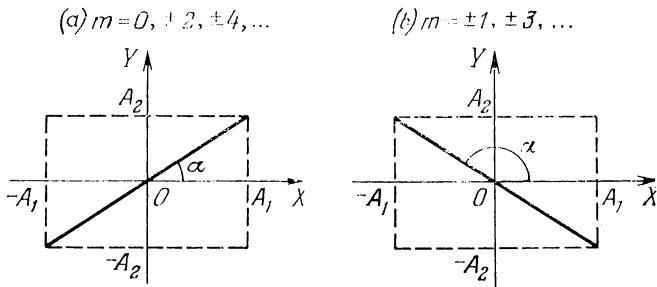
Fig. 28.11

opposition (antiphased) (Fig. 28.11b). In either case, point M executes *linearly polarized oscillation*. It oscillates with harmonic motion at the frequency of the oscillations being combined and with the amplitude $A = \sqrt{A_1^2 + A_2^2}$ along a straight line that makes the angle $\alpha = \arctan(A_2 \cos m\pi / A_1)$ with the axis OX.

28.4.10 In the addition of mutually perpendicular oscillations having the angular frequencies $p\omega$ and $q\omega$, where p and q are whole numbers, we obtain

$$x = A_1 \sin(p\omega t + \varphi_1) \text{ and } y = A_2 \sin(q\omega t + \varphi_2).$$

The values of coordinates x and y of the oscillating point M are simultaneously repeated after equal periods T_0 of time, equal

to the least common multiple of $T_1 = 2\pi/p\omega$ and $T_2 = 2\pi/q\omega$, which are the periods of oscillation along the OX and OY axes. Hence, the path of point M is a closed curve whose shape depends upon the ratios of the amplitudes, frequencies and initial phases of the oscillations being combined. Such closed paths of point M, which has simultaneous harmonic oscillations in two mutually perpendicular directions, are called *Lissajous figures*.
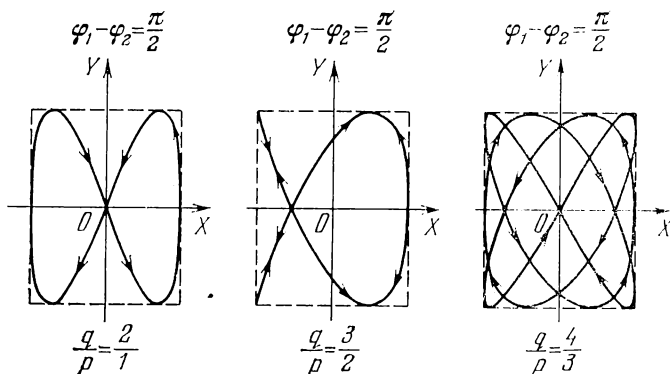
Fig. 28.12

Lissajous figures are inscribed into rectangles whose centre coincides with the origin of coordinates and whose sides are parallel to coordinate axes OX and OY. The distances of these sides from the axes are equal, respectively, to A_2 and A_1. The ratio of the frequencies $q\omega$ and $p\omega$ of the combined oscillations is equal to the ratio of the number of times the corresponding Lissajous figure touches a side of the rectangle parallel to the OX axis to the number of times it touches a side parallel to the OY axis. Shown in Fig. 28.12 are the Lissajous figures for three different values of the ratio q/p (2:1, 3:2 and 4:3) and for initial phase differences of $\Delta\varphi = \varphi_1 - \varphi_2 = \pi/2$.

CHAPTER 29   DAMPED AND FORCED OSCILLATIONS

29.1 Damped Oscillations

29.1.1 *The damping of an oscillation or vibration* is its gradual
weakening in the course of time due to the loss of energy of the
oscillatory (vibratory) system. Free oscillations (vibrations)
of real systems are always damped. The damping of free mecha-
nical vibrations is due chiefly to friction and the excitation of
elastic waves (30.1.3) in the surrounding medium. Damping in
electrical oscillatory systems is due to thermal losses in the
conductors making up the system or located in its variable
electric field, energy losses on the radiation of electromagnetic
waves (31.1.1), and heat losses in dielectrics and ferromagnetic
materials as a result of electrical and magnetic hysteresis
(18.4.5 and 26.5.2).
The damping of oscillations depends upon the properties of
the oscillatory system. A system is said to be *linear* if the para-
meters characterizing physical properties of the system, essential
in the process being considered, do not change in the course
of the process. Linear systems are described by linear differential
equations. A spring pendulum (28.2.3), for instance, moving in
a viscous medium, is a linear system if the resistance coefficient
of the medium and the elasticity of the spring are independent
of the velocity and displacement of the pendulum. An electrical
oscillatory circuit (28.3.1) can be considered linear if its elec-
trical resistance R, capacitance C and inductance L are indepen-
dent of the current in the circuit, and the voltage. In the majority
of cases real oscillatory or vibratory systems are sufficiently
close, with respect to their properties, to linear ones.
29.1.2 *The differential equation for the free damped oscillation of
a linear system* is of the form:

$$\frac{d^2s}{dt^2} + 2\beta \frac{ds}{dt} + \omega_0^2 s = 0.$$

Here s is the physical characteristic of the system that varies
in the oscillation, $\beta = \mathrm{const} > 0$ is the *damping factor*, and
ω_0 is the angular frequency of free *undamped* oscillation of the
same system, i.e. in the absence of energy losses (at $\beta = 0$).

Example 1. *Free damped oscillation of a spring pendulum* **(28.2.3).**
A pendulum of mass m, oscillating in a straight line along the OX
axis due to the elastic force of its spring, is also subject to a
resisting force $F_{res} = -bv$, where v is the velocity of the pen-
dulum and $b = \mathrm{const} > 0$ is the *resistance coefficient*. The differ-
ential equation of free damped oscillation of the pendulum is

$$m \frac{d^2x}{dt^2} = -b \frac{dx}{dt} - kx \quad \text{or} \quad \frac{d^2x}{dt^2} + 2\beta \frac{dx}{dt} + \omega_0^2 x = 0,$$

where $\beta = b/2m$, $\omega_0 = \sqrt{k/m}$, and k is a factor specifying the
elastic properties of the spring.

Example 2. *Free damped oscillation in an electric oscillatory
circuit.* The electrical resistance of a real circuit $R \neq 0$, and the
differential equation of the oscillation in the circuit **(28.3.1)** is
of the form

$$\frac{d^2q}{dt^2} + 2\beta \frac{dq}{dt} + \omega_0^2 q = 0,$$

where $\beta = R/2L$ and $\omega_0 = 1/\sqrt{LC}.$

29.1.3 If damping is small $(\beta < \omega_0)$ the dependence of s on t
that satisfies the equation of damped oscillation **(29.1.2)** is of
the form

$$s = A_0 e^{-\beta t} \sin(\omega t + \psi_0).$$

Here $\omega = \sqrt{\omega_0^2 - \beta^2}$ and the constant quantities A_0 and ψ_0
depend upon the initial conditions, i.e. upon the values of s
and ds/dt at the initial instant of time $(t = 0)$. The curve of
the dependence of s on t at $\psi_0 = 0$ is shown in Fig. 29.1.
Damped oscillations are not periodical ones **(28.1.2)**. For exam-
ple, the maximum value of the varying quantity s, reached
at a certain instant t_1 of time, is never repeated subsequently
(at $t > t_1$). But the quantity s becomes zero in damped oscil-
lation as it varies in a single direction (decreases, for instance)
and also reaches its maximum and minimum values after
equal time intervals. Thus

$$T = \frac{2\pi}{\omega} = \frac{2\pi}{\sqrt{\omega_0^2 - \beta^2}}.$$

For this reason, the quantities T and ω are conditionally called the *period* (*conditional period*) and *angular frequency* (*conditional angular frequency*) *of damped oscillation.*
The quantity

$$A = A_0 e^{-\beta t}$$

is called the *amplitude of damped oscillation*, where A_0 is the *initial amplitude*. The amplitude of damped oscillation de-
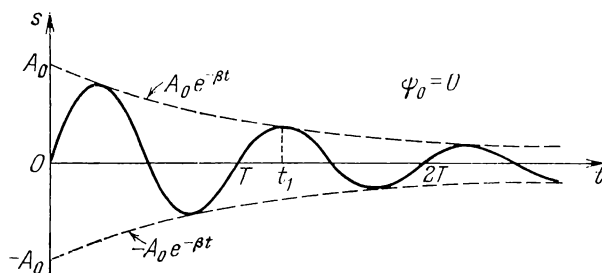
Fig. 29.1

creases in the course of time; the greater the damping factor β, the more rapidly the amplitude decreases.
The time interval $\tau = 1/\beta$ during which there is an *e*-fold decrease in the amplitude of damped oscillation, is called the *relaxation time.*
29.1.4 *The logarithmic decrement of damping* is the dimensionless quantity δ, equal to the natural logarithm of the ratio of the amplitudes of damped oscillation at the instants of time t and $t + T$ (where T is the conditional period of oscillation). Thus

$$\delta = \ln \frac{A(t)}{A(t+T)} = \beta T = \frac{T}{\tau} = \frac{1}{N} \, ,$$

where N is the number of oscillations in which the amplitude is reduced *e*-fold.

The relation between the angular frequency ω of damped oscillation of a system and the logarithmic decrement of damping δ is

$$\omega = \omega_0 \sqrt{1 - \left(\frac{\omega}{\omega_0}\right)^2 \left(\frac{\delta}{2\pi}\right)^2}.$$

29.1.5 *The Q factor* (or *quality factor* or *storage factor*) of an oscillatory system is the dimensionless quantity Q, equal to the product of 2π by the ratio of the energy of oscillation $E(t)$ of the system at the arbitrary time instant t to the decrease of this energy during the length of time from t to $t + T$, i.e. during one conditional period of damped oscillation. Thus

$$Q = 2\pi \frac{E(t)}{E(t) - E(t+T)}.$$

Since the energy $E(t)$ is proportional to the square of the amplitude $A(t)$ of the oscillation,

$$Q = 2\pi \frac{A^2(t)}{A^2(t) - A^2(t+T)} = \frac{2\pi}{1 - e^{-2\beta T}} = \frac{2\pi}{1 - e^{-2\delta}}.$$

At low values of the logarithmic decrement of damping δ, the Q factor of an oscillatory system is $Q = \pi/\delta$. Here the conditional period of damped oscillation T is practically equal to the period T_0 of free undamped oscillation and, consequently, $Q = \pi/\beta T_0 = \omega_0/2\beta$. For example, the Q factor of an electrical oscillatory circuit (28.3.1) is $Q = \sqrt{L/C}/R$, and that of a spring pendulum (29.1.2) is $Q = \sqrt{km}/b$.

29.1.6 The conditional period of damped oscillation increases with the damping factor β and becomes infinite at $\beta = \omega_0$. If $\beta > \omega_0$ the differential equation of motion of the system

$$\frac{d^2 s}{dt^2} + 2\beta \frac{ds}{dt} + \omega_0^2 s = 0$$

has the following general solution:

$$s = C_1 e^{-\alpha_1 t} + C_2 e^{-\alpha_2 t},$$

where $\alpha_1 = \beta + \sqrt{\beta^2 - \omega_0^2}$, $\alpha_2 = \beta - \sqrt{\beta^2 - \omega_0^2}$, and C_1 and C_2 are constant factors depending upon the initial con-

25 *

ditions. If the initial values (at the instant of time $t = 0$) equal: $s = s_0$ and $ds/dt = v_0$, then

$$C_1 = -\frac{\alpha_2 s_0 + v_0}{\alpha_1 - \alpha_2} \quad \text{and} \quad C_2 = \frac{\alpha_1 s_0 + v_0}{\alpha_1 - \alpha_2}.$$

Such motion of the system is not of an oscillatory nature and is said to be *aperiodic*. Depending upon the initial conditions,
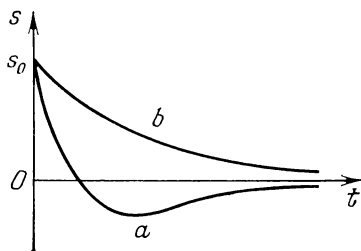
Fig. 29.2

aperiodic motion of the system may be either of two kinds (Fig. 29.2). Type a motion occurs when s_0 and v_0 are opposite in sign and $|v_0| > \alpha_1 |s_0|$. Otherwise, the aperiodic motion is of the b type

29.2 Forced Mechanical Vibration

29.2.1 A variable external force, applied to a system and causing its forced mechanical vibration, is called the *disturbing*, or *driving*, force.
The differential equation of forced vibration (or *oscillation*) *of the* simplest *linear system* (29.1.1), the spring pendulum (29.1.2), executed along the OX axis due to the action of the variable external force $\mathbf{F}$ (t) is:

$$\frac{d^2 r}{dt^2} + 2\beta \frac{dx}{dt} + \omega_0^2 x = \frac{1}{m} F_x (t).$$

If $F_x(t)$ is a periodic function of time, then, after application of the force to the pendulum, the *forced vibration* begins *with a transient condition*. The pendulum simultaneously participates in two vibrations :

$$x = x_1(t) + x_2(t).$$

The first term corresponds to free damped oscillation* of the pendulum (29.1.3):

$$x_1(t) = A_0 e^{-\beta t} \sin(\omega t + \psi_0),$$

where $\omega = \sqrt{\omega_0^2 - \beta^2}$.

The second term corresponds to the undamped periodic oscillation of the pendulum at a frequency equal to that of the driving force $F_x(t)$.

The amplitude value of $x_1(t)$, equal to $A_0 \exp(-\beta t)$, is reduced at a more or less rapid rate after the beginning of the forced vibration (or oscillation). During the time $\tau = 4.6/\beta$ the amplitude of $x_1(t)$ decreases by a factor of 100. Consequently, after a certain time τ following the beginning of vibration ($\tau \approx \tau_0$), free vibration of the pendulum practically ceases: $x(t) \approx x_2(t)$. The pendulum reaches a state of *steady forced vibration* (or oscillation) at the frequency of the driving force.

29.2.2 When the driving force varies according to a harmonic law, i.e. $F_x = F_0 \cos \Omega t$, the steady-state forced vibration of the pendulum is also harmonic at the same frequency. Thus

$$x = A \cos(\Omega t + \varphi_0).$$

The amplitude A of this vibration and the phase shift φ_0 between the displacement and the driving force depend upon the relation between the angular frequencies of the forced vibration (Ω) and the free undamped vibration (ω_0) of the pendulum. Thus

$$A = \frac{F_0}{m\sqrt{(\omega_0^2 - \Omega^2)^2 + 4\beta^2\Omega^2}} \quad \text{and} \quad \tan \varphi_0 = -\frac{2\beta\Omega}{\omega_0^2 - \Omega^2}.$$

---

* It is assumed that $\beta < \omega_0$. Otherwise, the free motion of the pendulum would be aperiodic (29.1.6), i.e.

$$x_1(t) = C_1 \exp\left[-(\beta + \sqrt{\beta^2 - \omega_0^2})\, t\right]$$
$$+ C_2 \exp\left[-(\beta - \sqrt{\beta^2 - \omega_0^2})\, t\right].$$

At $\Omega = 0$ we obtain $\varphi_0(0) = 0$ and $A(0) = A_0 = F_0/m\omega_0^2 = = F_0/k$, which is the *static displacement* of the pendulum from
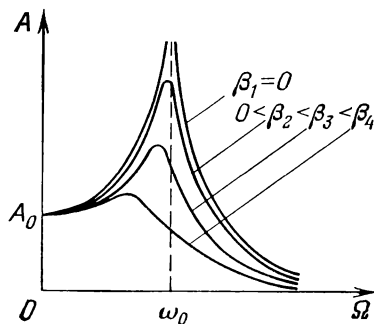
Fig. 29.3

its equilibrium position by the constant force $F_x = F_0$. At $\Omega \to \infty$, the amplitude $A(\Omega) \to 0$ and $\tan \varphi_0 \to 0$, whereas
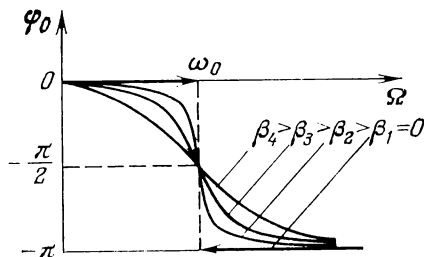
Fig. 29.4

$\varphi_0 \to -\pi$. The curves of the functions $A(\Omega)$ and $\varphi_0(\Omega)$ are shown in Figs. 29.3 and 29.4 for various values of the damping factor β.

29.2.3 The amplitude of the displacement in the case of steady-state forced harmonic vibration (oscillation) of the pendulum reaches its maximum value at the angular frequency

$$\Omega_r = \sqrt{\omega_0^2 - 2\beta^2} = \sqrt{\omega^2 - \beta^2},$$

where ω is the angular frequency of free damped vibration of the pendulum (29.1.3). The frequency Ω_r is called the *resonance frequency*.

The maximum amplitude is

$$A_{max} = A(\Omega_r) = \frac{F_0}{2m\beta\omega} = \frac{\pi F_0}{m\delta\omega^2},$$

where δ is the logarithmic decrement of damping (29.1.4). When $\beta \ll \omega_0$, $\Omega_r \approx \omega_0$, $\varphi_0(\Omega_r) \approx -\pi/2$ and $A_{max} \approx QA_0$, where $Q \approx \pi/\delta$ is the Q factor of the pendulum (29.1.5) and A_0 is the static displacement (29.2.2).

The drastic increase in the amplitude of forced mechanical vibration (or oscillation) as the angular frequency of the driving force approaches the value Ω_r is the phenomenon called *mechanical resonance*. Accordingly, the curves of the dependence of A on Ω, shown in Fig. 29.3 are called *resonance curves*.

As the damping factor β is increased, the peaks on the resonance curves are rapidly reduced (at small β values the amplitude $A_{max} \approx 1/\beta$) and the resonance frequency Ω_r is slowly decreased.

29.2.4 *The velocity of the pendulum* in steady-state forced harmonic vibration (oscillation) is

$$v_x = \frac{dx}{dt} = -A\Omega \sin(\Omega t + \varphi_0) = A_v \cos(\Omega t + \alpha).$$

Here $A_v = A\Omega$ and $\alpha = \varphi_0 + \pi/2$ are the amplitude of the velocity and the phase difference between the velocity and the driving force:

$$A_v = \frac{F_0 \Omega}{m\sqrt{(\omega_0^2 - \Omega^2)^2 + 4\beta^2\Omega^2}} = \frac{F_0}{m\sqrt{\dfrac{(\omega_0^2 - \Omega^2)^2}{\Omega^2} + 4\beta^2}}$$

and

$$\tan\alpha = -\cot\varphi_0 = \frac{\omega_0^2 - \Omega^2}{2\beta\Omega}.$$

The amplitude of the velocity is maximal at $\Omega = \omega_0$ and equals

$$(A_v)_{max} = A_v (\omega_0) = \frac{F_0}{2m\beta}\,.$$

In this case, $\alpha = 0$, i.e. the velocity of the pendulum oscillates in phase with the driving force. As $\Omega \to \infty$, the amplitude $A_v \to 0$ and $\alpha \to -\pi/2$, whereas as $\Omega \to 0$, the amplitude $A_v \to 0$ and $\alpha \to \pi/2$.

29.2.5 *The acceleration of the pendulum* in steady-state forced harmonic vibration (oscillation) is

$$a_x = \frac{d^2 x}{dt^2} = -A\Omega^2 \cos (\Omega t + \varphi_0) = A_a \cos (\Omega t + \gamma).$$

Here $A_a = A\Omega^2$ and $\gamma = \varphi_0 + \pi$ are the amplitude of the acceleration and the phase difference between the acceleration and the driving force:

$$A_a = \frac{F_0 \Omega^2}{m \sqrt{(\omega_0^2 - \Omega^2)^2 + 4\beta^2 \Omega^2}} = \frac{F_0}{m \sqrt{[(\omega_0/\Omega)^2 - 1]^2 + (2\beta/\Omega)^2}}\,.$$

The amplitude of acceleration is maximal at

$$\Omega = \frac{\omega_0}{\sqrt{1 + (\beta/\omega_0)^2}} \approx \sqrt{\omega_0^2 - \beta^2}.$$

At $\Omega = 0$, the amplitude $A_a = 0$, whereas as $\Omega \to \infty$ the amplitude of the acceleration tends to the value $A_a (\infty) = F_0/m$.

29.2.6 In steady-state forced vibration, the energy loss of the vibratory system, due to dissipative forces (1.3.7), is completely compensated by the work performed on the system by the driving force. The work, for example, done in one complete oscillation by the force of resistance acting on the spring pendulum is

$$W_1 = -b \int_0^T \left(\frac{dx}{dt}\right)^2 dt = -2m\beta A^2 \Omega^2 \int_0^T \sin^2 (\Omega t + \varphi_0)\, dt$$
$$= -m\beta A^2 \Omega^2,$$

where $b = 2m\beta$ is the resistance coefficient (29.1.2).
The work done during the same time by the driving force $F_x = F_0 \cos \Omega t$ is

$$W_2 = F_0 \int_0^T \frac{dx}{dt} \cos \Omega t \, dt = -A\Omega F_0 \int_0^T \cos \Omega t \sin (\Omega t + \varphi_0) \, dt$$

$$= -\frac{1}{2} A\Omega F_0 \sin \varphi_0 = -W_1,$$

since $\sin \varphi_0 = -2m\beta A\Omega/F_0$.

29.2.7 If the driving force acting on the spring pendulum varies periodically, but not according to a harmonic law, it can be represented as the sum of the harmonics of this force (28.4.6). These have different amplitudes, initial phases and angular frequencies that are multiples of $\Omega = 2\pi/T$, where T is the period of variation of the driving force. Since the pendulum is a linear vibratory (or oscillatory) system (29.1.1), each harmonic of the driving force acts on the system as if there were no other harmonics. Therefore, the steady-state forced oscillation of the pendulum, caused by an arbitrary periodic driving force, can be dealt with as the result of superposing steady-state forced oscillations due to each harmonic component of the driving force acting separately. The "contribution" of the various harmonic components to the resultant oscillation depends upon their frequency and amplitude. Owing to resonance phenomena, an essential role is played by only the harmonic components of the driving force that have an angular frequency close to the resonance frequency of the pendulum (29.2.3). If the damping factor β of the pendulum is small, the pendulum can execute steady-state forced oscillation close to the harmonic one, even if the driving force is far from harmonic. An example of such forced oscillation is that of a pendulum subject to short periodical external action in the form of impacts in a single direction and repeated over equal lengths of time, which are equal to the period of free oscillation of the pendulum.

29.3 Forced Electrical Oscillation

29.3.1 To obtain forced oscillation in an electrical oscillatory circuit (28.3.1) it is necessary to connect into it a source of electrical energy, whose emf $\mathscr{E}$ varies with time (Fig. 29.5).

In electrical engineering a source of electrical energy, characterized by its emf and internal electrical resistance, is called a *seat*, or *source*, *of emf* (*voltage source*). According to Ohm's law for the portion *1-R-L-2* (21.2.3), the quasi-stationary current (28.3.1), produced in the circuit in forced oscillation, can be found from the formula
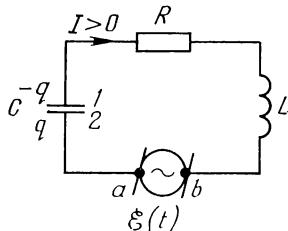
$$IR = V_1 - V_2 - L\frac{dI}{dt} + \mathscr{E}(t).$$

Here $V_2 - V_1 = q/C$ is the potential difference across the capacitor plates, q is the capacitor charge, and the internal electrical resistance of the source of emf is assumed to be negligibly small compared to R (such a source of emf is said

Fig. 29.5

to be *ideal*). From the law of conservation of electrical charge (14.1.3) it follows that $I = dq/dt$. Hence, the *differential equation of forced electrical oscillation in a circuit* can be presented in a form similar to the equation for forced mechanical vibration (29.2.1):

$$\frac{d^2q}{dt^2} + 2\beta\frac{dq}{dt} + \omega_0^2 q = \frac{1}{L}\mathscr{E}(t).$$

Here $\beta = R/2L$ is the damping factor for free oscillation in a circuit, and $\omega_0 = 1/\sqrt{LC}$ is the angular frequency of free undamped oscillation (i.e. at $R = 0$).
29.3.2 If the driving emf $\mathscr{E}(t)$ varies according to the harmonic law $\mathscr{E}(t) = \mathscr{E}_0 \cos \Omega t$, then at steady-state forced oscillation (29.2.1), the capacitor charge varies harmonically at the same angular frequency Ω:

$$q = q_0 \cos(\Omega t + \varphi_0).$$

The amplitude q_0 and the initial phase φ_0 are determined by the formulas

$$q_0 = \frac{\mathscr{E}_0}{\Omega\sqrt{R^2 + (\Omega L - 1/\Omega C)^2}} = \frac{\mathscr{E}_0}{L\sqrt{(\omega_0^2 - \Omega^2)^2 + 4\beta^2\Omega^2}}$$

and

$$\tan \varphi_0 = \frac{R}{\Omega L - 1/\Omega C} = -\frac{2\beta\Omega}{\omega_0^2 - \Omega^2} \, .$$

At $\Omega = 0$, the phase $\varphi_0 (0) = 0$ and $q_0 (0) = \mathscr{E}_0 C$, which is the charge of the capacitor at a constant potential difference between the plates equal to $\mathscr{E}_0$. At $\Omega \to \infty$, the amplitude $q_0 \to 0$, and $\varphi_0 \to -\pi$. The curves showing the dependence of φ_0 on Ω are shown in Fig. 29.4, and those of the dependence of q_0 on Ω, in Fig. 29.3, where $A = q_0$ and $A_0 = q_0 (0) = \mathscr{E}_0 C$.

29.3.3 With steady-state forced oscillation in the circuit, the current is

$$I = \frac{dq}{dt} = -q_0\Omega \sin (\Omega t + \varphi_0) = I_0 \cos (\Omega t - \varphi).$$

The amplitude of the current $I_0 = q_0\Omega$ and the initial phase $-\varphi = \varphi_0 + \pi/2$ are determined by the formulas:

$$I_0 = \frac{\mathscr{E}_0}{\sqrt{R^2 + (\Omega L - 1/\Omega C)^2}} \quad \text{and} \quad \tan \varphi = \frac{\Omega L - 1/\Omega C}{R} \, .$$

Curves showing the dependence of I_0 on Ω for various values of R are called *resonance curves of the oscillatory circuit* and are
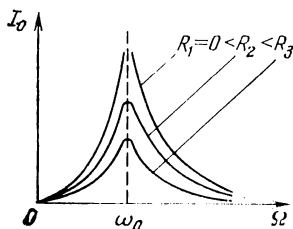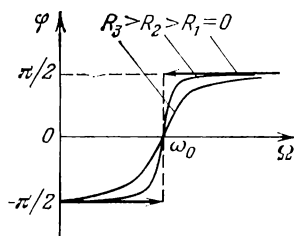
Fig. 29.6   Fig. 29.7

shown in Fig. 29.6; curves of the dependence of φ on Ω are shown in Fig. 29.7. The resonance angular frequency Ω_r, corresponding

to the maximum amplitude of the current in the circuit in forced oscillation, is independent of R. Thus

$$\Omega_r = \omega_0 = \frac{1}{\sqrt{LC}} .$$

The current amplitude at resonance is $I_0 (\Omega_r) = \mathscr{E}_0/R$, and the phase difference between the current and emf is $\varphi (\Omega_r) = 0$. When $\Omega < \omega_0$, $\varphi < 0$, i.e. the current leads the emf in phase; the lower the frequency Ω, the greater this phase lead ($\varphi = -\pi/2$ at $\Omega = 0$). When $\Omega > \omega_0$, $\varphi > 0$, i.e. the current lags in phase behind the emf; the higher the frequency Ω, the greater this phase lag ($\varphi \rightarrow \pi/2$ as $\Omega \rightarrow \infty$).

29.3.4 The potential difference across the terminals a and b of an ideal source of harmonic emf (Fig. 29.5) is equal to its emf. Thus

$$u = V_a - V_b = \mathscr{E}_0 \cos \Omega t.$$

The potential drops across various portions of the circuit illustrated in Fig. 29.5 and having the variable sinusoidal (harmonic) current $I = I_0 \cos (\Omega t - \varphi)$, a capacitor of capacitance C, resistor of resistance R and a coil of inductance L are as follows:

$$u_C = V_2 - V_1 = \frac{q}{C} = U_C \cos \left(\Omega t - \varphi - \frac{\pi}{2} \right) ,$$

$$u_R = IR = U_R \cos (\Omega t - \varphi),$$

$$u_L = L \frac{dI}{dt} = U_L \cos \left(\Omega t - \varphi + \frac{\pi}{2} \right) .$$

The variation of u_R is in phase with that of current I in the circuit, u_L leads the current in phase by $\pi/2$, and u_C lags behind the current in phase by $\pi/2$. Here

$$u_C + u_R + u_L = u = \mathscr{E}_0 \cos \Omega t.$$

29.3.5 The amplitude values of u_C, u_L and u_R are equal, respectively, to

$$U_C = X_C I_0, \quad U_L = X_L I_0 \text{ and } U_R = RI_0,$$

where $X_C = 1/\Omega C$ is the *capacitive reactance of the circuit*, and $X_L = \Omega L$ is the *inductive reactance of the circuit*. The quantity $X = X_L - X_C = \Omega L - 1/\Omega C$ is called the *reactance of the circuit*, R is called the *resistance* and $Z = \sqrt{R^2 + (\Omega L - 1/\Omega C)^2}$ is called its *impedance*.

The formulas in Sect. 29.3.3 for the amplitude of the sinusoidal current in the circuit and its initial phase can be rewritten in the form: $I_0 = \mathscr{E}_0/Z$ and $\tan \varphi = X/R$, with $\cos \varphi = R/Z$ and $\sin \varphi = X/Z$.

At resonance $\Omega = 1/\sqrt{LC}$ and $X_L = X_C$, so that the reactance of the circuit becomes zero and the impedance reaches its minimum value, which is equal to the resistance R. Thus

$$X\ (\Omega_r) = 0 \text{ and } Z\ (\Omega_r) = Z_{\min} = R.$$

In this case $U_R = \mathscr{E}_0$ and $U_C = U_L = \sqrt{L/C}\,\mathscr{E}_0/R$.

29.3.6 *The effective value of a periodic current* (and, accordingly, the emf, voltage, etc.) is the root-mean-square value of the current during the time T of its variation. Thus

$$I_{\text{eff}} = \sqrt{\frac{1}{T} \int_0^T I^2\, dt}.$$

For a sinusoidal current and a sinusoidal emf

$$I_{\text{eff}} = \frac{I_0}{\sqrt{2}} \quad \text{and} \quad \mathscr{E}_{\text{eff}} = \frac{\mathscr{E}_0}{\sqrt{2}}.$$

An element of work done by sinusoidal current during the short time dt in the circuit illustrated in Fig. 29.5 is equal to

$$\delta W = Iu\, dt = I_0 \mathscr{E}_0 \cos (\Omega t - \varphi) \cos \Omega t\, dt.$$

The instantaneous power of the current in the circuit is

$$P_{\text{ins}} = \frac{\delta W}{dt} = Iu = I_0 \mathscr{E}_0 \cos (\Omega t - \varphi) \cos \Omega t.$$

The average value of the instantaneous power during a period is called the *active*, or *true*, *power* P of the current in the electric circuit. Thus

$$P = \frac{1}{T} \int_0^T P_{1ns}\, dt = \frac{1}{2}\, I_0 \mathscr{E}_0 \cos \varphi = I_{eff} \mathscr{E}_{eff} \cos \varphi.$$

The factor $\cos \varphi$ is called the *power factor*. Since $I_{eff} = \mathscr{E}_{eff}/Z$ and $\cos \varphi = R/Z$,

$$P = \frac{R}{Z^2}\, \mathscr{E}_{eff}^2.$$

At resonance $Z = R$ and the active power reaches its maximum value

$$P = \frac{\mathscr{E}_{eff}^2}{R} = \frac{\mathscr{E}_0^2}{2R}.$$

CHAPTER 30   ELASTIC WAVES

30.1 Longitudinal and Transverse Waves in an Elastic Medium

30.1.1 A body is said to be *elastic* and its deformation, due to the application of external forces, is *elastic deformation* if it completely disappears after the external forces are removed. According to *Hooke's law* elastic deformation is directly proportional to the external effects that cause it, i.e. it has a linear dependence on these effects. At sufficiently small deformation, all bodies can be assumed, practically, to be elastic.

The elastic properties of bodies depend upon the nature of the thermal motion of their molecules and their forces of interaction. For example, a gaseous body changes its shape without hindrance to suit the shape of the vessel it is contained in; gas has no *elasticity of shape*. But a gas has *elasticity of volume*, i.e. the capacity to resist changes in its volume. This property of a gas is due to the thermal motion of its molecules and is manifested

in the change in the pressure p of the gas when its volume V is changed. According to *Hooke's law for volume deformation*, the change dp in the pressure of a gas for a small change dV in its volume is directly proportional to the relative volume deformation, called the volume strain. Thus

$$dp = -K \frac{dV}{V} \, ,$$

where K is called the *bulk modulus of elasticity of* the gas. For an ideal gas (8.4.1) the value of K depends upon the kind of thermodynamic process of compression (or expansion) being carried out. At a very slow change in the volume of the gas, the process can be considered isothermal (8.3.7), whereas at a rapid change it can be considered adiabatic (8.3.7). In the former case $K = p$ and in the latter $K = \varkappa p$, where $\varkappa$ is the adiabatic exponent (9.5.11).

30.1.2 The elasticity of a crystalline solid is due to the forces of mutual attraction and repulsion of its particles (ions, atoms or molecules) that make up the body and that have random thermal vibration about the points of its crystal lattice. The forces of interaction of the particle impede deformation of the crystal lattice that is associated with both the volume and shape of the body. Therefore, in addition to their bulk elasticity, solids also have elasticity of form, which is manifested by their resistance to shear deformation (40.3.9).

The elasticity of liquids is also due to the forces of intermolecular interaction. But, owing to the fact that the average time τ that molecules of a liquid remain in a temporarily settled state (13.2.4) is very short, liquids, like gases, possess only bulk elasticity. They manifest elasticity of form only with respect to variable deformation of ultrahigh frequency with a period that is less than or of the order of τ.

30.1.3 *Elastic waves* are *mechanical disturbances* (deformations) propagated in an elastic medium. Bodies that act on the medium and cause these disturbances are called *sources of the waves*. A member of the audience in a theatre, for instance hears the speech and singing of the actors and the sounds of the musical instruments due to the variations in the pressure of the air, reaching the audience and caused by the sources of sound.

Sonic, or *acoustic*, *waves* are elastic waves of low intensity, i.e. weak mechanical disturbances propagated in an elastic medium.

Sonic waves, acting on the organs of hearing of people, are
capable of producing auditory sensations if the frequency ν of
their corresponding vibrations lies in the range from 16 to
2×10^4 Hz (*audible sound*). Elastic waves with a frequency
$\nu < 16$ Hz are said to be *infrasonic*, whereas those with a fre-
quency $\nu > 2 \times 10^4$ Hz are *ultrasonic* (elastic waves with
$\nu > 10^9$ Hz are often said to be *hypersonic*).

30.1.4 The propagation of elastic waves in a medium does not
involve the transport of matter*. In a boundless medium, it
consists in exciting forced vibrations of parts of the medium
that are farther and farther away from the source of the waves.
It is feasible here to digress from the discrete (molecular) struc-
ture of the medium, dealing with it as a *continuous medium*,
distributed continuously in space and possessing definite elastic
properties. A *particle* of such a medium, having forced vibration,
is understood to be a small element of its volume whose size,
however, is many times greater than the intermolecular distances
so that it contains an enormous number of molecules. Practical-
ly, such a particle can be dealt with as a point because even in
a gas the intermolecular distances are extremely small (of the
order of 10^{-8} m under standard conditions).

30.1.5 An elastic wave is said to be *longitudinal* if the particles
of the medium vibrate in the direction of wave propagation.
Longitudinal waves are associated with bulk deformation of the
elastic medium and can therefore propagate in any kind of
medium: solid, liquid or gaseous. An example of such waves
are the sonic waves in air.

An elastic wave is said to be *transverse* if the particles of the
medium vibrate, remaining in planes perpendicular to the direc-
tion of wave propagation. Transverse waves are associated with
shear deformation of the elastic medium, and, consequently,
can be generated and propagated only in media having elasticity
of shape, i.e. in solids. An example of transverse waves are
those propagated along the strings of musical instruments.

Special properties are found in *surface waves*, which propagate
along the free surfaces of liquids (or the interface between two
immiscible liquids), being a perturbance of the surface due to

---

* A certain transport of matter (mass transfer) may occur in the
propagation in a medium of strong disturbances (for example,
shock waves in explosions), when the vibration of the particles
of the medium becomes nonlinear.

external effects (dropping of bodies into the liquid, the motion of ships, the wind, etc.). The forces of gravity and surface tension play a decisive role in the generation and propagation of these waves. In surface waves the particles of liquid have transverse and longitudinal vibrations simultaneously, describing elliptical or more complex paths.

30.1.6 A medium is said to be *uniform* if its physical properties, essential in the problem being considered, do not change from point to point. A medium that is uniform with respect to certain physical properties may be nouniform with respect to others. A monocrystalline body, for example, is uniform with respect to its elastic properties and, at the same time, optically non-uniform for X-rays.

A medium is *isotropic* if its physical properties that are essential in the problem being considered are the same in all directions. A medium that is isotropic with respect to certain physical properties may be anisotropic with respect to others. Cubic crystals, for instance, are optically isotropic, but are anisotropic with respect to other properties. Gases and liquids are isotropic with respect to any physical properties.

30.1.7 A medium is said to be *linear* if there is a directly proportional relation between the quantities characterizing the external effects being considered that act on the medium and the changes they lead to in the state of the medium. An elastic medium, for example, obeying Hooke's law (30.1.1) is linear with respect to its mechanical properties. A dielectric is a linear medium with respect to its electrical properties if its dielectric constant (14.2.4) is independent of the electric field strength. By analogy, a magnetic material is a linear medium with respect to its magnetic properties if its relative magnetic permeability (26.4.5) is independent of the magnetic induction of the field.

30.2 Travelling Wave Equation

30.2.1 *Travelling waves* are ones which, in contrast to standing waves (30.5.4), transport energy from one part of space to another.

The *travelling wave equation* is the dependence on the coordinates and time of the scalar or vector quantities that characterize the vibration of the medium as the wave being considered is propagated through it. For waves in a solid medium, for instance,

such a quantity may be the vector of displacement of the par-
ticles of the medium (30.1.4) from their equilibrium position or
the three projections of the vector on the coordinate axes. Longi-
tudinal waves in a gas or liquid are usually characterized by the
surplus pressure of the vibrating medium, equal to the differ-
ence between its variable and equilibrium pressures.

A *ray* is a curve whose tangent at each point coincides with the
direction of propagation of a wave, i.e. with the direction of
energy transport by the wave (30.3.5). In a uniform medium
(30.1.6) rays are straight lines.

30.2.2 An elastic wave is said to be *sinusoidal*, or *harmonic*, if
the corresponding vibration of particles of the medium is har-
monic (28.1.3). The frequency of this vibration is called the
wave frequency. Variations in pressure in a gaseous or liquid
medium through which a sinusoidal wave propagates also take
place according to a harmonic law at a frequency equal to the
wave frequency. In a transverse sinusoidal wave, particles of the
medium can simultaneously vibrate with harmonic motion at
the wave frequency in two mutually perpendicular directions,
each of which is perpendicular to the direction of wave prop-
agation. Depending upon the kind of polarization of the resultant
vibration (28.4.9), distinction is made between the following
types of polarization of transverse sinusoidal waves: *elliptical*,
circular and *linear* (or *plane*).

30.2.3 A mechanical disturbance (deformation) propagates in
an elastic medium at the finite velocity v. Therefore, distur-
bances, initiated by the wave source at the instant t_0 of time,
reach an arbitrary point M of the medium at the instant $t > t_0$.
The greater the path l that the wave travels from its source to
point M, the greater the difference $t - t_0 = l/v$. Consequently,
the vibration at point M lags in phase behind that at the source
of the waves.

The *wave surface*, or *wavefront*, is the locus of all the points at
which the phase of the vibration has the same value at the
instant of time being considered. The difference $t - t_0$ is the
same for all points in a wavefront. Through each point of a
medium subject to wave motion one wavefront can be passed
that corresponds to the value of the phase of the vibration at
this point at the instant of time being considered. Many different
values of the phase of a vibration correspond to a family of
wavefronts. In a uniform isotropic medium (30.1.6) the wave-
fronts are orthogonal to the rays (30.2.1).

30.2.4 A wave is said to be *plane* if its wavefronts constitute a system of parallel planes. In a plane wave, propagating along the OX axis, all values of s, characterizing the vibratory motion of the medium, depend only upon the time t and the coordinate x of the point M being considered in the medium. If there is no absorption of waves in the medium (30.3.7), the vibration at point M differs from that at the origin of coordinates O only in that it is shifted in time by the amount x/v, where v is the wave velocity. Hence, in a plane wave, propagating along the *positive* direction of the OX axis, s is a function of the difference $(t - x/v)$ and the *equation* of such a *plane* wave is

$$s = f\left(t - \frac{x}{v}\right).$$

Accordingly, the equation for a plane wave propagating in the opposite direction is

$$s = f\left(t + \frac{x}{v}\right).$$

30.2.5 The *equation of a plane sinusoidal*, or simply *sine*, *wave*, propagating in a nonabsorbing medium along the positive direction of the OX axis, is

$$s = A \sin\left[\omega\left(t - \frac{x}{v}\right) + \varphi_0\right] = A \sin\left(\omega t - \frac{\omega}{v}\, x + \varphi_0\right),$$

or

$$s = A \sin\left(\frac{2\pi}{T}\, t - \frac{2\pi}{vT}\, x + \varphi_0\right),$$

where $A = $ const is the amplitude of vibration and is called the *wave amplitude*, $\omega = 2\pi/T$ is the *angular frequency of the wave*, T is the period of vibration, and φ_0 is the initial phase of the vibration (at the instant of time $t = 0$) in points of the coordinate plane $x = 0$. The quantity $\Phi = \omega t - \omega x/v + \varphi_0$, equal to the phase of the vibration at an arbitrary point with the coordinate x, is called the *phase of the plane wave* (see also Sect. 30. 2.6).
30.2.6 The distance $\lambda = vT$ that a sine wave travels during the time equal to the period of vibration is called the *wavelength*. The wavelength is equal to the distance between two closest

26*

points in a medium in which the phase difference of the vibration equals 2π.

Along with the wavelength, use is made of another characteristic of a sine wave. It is the *wavenumber*

$$k = \frac{2\pi}{\lambda} = \frac{2\pi}{vT} = \frac{\omega}{v} \ .$$

Hence, the equation of a plane sine wave (30.2.5) can be written in the form

$$s = A \sin \left(\omega t - \frac{2\pi}{\lambda} x + \varphi_0 \right) = A \sin (\omega t - kx + \varphi_0).$$

Accordingly, the phase of this plane wave is $\Phi = \omega t - kx + \varphi_0$.

30.2.7 The *wave vector* is the vector **k**, equal in magnitude to the *wavenumber* k and having the direction along the ray (30.2.1) at the point M being considered in the medium. The wave vector of a plane sine wave does not depend on the choice of point M, and the equation for such a wave can be written in the form

$$s = A \sin (\omega t - \mathbf{kr} + \alpha),$$

where **r** is the radius vector of point M, and α is the initial phase of vibration at the origin of coordinates, i.e. at the point $\mathbf{r} = 0$.

On the basis of the Euler formula (28.1.7), the equation of a plane sine wave can be written in exponential form, which is convenient for differentiation:

$$\widetilde{s} = A e^{i(\omega t - \mathbf{kr} + \delta)},$$

where $\delta = \alpha - \pi/2$. Only the real part of the complex quantity $\widetilde{s}$ is of physical meaning, i.e. the quantity $s = \text{Re} \{\widetilde{s}\}$. Therefore, in making use of $\widetilde{s}$ to determine some characteristic of a wave, it is necessary to discard the imaginary part of the final complex expression.

30.2.8 A wave is said to be *spherical* if its wavefronts are concentric spheres. The centre of these spheres is called the *wave centre*. Such waves are generated in a uniform isotropic medium

by a solitary point source. The *equation of a diverging spherical wave* is of the form

$$\frac{1}{r} f \left(t - \frac{r}{v} \right) ,$$

where r is the distance from the wave centre to the point M being considered in the medium, and v is the wave velocity.
In the case of a *spherical sine wave*

$$\frac{a_0}{r} \sin (\omega t - kr + \alpha),$$

where $a_0/r = A\ (r)$ is the wave amplitude (30.3.6), a_0 is a physical quantity, numerically equal to the wave amplitude at unit distance from the wave centre, α is the initial phase of vibration at the wave centre, and $\Phi = \omega t - kr + \alpha$ is the phase of the spherical wave.
In the exponential form, the equation of a spherical sine wave is

$$\mathit{s} \quad \frac{a_0}{r} e^{i(\omega t - \mathbf{k}\mathbf{r} + \delta)} = \frac{a_0}{r} e^{i(\omega t - kr + \delta)},$$

where $\delta = \alpha - \pi/2$, $\mathbf{r}$ is the radius vector from the wave centre to the point M being considered, and wave vector $\mathbf{k}$ at point M is in the direction radially away from the wave centre.
Real sources of waves are always of finite size. But they can be dealt with as point sources and the waves they generate in a uniform isotropic medium as spherical ones if the distance r from the source to the points being considered in the medium is considerably greater than the dimensions of the source. When r is extremely great, any small portions of a wavefront can be regarded as being plane surfaces.
30.2.9 The propagation of waves in a uniform isotropic medium can be described by a partial differential equation, called the *wave equation*. Thus

$$\frac{\partial^2 s}{\partial x^2} + \frac{\partial^2 s}{\partial y^2} + \frac{\partial^2 s}{\partial z^2} = \frac{1}{v^2} \frac{\partial^2 s}{\partial t^2} \quad \text{or} \quad \Delta s = \frac{1}{v^2} \frac{\partial^2 s}{\partial t^2} .$$

Here s is a physical quantity that characterizes a disturbance propagating in the medium at the velocity v, $\Delta = \partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2$ is the Laplace operator, or Laplacian. This

equation is satisfied, among others, by a plane wave (30.2.7) and a diverging spherical wave (30.2.8).

Function s, characterizing a *sine wave* in a uniform isotropic medium, satisfies two partial differential equations at the same time:

$$\Delta s = -k^2 s,$$

where k is the wavenumber (30.2.6), and

$$\frac{\partial^2 s}{\partial t^2} = -\omega^2 s,$$

where ω is the angular frequency of the wave.

30.2.10 The velocity v of propagation of a *sine wave* is called the *phase velocity*. It is equal to the velocity of displacement in space of the points in a surface that corresponds to any fixed value of the phase of a sine wave. For example, in the case of a plane sine wave (30.2.6) it follows from the condition $\omega t - kx + \varphi_0 = \text{const}$ that $dx/dt = \omega/k = v$. In the same way, in the case of a spherical sine wave (30.2.8), it follows from the condition $\omega t - kr + \alpha = \text{const}$ that $dr/dt = \omega/k = v$.

30.3 Phase Velocity and Energy of Elastic Waves

30.3.1 The *phase velocity of sonic waves (velocity of sound) in a liquid or gas* is

$$v = \sqrt{\frac{K}{\rho}},$$

where ρ is the density of the undisturbed medium, and K is the bulk modulus of elasticity of the medium (30.1.1). The frequency of audible sonic waves $\nu > 16$ Hz, and the process of deforming the medium can be assumed to be adiabatic (8.3.7), i.e. $K = -V \left(dp/dV \right)_{\text{ad}}$. The relation between the pressure p and volume V for an ideal gas (8.4.1) in an adiabatic process is: $pV^\varkappa = \text{const}$, where $\varkappa$ is the adiabatic exponent (9.5.11), so that $K = \varkappa p$. Therefore, the *velocity of sound in an ideal gas* is equal to

$$v = \sqrt{\frac{\varkappa p}{\rho}} = \sqrt{\frac{\varkappa RT}{\mu}},$$

where μ is the molar mass of the gas, T is its absolute temperature, and R is the universal gas constant.

30.3.2 The *phase velocity of transverse elastic waves* in a uniform isotropic solid medium is

$$v = \sqrt{\frac{G}{\rho}} ,$$

where G is the shear modulus of the medium (40.3.9) and ρ is its density.

The propagation of longitudinal waves in a *thin long rod* is associated with its longitudinal tension and compression. Hence, the phase velocity of such waves is

$$v = \sqrt{\frac{E}{\rho}} ,$$

where E is Young's modulus (40.3.6) for the rod material.

The *velocity of propagation of transverse waves along a stretched string (a thin flexible cord)* is

$$v = \sqrt{\frac{F}{\rho S}} , \qquad \cdot$$

where F is the stretching force (tension) in the string, and ρ and S are the density of the string material and the cross-sectional area of the string.

30.3.3 An elastic medium in which mechanical waves propagate has kinetic energy of vibratory motion of its particles and potential energy due to deformation (30.1.1). If v_1 is the velocity of particles of the medium, then the *volume kinetic energy density of the medium* is

$$E_k^d = \frac{dE_k}{dV} = \frac{1}{2} \rho v_1^2,$$

where ρ is the density of the medium, and dE_k is the kinetic energy of all the particles in the small volume dV of the medium. This volume has been selected so that throughout it the velocity v_1 is the same.

The *volume potential energy density of* an elastically deformed *medium* is

$$E_p^d = \frac{dE_p}{dV} = \frac{1}{2} \rho v^2 \varepsilon^2,$$

where dE_p is the potential energy of a uniformly deformed small portion of the medium of volume dV, v is the phase velocity in the medium, and ε is the relative deformation.

The *volume energy density of elastic waves* is understood to be the volume density E^d of the mechanical energy of a medium due to the propagation of these waves. Thus

$$E^d = E_k^d + E_p^d = \frac{1}{2} \rho (v_1^2 + v^2 \varepsilon^2).$$

30.3.4 If a longitudinal plane travelling wave (30.2.4) propagates in a medium, then $v_1 = \partial s/\partial t$, where s is the displacement of the particles, and $\varepsilon = \partial s/\partial x = -v_1/v$, so that

$$E_p^d = E_k^d \quad \text{and} \quad E^d = \rho v_1^2 = \rho \left(\frac{\partial s}{\partial t} \right)^2.$$

At each point of the medium subject to wave motion, E_k^d and E_p^d are *alike* functions of time. Consequently, E^d also varies with time. This principle is valid for any travelling waves in an elastic medium, regardless of the shape of their wavefronts and the type of deformation of the medium. It follows from the law of conservation of energy as applied to the process of the propagation of vibrations in an elastic medium. To set regions of the medium more and more distant from the wave source into vibratory motion it is necessary to expend energy, which is supplied to the medium by the source. Consequently, the propagation of elastic waves is inseparably associated with the transport of energy from certain parts of the medium to others. This is precisely why the volume energy density E^d of the waves depends both on the coordinates and on the time as well.

For a plane travelling sine wave in a nonabsorbing medium
(30.2.5)

$$E^d = \rho A^2 \omega^2 \cos^2 (\omega t - kx + \varphi_0)$$

$$= \frac{1}{2} \rho A^2 \omega^2 [1 + \cos 2 (\omega t - kx + \varphi_0)],$$

where $A = $ const is the wave amplitude.
In the case of a diverging spherical sine wave in a nonabsorbing
medium (30.2.8)

$$E^d = \rho A^2 \omega^2 \cos^2 (\omega t - kr + \alpha),$$

where $A = a_0/r$ is the amplitude of
the wave.
The average volume energy density
over one period is

$$\langle E^d \rangle = \frac{1}{2} \rho A^2 \omega^2.$$
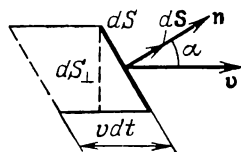
Fig. 30.1

30.3.5 The *velocity of energy transfer
by a wave* is equal to the velocity in
space of a surface corresponding to the maximum value of
the volume density E^d of the wave's energy. For sine waves
this velocity is equal to the phase velocity v.
The *energy flux* $d\Phi_E$ through the small surface dS is the ratio
of the energy dE transferred through this surface in a short
time interval to the magnitude of this interval dt. Thus

$$d\Phi_E = dE/dt.$$

If $\mathbf{v}$ is the velocity vector of energy transfer by the wave
(Fig. 30.1), then

$$dE = E^d v \, dt \, dS \cos \alpha = E^d (\mathbf{v} \, d\mathbf{S}) \, dt$$

and

$$d\Phi_E = E^d (\mathbf{v} \, d\mathbf{S}) = (\mathbf{U} \, d\mathbf{S}),$$

where E^d is the volume energy density of the wave, $d\mathbf{S} = \mathbf{n} \, dS$
is the vector of surface dS, $\mathbf{n}$ is a unit vector of the normal to
the surface, and α is the angle between $\mathbf{v}$ and $d\mathbf{S}$.

The vector $U = E^d v$, having the direction of energy transfer by the wave, is called the *Umov vector* (*energy flux density vector of the wave*). In magnitude it equals the ratio of the energy flux $d\Phi_E$ through the small surface dS to the area $dS_\perp = d\vec{S} \cos\alpha$ of the projection of this surface on a plane perpendicular to the direction of energy transfer: $U = d\Phi_E/dS_\perp$.

30.3.6 The *wave intensity* I is the magnitude of the average value of the Umov vector. Wave intensity is numerically equal to the energy transferred by the wave in unit time through unit area of a surface perpendicular to the direction of wave propagation. The intensity of a travelling sine wave is proportional to the square of its amplitude. For plane and spherical sine waves

$$I = |\langle U \rangle| = v \langle E^d \rangle = \frac{1}{2} \rho v \omega^2 A^2.$$

When a spherical wave propagates in a nonabsorbing medium, the same amount of energy is transferred in unit time through any spherical surface of radius r with its centre at the wave centre. This energy is equal to that expended during the same time by the wave source: $4\pi r^2 I = $ const. Hence, the intensity and amplitude of a spherical wave decrease with an increase in the distance from the wave centre according to the relations

$$I(r) = \frac{i_0}{r^2} \quad \text{and} \quad A(r) = \frac{a_0}{r},$$

where i_0 and a_0 are physical quantities numerically equal to the intensity and amplitude of the wave at the distance $r = 1$ m from the wave centre.

It can be shown in the same way that in the case of a plane sine wave in a nonabsorbing medium, the amplitude A of the wave is independent of the coordinates.

30.3.7 The conversion of wave energy into other forms of energy, taking place in the propagation of waves through a medium, is called *wave absorption*. In a uniform medium the absorption of elastic waves is due chiefly to the processes of internal friction, or viscosity (10.8.4), and heat conduction (10.8.5). The amplitude A and intensity I of a plane wave, propagating in an absorbing

medium along the positive direction of axis OX, varies according to the exponential law

$$A\,(x) = A_0 e^{-\alpha x} \quad \text{and} \quad I\,(x) = I_0 e^{-2\alpha x}.$$

Here A_0 and I_0 are the amplitude and intensity of the wave at the points $x = 0$, and α is the *linear absorption coefficient for elastic waves* and depends upon the properties of the medium and the wave frequency.

30.3.8 *The dispersion of waves* is the dependence of the phase velocity of sine waves in a medium on their frequency. A medium in which this phenomenon is observed is said to be a *dispersive medium*. The dispersion of acoustic waves in a boundless medium depends upon the properties of the medium and is always accompanied by the absorption of sound.

30.4 Principle of Superposition of Waves. Group Velocity

30.4.1 *The principle of superposition of waves* states that in a linear medium (30.1.7) waves are propagated independently of one another, so that the resultant disturbance at any point in the medium upon the simultaneous propagation of several waves is equal to the sum of the disturbances corresponding to each of the waves separately.

For example, if n different mechanical waves are propagated simultaneously through a linear medium the resultant displacement s, velocity v and acceleration a of the particles of the medium at the arbitrary instant t of time are

$$\mathrm{s} = \sum_{i=1}^{n} \mathrm{s}_i, \quad \mathrm{v} = \sum_{i=1}^{n} \mathrm{v}_i \quad \text{and} \quad \mathrm{a} = \sum_{i=1}^{n} \mathrm{a}_i,$$

where s_i, v_i and a_i are the values of the displacement, velocity and acceleration that the particles being considered would have at the instant of time t if only the ith wave was propagated in the medium.

30.4.2 On the basis of the principle of wave superposition and a Fourier expansion (28.4.6 and 28.4.7) any nonsine wave can be replaced by an equivalent system of sine waves, i.e. it can be represented in the form of a *group of waves*, called a *wave train*,

or packet. The totality of frequency values for these sine waves
is called the *frequency spectrum* (or simply *spectrum*) of the
nonsine wave. In accordance with the nature of the oscillation
excited by the wave, the wave's frequency spectrum can be
either discrete (28.4.6) or continuous (28.4.7).

The law of propagation of an arbitrary disturbance (signal), which
is a nonsine wave, in a linear medium is simple only provided
that the medium is nondispersive (30.3.8). In this case the signal
travels in the medium without changing its "shape" because all
the sine waves making up the wave train have the same phase
velocity, which is equal to the velocity of the signal.

In a dispersive medium, the sine components of the wave train,
corresponding to the nonsine wave, propagate at different
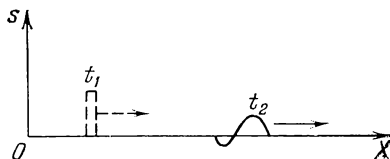
Fig. 30.2

velocities. Therefore, this wave train "spreads" as it propagates,
thereby changing the "shape" of the signal. If, for example, a
signal propagating along the OX axis in a dispersive medium
had the "shape" shown by dash lines in Fig. 30.2 at the instant
of time t_1, then at the instant $t_2 > t_1$ it could have the different
"shape" shown by a solid line.

30.4.3 The simplest wave train is the *plane quasi-sine wave*
obtained by the superposition of two plane waves of the same
amplitude and with frequencies and wavenumbers close in
value propagating along the OX axis. Thus

$$s = A_0 \sin (\omega t - kx) + A_0 \sin [(\omega + d\omega)\, t - (k + dk)\, x]$$
$$= 2A_0 \cos \left(\frac{t\, d\omega - x\, dk}{2} \right) \sin (\omega t - kx).$$

The dependence of s on x at some fixed instant of time is shown in Fig. 30.3. This wave differs from a sine wave in that its amplitude

$$A = 2A_0 \left| \cos \left(\frac{t\, d\omega - x\, dk}{2} \right) \right|$$

is a slowly varying function of the coordinate x and time t. Taken as the velocity of propagation of this nonsine wave is a
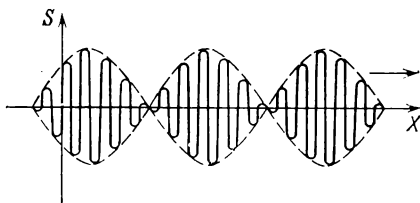
Fig. 30.3

velocity u of point M in which the amplitude A has some fixed value (for example, $A = 0$ or $A = 2A_0$). Consequently, the law of motion of point M is: $t\, d\omega - x\, dk = $ const. From this

$$u = \frac{dx}{dt} = \frac{d\omega}{dk}.$$

Velocity u is called the *group velocity*. It is equal to the velocity of energy transfer by a quasi-sine wave. The group velocity $u = d\omega/dk$ is suitable for describing energy transfer (signal transmission) by nonsine waves having a different frequency spectrum provided that the spectrum is not very wide and that the dispersion of waves in the medium at these frequencies is not too great.

The relation between the group velocity ($u = d\omega/dk$) and the phase velocity ($v = \omega/k$) of waves is of the form:

$$u = v - \lambda\, \frac{dv}{d\lambda},$$

where λ is the wavelength (30.2.6). In a nondispersive medium $dv/d\lambda = 0$ and the group and phase velocities coincide.

30.5 Interference of Waves. Standing Waves

30.5.1 Two waves are said to be *coherent* if their phase difference is independent of time. Coherent waves correspond to coherent

Fig. 30.4

oscillations (or vibrations) (28.4.3). The sources of coherent waves are called *coherent sources*. Sine waves of equal frequency are always coherent. Waves of different frequencies are coherent only for the length of time that the oscillations (or vibrations) they excite are coherent (28.4.4). In superposing incoherent sine waves, generated by point sources S_1 and S_2 (Fig. 30.4), the square of the amplitude A of the resultant nonharmonic vibration at arbitrary point M varies periodically with time t according to the relation (see Sects. 28.4.2 and 30.2.8):
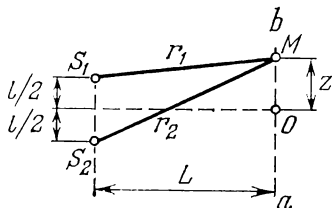
$$A^2 = A_1^2 + A_2^2 + 2A_1 A_2 \cos\left[(\omega_2 - \omega_1)\, t\right.$$
$$\left. - (k_2 r_2 - k_1 r_1) + (\alpha_2 - \alpha_1)\right].$$

Here A_1 and A_2, ω_1 and ω_2, k_1 and k_2, and α_1 and α_2 are the amplitudes at point M, angular velocities, wavenumbers and the initial phases of the two spherical waves being superposed. The period of variation of A^2 is equal to $T = 2\pi/|\,\omega_2 - \omega_1\,|$. The average value of the square of the amplitude during the period is

$$\langle A^2 \rangle = A_1^2 + A_2^2.$$

In the superposition of incoherent waves, the squares of their amplitudes are added together.

30.5.2 The *interference of waves* is the phenomenon, in superposing waves, of their mutual strengthening, stable with time, at certain points in space and their weakening at others, depending upon the relations between the phases of these waves.

Only coherent waves display interference, provided that they correspond to vibrations in the same or almost the same direction.

In superposing coherent spherical waves, generated by the point sources S_1 and S_2 (Fig. 30.4),

$$s_1 = A_1 \sin (\omega t - kr_1 + \alpha_1) = A_1 \sin \Phi_1$$

and

$$s_2 = A_2 \sin (\omega t - kr_2 + \alpha_2) = A_2 \sin \Phi_2;$$

amplitude A and phase Φ of the resultant harmonic vibrations at point M (where $s = s_1 + s_2 = A \sin \Phi$) are determined from the relations (28.4.3):

$$A^2 = A_1^2 + A_2^2 + 2A_1A_2 \cos [k (r_2 - r_1) - (\alpha_2 - \alpha_1)],$$

$$\tan \Phi = \frac{A_1 \sin \Phi_1 + A_2 \sin \Phi_2}{A_1 \cos \Phi_1 + A_2 \cos \Phi_2}.$$

Since the difference of the initial phases $\alpha_2 - \alpha_1 = $ const for the vibration of the coherent sources S_2 and S_1, the resulting interference of the two waves at various points M depends on the quantity $\Delta = r_2 - r_1$, called the *path difference of the waves*. At interference maxima, the amplitude of the resultant vibration is $A = A_1 + A_2$, whereas at interference minima $A = | A_1 - A_2 |$. Maxima are observed at the points M that comply with the condition

$$k \Delta - (\alpha_2 - \alpha_1) = \pm 2m\pi,$$

where $m = 0, 1, 2, \ldots$ is the *order of the interference maximum*. The condition for interference minima is of the form

$$k \Delta - (\alpha_2 - \alpha_1) = \pm (2m - 1) \pi,$$

where $m = 1, 2, 3, \ldots$ is the *order of the interference minimum*. Since the wavenumber $k = 2\pi/\lambda$, where λ is the wavelength in the given medium, the *conditions for interference maxima and minima* can be presented in the form:

$$\Delta = \pm m\lambda + \frac{\alpha_2 - \alpha_1}{2\pi} \lambda \quad \text{(maxima)},$$

$$\Delta = \pm (2m - 1) \frac{\lambda}{2} + \frac{\alpha_2 - \alpha_1}{2\pi} \lambda \quad \text{(minima)}.$$

Finally, when $\alpha_2 = \alpha_1$, the conditions take the form: $\Delta = \pm m\lambda$ (maxima) and $\Delta = \pm (2m - 1) \lambda/2$ (minima).

On the straight line ab, parallel to and at the distance L from the line passing through the sources S_1 and S_2, (Fig. 30.4), the central maximum of zero order is at point O, equidistant from sources S_1 and S_2. If the distance between the sources $l \ll L$, then for point M on line ab, at the distance $z \ll L$ from point O, the path difference of the waves is

$$\Delta = \frac{lz}{L}.$$

Maxima of the mth and $(m + 1)$th orders correspond to the values

$$z_m = \frac{m\lambda L}{l} \quad \text{and} \quad z_{m+1} = \frac{(m+1)\,\lambda L}{l},$$

so that the difference between adjacent maxima equals $\lambda L/l$.

30.5.3 When waves interfere, their energies cannot be simply added together. At interference maxima, the intensity of the resultant wave is greater than the sum of the intensities of the superposed waves, and at minima, it is less than their sum. Wave interference leads to a redistribution of vibrational energy between neighbouring regions of the medium. But on an average, for a large region of space, the energy of the resultant wave is equal to the sum of the energies of the interfering waves. This follows from the laws of conservation and conversion of energy.

30.5.4 A special case of wave interference is standing waves. A *standing wave* is one that is formed by the superposition of two travelling sine waves that are propagating toward each other and have the same frequency and amplitude. In the case of transverse waves, they also have the same polarization (30.2.2). A standing wave is formed, for instance, in a stretched flexible string when one end is fixed and vibratory motion is imparted to the other end.

In the superposition of two coherent travelling plane waves of the form

$$s_1 = A \sin (\omega t - kx) \quad \text{and} \quad s_2 = A \sin (\omega t + kx + \alpha),$$

where α is the phase difference at points $x = 0$, a *plane standing wave* is formed. It is described by the equation

$$s = s_1 + s_2 = 2A \cos \left(kx + \frac{\alpha}{2} \right) \sin \left(\omega t + \frac{\alpha}{2} \right).$$

In contrast to the amplitude A of travelling waves, the *amplitude A_{st} of a standing wave* is a periodic function of the coordinate x:

$$A_{st} = 2A \left| \cos \left(kx + \frac{\alpha}{2} \right) \right|.$$

30.5.5 Points of a standing wave at which $A_{st} = 0$ are called its *nodes*; points at which the amplitude A_{st} reaches its maximum value $(A_{st} = 2A)$ are called the *antinodes*. The positions of the nodes and antinodes are determined from the conditions:

$$kx + \frac{\alpha}{2} = (2m + 1) \frac{\pi}{2} \quad \text{(nodes)},$$

$$kx + \frac{\alpha}{2} = m\pi \quad \text{(antinodes)},$$

where $m = 0, 1, 2, \ldots$.
The distances between two adjacent nodes and between two adjacent antinodes are the same and equal to half the wavelength λ of the travelling waves. This distance is called the *length of the standing wave*: $\lambda_{st} = \lambda/2$. The distance between adjacent nodes and antinodes of a standing wave is equal to $\lambda_{st}/2$.
30.5.6 In a travelling wave the phase of the vibration depends upon the coordinate x of the point being considered. In a standing wave all the points between two nodes vibrate with different amplitudes, but in the *same* phase (cophased) because the argument of the sine in the equation of a standing wave (30.5.4) is independent of the coordinate x. In passing through a node, the phase of the vibration changes abruptly by π, so that the sign of $\cos (kx + \alpha/2)$ is reversed.
Illustrated in Fig. 30.5 is the nature of the motion of various points on a stretched flexible string of length l when a transverse standing wave is established in it. The left end O of the string is set into harmonic motion and the right end N is fixed.

27--0186

In this case, as the wave is reflected from the place where the
string is fixed, its phase is changed by π, so that a node of the
standing wave is formed at this point. The phase difference
of the reflected and incident waves at point O (where $x = 0$) is
$\alpha = -(2kl + \pi)$. The nodes of the standing wave are designated
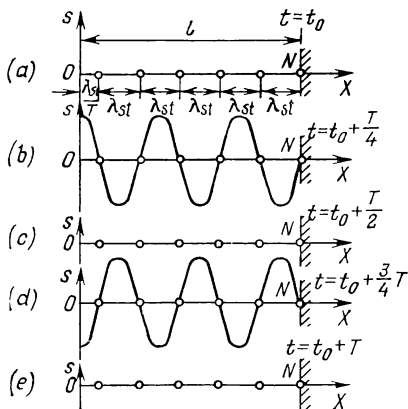
Fig. 30.5

by small circles in Fig. 30.5, and the instant of time t_0 is selected
so that $\sin(\omega t_0 + \alpha/2) = 0$.

30.5.7 In a standing wave (30.5.4), the velocity of vibratory
motion of the particles of the medium is

$$v_1 = \frac{\partial s}{\partial t} = 2A\omega \cos\left(kx + \frac{\alpha}{2}\right) \cos\left(\omega t + \frac{\alpha}{2}\right),$$

and the relative deformation of the medium is

$$\varepsilon = \frac{\partial s}{\partial t} = -2Ak \sin\left(kx + \frac{\alpha}{2}\right) \sin\left(\omega t + \frac{\alpha}{2}\right)$$

$$= 2Ak \sin\left(kx + \frac{\alpha}{2}\right) \cos\left(\omega t + \frac{\alpha}{2} + \frac{\pi}{2}\right).$$

Thus, in contrast to a travelling wave (30.3.4), in a standing wave ε has a phase lead of $\pi/2$ over v_1. Consequently, at the instants of time when v_1 reaches its amplitude value, ε becomes zero and vice versa. Moreover, the amplitudes of v_1 and ε depend upon the coordinate x, but in different ways. Thus, the antinodes of the particle velocity and nodes of deformation of the medium are at the antinodes of the standing wave (30.5.5), whereas the antinodes of the deformation and the nodes of the velocity are at the nodes of the standing wave.

In an elastic standing wave, the energy is periodically converted from potential energy, localized mainly near the deformation antinodes, into kinetic energy, localized mainly near the velocity antinodes, and back again. Therefore, energy periodically migrates from the nodes of a standing wave to its antinodes and back again. But in the nodes and antinodes themselves the energy flux density is identically equal to zero. The average value of the energy flux density over one period is equal to zero at any point on a standing wave, because the two travelling waves that form the standing wave transmit equal amounts of energy over a period in opposite directions. This is why they have been called standing waves.

30.5.8 In the free vibrations of strings, rods and columns of gas, standing waves are set up in them. The frequencies of these waves comply with definite conditions, i.e. they can have only certain discrete values, which are called the *natural frequencies of vibration* of the corresponding vibratory system. Displacement nodes (deformation antinodes) are located at the rigidly fixed ends of strings or rods, and displacement antinodes (deformation nodes) are located at the free ends of the rods. In the vibrations of a cylindrical column of gas in a pipe, a pressure antinode is located at the closed end of the pipe and a pressure node at the open end.

If l is the length of the string, rod or gas column, v is the phase velocity of the wave and λ is the wavelength, then for strings and rods fixed at both ends and for gas columns in pipes that are open or closed at both ends, a whole number of lengths $\lambda_{st} = \lambda/2$ of the standing wave fits into the length l, i.e.

$l = m\lambda_{st} = m\lambda/2$, where $m = 1, 2, 3, \ldots$

The natural frequency of vibrations of such systems is

$$\nu = \frac{mv}{2l} \cdot$$

27*

For rods fixed at one end and free at the other, or for pipes closed at one end and open at the other,

$$l = (2m-1)\frac{\lambda_{st}}{2} = (2m-1)\frac{\lambda}{4}$$

and the natural frequency of vibration is

$$\nu = \frac{(2m-1)\,v}{4l}\,.$$

30.6 Doppler Effect in Acoustics

30.6.1 The *Doppler effect* refers to the change in the frequency of waves, observed or recorded by a receiver, or detector, which is due to motion of the source of the waves and the detector (or observer). For example, as a rapid train approaches a stationary observer, the pitch of the locomotive whistle is higher, and as the train passes the observer and travels away it is lower, than the pitch of the same whistle when the train is standing still.

Assume that the detector D of acoustic waves in a gaseous (or liquid) medium is fixed with respect to the medium, and that the source S is moving away at the velocity v_1 along a straight line joining them (Fig. 30.6a). During the period T_0 of its vibration, the source travels in the medium a distance of $v_1 T_0 = v_1/\nu_0$, where ν_0 is the source's frequency of vibration. There-
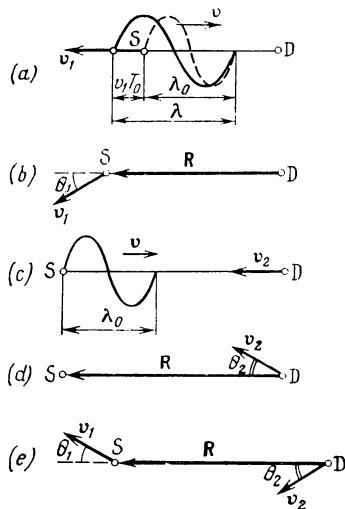
Fig. 30.6

fore, upon motion of the source the wavelength λ in the medium differs from its value λ_0 when the source is stationary:

$$\lambda = \lambda_0 + v_1 T_0 = (v + v_1)\,T_0 = (v + v_1)/\nu_0,$$

where v is the phase velocity of the wave in the medium. The frequency of the wave recorded by the detector is

$$\nu = \frac{v}{\lambda} = \frac{\nu_0}{1 + v_1/v} \; .$$

If the velocity vector $\mathbf{v}_1$ of the source makes the arbitrary angle θ_1 with the radius vector $\mathbf{R}$, connecting the stationary detector to the source (Fig. 30.6b), then

$$\nu = \frac{\nu_0}{1 + v_1 \cos \theta_1/v} \; .$$

30.6.2 When the source is stationary and the detector approaches it at the velocity $\mathbf{v}_2$ along a straight line joining them (Fig. 30.6c), the wavelength in the medium is $\lambda = \lambda_0 = v/\nu_0$. But the velocity of propagation of the wave with respect to the detector is $v + v_2$. Hence, the wave frequency recorded by the detector is

$$\nu = \frac{v + v_2}{\lambda_0} = \nu_0 \left(1 + \frac{v_2}{v} \right).$$

In the case when the velocity vector $\mathbf{v}_2$ makes the arbitrary angle θ_2 with radius vector $\mathbf{R}$, connecting the travelling detector to the stationary source (Fig. 30.6d),

$$\nu = \nu_0 \left(1 + \frac{v_2}{v} \cos \theta_2 \right).$$

30.6.3 In the most general case, when both the detector and source of acoustic waves are travelling with respect to the medium at arbitrary velocities (Fig. 30.6e),

$$\nu = \nu_0 \frac{1 + \dfrac{v_2}{v} \cos \theta_2}{1 + \dfrac{v_1}{v} \cos \theta_1} \; .$$

This formula can also be presented in the form

$$\nu = \nu_0 \left\{ 1 - \frac{V \cos \theta}{v} \left[1 - \frac{v_1}{v} \cos \theta_1 + \left(\frac{v_1}{v} \cos \theta_1 \right)^2 - \ldots \right] \right\} ,$$

where $V = v_1 - v_2$ is the velocity of the wave source with respect to the detector, and θ is the angle between vectors V and R. The quantity $V \cos \theta$, equal to the projection of V on the direction of R, is called the *ray velocity of the source*. When $v_1 \ll v$,

$$\nu \approx \nu_0 \left(1 - \frac{V \cos \theta}{v} \right).$$

CHAPTER 31   ELECTROMAGNETIC WAVES

31.1 Properties of Electromagnetic Waves

31.1.1 *Electromagnetic waves* are disturbances of an electromagnetic field (i.e. a variable electromagnetic field) that propagate through space. The statement of the existence of electromagnetic waves follows directly from Maxwell's equations (27.4.3). For an electromagnetic field, distant from the free electric charges (18.3.1) and macroscopic currents (26.4.1) that it was set up by, these equations are of the form*

$$\operatorname{curl} E = -\frac{\partial B}{\partial t}, \quad \operatorname{curl} H = \frac{\partial D}{\partial t},$$

$$\operatorname{div} D = 0, \qquad \operatorname{div} B = 0.$$

31.1.2 If the medium is a uniform and isotropic dielectric (30.1.6), having no ferroelectric (18.4.1) or ferromagnetic (26.5.1) properties, then $D = \varepsilon_r \varepsilon_0 E$ and $B = \mu_r \mu_0 H$, where ε_r and μ_r are constant scalar quantities independent both of the coordinates and of time. Here Maxwell's equations (31.1.1) can be rewritten in the form

$$\operatorname{curl} E = -\mu_r \mu_0 \frac{\partial H}{\partial t}, \quad \operatorname{curl} H = \varepsilon_r \varepsilon_0 \frac{\partial E}{\partial t}, \quad \operatorname{div} E = 0$$

and $\operatorname{div} H = 0$,

---

* All equations in Chap. 31 are written in SI units.

or in projections on the axes of Cartesian coordinates:

$$\frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z} = -\mu_r\mu_0\frac{\partial H_x}{\partial t}\ , \qquad \frac{\partial H_z}{\partial y} - \frac{\partial H_y}{\partial z} = \varepsilon_r\varepsilon_0\frac{\partial E_x}{\partial t}\ ,$$

$$\frac{\partial E_x}{\partial z} - \frac{\partial E_z}{\partial x} = -\mu_r\mu_0\frac{\partial H_y}{\partial t}\ , \qquad \frac{\partial H_x}{\partial z} - \frac{\partial H_z}{\partial x} = \varepsilon_r\varepsilon_0\frac{\partial E_y}{\partial t}\ ,$$

$$\frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y} = -\mu_r\mu_0\frac{\partial H_z}{\partial t}\ , \qquad \frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} = \varepsilon_r\varepsilon_0\frac{\partial E_z}{\partial t}\ ,$$

$$\frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} = 0, \qquad \frac{\partial H_x}{\partial x} + \frac{\partial H_y}{\partial y} + \frac{\partial H_z}{\partial z} = 0.$$

31.1.3 It follows from Maxwell's equations (31.1.2) that the strength vectors **E** and **H** of a variable electromagnetic field and their projections onto the axes of Cartesian coordinates comply in a uniform, isotropic, nonconducting medium, with the wave equation (30.2.9):

$$\Delta\mathbf{E} - \varepsilon_r\varepsilon_0\mu_r\mu_0\ \frac{\partial^2\mathbf{E}}{\partial t^2} = 0, \qquad \Delta\mathbf{H} - \varepsilon_r\varepsilon_0\mu_r\mu_0\ \frac{\partial^2\mathbf{H}}{\partial t^2} = 0,$$

$$\Delta E_x - \varepsilon_r\varepsilon_0\mu_r\mu_0\ \frac{\partial^2 E_x}{\partial t^2} = 0, \qquad \Delta E_y - \varepsilon_r\varepsilon_0\mu_r\mu_0\ \frac{\partial^2 E_y}{\partial t^2} = 0,$$

$$\Delta E_z - \varepsilon_r\varepsilon_0\mu_r\mu_0\ \frac{\partial^2 E_z}{\partial t^2} = 0,$$

$$\Delta H_x - \varepsilon_r\varepsilon_0\mu_r\mu_0\ \frac{\partial^2 H_x}{\partial t^2} = 0, \qquad \Delta H_y - \varepsilon_r\varepsilon_0\mu_r\mu_0\ \frac{\partial^2 H_y}{\partial t^2} = 0,$$

$$\Delta H_z - \varepsilon_r\varepsilon_0\mu_r\mu_0\ \frac{\partial^2 H_z}{\partial t^2} = 0.$$

Thus a variable electromagnetic field really propagates in space in the form of waves with the phase velocity

$$v = \frac{c}{\sqrt{\varepsilon_r\mu_r}}\ , \quad \text{where} \quad c = \frac{1}{\sqrt{\varepsilon_0\mu_0}} = 3 \times 10^8 \text{ m/s}.$$

In vacuum $\varepsilon_r = \mu_r = 1$. Hence, c is the *velocity of electromagnetic waves* in vacuum (in free space).

31.1.4 Electromagnetic waves are *transverse* waves: vectors **E** and **H** of the field lie in a plane perpendicular to the direction of wave propagation, i.e. to the wave velocity vector **v** at the point being considered in the field. This can be most readily demonstrated by the example of a plane wave, propagating along the positive direction of axis OX (30.2.4):

$$\mathbf{E} = f\left(t - \frac{x}{v}\right) \quad \text{and} \quad \mathbf{H} = \varphi\left(t - \frac{x}{v}\right).$$

Vector **E** and **H** and their projections on the coordinate axes are independent of y and z:

$$\frac{\partial E_x}{\partial y} = \frac{\partial E_x}{\partial z} = \frac{\partial E_y}{\partial y} = \frac{\partial E_y}{\partial z} = \frac{\partial E_z}{\partial y} = \frac{\partial E_z}{\partial z} = 0$$

and

$$\frac{\partial H_x}{\partial y} = \frac{\partial H_x}{\partial z} = \frac{\partial H_y}{\partial y} = \frac{\partial H_y}{\partial z} = \frac{\partial H_z}{\partial y} = \frac{\partial H_z}{\partial z} = 0.$$

It follows from Maxwell's equations (31.1.2) that for the field of a plane wave,

$$\frac{\partial E_x}{\partial} = \frac{\partial E_x}{\partial t} = 0 \quad \text{and} \quad \frac{\partial H_x}{\partial x} = \frac{\partial H_x}{\partial t} = 0,$$

i.e. E_x and H_x are independent of both coordinates and time.

Fig. 31.1

Hence, for a *variable* field of a plane wave $E_x = H_x = 0$, and the vectors **E** and **H** are perpendicular to the direction of wave propagation:

$$\mathbf{E} = E_y \mathbf{j} + E_z \mathbf{k} \quad \text{and} \quad \mathbf{H} = H_y \mathbf{j} + H_z \mathbf{k},$$

where **j** and **k** are unit vectors of the coordinate axes.

31.1.5 Vectors **E** and **H** of the field of an electromagnetic wave are mutually perpendicular, so that **v**, **E** and **H** form a right-hand set of three vectors (Fig. 31.1). As a matter of fact, for the field of a plane wave (31.1.4),

$$E_y = f_1\left(t - \frac{x}{v}\right), \quad E_z = f_2\left(t - \frac{x}{v}\right),$$
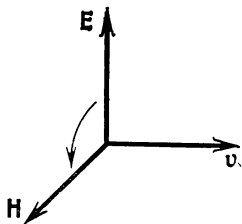
$$H_y = \varphi_1 \left(t - \frac{x}{v} \right) \quad \text{and} \quad H_z = \varphi_2 \left(t - \frac{x}{v} \right).$$

It follows from Maxwell's equations (31.1.2) that

$$\sqrt{\varepsilon_r \varepsilon_0} \frac{dE_y}{d\xi} = \sqrt{\mu_r \mu_0} \frac{dH_z}{d\xi} \quad \text{and}$$

$$\sqrt{\varepsilon_r \varepsilon_0} \frac{dE_z}{d\xi} = - \sqrt{\mu_r \mu_0} \frac{dH_y}{d\xi},$$

where $\xi = t - x/v$. Therefore, for a *variable* field of a plane wave

$$H_y = - \sqrt{\frac{\varepsilon_r \varepsilon_0}{\mu_r \mu_0}} E_z, \qquad H_z = \sqrt{\frac{\varepsilon_r \varepsilon_0}{\mu_r \mu_0}} E_y \quad \text{and} \quad (\mathbf{EH}) = 0.$$

The mutually perpendicular vectors $\mathbf{E}$ and $\mathbf{H}$ oscillate in phase: they become zero simultaneously and reach their maximum values simultaneously. The relation between their magnitudes is

$$\sqrt{\varepsilon_r \varepsilon_0} \, E = \sqrt{\mu_r \mu_0} \, H,$$

which is valid for any travelling electromagnetic wave (30.2.1) regardless of the shape of its wavefronts (30.2.3).

31.1.6 A *monochromatic wave* is an electromagnetic wave of a single definite frequence ν, i.e. an electromagnetic sine wave. At each point of the electromagnetic field of a monochromatic wave, the projections of vectors $\mathbf{E}$ and $\mathbf{H}$ on the coordinate axes of an inertial frame of reference have harmonic oscillations (28.1.3) of the same frequency, equal to the wave frequency ν. For example, in the case of a *plane monochromatic wave*, propagating along the positive direction of axis OX,

$$E_y = A_1 \sin (\omega t - kx), \qquad H_y = - \sqrt{\frac{\varepsilon_r \varepsilon_0}{\mu_r \mu_0}} E_z,$$

$$E_z = A_2 \sin (\omega t - kx + \varphi), \qquad H_z = \sqrt{\frac{\varepsilon_r \varepsilon_0}{\mu_r \mu_0}} E_y.$$

where $\omega = 2\pi\nu$ is the angular frequency of the wave, k is the wavenumber (30.2.6), A_1 and A_2 are the amplitudes of E_y and E_z, and φ is the phase difference between the oscillations of E_z and E_y.

31.1.7 For an arbitrary value of φ (31.1.6), a plane monochromatic wave is *elliptically polarized*, i.e. at each point of the wave's field, vectors E and H, while remaining mutually perpendicular, vary with time so that their heads describe ellipses lying in a plane perpendicular to the direction of wave propagation:

$$\frac{E_y^2}{A_1^2} + \frac{E_z^2}{A_2^2} - \frac{2E_yE_z}{A_1A_2} \cos \varphi = \sin^2 \varphi,$$

and

$$\frac{H_y^2}{A_2^2} + \frac{H_z^2}{A_1^2} + \frac{2H_yH_z}{A_1A_2} \cos \varphi = \frac{\varepsilon_r\varepsilon_0}{\mu_r\mu_0} \sin^2 \varphi.$$

In particular, when $A_1 = A_2$ and $\varphi = \pm (2m + 1) \pi/2$, where $m = 0, 1, 2, \ldots$, the ellipses become circles:

$$E_y^2 + E_z^2 = A_1^2 \quad \text{and} \quad H_y^2 + H_z^2 = \frac{\varepsilon_r\varepsilon_0}{\mu_r\mu_0} A_1^2.$$

Such a wave is said to be *circularly polarized*.
When $\varphi = \pm m\pi$, where $m = 0, 1, 2, \ldots$, the ellipses degenerate into straight lines:

$$\frac{E_y}{A_1} \pm \frac{E_z}{A_2} = 0 \quad \text{and} \quad \frac{H_y}{A_2} \mp \frac{H_z}{A_1} = 0.$$

Such a wave is said to be *linearly*, or *plane*, *polarized*. Shown in Fig. 31.2 are the values of E and H of the field of a plane, linearly polarized, monochromatic wave at various points along the ray (OX axis) at the same instant of time. Axes OY and OZ are in the directions of oscillation of vectors E and H, respectively, so that $E_z = H_y = 0$. According to accepted terminology, the plane passing through the electric vector E and the ray is called the *plane of polarization* of the linearly polarized wave. Previously, this plane was called the *plane of wave oscillation*, and the plane of polarization was called the one passing through the magnetic vector H and the ray (this obsolete terminology may still be found in the literature).

31.1.8 An arbitrary plane monochromatic wave can be represented as a combination of two plane monochromatic waves of the same frequency as the first wave and simultaneously propagating in the same direction as the first wave, but linearly
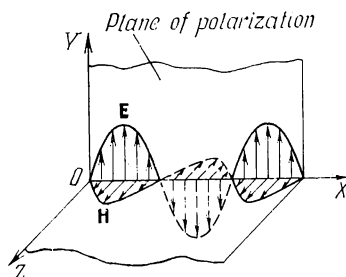
Fig. 31.2

polarized in mutually perpendicular planes. For example, a plane monochromatic wave, propagating along the OX axis (31.1.6) can be dealt with as the result of superposition of a y-wave ($\mathbf{E_1} = \mathbf{E}_y$) and a z-wave ($\mathbf{E_2} = \mathbf{E}_z$).

31.2 Energy of Electromagnetic Waves

31.2.1 The *volume energy density of an electromagnetic field* in an isotropic medium is

$$E^{\mathrm{d}} = \frac{\varepsilon_r \varepsilon_0 E^2}{2} + \frac{\mu_r \mu_0 H^2}{2} \,,$$

where ε_r and μ_r are the relative permittivity (dielectric constant) and relative permeability of the medium. From the relations between the magnitudes of vectors $\mathbf{E}$ and $\mathbf{H}$ of the field set up by an electromagnetic wave (31.1.5) it follows that the *volume energy density of electromagnetic waves* is equal to

$$E^{\mathrm{d}} = \varepsilon_r \varepsilon_0 E^2 = \mu_r \mu_0 H^2 = \sqrt{\varepsilon_r \varepsilon_0 \mu_r \mu_0} \; EH = \frac{\sqrt{\varepsilon_r \mu_r}}{c} \, EH,$$

where c is the velocity of electromagnetic waves in free space (31.1.3).

31.2.2 In the case of a plane, linearly polarized, monochromatic wave, propagating along the positive direction of axis OX, the field strength $E = A \sin (\omega t - kx)$. Consequently, the volume energy density of this wave is

$$E^{\mathrm{d}} = \varepsilon_{\mathrm{r}} \varepsilon_0 A^2 \sin^2 (\omega t - kx).$$

The value of E^{d} at each point of this field varies periodically with the frequency ω/π in a range from 0 to $E^{\mathrm{d}}_{\max} = \varepsilon_{\mathrm{r}} \varepsilon_0 A^2$. The average value of E^{d} over one period is proportional to the square of the amplitude of the field strength:

$$\langle E^{\mathrm{d}} \rangle = \frac{\omega}{\pi} \int\limits_0^{\pi/\omega} E^{\mathrm{d}} \, dt = \frac{1}{2} \varepsilon_{\mathrm{r}} \varepsilon_0 A^2.$$

If the plane monochromatic wave has an arbitrary (elliptical) polarization (31.1.7), then, according to Sect. 31.1.8,

$$E^{\mathrm{d}} = \varepsilon_{\mathrm{r}} \varepsilon_0 \left[A_1^2 \sin^2 (\omega t - kx) + A_2^2 \sin^2 (\omega t - kx + \varphi) \right]$$

and

$$\langle E^{\mathrm{d}} \rangle = \frac{1}{2} \varepsilon_{\mathrm{r}} \varepsilon_0 (A_1^2 + A_2^2).$$

31.2.3 The energy flux density vector (30.3.5) of an electromagnetic wave is called the *Umov-Poynting vector* **P** (more frequently, it is called the *Poynting vector*). The velocity of energy transfer by a monochromatic wave is equal to the phase velocity $v = c/\sqrt{\varepsilon_{\mathrm{r}} \mu_{\mathrm{r}}}$ of this wave (30.3.5). Therefore, the Umov-Poynting vector for such a wave is

$$\mathbf{P} = E^{\mathrm{d}} \mathbf{v} = [\mathbf{EH}].$$

In the case of a plane, linearly polarized, monochromatic wave (31.2.2) the Umov-Poynting vector is in the direction of wave propagation and is numerically equal to

$$P = \sqrt{\frac{\varepsilon_{\mathrm{r}} \varepsilon_0}{\mu_{\mathrm{r}} \mu_0}} A^2 \sin^2 (\omega t - kx).$$

When a plane monochromatic wave is elliptically polarized (31.1.7),

$$P = \sqrt{\frac{\varepsilon_r \varepsilon_0}{\mu_r \mu_0}} \left[A_1^2 \sin^2 (\omega t - kx) + A_2^2 \sin^2 (\omega t - kx + \varphi) \right].$$

31.2.4 The *intensity I of an electromagnetic wave* is a physical quantity numerically equal to the energy transferred by the wave in unit time through unit area of a surface located perpendicular to the direction of wave propagation. The intensity of an electromagnetic wave is equal to the absolute value of the average Umov-Poynting vector during a period of its complete oscillation:

$$I = |\langle P \rangle|.$$

The intensity of a travelling monochromatic wave is

$$I = \langle E^d \rangle v,$$

where v is the phase velocity of the wave, and $\langle E^d \rangle$ is the average value of the volume energy density of the field of the wave.
The intensity of a plane, linearly polarized, travelling mono-chromatic wave (31.2.2) is proportional to the square of the amplitude A of oscillation of the wave's field strength vector E. Thus

$$I = \frac{1}{2} \sqrt{\frac{\varepsilon_r \varepsilon_0}{\mu_r \mu_0}} A^2.$$

When a plane monochromatic wave is elliptically polarized (31.1.7), its intensity is equal to the sum of the intensities of the y- and z-waves that form the wave being considered (31.1.8):

$$I = I_y + I_z = \frac{1}{2} \sqrt{\frac{\varepsilon_r \varepsilon_0}{\mu_r \mu_0}} (A_1^2 + A_2^2).$$

Note. The *intensity of light*, i.e. the electromagnetic waves dealt with in optics, is commonly understood to be the square of the amplitude of oscillation of the light wave's field strength vector E.
31.2.5 The intensity I of a spherical, linearly polarized, monochromatic wave is related to the amplitude A of oscillation of vector E in the same way as for a plane wave (31.2.4). But

the amplitude and intensity of a spherical wave decrease as the distance r increases from the wave centre (30.3.6):

$$A = a_0/r \quad \text{and} \quad I = i_0/r^2.$$

31.2.6 Electromagnetic waves exert pressure on the obstacles in their path that absorb and reflect the waves. The *pressure of electromagnetic waves* is explained by the fact that due to the effect of the waves' electric field charged particles of matter are put into ordered motion and are subject to the action of Lorentz forces exerted by the magnetic field (24.1.1).

According to Maxwell's theory the pressure of electromagnetic waves is equal to

$$p = \frac{I}{v} (1 + R) \cos^2 i = \langle E^{\mathrm{d}} \rangle (1 + R) \cos^2 i,$$

where I is the intensity of the wave incident on the obstacle, v is its velocity, $\langle E^{\mathrm{d}} \rangle$ is the average value of the volume energy density of the wave, i is the angle of incidence (31.5.2), and R is the reflection coefficient (31.5.6).

31.3 Electromagnetic Radiation

31.3.1 The process of exciting electromagnetic waves and emitting them into surrounding space by some system is called the *radiation* of these *waves*. The system itself is called the *radiating system*. The field set up by these electromagnetic waves is called the *radiation field*. According to the concepts of classical electrodynamics, electromagnetic waves are excited by electric charges travelling with acceleration (in particular, an electric circuit in which the current varies). Vavilov-Cherenkov radiation can also occur in matter (34.6.1).

The simplest radiating system is the electric dipole (15.2.3) whose dipole moment $\mathbf{p_e}$ varies with time. Such an "oscillating" dipole is called an *oscillator*, or *elementary vibrator*. Oscillators are extensively used in physics for simulating and calculating the radiation fields of real systems. If the radiating system is electrically neutral and its size is small compared to the length λ of the waves it radiates, then in the wave zone of the system, i.e. at points whose distance from the system equals $r \gg \lambda$, the radiation field is almost the same as that of an oscillator having

the same electric dipole moment as the whole radiating system.

31.3.2 A *linear harmonic oscillator* is an electric dipole whose moment $\mathbf{p_e}$ varies according to the harmonic law:

$\mathbf{p_e} = \mathbf{p_0} \sin \omega t$,   where $\mathbf{p_0} = \text{const.}$

At point M of the wave zone of a linear harmonic oscillator, located in vacuum, the vectors $\mathbf{E}$ and $\mathbf{H}$ of the radiation field at the instant of time t are equal to:

$$\mathbf{E}\,(t) = - \frac{\mu_0 \omega^2}{4\pi r^3}\, [[\mathbf{p_0 r}]\,\mathbf{r}] \sin(\omega t - kr),$$

$$\mathbf{H}\,(t) = - \frac{\omega^2}{4\pi c r^2}\, [\mathbf{p_0 r}] \sin(\omega t - kr),$$

where $\mathbf{r}$ is the radius vector from point O, where the dipole is located, to point M (Fig. 31.3). Vectors $\mathbf{E}$ and $\mathbf{H}$ are mutually perpendicular and lie in a plane perpendicular to the radius
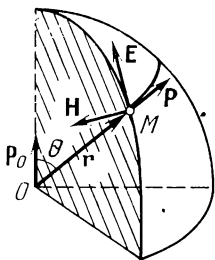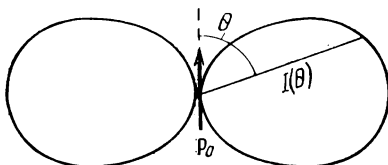
Fig. 31.3   Fig. 31.4

vector $\mathbf{r}$. Hence, $\mathbf{E}$, $\mathbf{H}$ and $\mathbf{r}$ form a right-hand set of three vectors. Vector $\mathbf{E}$ is tangent to the meridian, and vector $\mathbf{H}$ is tangent to the latitudinal circle passing through point M of a sphere of radius r having its centre at point O.

31.3.3 The Umov-Poynting vector at point M of the wave zone of a linear harmonic oscillator is

$$\mathbf{P} = \frac{\mu_0 \omega^4 p_0^2 \sin^2 \theta}{16\pi^2 c r^3}\, \mathbf{r} \sin^2(\omega t - kr),$$

where θ is the angle between vectors $\mathbf{p_0}$ and $\mathbf{r}$.

The intensity of the electromagnetic wave at point M is

$$I = \frac{\mu_0 \omega^4 p_0^2 \sin^2 \theta}{32\pi^2 c r^2}.$$

The dependence of I on θ for a fixed value of r, shown in polar coordinates (Fig. 31.4), is called the *polar directional diagram of oscillator (dipole) radiation*. The dipole radiates most intensively in directions that make an angle $\theta = \pi/2$ with its axis, i.e. in a plane passing through the middle of the dipole and perpendicular to its axis. Along its axis ($\theta = 0$ and $\theta = \pi$) the dipole does not radiate at all.

The average power of radiation of a linear harmonic oscillator is

$$\langle P^r \rangle = 2\pi r^2 \int_0^\pi I \sin 0 \, d\theta = \frac{\mu_0 \omega^4 p_0^2}{12\pi c}.$$

31.3.4 The instantaneous power of radiation of charge q, travelling at the acceleration $\mathbf{a}$, is

$$P^r_{\text{ins}} = \frac{\mu_0}{6\pi c} q^2 a^2.$$

If the charge executes harmonic oscillation of the amplitude l_0 and angular frequency ω, the average power of its radiation is

$$\langle P^r \rangle = \frac{\mu_0 q^2 \omega^4 l_0^2}{12\pi c}.$$

According to classical theory the radiation of light by atoms is due to the vibration of electrons in them. Owing to the consumption of energy to produce radiation this vibration is gradually damped, i.e. its amplitude varies according to the law (29.1.3): $l_0 = l_{00} \exp(-\beta t)$, where l_{00} is the initial amplitude and β is the damping factor. The energy of vibration of an electron (28.2.2) is $E = m_e \omega^2 l_0^2 / 2$, where m_e is the mass of the electron. Since $-dE = \langle P^r \rangle \, dt = 2\beta E \, dt$ and $q = -e$, the relaxation time (29.1.3) of this vibration, called the *average de-excitation time of the atom*, is

$$\tau = \frac{1}{\beta} = \frac{12\pi c m_e}{\mu_0 e^2 \omega^2}.$$

For visible light $\omega \approx 4 \times 10^{15}$ s^{-1} and $\tau \approx 2 \times 10^{-8}$ s.

31.4 Electromagnetic Spectrum

31.4.1 Depending upon the frequency ν (or wavelength $\lambda = c/\nu$ in vacuum, where c is the velocity of electromagnetic waves in free space), as well as on the way in which they are radiated and recorded (or detected), distinction is made between several kinds of electromagnetic waves: radio waves, optical radiation, X-rays and gamma rays.

Radio waves are electromagnetic waves of a length λ in free space that exceeds 5×10^{-5} m (correspondingly, $\nu < 6 \times 10^{12}$ Hz). In connection with special features of their propagation and generation, the full range of radio waves is commonly divided into nine bands (Table 31.1).

TABLE 31.1

Band name	Wavelength, m	Frequency, Hz
Very-long (myriametric) waves	Over 10^4	Less than 3×10^4
Long (kilometric) waves	10^4 to 10^3	3×10^4 to 3×10^5
Medium (hectometric) waves	10^3 to 10^2	3×10^5 to 3×10^6
Short (decametric) waves	10^2 to 10	3×10^6 to 3×10^7
Metric waves	10 to 1	3×10^7 to 3×10^8
Decimetric waves	1 to 0.1	3×10^8 to 3×10^9
Centimetric waves	0.1 to 0.01	3×10^9 to 3×10^{10}
Millimetric waves	10^{-2} to 10^{-3}	3×10^{10} to 3×10^{11}
Submillimetric waves	10^{-3} to 5×10^{-5}	3×10^{11} to 6×10^{12}

31.4.2 Radio frequencies are divided into twelve bands according to the international regulations for radio communications. These bands and their corresponding wavelength in free space are listed in Table 31.2.

28—0186

TABLE 31.2

Radio-frequency band designation			Radio-wave band designation		
Basic designation	Optional designation	Band limits	Basic designation	Optional designation	Band limits
Band 1	Extremely low frequency* ELF	3 to 30 Hz	Band 1	Dekamegametric waves	100 to 10 Mm
Band 2	Super low frequency SLF	30 to 300 Hz	Band 2	Megametric waves	10 to 1 Mm
Band 3	Voice frequency VF	0.3 to 3 kHz	Band 3	Hectokilometric waves	1000 to 100 km
Band 4	Very low frequency VLF	3 to 30 kHz	Band 4	Myriametric waves	100 to 10 km
Band 5	Low frequency LF	30 to 300 kHz	Band 5	Kilometric waves	10 to 1 km
Band 6	Medium frequency MF	0.3 to 3 MHz	Band 6	Hectometric waves	1 to 0.1 km
Band 7	High frequency HF	3 to 30 MHz	Band 7	Dekametric waves	100 to 10 m
Band 8	Very high frequency VHF	30 to 300 MHz	Band 8	Metric waves	10 to 1 m
Band 9	Ultra high frequency UHF	0.3 to 3 GHz	Band 9	Decimetric waves	1 to 0.1 m
Band 10	Super high frequency SHF	3 to 30 GHz	Band 10	Centimetric waves	10 to 1 cm
Band 11	Extremely high frequency EHF	30 to 300 GHz	Band 11	Millimetric waves	10 to 1 mm
Band 12	Hyper high frequency HHF	0.3 to 3 THz	Band 12	Decimillimetric waves	1 to 0.1 mm

* In the terminology accepted in the U.S.A. the extremely low frequency (ELF) band includes frequencies from 3 Hz to 3 kHz, i.e. the first three bands of the list given here.

31.4.3 *Optical radiation*, or *light*, refers to electromagnetic waves (electromagnetic radiation) of a wavelength in vacuum ranging from 10 nm to 1 mm (by convention). Optical radiation includes infrared, visible and ultraviolet radiation.*

Infrared (IR) *radiation* is the electromagnetic radiation emitted by heated bodies. Its wavelength in vacuum is within the range from 1 mm to 770 nm (1 nm = 10^{-9} m).

Visible radiation, or *visible light*, is electromagnetic radiation of a wavelength in vacuum from 770 to 380 nm. It is capable of directly producing a visible sensation in the human eye.

Ultraviolet (UV) *radiation* is electromagnetic radiation of a wavelength in vacuum from 380 to 10 nm.

31.4.4 *X-radiation*, or *X-rays* (also called *Roentgen rays*), is the electromagnetic radiation produced in the interaction of charged particles and photons (37.1.4) with atoms of matter. Its wavelength in vacuum ranges within the wide conditional limits from 10 or 100 nm to 0.01 or 1 pm (1 pm = 10^{-12} m).

Gamma radiation, or *gamma rays*, is electromagnetic radiation of a wavelength in vacuum less than 0.1 nm. It is emitted by excited atomic nuclei upon radioactive transformations and nuclear reactions. It is also formed in particle decay, in particle-antiparticle annihilation (43.5.5) and other processes.

31.5 Reflection and Refraction of Electromagnetic Waves at the ·Interface Between Two Dielectric Media

31.5.1 The *refractive index* (*absolute refractive index*) *of a medium* is the quantity n, which is equal to the ratio of the velocity c of the electromagnetic waves in vacuum to their phase velocity v in the medium: $n = c/v$. From Sect. 31.1.3 it follows that $n = \sqrt{\varepsilon_r \mu_r}$, where ε_r and μ_r are the relative dielectric permittivity and relative magnetic permeability of the medium.

---

* According to the recommendation of the Committee for Scientific and Engineering Terminology of the USSR Academy of Sciences [Physical Optics (Terminology), Collection of Recommended Terms, Issue 74, Nauka Publishers, 1968, Moscow] optical radiation also includes X-rays, the range of wavelengths of optical radiation being established from 0.1 Å (10^{-11} m) to 1 cm.

For media having no ferromagnetic properties, $\mu_r \approx 1$ and $n \approx \sqrt{\varepsilon_r}$.

The *relative refractive index of two media* (of the second medium with respect to the first) is the quantity n_{21}, equal to the ratio of the refractive indices of these media: $n_{21} = n_2/n_1$. For non-ferromagnetic media $n_{21} = \sqrt{\varepsilon_{r2}/\varepsilon_{r1}}$.

31.5.2 When an electromagnetic wave falls on the interface between two media, it is partly reflected from the interface and partly refracted as it enters the second medium. In Fig. 31.5 the line AB is the plane interface between the media. Rays *1, 1'* and *2* represent the directions of propagation of the incident, reflected and refracted plane waves. They are called the *incident, reflected* and *refracted rays*, respectively, and the angles between them and the angles they make with the perpendicular ab to the interface at the point of incidence O are: i is the *angle of incidence*, i' is the *angle of reflection* and r is the *refractive angle*.
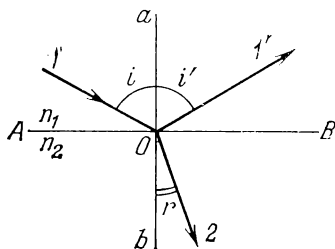
Fig. 31.5

The *plane of incidence* is the plane passing through the incident ray and perpendicular to the interface at the point of incidence.

31.5.3 The laws of reflection and refraction of electromagnetic waves at the interface between two dielectric media can be obtained on the basis of the boundary conditions for an electromagnetic field (27.4.4). In the first medium the field of the reflected wave (E^{rfl} and H^{rfl}) is superposed on the field of the incident wave (E^0 and H^0). In the second medium there is only the field (E^{rfr} and H^{rfr}) of the refracted wave (propagating in this medium). Hence, the boundary conditions are of the form (assuming that $\mu_{r2} = \mu_{r1} = 1$):

$$E_\tau^0 + E_\tau^{rfl} = E_\tau^{rfr}, \quad \varepsilon_{r1}(E_n^0 + E_n^{rfl}) = \varepsilon_{r2}E_n^{rfr}$$

$$H_\tau^0 + H_\tau^{rfl} = H_\tau^{rfr}, \qquad H_n^0 + H_n^{rfl} = H_n^{rfr}.$$

Here E_τ, H_τ, E_n and H_n are the projections of vectors **E** and **H** on the tangential plane and the normal to the interface between the media. It follows from these relations that in the incidence of a plane monochromatic wave on a smooth plane interface between media, the following laws are complied with (regardless of the kind of polarization of this wave):

(a) The reflected and refracted waves are also monochromatic waves of the same frequency as the incident one.

(b) The *law of reflection* states that the reflected ray lies in the plane of incidence, and the angle of reflection is equal to the angle of incidence $(i' = i)$.

(c) The *law of refraction* states that the refracted ray lies in the plane of incidence and that the refractive angle is related to the angle of incidence by *Snell's law*:

$$\frac{\sin i}{\sin r} = \frac{n_2}{n_1} = n_{21}.$$

31.5.4 Making use of the boundary conditions (31.5.3) the relations can be found between the phases, amplitudes and intensities of incident, reflected and refracted monochromatic waves.
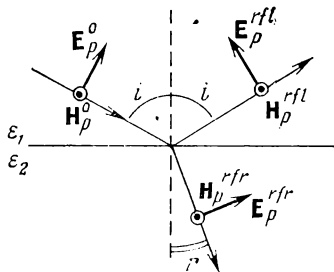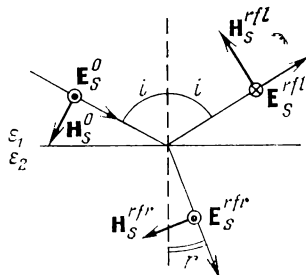
Fig. 31.6   Fig. 31.7

It is sufficient for this purpose, according to Sect. 31.1.8, to know the above-indicated relations for linearly polarized waves of two types: *p-waves*, whose vector $E = E_p$ lies in the plane of incidence and whose vector $H = H_p$ is perpendicular to this plane (Fig. 31.6), and *s-waves*, whose vector $E = E_s$ is perpendicular to the plane of incidence and whose vector $H = H_s$ lies in this plane (Fig. 31.7).

The relations between the amplitudes of oscillation of vector **E** in the incident (A^0), reflected (A^{rfl}) and refracted (A^{rfr}) waves in the cases of p- and s-waves are expressed by the *Fresnel formulas*:

$$A_p^{\text{rfl}} = -A_p^0 \frac{\tan(i-r)}{\tan(i+r)}\,, \quad A_p^{\text{rfr}} = A_p^0 \frac{2\cos i \sin r}{\sin(i+r)\cos(i-r)}\,,$$

$$A_s^{\text{rfl}} = -A_s^0 \frac{\sin(i-r)}{\sin(i+r)}\,, \quad A_s^{\text{rfr}} = A_s^0 \frac{2\cos i \sin r}{\sin(i+r)}\,.$$

In particular, with normal incidence of waves on the interface between media $(i = r = 0)$,

$$A_p^{\text{rfl}} = -A_p^0 \frac{n_{21}-1}{n_{21}+1}\,, \quad A_p^{\text{rfr}} = A_p^0 \frac{2}{n_{21}+1}$$

$$A_s^{\text{rfl}} = -A_s^0 \frac{n_{21}-1}{n_{21}+1}\,, \quad A_s^{\text{rfr}} = A_s^0 \frac{2}{n_{21}+1}\,.$$

In the Fresnel formulas A_p^0 and A_s^0 are positive quantities, whereas A_p^{rfr} and A_s^{rfr} are also positive for any possible angles of incidence and refractive angles. This indicates that the refracted and incident waves are in phase. Quantities A_p^{rfl} and A_s^{rfl} can be either negative or positive. In the first case, illustrated in Figs. 31.6 and 31.7, the phase of oscillation of vector **E** is changed by π upon reflection (the phase of oscillation of vector **H** does not change). In the second case, the phase of oscillation of vector **E** remains unchanged (and, accordingly, the phase of oscillation of vector **H** changes by π).

31.5.5 The values of the phase difference in the oscillation of vector **E** upon the reflection of p- and s-type electromagnetic waves are listed in Table 31.3, in accordance with the conditions (i.e. depending on the angle of incidence i and the relative refractive index n_{21} of the media).

The angle of incidence i_{Br} at which the reflected and refracted rays are perpendicular to each other is called *Brewster's angle*. When $i = i_{\text{Br}}$, then $i + r = \pi/2$ and it follows from the law of wave refraction (31.5.3) that $\tan i_{\text{Br}} = n_{21}$. It is evident from the Fresnel formulas (31.5.4) that at $i = i_{\text{Br}}$, the amplitude $A_p^{\text{rfl}} = 0$, i.e. the p-wave is not reflected from the interface

TABLE 31.3

Conditions	$i + r < \pi/2$ $(i < i_{Br})$		$i + r > \pi/2$ $(i > i_{Br})$	
Type of wave	$n_{21} > 1$ $(i > r)$	$n_{21} < 1$ $(i < r)$	$n_{21} > 1$ $(i > r)$	$n_{21} < 1$ $(i < r)$
p-wave	π	0	0	π
s-wave	π	0	π	0

between the media, but completely passes from medium **1** into medium **2**.

31.5.6 The *coefficient of reflection* (or *reflection coefficient*) R of an electromagnetic wave from the interface of two media is the ratio of the intensities (31.2.4) of the reflected and incident waves:

$$R = \frac{I^{\mathrm{rfl}}}{I^0} = \left(\frac{A^{\mathrm{rfl}}}{A^0} \right)^2.$$

The reflection coefficients of the p- and s-waves are equal to

$$R_p = \frac{\tan^2(i-r)}{\tan^2(i+r)}, \quad R_s = \frac{\sin^2(i-r)}{\sin^2(i+r)}.$$

In particular, with normal incidence of waves on the interface between two media $(i = r = 0)$

$$R_p = R_s = \left(\frac{n_{21} - 1}{n_{21} + 1} \right)^2.$$

When the incident wave is arbitrarily polarized, the reflection coefficient is

$$R = \frac{I_p^{\mathrm{rfl}} + I_s^{\mathrm{rfl}}}{I_p^0 + I_s^0} = \frac{R_p I_p^0 + R_s I_s^0}{I_p^0 + I_s^0},$$

where I_p^0 and I_s^0 are the intensities of the p- and s-components of the incident wave whose intensity is $I^0 = I_p^0 + I_s^0$.

31.5.7 The *transmission factor*, or transmittance, T is the ratio of the intensity of the transmitted (refracted) wave to that of the incident wave. Thus

$$T = \frac{I^{\text{rfr}}}{I^0} = \sqrt{\frac{\varepsilon_{r2}}{\varepsilon_{r1}}} \left(\frac{A^{\text{rfr}}}{A^0} \right)^2 = n_{21} \left(\frac{A^{\text{rfr}}}{A^0} \right)^2 .$$

The transmission factors for the p- and s-waves are

$$T_p = \frac{4 \cos^2 i \sin i \sin r}{\sin^2 (i+r) \cos^2 (i-r)} , \qquad T_s = \frac{4 \cos^2 i \sin i \sin r}{\sin^2 (i+r)} .$$

In particular, upon normal incidence of waves on the interface between media $(i = r = 0)$,

$$T_p = T_s = \frac{4n_{21}}{(n_{21} + 1)^2} .$$

31.5.8 When $n_{21} = (n_2/n_1) < 1$, the refractive angle is greater than the angle of incidence: $\sin r = \sin i/n_{21}$ and $r > i$. The angle of incidence at which the refractive angle becomes equal to $\pi/2$ is called the *critical angle* i_{cr}. Angle $i_{\text{cr}} = \arcsin n_{21}$. When $i \geqslant i_{\text{cr}}$, the intensities of the reflected and incident waves are the same, i.e. the wave is completely reflected from the interface between the media $(R = 1)$. This phenomenon is called *total internal reflection*.

31.6 Doppler Effect

31.6.1 Upon motion of the source and detector of electromagnetic waves toward each other the Doppler effect is observed (30.6.1). The laws governing this phenomenon for electromagnetic waves can be established on the basis of the special theory of relativity. Assume the detector D to be stationary with respect to an inertial frame of reference K, and that the source S travels with respect to K along the positive direction of axis OX at the velocity **V**

(Fig. 31.8). Source S is stationary in frame of reference K' and is located at its origin of coordinates. The axes of coordinate systems K and K' are parallel pairwise (axis $O'X'$ coincides with axis OX). Illustrated in Fig. 31.8 is the position of source S at the instant of time $t = t' = 0$. According to Einstein's rela-
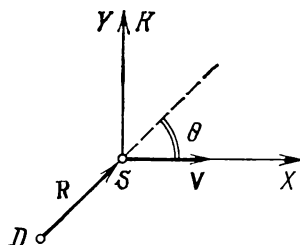
Fig. 31.8

tivity principle (5.1.2), the equation of a spherical monochromatic wave (30.2.8), sent by the source at this instant of time in the direction of the detector D, is of identical form in frames of reference K and K':

$$\widetilde{s} = \frac{a_0}{r}\, e^{i(\omega t + kx \cos \theta + ky \sin \theta + \delta)} \tag{a}$$

and

$$\widetilde{s'} = \frac{a_0'}{r'}\, e^{i(\omega' t' + k'x' \cos \theta' + k'y' \sin \theta' + \delta')}. \tag{b}$$

Here $\omega' = \omega_0$ and ω are the angular frequencies of vibration of the source and detector, $k = \omega/c$ and $k' = \omega'/c$ are the wavenumbers (it is assumed that the wave propagates in vacuum), and θ and θ' are the angles between the directions of observation and of velocity V (OX axis), measured in the frames of reference of the detector (K) and for the source (K').

Expression (b) should be obtained from expression (a) by replacing the variables x, y and t by x', y' and t' in accordance with Lorentz's transformations (5.3.2):

$$\omega' t' + k' x' \cos \theta' + k' y' \sin \theta' + \delta'$$

$$\equiv \omega \left(\frac{t' + V x'/c^2}{\sqrt{1 - (V/c^2)}} \right) + k \left(\frac{x' + V t'}{\sqrt{1 - (V/c)^2}} \right) \cos \theta + k y' \sin \theta + \delta.$$

Consequently,

$$\omega' = \frac{\omega}{\sqrt{1 - (V/c^2)}} \left(1 + \frac{V}{c} \cos \theta \right) ,$$

$$k' \cos \theta' = \frac{k}{\sqrt{1 - (V/c)^2}} \left(\cos \theta + \frac{V}{c} \right) ,$$

$$k' \sin \theta' = k \sin \theta \quad \text{and} \quad \delta' = \delta.$$

Hence, the relations describing the *Doppler effect for electromagnetic waves* in vacuum are of the form:

$$\omega = \frac{\omega_0 \sqrt{1 - (V/c)^2}}{1 + \frac{V}{c} \cos \theta} \quad \text{and} \quad \nu = \frac{\nu_0 \sqrt{1 - (V/c)^2}}{1 + \frac{V}{c} \cos \theta} .$$

31.6.2 At low velocities of the wave source with respect to the detector ($V \ll c$), the relativistic formula for the Doppler effect (31.6.1) coincides with the classical formula (30.6.3):

$$\nu \approx \nu_0 \left(1 - \frac{V}{c} \cos \theta \right) .$$

When the source travels with respect to the detector along the straight line that joins them ($\theta = 0$ or π), the *longitudinal Doppler effect* is observed. If the source and detector approach each other ($\theta = \pi$)

$$\nu = \nu_0 \sqrt{\frac{1 + V/c}{1 - V/c}} > \nu_0,$$

whereas if they travel away from each other ($\theta = 0$)

$$\nu = \nu_0 \sqrt{\frac{1 - V/c}{1 + V/c}} < \nu_0.$$

31.6.3 Following from relativistic theory of the Doppler effect (31.6.1) is the existence of the *transverse Doppler effect*. It is observed when $\theta = \pi/2$ and $\theta = 3\pi/2$, i.e. when the source travels perpendicular to the line of observation. Thus

$$\nu = \nu_0 \sqrt{1 - (V/c)^2}.$$

The transverse Doppler effect is substantially weaker than the longitudinal effect (due to the smallness of V/c). This effect is due to the different passage of time in the frames of reference in which the detector and source of waves are fixed. Therefore, the experimental detection of the transverse Doppler effect was one of the most vital experimental proofs of the special theory of relativity.

31.6.4 The Doppler effect, associated with the thermal motion of the gas atoms that radiate light, leads to *Doppler broadening of spectral lines*. The range of frequencies of light recorded by the detector extends from $\nu_0 - \Delta\nu_D/2$ to $\nu_0 + \Delta\nu_D/2$, where ν_0 is the frequency of monochromatic radiation of a stationary atom. The amount of Doppler broadening $\Delta\nu_D \propto \sqrt{T/m}$, where T is the absolute temperature and m is the mass of the atom.

PART FIVE OPTICS

CHAPTER 32 INTERFERENCE OF LIGHT

32.1 Monochromaticity and Time Coherence of Light

32.1.1 *Optics* is the branch of physics concerned with the study of the nature of light (31.4.3), the laws of its production, propagation and interaction with matter. *Wave optics* deals with optical phenomena in which the wave nature of light is manifested (for instance, interference, diffraction, polarization and dispersion of light). Since light comprises electromagnetic waves, wave optics is based on Maxwell's equations (27.4.3) and the relations that follow from them for electromagnetic waves (31.1). Classical wave optics deals with media that are linear with respect to their optical properties (30.1.7), i.e. media in which the dielectric constant ε_r and the magnetic permeability μ_r are independent of the intensity of the light (31.2.4). Consequently, the principle of superposition (30.4.1) is valid in wave optics. Phenomena that are observed in the propagation of light in nonlinear media are investigated in *nonlinear optics*. Nonlinear optical effects become appreciable at the very high intensities of light radiated by powerful lasers (39.6.8).

32.1.2 It has been experimentally established that the effect of light on a photocell, photographic film, fluorescent screen or other devices that record light is determined by the *electric vector* E of the electromagnetic field set up by the light wave. The same conclusion is reached by classical electronic theory, according to which the processes caused in matter by light are associated with the effect of the field of the light wave on charged particles of matter, i.e. on the electrons and ions. The frequency of visible and more short-waved light is so high ($\nu \geqslant 10^{15}$ Hz) that only electrons can execute forced vibration anywhere near appreciable in amplitude. The force exerted by the electromagnetic field (24.1.5) on an electron is

$$\mathbf{F} = -e\,\{\mathbf{E} + [\mathbf{v_1 B}]\} = -e\,\{\mathbf{E} + \mu_r\mu_0\,[\mathbf{v_1 H}]\},$$

Here $-e$ and v_1 are the charge and velocity of the electron, and $B = \mu_r \mu_0 H$ is the vector of magnetic induction. It follows from Sect. 31.1.5 that the absolute value of the magnetic component of force F is substantially less than its electrical component. Thus

$$\mu_r \mu_0 |[v_1 H]| \leqslant \mu_r \mu_0 v_1 H = \frac{v_1}{v} E \ll E \quad \text{and} \quad F \approx -eE,$$

because the velocity of electromagnetic waves is $v \approx 10^8$ m/s, and the velocity of the electron in the atom upon forced vibrations set up by light is $v_1 \approx 10^6$ m/s.

32.1.3 In superposing light from two nonlaser sources* (for instance, two identical gas-discharge lamps), or even from different portions of the same source, no interference (30.5.2) is observed. Consequently, independent sources of light are incoherent (30.5.1) and their radiation is nonmonochromatic (31.1.6). The reason for this resides in the very mechanism by which light is radiated by the atoms (or molecules or ions) of the source. The excited atom radiates during an extremely short de-excitation time, of the order of $\tau \approx 10^{-8}$ s (31.3.4), after which it returns to its normal (unexcited) state, having expended its surplus energy on radiation. Following a certain period of time the atom may again be excited, obtaining energy externally, and begin to emit radiation again. Such intermittent radiation of light by atoms in the form of separate short impulses—*wave trains*, or *packets*—is typical of any light source, regardless of the specific processes that occur in the source and excite its atoms. In spontaneous radiation (39.6.6), the atoms emit radiation independently of one another with random initial phases that vary without any order from one event of emission to another. Therefore, spontaneously radiating atoms are incoherent sources of light.

Induced, or stimulated, emission of radiation is quite another matter. It is produced in a nonequilibrium (active) medium by the action of a variable electromagnetic field (39.6.1). Stimulated emission of radiation by all the particles of a system

---

* This refers to ordinary light sources (incandescent lamps, discharge lamps, electric arcs, etc.), based on the principle of spontaneous radiation (39.6.6).

is coherent with the monochromatic radiation that it is produced by, having the same frequency, polarization and direction of propagation. These features of stimulated emission of radiation are made use of in lasers and masers (39.8.8).

32.1.4 A real wave, radiated during a limited period of time and encompassing a limited region of space, is not monochromatic. The spectrum of its angular frequencies (28.4.7) is of finite width $\Delta\omega$, i.e. it includes angular frequencies from $\omega - \Delta\omega/2$ to $\omega + \Delta\omega/2$. During the time interval $\Delta t \ll \tau_{coh} = \pi/\Delta\omega$ such a wave can be considered as an approximately monochromatic wave with the angular frequency ω. The quantity τ_{coh} is called the *time coherence of a nonmonochromatic wave.* During the time interval τ_{coh} the phase difference of the vibrations corresponding to waves of the frequencies $\omega + \Delta\omega/2$ and $\omega - \Delta\omega/2$ changes by π. A wave with the angular frequency ω and phase velocity v propagates during this time over the distance $l_{coh} = v\tau_{coh} = \pi v/\Delta\omega$.

The quantity l_{coh} is called the *coherence length,* or the *harmonic wave train length,* of the corresponding nonmonochromatic wave being considered. The closer a given wave is to a monochromatic one, the less the width $\Delta\omega$ of its frequency spectrum and the greater its coherence time and coherence length. For example, for visible sunlight, having a continuous frequency spectrum from 4×10^{14} to 8×10^{14} Hz, $\tau_{coh} \approx 10^{-14}$ s and $l_{coh} \approx \approx 10^{-6}$ m. The coherence time of stimulated radiation is substantially longer than the de-excitation time of the atom (31.3.4). For example, τ_{coh} reaches 10^{-5} s for a continuous-wave laser and the coherence length $l_{coh} \approx 10^3$ m.

32.2 Interference of Light. Spatial Coherence of Light

32.2.1 To obtain coherent light waves by means of conventional (nonlaser) light sources, light from a single source is split into two or more systems of waves. Each system represents the radiation of the same atoms of the source. Hence, by virtue of their common origin, these systems of waves are coherent with one another and interfere when superposed. Light can be split into coherent systems of waves by its reflection or refraction.

Illustrated as an example in Fig. 32.1 is the principle of the *Fresnel mirrors.* Light from point source S is incident on two plane mirrors A_1O and A_2O arranged perpendicular to the plane of the drawing and joined along line O. The angle α be-

tween the planes of the mirrors is very small. Light from source S propagates after being reflected as two beams having their centres at points S_1 and S_2, which are the virtual images of source S in the mirrors. These beams are coherent and upon being superposed produce an interference pattern on screen E (the region BC is sometimes called the *field of interference*). The result of the interference at some point M on the screen depends upon the wavelength λ of the light and the path difference of the waves (30.5.2) from the coherent virtual sources S_1 and S_2 to point M. Thus
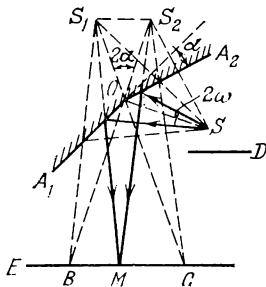
Fig 32.1

$$\Delta = r_2 - r_1 = \overline{MS_2} - \overline{MS_1}.$$

The initial phases of vibration of sources S_1 and S_2 are the same. Hence, the *conditions for interference maxima and minima* (30.5.2) are of the form

$r_2 - r_1 = \pm m\lambda$ for a maximum of the mth order $(m = 0, 1, 2,...)$,

$r_2 - r_1 = \pm(2m - 1)\dfrac{\lambda}{2}$ for a minimum of the mth order $(m = 1, 2,...)$.

The angle 2ω at apex S between the two rays of light that converge at point M of the interference pattern after being reflected from mirrors A_1O and A_2O is called the *aperture angle of the interference*. This angle usually varies very little upon changing the position of point M within the limits of the field of interference.

32.2.2 The principles of the techniques used for observing interference by means of the *Fresnel biprism* (Fig. 32.2) and the *Billet split lens* (Fig. 32.3) are similar to those of the Fresnel mirrors. The biprism consists of two identical triangular prisms joined together at their bases and prepared as a whole. The refracting angles α at the upper and lower edges of the biprism are very small (of the order of a fraction of one degree). Light from source S is refracted in the biprism and propagates in it

in the form of two systems of waves, corresponding to the coherent virtual sources of light S_1 and S_2. The interference of these waves is observed in the region of their overlapping on screen E.

The Billet split lens consists of two halves L_1 and L_2 of a collecting lens, cut along a diameter. The two halves are slightly displaced laterally (spread apart). In consequence, they produce two noncoinciding real images S_1 and S_2 of the point source of
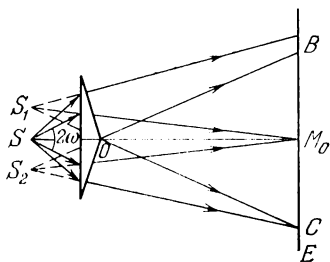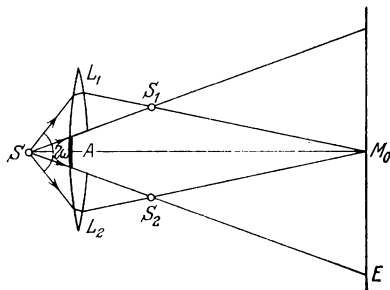
Fig. 32.2   **Fig.** 32.3

light S. The interference of light from these coherent secondary sources is observed on screen E. The space between parts L_1 and L_2 of the split lens is covered by opaque screen A.

The value of the aperture angle 2ω of interference is shown in Figs. 32.2 and 32.3 for the central point M_0 of the interference pattern obtained by means of a biprism and a split lens.

32.2.3 The *width of an interference fringe* (as the bands are called) is the distance between two adjacent interference maxima (or minima). When the Fresnel mirrors or similar devices (biprisms or split lenses) are employed to obtain interference, the width of an interference fringe (30.5.2) is $\lambda L/l$. Here l is the distance between sources S_1 and S_2, and L is the distance from them to screen E. The wavelength of visible light is extremely short ($\lambda \approx 5 \times 10^{-7}$ m). Therefore, to obtain an interference fringe of a width that can be resolved by the eye, it is necessary to comply with the condition: $l \ll L$. Consequently, the angle α in the Fresnel mirrors and the refracting angles α of the biprism must be very small.

The possibility of observing interference fringes also depends on the *degree of contrast*, i.e. on the degree of different illumination of the screen at the maxima and minima. The illuminance is proportional to the intensity I of the incident light. A quantitative characteristic of the degree of contrast of the interference pattern is the dimensionless quantity called the *visibility of fringes*

$$V = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}},$$

where $I_{\max}$ and $I_{\min}$ are the values of the light intensities in the interference maxima and minima on the screen. The human eye reliably distinguishes fringes if their visibility $V > 0.1$, i.e. when $I_{\min} < 0.82\ I_{\max}$.
In superposing two likewise polarized coherent monochromatic waves, with the amplitudes and intensities A_1, I_1, A_2 and I_2, the visibility of the fringes is

$$V = \frac{2A_1 A_2}{A_1^2 + A_2^2} = \frac{2\sqrt{I_1 I_2}}{I_1 + I_2}.$$

The visibility of the fringes is a maximum ($V = 1$) when $A_1 = A_2$.
32.2.4 In an interference arrangement of the type of the Fresnel mirrors, illuminated by a point source S (Fig. 32.1), the superposed waves are never ideally monochromatic (32.1.3). Hence, these waves are only *partly coherent*. They are capable of interference only under the condition that the vibration, excited by them at point M being considered on the screen, corresponds to one and the same harmonic train, or packet, of radiation of source S (32.1.4) i.e. if

$$|r_2 - r_1| < v\tau_{\mathrm{coh}} \ \text{ or }\ |r_2 - r_1| < l_{\mathrm{coh}}.$$

Here $r_2 - r_1$ is the path difference of the superposed waves, v is their velocity, and τ_{coh} and l_{coh} are the coherence time and coherence length of light from source S (32.1.4). *Partly coherent vibrations are combined* at point M. These vibrations are excited by the same source S at *different* instants of time, t and $t + \tau$, where $\tau = |r_2 - r_1|/v$. Therefore, the visibility of the interference pattern in such arrangements essentially depends on the *time-coherence of the vibrations*, which is limited by the mono-

chromaticity of the light from source S, i.e. its coherence time τ_{coh} (32.1.4). At $\tau \ll \tau_{coh}$, the superposed vibrations are practically fully coherent and the visibility of the interference fringes (at equal intensity of the superposed waves) $V \approx 1$. If, however, $\tau \geqslant \tau_{coh}$, the superposed vibrations are incoherent and there is no interference ($V = 0$).

Thus, to observe the interference of light at large path differences $r_2 - r_1$ (and corresponding large values of τ) it is necessary for the light to have a sufficiently long coherence time, i.e. a sufficiently high degree of monochromaticity.

32.2.5 The positions on the screen of all the interference maxima, with the exception of the zero-order maximum, depend upon the wavelength of the light. For two wavelengths λ_1 and λ_2, the higher the order m to be observed, the more their maxima of the mth order are displaced from each other. Therefore, as m is increased, the visibility of interference fringes, obtained by the illumination of Fresnel mirrors with nonmonochromatic light, is impaired. Fringes, corresponding to different values of λ, are superposed on one another, and the interference pattern is blurred.

Assume that the wavelength of light is within the limits from λ to $\lambda + \Delta\lambda$, and the angular frequency, from ω to $\omega - \Delta\omega$, where $\Delta\omega = 2\pi v \Delta\lambda/\lambda^2$. Then, according to the *Rayleigh criterion*, the interference pattern is still distinguishable up to a maximum of the m_0 order for light with the wavelength $\lambda + \Delta\lambda$ (where $\Delta\lambda > 0$), which is superposed on the screen on the nearest interference minimum for light of wavelength λ:

$$m_0 (\lambda + \Delta\lambda) = (2m_0 + 1) \frac{\lambda}{2}, \quad \text{from which} \quad m_0 = \frac{\lambda}{2\Delta\lambda}.$$

Thus, interference can be observed at path differences of the waves that comply with the condition

$$|r_2 - r_1| \leqslant \frac{\lambda^2}{2\Delta\lambda} = \frac{\pi v}{\Delta\omega} = v\tau_{coh}.$$

This result agrees with the assessment made in Sect. 32.2.4 on the basis of the concept of time coherence of vibrations.

32.2.6 Partly coherent light, of a total intensity I, can be dealt with as the combination of two components: the coherent component of intensity γI, where γ is the *degree of coherence of the light*, and the incoherent component of the intensity $(1 - \gamma) I$.

When partly coherent waves are superposed, only their coherent components interfere. The incoherent components produce the uniformly illuminated background of the interference pattern. Therefore, as the degree of coherence of the light is reduced, the visibility V of the interference fringes decreases. Thus

$$V = \frac{2 \sqrt{I_1 I_2}}{I_1 + I_2} \, \gamma.$$

When the intensities of partly coherent waves are the same, $V = \gamma$.

32.2.7 As a rule, in an interference device with mirrors (or a biprism), a brightly illuminated narrow slit, parallel to the joint O between the mirrors, is used instead of a point source of light. In this case, the interference patterns, obtained on the screen from the various portions along the length of the slit, are displaced with respect to one another in the direction of the slit S. Accordingly, a system of interference fringes, parallel to joint O of the mirrors, is observed on the screen. The visibility of the interference fringes is reduced as the width of the slit S is increased. This is due to the fact that the interference fringes that could be obtained on the screen from various narrow slits into which slit S could be conceivably divided would be displaced with respect to one another. An interference pattern in monochromatic light of wavelength λ is sharp and distinct if the following approximate condition is complied with:

$$b \, \sin \, \omega \leqslant \lambda/4,$$

where b is the slit width and 2ω is the aperture angle of interference (32.2.1).

32.2.8 The principle of *Young's two-slit interference method* is illustrated in Fig. 32.4. The source of light is a brightly illuminated narrow slit S in screen A_1. Light from slit S falls on a second opaque screen A_2 having two identical narrow slits S_1 and S_2, parallel to slit S. Two systems of cylindrical waves are propagated in the space beyond screen A_2; their interference is observed on screen E. The visibility of the interference fringes for small path differences is determined mainly by the degree of agreement in the way the vibrations are executed at points of

29*

slits S_1 and S_2. These slits can be considered as the "sources" of the waves interfering on the screen.

32.2.9 The coherence of vibrations executed at the same instant of time at *various* points of a plane Q, perpendicular to the direction of wave propagation, is called *spatial coherence* (as distin-

Fig. 32.4

guished from time coherence of vibrations, executed at one and the same point, but at different instants of time). Spatial coherence depends upon the conditions for the radiation and formation of light waves. For instance, a light wave emitted by a point source has complete spatial coherence. In the case of an ideal plane wave, the amplitude and phase of the vi-bration are the same at all points of plane Q, i.e. there is also complete spatial coherence. Spatial coherence is also maintained over the whole cross section of a light beam emitted by a laser. In a real wave, emitted by a host of independent atoms of an extended nonlaser light source, the phase difference of the vibrations at two points K_1 and K_2 of plane Q is a random function of time. The random variations of the phase difference increase with the distance between the points. The *length of spatial coherence* is the distance l_{sc} between points K_1 and K_2 of plane Q at which the random changes of phase difference reach a value equal to π. If in Young's two-slit interference arrangement the distance l between slits S_1 and S_2 is greater than or equal to l_{sc} the visibility of the interference fringes equals zero. To ensure spatial coherence in the illumination of slits S_1 and S_2, the width b of inlet slit S should be sufficiently narrow. Thus

$$b < \lambda d/l \text{ and } \theta < \lambda/l,$$

where d is the distance between screens A_1 and A_2, and $\theta = b/d$ is the angular dimension of the light source, i.e. slit S.

The length of spatial coherence $l_{sc} = \lambda/\theta$ increases with the distance from the light source. For example, for a star of diameter D, located at the distance r, $\theta = D/r$ and $l_{sc} = \lambda r/D$.

The area of a circle of radius l_{sc} is called the *dimension of spatial*

coherence, and the volume of a right cylinder having such a circle
as its base and a height equal to length of the harmonic wave
train (32.1.4) $l_{coh} = v\tau_{coh}$, is called the *volume of coherence*.

32.3 Interference of Light in Thin Films

32.3.1 Examples of light interference observed under everyday
conditions are the iridescent colouring of thin films (soap bubbles,
films of petroleum or other oils on the surface of water, trans-
parent films of oxides on the surfaces of heat-treated metal com-
ponents, called temper colours, etc.). The formation of partly
coherent waves, displaying interference upon superposition,
occurs in these cases due to the reflection of light, incident on the
film, from its upper and lower surfaces. This interference depends
upon the phase difference acquired by the superposed waves in
the film and depending upon their optical path difference.

32.3.2 The *optical path*, or *distance*, s of light is the product of
the geometric path length l travelled by light in the medium
by the refractive index of the medium (31.5.1): $s = nl$. The
quantity s is equal to the path travelled by light in vacuum
during the time it covers the path l in the given medium.

The *optical-path difference* of two waves is the difference in the
optical path of these waves: $\Delta s = s_2 - s_1$ (the optical-path
difference is frequently denoted by Δ or δ). An optical-path
difference Δs corresponds to a change in the phase by $\Delta\Phi =$
$= 2\pi\Delta s/\lambda_0$, where $\lambda_0 = n\lambda$ is the *wavelength of the light in vacuum*,
λ is the light wavelength in a medium with the refractive index n.
Paths of propagation of waves, having the same optical length,
are said to be *tautochronous*. Light requires equal amounts
of time to travel them. In an optical system (microscope, tele-
scope, etc.), for example, all possible paths of the rays of light
from some point on the object to the corresponding point on its
image are tautochronous.

32.3.3 Assume that a plane monochromatic light wave *1* is
incident at the angle i on a plane-parallel homogeneous plate of
thickness d and transparent to light (Fig. 32.5). Owing to the
reflection of light from the upper and lower surfaces of the plate,
two plane waves propagate in the directions of the reflected rays
1′ and *1″*. Their optical-path difference is

$$\Delta s = n\,(\overline{AB} + \overline{BC}) - n_1\left(\overline{AD} + \frac{\lambda_1}{2}\right) = 2\,dn\cos r - \frac{\lambda_0}{2}\ ,$$

where n_1 and n are the refractive indices of the surrounding medium and of the plate, D is the foot of a perpendicular erected on ray $1'$ from point C, member $\lambda_1/2$ takes into account the phase difference of π as light is reflected from point A (31.5.5),*
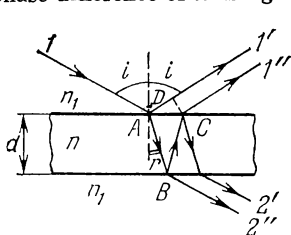
Fig. 32.5

r is the refractive angle, and λ_0 is the wavelength of the light in vacuum. The conditions for interference maxima of reflection are

$$2dn \cos r = (2m+1) \frac{\lambda_0}{2}, \quad \text{or}$$

$$2d \cos r = (2m+1) \frac{\lambda}{2}, \quad \text{where}$$

$m = 0, 1, 2, \ldots$ is the order of the interference maximum.

The conditions for interference minima of reflection are

$2dn \cos r = m\lambda_0$ or $2d \cos r = m\lambda$,

where $m = 1, 2, \ldots$ is the order of the interference minimum. The optical-path difference for the waves passing through the plate (rays $2'$ and $2''$) is $\Delta s = 2dn \cos r$, i.e. it differs from Δs for the reflected light by $\lambda_0/2$. Therefore, the maxima of reflection correspond to the minima of transmission of light and vice versa. If the plate is illuminated with white light, it has additional colouring in the reflected and transmitted light.

The maximum thickness d of the plate at which it is still possible to observe interference fringes is limited by the coherence time τ_{coh} of the light (32.1.4). Thus

$2dn \cos r < c\tau_{\text{coh}}$ or $2d \cos r < l_{\text{coh}}$,

where $l_{\text{coh}} = v\tau_{\text{coh}} = c\tau_{\text{coh}}/n$ is the coherence length.

32.3.4 In the calculation of the optical-path difference of interfering waves in a plate (32.3.3) only two waves were taken into

* It is assumed that $i < i_{\text{Br}}$ and that $n > n_1$. When $n < n_1$, a phase difference of π occurs upon reflection of light from point B and $\Delta s = 2dn \cos r + \lambda_0/2$, i.e. it differs from the value of Δs for the case when $n > n_1$ by λ_0, but this has no effect on the result of the interference.

account. They correspond to the first reflection from the upper
and from the lower surfaces of the plate; the possibility of mul-
tiple reflection of light was neglected. Such simplification is
justified only if the intensity I_2 of the wave corresponding to the
second reflection from the lower surface of the plate is substanti-
ally less than the intensity I_1 of the wave corresponding to the
first reflection. If R is the coefficient of reflection (31.5.6) from
the upper and lower surfaces of the plate, then $I_2 = R^2 I_1$. As
a rule $R^2 \ll 1$. At the interface, for instance between air and
glass ($n_{21} = 1.5$), at angles
of incidence $i < 50°$, the
coefficient of reflection $R <$
< 0.05. In certain special
cases, when I_2 is compar-
able to I_1, it is necessary
to deal with multiwave inter-
ference (32.4).

32.3.5 In considering the
interference of light in thin
films, distinction is made
between interference fringes
of equal inclination (iso-
clinic fringes) and those of
equal thickness (isopachic

Fig. 32.6

fringes). *Equal-inclination fringes* are observed when a diverging
(or converging) beam of light is incident at various angles i on
a thin plane-parallel film. These, for instance, are the conditions
that prevail when the film is illuminated by an extended source
or scattered sunlight. Since d and n are the same all over the
film, the optical-path difference of the interfering waves varies
along the surface of the film only owing to the variation in the
angle of incidence i. The conditions of interference are the same
for all rays incident on the surface of the film and reflected from
it at the same angle. Consequently, the interference pattern in
this case is said to be made up of equal-inclination fringes.
Equal-inclination fringes are observed on screen E arranged in
the focal plane of a convergent lens L (Fig. 32.6). With no lens
available the interference pattern could only be observed at
infinity, at the point of intersection of the parallel rays $1'$ and
$1''$, $2'$ and $2''$ etc. Therefore, equal-inclination fringes are said to
be localized at infinity. For their visual observation it would be
necessary for the eyes to accommodate to infinity.
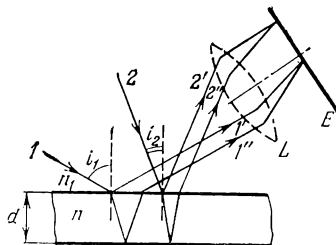
32.3.6 *Equal-thickness fringes* are observed in reflecting a parallel or almost parallel beam of light rays ($i =$ const) from a thin transparent film whose thickness d is not the same at various places. The optical-path difference of the interfering waves varies in going from certain points on the surface of the film to other points with a corresponding change in the film thickness d. Hence the conditions of interference are the same for points having the same values of d. This is why the interference pattern being considered is said to be made up of equal-thickness fringes. Equal-thickness fringes are localized close to the surface of the film. Hence, they can be observed by accommodation of the eyes practically to the film's surface.

When light interferes in a thin transparent wedge with a small wedge angle α, equal-thickness fringes have the form of straight bands parallel to the edge of the wedge. If the wedge is illuminated by monochromatic light with the wavelength λ_0 in vacuum, normally incident on the wedge surface ($i = 0$), the width of the interference fringes (32.2.3) is equal to $\lambda_0/2n\alpha$, where n is the absolute refractive index of the wedge material.

32.3.7 Equal-thickness fringes having the form of concentric rings, and called *Newton's rings*, are observed in the interference of light in the thin air clearance between a flat glass plate A and a flat-convex lens L pressed tightly against the plate (Fig. 32.7). The flat surface of the lens is parallel to the plate surface and light is normally incident to the lens surface. The centre of Newton's rings coincides with point O of contact between the lens and plate. At a small distance r from point O the optical-path difference of the waves reflected from the upper and lower surfaces of the air clearance is
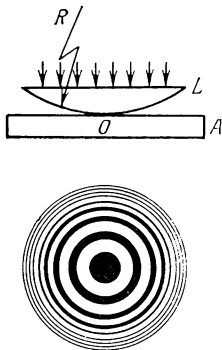
Fig. 32.7

$$\Delta s \approx n\, \frac{r^2}{R} + \frac{\lambda_0}{2}\,,$$

where $n \approx 1$ (for air), and R is the radius of the convex surface on the lens.

Note. In calculations concerning Newton's rings, the wave reflected from the upper (flat) surface of the lens need not be taken into account because the optical-path difference between this wave and the waves reflected from the boundaries of the air clearance is greater than the coherence length for nonlaser light (32.3.3).

In reflected monochromatic light with the wavelength in air $\lambda = \lambda_0/n \approx \lambda_0$, the radii of the dark and bright Newton's rings are

$$r_{\text{dark}} = \sqrt{mR\lambda} \quad \text{and} \quad r_{\text{bright}} = \sqrt{(2m + 1)\frac{R\lambda}{2}} \, ,$$

where $m = 0, 1, 2, \ldots$. The dark spot in the centre of the pattern corresponds to the change in the phase of the wave by π when it is reflected from the lower surface of the air clearance. If white light is incident on the lens, a central dark spot is observed in the reflected light, surrounded by a system of coloured rings. These rings correspond to the interference maxima of the reflection of light with various values of wavelength λ.
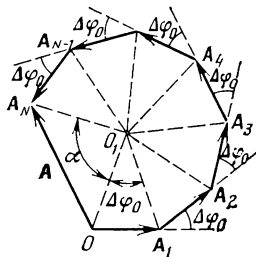
Fig. 32.8

32.4 Multiwave Interference

32.4.1. Special interference instruments, such as the diffraction grating (33.3.4), the Fabry-Perot etalon and others are used to obtain interference of many light waves with close or equal amplitudes. The amplitude A of the resultant vibration and its intensity $I = A^2$ at arbitrary point M can be found by applying the method of vector diagrams for adding oscillations or vibrations having the same direction (28.4.2).

32.4.2 Shown in Fig. 32.8 is a vector diagram of the addition of vibrations in interference of N waves. At the point M being considered, these vibrations set up coherent vibrations in the

same direction with equal amplitudes $A_i = A_1$ and a phase difference, independent of i, between $(i + 1)$th and ith vibrations. Thus $\Phi_{i+1}(t) - \Phi_i(t) = \Delta\varphi_0$. The amplitude of the resultant vibration is

$$A = 2 \times \overline{00}_1 \left| \sin \frac{\alpha}{2} \right|, \quad \text{where} \quad \alpha = 2\pi - N\Delta\varphi_0$$

and $\quad \overline{00}_1 = \dfrac{A_1}{2 \left| \sin \dfrac{\Delta\varphi_0}{2} \right|}$.

Hence,

$$A = A_1 \left| \frac{\sin \dfrac{N\Delta\varphi_0}{2}}{\sin \dfrac{\Delta\varphi_0}{2}} \right| \quad \text{and} \quad I = I_1 \frac{\sin^2 \dfrac{N\Delta\varphi_0}{2}}{\sin^2 \dfrac{\Delta\varphi_0}{2}} ,$$

where $I_1 = A_1^2$ is the intensity of the vibration set up at point M by each of the N interfering waves separately.

32.4.3 The *principal maxima in the interference* of N waves (32.4.2) are at points M that comply with the condition: $\Delta\varphi_0 = = \pm 2n\pi$, where $n = 0, 1, 2, \ldots$ is the *order of the principal*
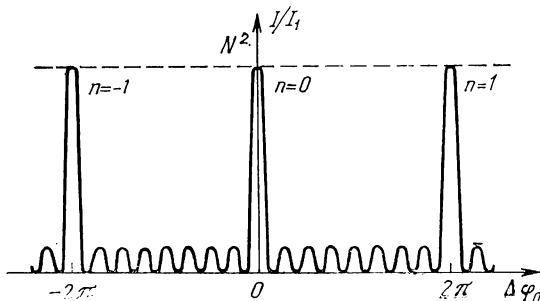
Fig. 32.9

maximum. The amplitude **and** intensity of vibration at the principal maxima are:

$$A_{\max} = NA_1 \quad \text{and} \quad I_{\max} = N^2 I_1.$$

Interference minima ($A = 0$) comply with the condition: $\Delta\varphi_0 = \pm 2\pi p/N$, where p may be any positive whole number except a multiple of N.

The nature of the dependence of I/I_1 on $\Delta\varphi_0$ (32.4.2) is shown in Fig. 32.9. Between each pair of adjacent interference minima there is a maximum, either a *principal* or a *subordinate* maximum. At high N values the intensities of the subordinate maxima are negligibly small compared to that of the principal maxima. The values $\Delta\varphi_0 = \pm (2\pi n \pm 2\pi/N)$ correspond to two minima that bound a principal maximum of the nth order. Therefore, the "width" of the principal maximum, equal to $4\pi/N$, is inversely proportional to the number N of interfering waves, and its intensity is proportional to N^2. The fact that the interference pattern changes in such a manner as N is varied fully complies with the law of conservation of energy. The total energy of vibration at all points of the screen on which the interference pattern is observed is proportional to N.

32.4.4 If the number N of interfering waves (32.4.2) is increased indefinitely, and their amplitudes A_1 and phase differences $\Delta\varphi_0$ are correspondingly reduced so that NA_1 and $N\Delta\varphi_0$ remain

Fig. 32.10

finite quantities, equal to A_0 and $\Delta\varphi$, then, in the limit, the vector diagram (Fig. 32.8) assumes the form illustrated in Fig. 32.10. Vector A, the amplitude of the resultant vibration, closes the arc BC of a circle. The length of this arc is equal to A_0 and its central angle $\angle BOC = \Delta\varphi$. Therefore, the radius of the circle $\overline{OB} = A_0/\Delta\varphi$, and the amplitude A and intensity I of the resultant vibration equal
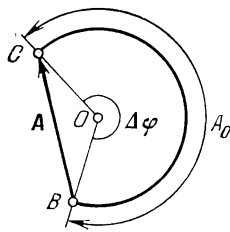
$$A = A_0 \left| \frac{\sin \dfrac{\Delta\varphi}{2}}{\dfrac{\Delta\varphi}{2}} \right| \quad \text{and} \quad I = I_0 \frac{\sin^2 \dfrac{\Delta\varphi}{2}}{\left(\dfrac{\Delta\varphi}{2}\right)^2} \, ,$$

where $I_0 = A_0^2$.

Interference minima are at points of the interference pattern for which

$$\Delta\varphi = \pm\ 2m\pi \quad \text{(where} \quad m = 1,\ 2,\ \ldots).$$

Interference maxima are at points for which $\Delta\varphi = \pm 2k_m\pi$, where $m = 0, 1, 2, \ldots$ is the *order of the maximum*. The values
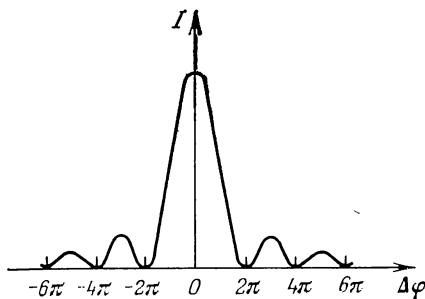
Fig. 32.11

of factor k_m are determined from the transcendental equation $\tan k_m\pi = k_m\pi$. For the central, zero-order maximum $k_0 = 0$ and $\Delta\varphi = 0$. The amplitude and intensity of vibration at a zero-order maximum are equal to A_0 and I_0. For all other maxima $(m \geqslant 1)$ it can be assumed approximately that

$$k_m = (2m + 1)/2 \quad \text{and} \quad \Delta\varphi = \pm\ (2m + 1)\ \pi.$$

Accordingly, the ratio of the intensities for the mth and zero-order maxima equals

$$\frac{I_m}{I_0} = \frac{4}{(2m + 1)^2\ \pi^2}\ .$$

TABLE 32.1

Order of the maximum	0	1	2	3	4
I_m/I_0	1	0.045	0.016	0.008	0.005

This ratio decreases rapidly with an increase of m (Table 32.1). The shape of the curve showing the dependence of I on $\Delta\varphi$ is shown in Fig. 32.11.

CHAPTER 33   DIFFRACTION OF LIGHT

33.1 Huygens-Fresnel Principle

33.1.1 If the position of the wavefront (30.2.3) is known at a certain instant t of time, as well as the velocity v of the wave, the position of the wavefront at the subsequent instant $t + \Delta t$ can be determined on the basis of the *Huygens principle*. According to this principle, each point on the surface S (t), through which the wavefront passes at the instant t can be regarded as a source of secondary waves, and the sought-for position S $(t + \Delta t)$ of the wavefront at the instant of time $t + \Delta t$ coincides with the envelope of all the secondary waves at this later time. It is assumed here that in a homogeneous medium the secondary waves radiate only forward, i.e. in directions making acute angles with the outward normals to the wavefront. In a homogeneous isotropic medium, the secondary waves are spherical (Fig. 33.1).
33.1.2 The Huygens principle can be used to derive laws for the reflection and refraction of light at the interface between two media. The line MN in Fig. 33.2 represents the plane interface between two media in which the velocities of light equal v_1 and v_2. Incident on this interface at the angle i is a plane wave (rays 1 and 2). At the instant t of time the wavefront (plane AB) reaches the interface at point A. Hence point A begins to radiate secondary waves, propagating both in the first medium (reflected wave) and in the second medium (transmitted wave). During the time Δt required for the incident wave to travel the

distance $\overline{BC}$ (wher $\Delta t = \overline{BC}/v_1$), the front of the secondary wave, radiated fr $\vert$ ι point A, reaches points of a semisphere in the first medium having the radius $R_1 = v_1\Delta t = \overline{BC}$, and points of a semisphere in the second medium having the radius $R_2 = v_2\Delta t = v_2\overline{BC}/v_1$. The front of the reflected wave (rays $1'$
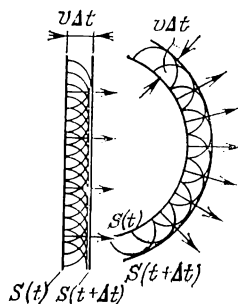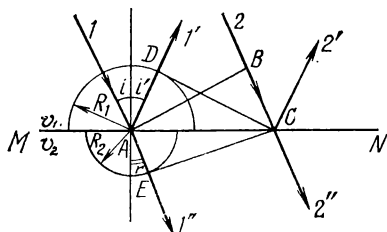
Fig. 33.1   Fig. 33.2

and $2'$), propagating at the angle i' of reflection is the plane DC, tangent to a sphere of radius R_1 with its centre at point A. Correspondingly, the front of the transmitted (refracted) wave (rays $1''$ and $2''$), propagating at the refraction angle r, is the plane CE, tangent to a semisphere of radius R_2 with its centre at point A. The *law of light reflection*: $i = i'$ follows from the equality of triangles ACD and ACB. The *law of light refraction* follows from the rectangular triangles ACB and ACE, which have a common hypotenuse. Thus

$$\frac{\sin i}{\sin r} = \frac{\overline{BC}}{\overline{AE}} = \frac{v_1}{v_2} = n_{21},$$

where n_{21} is the relative refractive index of the second and first media (31.5.1).

33.1.3 The Huygens principle is a purely geometrical one. It indicates no method of calculating the amplitude of the wave that is an envelope of the secondary waves. Therefore, the Huygens principle is insufficient for deriving the laws of propaga-

tion of light waves. An approximate method of solving this problem, which is a development of the Huygens principle on the basis of the idea proposed by Fresnel on the coherence of secondary waves and their interference upon superposition is called the *Huygens-Fresnel principle*. This principle can be expressed by the following series of propositions:

(a) In calculating the amplitude of light vibration, set up by source S_0, at arbitrary point M, the source can be replaced by a system of secondary sources. These sources are small portions ds of any closed auxiliary surface S arranged so that it encloses the source S_0, but does not enclose point M being considered.

(b) The secondary sources are coherent with S_0 and with one another. Hence, the secondary waves they set up interfere upon superposition. Interference calculations are simplest when S is the wavefront for the light from source S_0 because, in this case, the phases of vibration of all secondary sources are the same.

(c) The amplitude dA of the vibration set up at point M by a secondary source is proportional to the ratio of the area ds of the corresponding portion of wavefront S to the distance r from this portion to point M and depends upon the angle α between the outward normal to the wavefront and the direction of a line from the element ds of area to point M. Thus

$$dA = f(\alpha) \frac{a\, ds}{r} ,$$

where a is a quantity proportional to the amplitude of the primary wave at points of the element ds of area; $f(\alpha)$ decreases monotonically from 1 at $\alpha = 0$ to 0 at $\alpha \geqslant \pi/2$ (secondary sources do not radiate backward)*.

(d) If part of surface S is an opaque screen, the corresponding secondary sources (closed by the screen) do not radiate, whereas the rest radiate in the same way as in the absence of screens.**

---

* As shown by Kirchhoff, $f(\alpha) \propto (1 + \cos\alpha)$, i.e. the function becomes zero only at $\alpha = \pi$; at small angles of diffraction α, however, this more exact relation is of no significance.

** Actually, the material of the opaque screen has an effect on the radiation of the open secondary sources that are close to the edge of the screen (at a distance of the order of the wavelength of the light).

33.1.4 On the basis of the Huygens-Fresnel principle, the *law of rectilinear propagation of light* in a homogeneous medium can be substantiated from the wave point of view. Assume that S_0 is a point source of monochromatic light (Fig. 33.3) and M is the point of observation. As the auxiliary surface S, a spherical surface of radius R is chosen so that the distance L from point M to this sphere ($L = \overline{OM}$) is of the order of R. Next, surface S
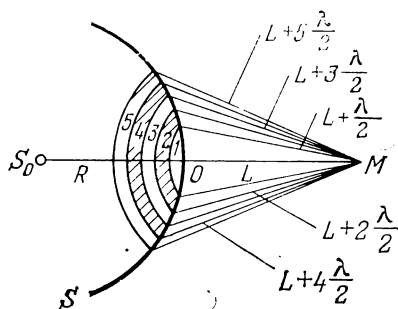
Fig. 33.3

is divided into annular portions of small area, called Fresnel zones, as shown in Fig. 33.3, where λ is the wavelength of the light.

Vibrations set up at point M by two adjacent zones are opposite in phase, because the path difference from like points of these zones to point M is equal to $\lambda/2$. Consequently, the amplitude of the resultant vibration at point M is equal to $A = A_1 - A_2 + A_3 - A_4 + \ldots$, where A_i is the amplitude of vibration set up at point M by secondary sources, located within the limits of the ith zone. At $i \ll L/\lambda$, the areas of all the zones are the same: $\sigma_i = \pi R L \lambda/(R + L)$, and very small (at $R = L = 10$ cm and $\lambda = 5 \times 10^{-5}$ cm we obtain $\sigma_i \approx 8 \times 10^{-5}$ cm^2). With an increase in i, the distance r_i from the zone to point M also increases, as does angle α_i between the normal to the surface of the zone and the direction of a line from the zone to point M.

Therefore, according to the Huygens-Fresnel principle, $A_1 >$ $> A_2 > A_3 > \ldots$ and $A_i \approx (A_{i-1} + A_{i+1})/2$. Consequently, the amplitude of vibration at point M is equal to $A \approx A_1/2$, i.e. the resultant effect of the whole open wavefront is equal to one-half of the effect of the first (central) Fresnel zone, whose radius is very small. Thus, it can be assumed that practically light is propagated from S_0 to M in a straight line.

33.1.5 If an opaque screen, covering all, except the first, Fresnel zones for the point M of observation is placed in the path of monochromatic light from source S_0, the amplitude and intensity of light at point M are doubled and quadrupled, respectively, compared to their values in the absence of the screen: $A = A_1$ and $I = A_1^2$. Light can be intensified to a substantially greater extent by means of a *zone plate*. This is a glass plate to which an opaque coating has been applied in the form of rings which enclose only the even (or only the odd) Fresnel zones, i.e. every other zone. A zone plate has the same effect on light as a converging lens.

33.2 Fresnel Diffraction

33.2.1 The *diffraction of light* is the whole set of phenomena due to the wave nature of light and observed in its propagation in a medium with sharply defined optical inhomogeneities (for example, in passing through apertures in opaque screens or close to the boundaries of opaque bodies). In its narrower sense, diffraction refers to the bending of light around small opaque obstacles, i.e. deviations from the laws of geometrical optics. Distinction is made between two kinds of light diffraction: *Fresnel diffraction*, or *diffraction with convergent rays*, and *Fraunhofer diffraction*, or *diffraction with parallel rays*. In the former, a spherical or plane wave is incident on the obstacle and the diffraction pattern is observed on a screen at a finite distance behind the obstacle. In the latter, a plane wave is incident on the obstacle and the diffraction pattern is observed on a screen located in the focal plane of a converging lens placed in the path of the light that passes by the obstacle. In Fresnel diffraction, a "diffraction image" of the obstacle is obtained on the screen; in Fraunhofer diffraction, the "diffraction image" is of the distant light source.

33.2.2 In the simplest Fresnel diffraction problems, the diffrac-

tion pattern can be investigated by using the method of Fresnel
zones (33.1.4).

Example 1. *Fresnel diffraction by a small circular aperture* in an
opaque screen AB (Fig. 33.4). When the aperture is illuminated
by monochromatic light of wavelength λ, a system of alternat-
ing dark and bright interference rings is observed on screen E,
which is parallel to AB. These rings have a common centre at
point O, located opposite the centre of the aperture. If for point O
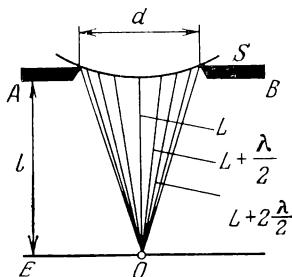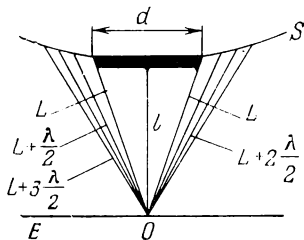
Fig. 33.4   Fig. 33.5

an even number $2k$ (where $k = 1, 2, \ldots$) of Fresnel zones fit into
the aperture, there is a dark spot at point O; the amplitude of
light at point O is less than in the absence of a screen:

$$A \approx \frac{1}{2}(A_1 - A_{2k}) < A_1/2,$$

where A_i is the amplitude corresponding to only the ith zone. If
the number of zones is odd $(2k + 1)$, there is a bright spot
at point O and

$$A \approx \frac{1}{2}(A_1 + A_{2k+1}) > A_1/2.$$

If the aperture is illuminated by white light, a system of con-
centric coloured rings is observed on screen E.

The number of Fresnel zones that fit into the aperture and the
degree of contrast of the interference pattern depend upon the
ratio of the diameter d of the aperture to the distance l between
screens AB and E. As d/l is increased, the amplitude of light at

the centre of screen E approaches $A_1/2$ and the degree of contrast of the interference rings is reduced.

33.2.3 Example 2. *Fresnel diffraction by a small disk* (opaque round screen). The method of constructing open Fresnel zones on the wavefront S of the incident monochromatic spherical wave is illustrated in Fig. 33.5. The interference pattern on screen E consists of alternating bright and dark rings with their centre at point O, where the interference maximum (*Poisson's spot*) is always observed. The amplitude of light at point O is one-half of the amplitude A_1, corresponding to the effect at this point of only the single first open Fresnel zone: $A = A_1/2$. When the disk is illuminated by white light, a white spot is observed at the centre of the screen. This spot is surrounded by a system of concentric coloured rings.

As the ratio of the diameter d of the disk to its distance l from screen E is increased, the brightness of Poisson's spot is gradually reduced and the next dark ring becomes broader, forming a region of shadow beyond the disk.

33.3 Fraunhofer Diffraction

33.3.1 Example 1. *Diffraction of light by a narrow long slit* in an opaque screen (Fig. 33.6). The slit width $\overline{BC} = b$, and its length in the direction perpendicular to the plane of ·the figure is $l \gg b$. Light is normally incident on the surface of the slit so that vibrations at all points of the slit are in phase. The diffraction pattern is observed on screen E, which is placed in the focal plane of converging lens L. Parallel rays BM and CN, travelling from the edges of the slit at the *diffraction angle* ψ to the direction of incident light, are collected by the lens at its secondary focus F_ψ. The lens are made so that the paths of the rays of light BMF_ψ and DNF_ψ, where D is the foot of the perpendicular dropped

Fig. 33.6

from point B to ray CN, are tautochronous paths (32.3.2). Therefore, the result of the interference at point F_ψ of the screen
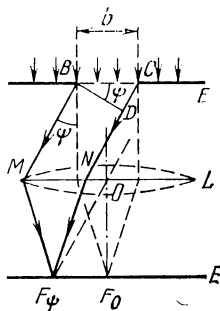
depends upon the path difference $\overline{CD} = b \sin \psi$ and the wavelength λ of the light.

(a) *Approximate solution.* The slit can be divided in width into Fresnel zones, having the form of strips parallel to edge B, such that the path difference from the edges of the strips is equal to $\lambda/2$. The number of such Fresnel zones that fit into the slit is equal to $2b|\sin \psi|/\lambda$. All the zones radiate light entirely equally in the direction being considered; and the vibrations set up at point F_ψ by two adjacent zones are equal in amplitude and opposite in phase. Therefore, when there is an even number of zones:

$$b \sin \psi = \pm 2m \frac{\lambda}{2}, \quad \text{where} \quad m = 1, 2, \ldots,$$

and a diffraction minimum is observed (complete darkness). When the number of zones is odd:

$$b \sin \psi = \pm (2m+1) \frac{\lambda}{2}, \quad \text{where} \quad m = 1, 2, \ldots,$$

and a diffraction maximum corresponding to one Fresnel zone, is observed. The brightest central maximum is observed at the principal focus F_0 of the lens ($\psi = 0$). With an increase in m the width of the Fresnel zones and the intensity of the maxima are rapidly reduced.

(b) *Exact solution.* The slit is divided into a very great number of identical very narrow strips, parallel to edge B. Secondary waves radiated by these elements of the slit, set up vibrations at point F_ψ that have the same small amplitudes, and their initial phases continuously fill an interval of width $\Delta \varphi = (2\pi b \sin \psi)/\lambda$. According to Sect. 32.4.4, the amplitude and intensity of light at point F_ψ equal:

$$A_\psi = A_0 \left| \frac{\sin \dfrac{\pi b \sin \psi}{\lambda}}{\dfrac{\pi b \sin \psi}{\lambda}} \right| \quad \text{and} \quad I_\psi = I_0 \frac{\sin^2 \dfrac{\pi b \sin \psi}{\lambda}}{\left(\dfrac{\pi b \sin \psi}{\lambda} \right)^2},$$

where A_0 and I_0 are the amplitude and intensity at the central maximum ($\psi = 0$).

The condition for diffraction minima of light is the same as in the approximate solution: $b \sin \psi = \pm 2m\lambda/2$, where $m = 1, 2, \ldots$. The condition for diffraction maxima is of the form

$$\tan \frac{\pi b \sin \psi}{\lambda} = \frac{\pi b \sin \psi}{\lambda}$$

and it differs only slightly from the condition obtained by means of the Fresnel zones method.

33.3.2 If white light, instead of monochromatic light, falls on the slit, the central maximum is white with iridescent colouring along its edges. All the other interference fringes are coloured because different angles ψ and different points F_ψ on the screen correspond to both minima and maxima of the same orders m, depending upon the wavelength λ.

As the width b of the slit is reduced, the width of the central maximum is increased because of the increase of angles $\psi_1 = \pm \mathrm{arc} \sin (\lambda/b)$, which correspond to minima of the first order, limiting the central maximum. At $b \leqslant \lambda$ the illumination of the screen is reduced monotonically from the centre (point F_0) toward the edges. When the slit is very wide ($b \gg \lambda$), a bright and sharp image of the light source, formed by lens L according to the laws of geometrical optics, is seen on the screen.

33.3.3 **Example 2.** *Diffraction of light by a circular aperture.* A plane monochromatic light wave is normally incident on the aperture, i.e. perpendicular to the plane of the aperture. The diffraction pattern is observed in the focal plane of a converging lens arranged beyond the aperture so that its optical axis is perpendicular to the plane of the aperture. The diffraction pattern has the form of a bright white spot at the principal focus F_0 of the lens and, concentric about this spot, alternating dark and bright rings. The intensity of the bright rings is very low compared to that I_0 of the central maximum, and decreases with an increase in their radius. For instance, the intensity of the nearest maximum of the first order to the central one is $I_1 < 0.02\ I_0$. The diffraction angle ψ_1, corresponding to the first dark ring, bounding the central maximum, complies with the condition

$$\sin \psi_1 = 1.22 \frac{\lambda}{D} \ ,$$

where D is the diameter of the aperture and λ is the wavelength of the light.

If light falls on the aperture at a small angle α to the normal to the plane of the aperture, the interference pattern remains practically unchanged, but its centre is displaced to a secondary focus of the lens, corresponding to the angle $\psi = \alpha$.

33.3.4 Example 3. *Diffraction of light by a one-dimensional grating.* A one-dimensional *diffraction grating* is a system of a large number N of parallel slits in a screen, the slits being identical in width and separated by opaque strips, also of equal width. Shown in Fig. 33.7 are only two adjacent slits of the grating. The quantity $d = a + b$, where $a = \overline{CD}$ is the width of an opaque strip between two slits and $b = \overline{BC}$ is the slit width, is called the *diffraction grating constant*, or *interval*.
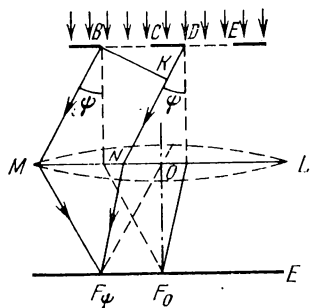
Fig. 33.7

In analysing the diffraction pattern on screen E, mounted in the focal plane of converging lens L, it is necessary to take into account the interference of the secondary waves not only from different portions of one slit, but from the different slits as well. If a plane monochromatic wave is normally incident on the grating, the vibrations at all points of the slits are in phase.

Vibrations set up at the arbitrary point F_ψ of the focal plane of lens L by each of the slits coincide in amplitude (A_1) and differ in phase. For each pair of adjacent slits the phase difference $\Delta\varphi_0$ between these vibrations is the same. It depends upon the wavelength λ and the path difference from like points of the slits (for instance, points B and D), i.e. on the length $\overline{KD} = d \sin \psi$, where K is the foot of the perpendicular, dropped from point B to the ray DN. Thus

$$\Delta\varphi_0 = \frac{2\pi d}{\lambda} \sin \psi.$$

Therefore, according to Sects. 32.4.2 and 33.3.1, the amplitude and intensity of the resultant vibration at point F_ψ equal

$$A = A_0 \left| \frac{\sin \dfrac{\pi b \sin \psi}{\lambda}}{\dfrac{\pi b \sin \psi}{\lambda}} \; \frac{\sin \dfrac{\pi N d \sin \psi}{\lambda}}{\sin \dfrac{\pi d \sin \psi}{\lambda}} \right| ,$$

$$I = I_0 \frac{\sin^2 \dfrac{\pi b \sin \psi}{\lambda}}{\left(\dfrac{\pi b \sin \psi}{\lambda} \right)^2} \; \frac{\sin^2 \dfrac{\pi N d \sin \psi}{\lambda}}{\sin^2 \dfrac{\pi d \sin \psi}{\lambda}} ,$$

where A_0 and I_0 are the amplitude and intensity of the vibration at point F_0 (i.e. at $\psi = 0$) due to the effect of a *single* slit.
33.3.5 The *principal minima* in the diffraction of light by a diffraction grating occur at diffraction angles ψ, corresponding to interference minima in diffraction by a single slit (33.3.1). Thus

$b \sin \psi = \pm m\lambda \; (m = 1, 2, \ldots).$

In these directions none of the slits provide light ("each slit extinguishes itself").
The *principal maxima* (32.4.3) correspond to diffraction angles ψ that comply with the condition

$d \sin \psi = \pm n\lambda,$

where $n = 0, 1, 2, \ldots$ is the *order of the principal maximum*.
Note. If certain values of ψ simultaneously comply with the conditions for both principal maxima and principal minima, then the principal maxima corresponding to these values of ψ are not observed. For example, if $d = 2b$, then all the even principal maxima ($n = 2, 4, 6$, etc.) are absent.
The intensity of a principal maximum of the nth order is

$$I_n = I_0 \left(\frac{Nd}{\pi n b} \right)^2 \sin^2 \frac{\pi n b}{d} .$$

33.3.6 Between each two principal maxima, there are $N - 1$ *additional minima* that comply with the condition

$$d \sin \psi = \pm \frac{p\lambda}{N} ,$$

where p may be any positive whole number with the exception of N, $2N$, $3N$, etc. Accordingly, there are $N-2$ additional maxima whose intensity is negligibly low compared to that of the principal maxima.

The *angular "width" of a principal maximum* of the nth order, i.e. the difference between the values of the angles ψ corresponding to the additional minima bounding this maximum, is

$$\Delta\psi_n = \frac{2\lambda}{Nd\cos\psi_n} = \frac{2\lambda}{L\cos\psi_n},$$

where $\psi_n = \arcsin(n\lambda/d)$, and $L = Nd$ is the length of the diffraction grating. For principal maxima of not particularly high orders, angles ψ_n are small and $\cos\psi_n \approx 1$, so that $\Delta\psi_n \approx \approx 2\lambda/L$.

33.3.7 With monochromatic light the diffraction pattern on screen E in the focal plane of lens L (Fig. 33.7) has, at large N values, the form of narrow and bright principal maxima separated by practically dark and wide intervals. If the ratio d/b is an irrational number, the intensities of the principal maxima (33.3.5) decrease monotonically with an increase in the order n of the maximum.
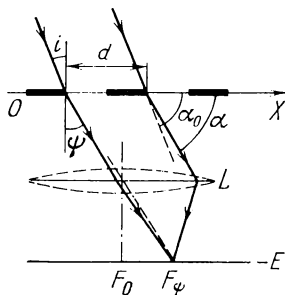
Fig. 33.8

When the grating is illuminated by white light, an uncoloured central zero-order maximum is observed on the screen. On both sides of this maximum, *diffraction spectra* of the 1st, 2nd, etc. orders are observed. These spectra appear as iridescent bands in which there is a continuous transition in colour from blue-violet at the inner edge of the spectrum to red at the outer edge.

33.3.8 Upon *inclined incidence of light* on a diffraction grating (Fig. 33.8), the condition for the principal maxima is of the form

$$d(\sin\psi - \sin i) = \pm n\lambda \quad (n = 0,\ 1,\ 2,\ \ldots),$$

where i is the angle of incidence of light on the surface of the grating. Frequently, the directions of the rays of light falling on the grating and diffracted from it are specified by the angles α_0 and α which these rays make with the axis OX of coordinates, arranged in the plane of the grating, perpendicular to the slits. In this case, the condition for the principal maxima can be rewritten in the form

$$d\,(\cos \alpha - \cos \alpha_0) = \pm\, n\lambda.$$

33.3.9 Two screens are said to be *complementary* if the apertures in one of them exactly coincide in shape, size and relative position with the opaque areas in the other screen, and vice versa. Complementary, for example, are an opaque screen in the form of a circle (disk) of radius R and an opaque screen having a circular aperture of the same radius R. On the basis of the Huygens-Fresnel principle, it is possible to prove *Babinet's theorem* (also called *Babinet's principle*). This theorem states that in Fraunhofer diffraction by any screen, the intensity of diffracted light in any direction, except that of the propagation of a plane wave falling on the screen, should be the same as in diffraction by the complementary screen.

33.3.10 Example 4. *Diffraction by a large number of identical and identically oriented obstacles.* The light intensity I at an arbitrary point M of the diffraction pattern, like in diffraction by a one-dimensional grating (33.3.4), can be presented in the form: $I = fI_1$. Here I_1 is the intensity at point M upon diffraction of the same incident plane wave by a *single* obstacle. The function f depends only upon the number and relative arrangement of the obstacles.

If the obstacles are arranged quite randomly with respect to one another and their number N is large, then $f \approx N$. In this case, the distribution of light intensity is the same as in diffraction by a single obstacle. But the intensity at each point of the diffraction pattern is N times as much. For example, in diffraction of light by a glass plate coated with spherical dust particles of diameter d, a system of bright concentric interference rings is observed. The sizes of these rings correspond to the diffraction of light by an opaque disk of diameter d or, in accordance with Babinet's principle (33.3.9), by a round aperture of the same diameter.

33.4 Diffraction by a Space Lattice

33.4.1 *A space*, or *three-dimensional*, *diffraction grating* is an optically nonhomogeneous medium in which the inhomogeneities are periodically repeated in varying all three spatial coordinates. An example of a space diffraction grating is the crystal lattice of a crystalline solid. Particles at the lattice points (atoms, molecules or ions) serve as orderly arranged centres that coherently scatter light falling on them. Assume that d_1, d_2 and d_3 are the lattice constants along the three coordinate axes ξ, η and ζ that are oriented along three edges of the lattice intersecting in one of its points. Then in Fraunhofer diffraction (33.2.1), the principal maxima comply with the *Laue conditions*

$$d_1 (\cos \alpha - \cos \alpha_0) = n_1\lambda, \quad d_2 (\cos \beta - \cos \beta_0) = n_2\lambda \quad \text{and}$$

$$d_3 (\cos \gamma - \cos \gamma_0) = n_3\lambda.$$

Here α_0, β_0 and γ_0, and α, β and γ are the angles between the coordinate axes ξ, η and ζ, and the directions of propagation of the incident and diffracted rays, respectively, n_1, n_2 and n_3 are whole numbers determining the order of the maximum, and λ is the wavelength of the light. The Laue conditions follow from the relations (33.3.8) for diffraction maxima in inclined incidence of light on a one-dimensional diffraction grating.

33.4.2 Of the three angles α, β and γ (and likewise for angles α_0, β_0 and γ_0) only two angles are independent because they must comply with one geometric relationship whose specific form depends upon the angles between the coordinate axes ξ, η and ζ. For example, if the coordinate axes are mutually perpendicular, i.e. if the lattice is orthogonal, the geometric relationship between α, β, and γ is of the form:

$$\cos^2\alpha + \cos^2\beta + \cos^2\gamma = 1.$$

Therefore, with an arbitrarily given direction of incidence on a space lattice of monochromatic light with a given wavelength λ, it is impossible, in general, to find values of α, β and γ that simultaneously comply with the geometric relationship and Laue's three conditions. The only exception is the zero-order maximum $(n_1 = n_2 = n_3 = 0)$, for which $\alpha = \alpha_0$, $\beta = \beta_0$ and $\gamma = \gamma_0$.

In order to observe a diffraction maximum of the order $(n_1$, n_2 and $n_3)$ at given values of the angles α_0, β_0 and γ_0, it is necessary

for the incident light to have a certain definite wavelength. For instance, in the case of an orthogonal lattice, the wavelength should be

$$\lambda = -2\ \frac{\dfrac{n_1}{d_1}\cos\alpha_0 + \dfrac{n_2}{d_2}\cos\beta_0 + \dfrac{n_3}{d_3}\cos\gamma_0}{\left(\dfrac{n_1}{d_1}\right)^2 + \left(\dfrac{n_2}{d_2}\right)^2 + \left(\dfrac{n_3}{d_3}\right)^2}.$$

If the wavelength λ of the incident light is a fixed value, the Laue conditions and the geometric relationship between angles α, β and γ can simultaneously be complied with by a suitable choice of the direction of light incidence on the diffraction grating, i.e. of angles α_0, β_0 and γ_0.

33.4.3 It follows from Laue's conditions that at $\lambda \geqslant 2d_{max}$, where d_{max} is the maximum of the values of d_1, d_2 and d_3, there should be no maxima, except the zero-order one ($n_1 = n_2 = n_3 = 0$). Light of such wavelengths propagates in a medium "without noticing" its inhomogeneities, i.e. without undergoing diffraction. Hence, the condition $\lambda \geqslant 2d_{max}$ is called the *condition for optical homogeneity of the medium.*

The crystal lattice constants of solids are considerably shorter than the wavelengths of visible light ($d_i \approx 5 \times 10^{-10}$ m, whereas $\lambda_{vis\ l} \approx 5 \times 10^{-7}$ m). Hence, crystals are optically homogeneous media for visible light.* At the same time, crystals are natural diffraction lattices for X rays (31.4.4).

33.4.4 The diffraction of X rays by crystals can be interpreted as the result of interference of X rays that are mirror reflected from a system of parallel planes passing through the crystal lattice points. They are called *planes of atoms,* or *lattice planes, of the crystal.* The distance d between two adjacent lattice planes is called the *interplanar spacing,* or *distance,* and the angle θ between the incident ray and the lattice plane (Fig. 33.9) is called the *grazing* angle. The path difference of rays reflected from two adjacent lattice planes** is $\Delta = \overline{BC} + \overline{BD} = 2d\sin\theta$. There-

---

* Molecular scattering of visible light (34.3.1) is possible in crystals. It is associated with the impairment of their optical homogeneity due to fluctuations in density.

** X rays are not refracted in crystals because the values of the refractive index (31.5.1) of all crystals for electromagnetic radiation of such high frequency are practically equal to unity.

fore, according to Sect. 32.2.1, reflection is observed only in the
directions, corresponding to diffraction maxima that comply
with the Wulff-Bragg condition:

$$2d \sin \theta = m\lambda,$$

where $m = 1, 2, \ldots$ is the order of the diffraction maximum.
33.4.5 The experimental method of investigating the atomic
structure of matter by analysing the diffraction of X radiation
as it passes through the sub-
stance being investigated is
called *X-ray structure analy-
sis.* This method proves most
efficient in studying the
structure of crystalline bod-
ies. The diffraction pattern,
registered on photographic
film, is called the *X-ray photo-
graph,* or *pattern, of the speci-
men.* An X-ray photograph,
obtained in the diffraction of a beam of "white" X radiation
(with a continuous frequency spectrum) by a *monocrystal* is
called a *Laue diffraction pattern.* It consists of discrete diffraction
(Laue pattern) spots whose position is determined by the Laue
conditions (33.4.1). A *Debye-Scherrer-Hull powder photograph,*
or *X-ray powder photograph,* is an X-ray photograph obtained in
the diffraction of a beam of monochromatic X radiation by
a *polycrystalline* specimen (for instance, a crystalline powder).
Such a photographic pattern consists of a system of concentric
diffraction rings. The radii of the rings $r = l \tan 2\theta$, where l is
the distance from the specimen to the plane of the X-ray powder
photograph, perpendicular to the incident ray, and angles θ
comply with the Wulff-Bragg condition (33.4.4).
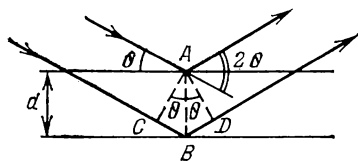
Fig. 33.9

33.5 Resolving Power of Optical Instruments

33.5.1 The image of the object in any optical system (telescope,
microscope, photographic camera, etc.) is obtained by means of
a restricted beam of light admitted into the instrument by the
so-called *aperture diaphragm.* Such a diaphragm may be that of
a photographic camera, lens mount of a telescope, etc. Reducing
the diameter of the aperture diaphragm helps to decrease vari-

ous distortions of the image resulting from the use of wide beams
of light and called *geometrical aberrations of the optical system.*
But, owing to diffraction of light in an optical instrument, the
image of a bright point is not a point, but a bright spot sur-
rounded by a system of concentric interference rings (alternating
dark and bright ones in the case of monochromatic light and
iridescent in the case of white light). This phenomenon limits the
resolving power (resolution) of the optical instrument, i.e. its
capacity to provide separate images of two points of the object
close to each other.

33.5.2 According to *Rayleigh's criterion,* the images of two iden-
tical point sources of light can still be seen separately if the
central maximum of the diffraction pattern of one falls exactly
on the first minimum of the diffraction pattern of the other. It
follows from Sect. 33.3.3 that in accordance with the Rayleigh
criterion, two close stars, observed in a telescope with mono-
chromatic light of the wavelength λ, can be resolved (viewed
separately) if the angular separation (distance) between them is

$$\Delta\varphi \geqslant 1.22\ \lambda/D,$$

where D is the diameter of the objective. The quantity $(\Delta\varphi)_0 =$
$= 1.22\ \lambda/D$ is called the *angular resolution limit of the telescope.*
The reciprocal quantity $1/(\Delta\varphi)_0$ is called the *resolving power of
the telescope.* The resolving power of a telescope increases in
proportion to the diameter of its objective. The condition for
the resolution of a terrestrial telescope and a photographic came-
ra in viewing and photographing distant items coincides with
that for an astronomical telescope.

The *angular resolution limit of the eye* is determined by the
diffraction of light by the pupil ($D \approx 2$ mm) and the granular
structure of the retina. This limit is about one minute of arc.

33.5.3 *The resolving power of a microscope* is characterized by
the quantity $(\Delta l)_0$, the shortest distance between two points of
the object that are separately visible on the image. In the case
of a self-luminous object, all of whose points can be assumed to
be incoherent sources,

$$(\Delta l)_0 = \frac{0.61\lambda_0}{A}\ ,$$

where λ_0 is the wavelength of the light in vacuum, $A = n \sin u$
is the *numerical aperture of the objective,* n is the refractive

index of the medium between the object and the objective, and
u is one half of the aperture angle of the beam of light issuing
from a point of the object and entering the microscope objective
(i.e. the angle between the outermost ray entering the objective
and the optical axis). For non-self-luminous objects, the value
of $(\Delta l)_0$ depends upon the illumination conditions. But in this
case as well, $(\Delta l)_0 \geqslant \lambda_0/A$.

The resolving power of a microscope can be improved either
by reducing the wavelength λ_0, or by increasing the numerical
aperture A (numerical aperture is abbreviated N.A.). The for-
mer method is employed in ultraviolet and electron microscopy,
and the latter, in the immersion microscope, in which the space
between the objective and the object is filled with a transparent
liquid having a refractive index $n > 1$.

33.6 Holography

33.6.1 *Holography* is a technique for obtaining a three-dimen-
sional optical image of objects and is based on the phenomenon
of wave interference. In contrast to the conventional photographic
method, light-sensitive fine-grained photographic emulsion is
used to record the relations, not only between the amplitudes
(or their squares, i.e. the intensities) of light waves scattered by
various small portions of the object's surface, but also between
the phases of these waves.

The essence of the holographic method is made clear in
Fig. 33.10. Photographic plate P (Fig. 33.10a) records the
interference pattern obtained in superposing wave *1*, scattered
by the object Q and called the *signal wave*, or *object beam*, and
wave *2*, coherent with wave *1* and having fixed values of ampli-
tude and phase. Wave *2*, called the *reference wave*, or *reference
beam*, is emitted by the same source of light that illuminates the
object and, after reflection from mirror B, falls directly on the
photographic plate P. The interference pattern recorded on the
photographic plate after it is developed is called the *hologram of
object Q*. It consists of a very fine and intricate pattern of alter-
nating interference maxima and minima in the blackening of
the photographic emulsion and, in contrast to a photographic
image, has no outward resemblance to the object.

Holograms are obtained by the interference of light at large path
differences, i.e. they require an extremely high degree of coher-

ence of the light (32.2.4). For this reason lasers (39.6.8) are used
in holography as the light source.

33.6.2 *Reconstructing the image of the object* from its hologram *C*
(Fig. 33.10*b*) is accomplished by passing reference wave *2*,
from the same laser used to make the hologram, through it as
in the projection of a slide (diapositive). In this operation, the
orientation of the plate with the hologram with respect to the
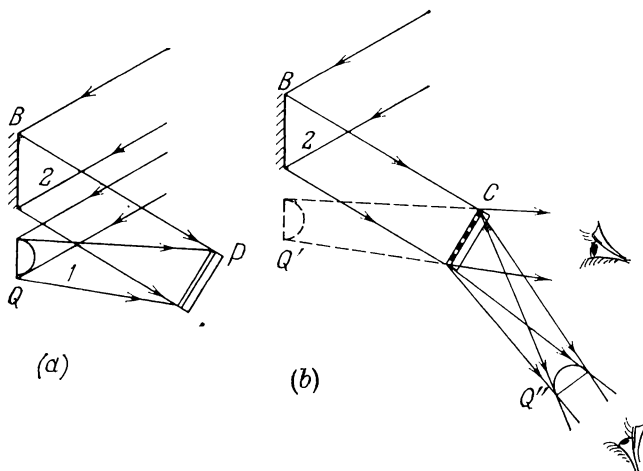
Fig. 33.10

reference wave should also be maintained. Wave *2* is diffracted
by the hologram. As a result of this diffraction, two three-
dimensional images of the object, a virtual and a real image, are
seen. The virtual image *Q'* is at the same place, with respect to
the hologram, where the object *Q* was in taking the photograph.
This image is visible in looking through the hologram as if it
was a window. The real image *Q''* is on the other side of the
hologram. It seems to be in the air in front of the hologram and
is a mirror image of the object.

Ordinarily, the virtual image of the object is used; in visual perception it is identical to the object itself. It not only has the property of being three-dimensional, but its perspective varies with the position of the eye of the observer with respect to the hologram. For example, by moving the head along the hologram, it is possible to see behind some item located in the foreground of the holographic image.

33.6.3 The *interference pattern* at each point of the hologram is determined by light scattered by all the points of the object. Hence, any portion of the hologram contains information about the whole object and can enable the image of the whole object to be restored if, in damaging the hologram, only this single portion remains intact. The smaller the size of the saved portion of the damaged hologram, the less the light diffracted by it in restoring the image. The luminance is correspondingly reduced and the sharpness of the holographic image of the object is impaired. Thus, a hologram has essential advantages with respect to reliability of storage of information over a conventional photograph or photographic negative, each element of which contains information only about the portion of the object depicted on it.

Holographic recording of information is distinguished for its high capacity and compact form. A great many holograms can be recorded on a single photographic plate. For instance, it is sufficient, for this purpose, to record each hologram with a different value of the angle of incidence of the reference wave on the photographic plate.

33.6.4 Holography can be used to obtain coloured three-dimensional images of objects. To make such a "coloured" hologram, use is made of the monochromatic light of three lasers with three primary colours (for instance, red, green and blue). The recording of the interference patterns, corresponding to light of three wavelengths, is accomplished simultaneously or consecutively on the same photographic plate. To reconstruct the coloured three-dimensional image of the object it is necessary to simultaneously direct onto the hologram the three reference beams of monochromatic light that were used in recording.

33.6.5 *Three-dimensional holograms* have special properties not possessed by the ordinary kind. They are obtained by using thick-layer photographic emulsions. Recorded on a three-dimensional hologram is a spatial interference pattern, instead of a plane one. This pattern is obtained in superposing the sig-

nal and reference waves. Such a hologram is similar to a spatial diffraction grating (lattice). It is capable of separating out of the incident white light, monochromatic light of the wavelength that was used in recording the hologram. Therefore, an image recorded as a three-dimensional hologram can be reconstructed by illuminating the hologram either by the corresponding monochromatic light or by white light. If a three-dimensional hologram is "coloured", the coloured three-dimensional image can be reconstructed by simply illuminating the hologram with white light.

CHAPTER 34   ABSORPTION, SCATTERING AND DISPERSION OF LIGHT. VAVILOV-CHERENKOV RADIATION

34.1 Interaction of Light With Matter

34.1.1 According to the concepts of classical electron theory, the variable electromagnetic field of a light wave, propagating in a dielectric medium, sets up forced vibrations of the bound charges (electrons and ions) included in the composition of the molecules of the medium. Consequently, each molecule of the medium can be considered as a system of oscillators with different angular frequencies of natural vibration (28.1.1). Ions are considerably more massive than electrons and execute appreciable vibration only as the effect of low-frequency (infrared) radiation. In the frequency region of visible and ultraviolet light, the deciding role is played by the forced vibration of the external, more weakly bound electrons of atoms and molecules. They are called *optical electrons.*

34.1.2 Electrons and ions that execute forced vibrations by the action of light radiate secondary light waves of the same frequency. The average distance between the molecules of a medium is only a small fraction of the coherence length (32.1.4) of the light. Therefore, the secondary waves radiated by a great number of neighbouring molecules are coherent and display interference when superposed.

If the medium is homogeneous and isotropic (30.1.6), interference results in the formation of a transmitted wave, whose phase velocity depends upon the frequency and whose direction of propagation coincides with that of the primary wave.

34.1.3 When the medium is optically nonhomogeneous, the superposition of the primary and secondary waves results in scattering of the light (34.3.1). Finally, when light falls on the interface between two different media, interference results in the formation of a reflected wave in addition to the transmitted wave. Thus, a more or less considerable layer of particles of the medium, adjoining the reflecting surface, participates in the formation of the reflected wave. Therefore, upon total internal reflection (31.5.8), the electromagnetic field of the light wave is not completely cut off at an interface with an optically less dense medium, but partly penetrates it. The field strength E, however, decreases very rapidly as the distance to the interface increases according to the equation

$$E \propto \exp\left[-\frac{2\pi z}{\lambda_2} \sqrt{(\sin^2 i/n_{21}^2) - 1} \right],$$

where z is the distance from the interface, i is the angle of incidence $(i > i_{cr})$, λ_2 is the wavelength of light in the medium, i_{cr} is the critical angle (31.5.8), and n_{21} is the relative refractive index of the medium.

34.2 Absorption of Light

34.2.1 The *absorption of light* refers to the reduction in the energy of a light wave as it is propagated in a substance. It results from the conversion of energy of the wave into internal energy (9.1.2) of the substance or into energy of secondary radiation having a different spectral composition and other directions of propagation (photoluminescence, Sect. 39.7.5). Light absorption may lead to heating of the substance, excitation and ionization of its atoms or molecules, photochemical reactions and other processes in the substance.

The absorption of light is described by the *Bouguer-Lambert law* (also called the *Lambert law*), which states that the intensity I of a plane wave of monochromatic light decreases as it passes

through an absorbing medium according to the exponential relation

$$I = I_0 e^{-\alpha x},$$

where I_0 and I are the intensities of the light at the entrance to and exit from a layer of the medium of thickness x, and α is the *absorption factor (linear coefficient of absorption)*, which depends on the chemical nature and state of the substance and on the wavelength λ of the light.

Beer's law is valid for dilute solutions of an absorbing substance in a nonabsorbing solvent: $\alpha = bc$, where c is the concentration of the solution and b is a proportionality factor that is independent of c. Beer's law does not hold for high concentrations because of the effect of the interaction between the closely spaced molecules of the absorbing substance.

34.2.2 According to the Bouguer-Lambert law, the equation of a plane, linearly polarized, monochromatic light wave, propagating in an absorbing medium along the positive direction of the OX axis, is of the form:

$$E = E_0 e^{-\alpha x/2} \cos(\omega t - kx).$$

Here E is the strength of the wave's electric field at points with the coordinate x, E_0 is the amplitude of E at points in the plane $x = 0$, ω is the angular frequency of the light, $k = 2\pi/\lambda = \omega n/c$ is the wavenumber, λ is the wavelength of the light in the medium, c is the velocity of light in free space, and n is the refractive index of the medium.

In the exponential form the equation of this wave is of the form (30.2.7):

$$\widetilde{E} = E_0 e^{-\alpha x/2} e^{i(\omega t - kx)} = E_0 e^{i(\omega t - \tilde{n}\omega x/c)},$$

where $\tilde{n} = n\left(1 - i\dfrac{\alpha\lambda}{4\pi}\right) = n(1 - i\varkappa)$ is the *complex refractive index of the medium* $(i = \sqrt{-1})$, and $\varkappa = \alpha/2k = \alpha\lambda/4\pi$ is the *absorptivity*, or *absorptive power*, which characterizes the decrease in the intensity and amplitude of a plane wave as it propagates in a medium.

34.2.3 The dependence of the absorption factor α of a dielectric on the light's wavelength λ, characterizing the *light absorption*

31*

spectrum in this medium, is related to the resonance phenomenon upon forced vibration of the electrons in the atoms and the atoms in the molecules of the dielectric. Dielectrics absorb light more or less selectively: absorption is great only in frequency regions close to the natural frequencies of vibrations of the electrons in the atoms and the atoms in the molecules. This phenomenon of *resonance absorption of light* is displayed most clearly in dilute monatomic gases (for example, the vapours of most metals), which have a *line spectrum of light absorption.* Discrete frequencies of intensive light absorption coincide with the frequencies of the self-radiation of excited atoms in these gases.

Systems of closely spaced lines, forming *absorption bands*, are observed in gases having polyatomic molecules. The structure of the absorption bands is determined by the composition and structure of the molecules. Liquid and solid dielectrics have *continuous absorption spectra*, consisting of comparatively wide absorption bands, within which the absorption factor α reaches a quite high value and varies smoothly in accordance with the wavelength λ. This kind of dependence of α on λ in condensed media is due to the strong interaction between the particles of the medium, leading to the appearance of a great number of additional resonance frequencies.

34.2.4 At sufficiently high intensities of light, deviations from the Bouguer-Lambert law (34.2.1) are observed: the absorption factor of a dielectric medium begins to depend upon I, decreasing with an increase in I. This phenomenon, inexplicable within the framework of classical light absorption theory, is readily interpreted in the quantum theory of the interaction of light with matter. In the absorption of light, a part of the molecules of the medium goes over into an excited state. The excited molecules cannot participate in further light absorption until they return, after expending their surplus energy, to the unexcited ("normal") state. The higher the intensity of the light and the longer the average lifetime $\langle \tau \rangle$ of a molecule in the excited state, the greater the fraction of excited molecules in the medium. If this fraction of excited molecules is negligible, light is absorbed in accordance with the Bouguer-Lambert law. Otherwise, α decreases with an increase in the intensity of the light.

It is feasible to obtain a nonequilibrium state of the medium, in which the fraction of excited molecules is so large that the absorption factor of the medium becomes negative. This phenom-

enon is made use of in a quantum oscillator of radio waves and
light (39.8.8).
34.2.5 Metals in a condensed state contain a huge amount of
conduction electrons and, therefore, have high electrical con-
ductance. When subjected to light, the conduction electrons
execute variable motion and radiate secondary waves. As
a result of the superposition of the primary wave, falling on the
surface of the metal, and the secondary wave, an intensive reflect-
ed wave is formed and a relatively weak wave that enters the
metal. The reflection coefficient (31.5.6) can reach 95 per cent
and even more. It depends upon the surface finish of the metal,
its electrical conductance and the light frequency. The refracted
wave is rapidly absorbed in the metal. Its energy is consumed
on the production of Joule heat, evolved by the conduction
currents initiated by the action of light in the thin layer of
metal adjacent to the surface. In the frequency region of infrared
radiation, the optical properties of metals are determined main-
ly by the conduction electrons. But in the region of visible
light and especially ultraviolet radiation, bound electrons, in
the ions of the metal, begin to play an appreciable role. This
leads to a reduction in the reflection coefficient and to its pro-
nounced dependence on the frequency.

34.3 Scattering of Light

34.3.1 *The scattering of light* is the phenomenon in which light
is transformed by a substance. It is accompanied by a change
in the direction of light propagation, and is manifested as an
extrinsic glow of the substance. This glow is due to the forced
vibration of the electrons in the atoms of the scattering medium
by the action of the incident light. Light scattering occurs in
an optically nonhomogeneous medium, whose refractive index
varies irregularly from point to point due either to fluctuations
in the density of the medium or to the presence in it of small
foreign particles. In the former case, *the light scattering* is said
to be *molecular* or *Rayleigh scattering*; in the second, *light scat-
tering in a turbid medium*. Examples of turbid media are aerosols
(smoke and fog), emulsions and colloidal solutions.
34.3.2 The scattering of light in turbid media by particles whose
size is small compared to the wavelength λ, is called the *Tyndal
effect*. A system of electrons, executing forced vibration in the

atoms of electrically isotropic particles of small size $r_0 \approx$ $\approx$ (0.1 to 0.2) λ, is equivalent to one oscillating electrical dipole (linear harmonic oscillator). This dipole oscillates at the frequency ν of the light incident on it, and the intensity of the light it radiates is proportional to ν^4 (31.3.3). Hence, *the Rayleigh law* is valid for scattered light. It states that the intensity I of scattered light is inversely proportional to the fourth power of the wavelength: $I \propto \lambda^{-4}$. When white light passes through a finely divided turbid medium, short-wave (dark-to-light blue) light predominates in the scattered light, and long-wave (yellow-to-red) light in the transmitted light. This explains, for instance, the light-blue colour of the sky and the yellowish-red colour of the rising and setting sun.

In scattering of natural light (34.1.1), the dependence of the intensity of the scattered light on the scattering angle θ is of the form

$$I_\theta = I_{\pi/2}\,(1 + \cos^2\theta).$$

Here I_θ and $I_{\pi/2}$ are the intensities of light scattered at angles of θ and $\pi/2$ to the direction of the primary light beam falling on a turbid medium. Light scattered at the arbitrary angle θ is partly polarized (35.1.1); that scattered at the angle $\pi/2$ is completely linearly polarized (31.1.7): vector E of the field set up by this light is perpendicular to the plane passing through the incident and scattered rays.

34.3.3 As the size r_0 of the inhomogeneities in a turbid medium is increased, the laws of light scattering are changed. At $r_0 > \lambda$, the dependence of I_θ on θ is of complex form, the intensity of the forward scattering of light (in the directions $0 < \pi/2$) being greater than for backward scattering. This phenomenon is called the *Mie effect*. In this case, light scattered at the angle $\theta = \pi/2$ is only partly polarized. The dependence of the scattered light intensity I on the wavelength λ is of the form: $I \propto \lambda^{-p}$, where $p < 4$ and decreases with an increase in r_0. At $r_0 \gg \lambda$, the spectral compositions of the scattered and incident light practically coincide. This explains, for instance, the white colour of the clouds.

34.3.4 Molecular scattering in pure media, not containing any foreign impurities, is due to inhomogeneities that occur in the process of random thermal motion of the particles of the medium. These inhomogeneities are associated with fluctuations in densi-

ty (11.6.1), whereas in media having anisotropic (polar) mole-
cules, the inhomogeneities are also due to fluctuations in the
orientation of these molecules (fluctuations of anisotropy). In
true solutions, light may be scattered by fluctuations in concen-
tration. Under ordinary conditions the size of the regions of the
medium, corresponding to any appreciable fluctuations, is much
less than the wavelength of visible light. Therefore, the depend-
ence of the scattered light intensity on the wavelength λ and
angle θ, as well as on the kind of polarization in molecular
scattering, is analogous to the corresponding laws for the Tyn-
dal effect. But, in contrast to the Tyndal effect, the intensity
of molecular scattering of light depends upon the temperature of
the medium, increasing with the temperature.

34.4 Normal and Anomalous Light Dispersion

34.4.1 The *dispersion of light* is the dependence of its phase
velocity v in a medium on its frequency v. According to Sect.
31.5.1, $v = c/n$, where c is the velocity of light in free space,
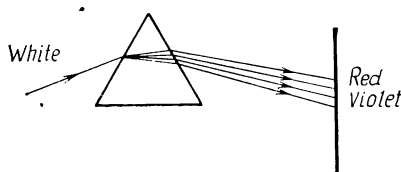
Fig. 34.1

and n is the refractive index of the medium. Since c is a univer-
sal constant, the same for electromagnetic waves of any frequen-
cy, the existence of the dispersion of light in a medium is due to
the fact that its refractive index n depends upon the frequen-
cy v. This dependence is readily detected, for example, in pas-
sing a beam of white light through a prism made of some trans-
parent medium. Observed on the screen, mounted behind the
prism, is an iridescent band (Fig. 34.1) called a *prismatic*, or
dispersion, spectrum.

34.4.2 The dependence of the refractive index n of a medium on the light frequency ν is neither linear nor monotonic. The region of ν values in which $dn/d\nu > 0$, i.e. n increases with ν, corresponds to *normal dispersion of light*. Normal dispersion is found for substances transparent to light. For example, ordinary glass is transparent to visible light and, in this frequency region, normal dispersion of light is observed in glass. A case of normal dispersion of light is illustrated in Fig. 34.1.

The dispersion of light is said to be *anomalous* if $dn/d\nu < 0$, i.e. the refractive index of the medium decreases with an increase in ν. Anomalous dispersion is found in frequency regions that correspond to intensive absorption of light in the given medium (34.2.1). In ordinary glass, for instance, these bands are in the infrared and ultraviolet parts of the spectrum.

34.4.3 Depending upon the kind of dispersion light undergoes, its group velocity u can be either more or less than its phase velocity v. According to Sect. 30.4.3, the group velocity is related to the angular velocity ω of the wave and to its wavenumber k by the equation: $u = d\omega/dk$. Since $\omega = 2\pi\nu$ and $k = 2\pi/\lambda = 2\pi n\nu/c$,

$$u = \frac{c}{n + \nu \dfrac{dn}{d\nu}} = \frac{v}{1 + \dfrac{\nu}{n} \dfrac{dn}{d\nu}}.$$

In normal dispersion the group velocity is less than the phase velocity ($u < v$). In the case of anomalous dispersion $u > v$ and, in particular, if $n + \nu\, dn/d\nu < 1$, then $u > c$. This does not contradict the statement of the special theory of relativity that the velocity of any signal (including a light signal) cannot exceed c (5.1.3). The concept of group velocity correctly describes only a signal whose "shape", i.e. the distribution of amplitudes and energies along its "length", does not change when the signal travels through the medium. But for light, this condition is complied with only approximately; the narrower the frequency spectrum of the signal and the less the light dispersion in the medium, the more exactly the condition is complied with. In the frequency regions corresponding to anomalous dispersion, the group velocity does not coincide with the velocity of the signal because, owing to considerable dispersion of light, the "shape" of the signal changes rapidly as it propagates in the medium.

34.5 Classical Electron Theory of Light Dispersion

34.5.1 Optically transparent media are nonmagnetic ($\mu_r \approx 1$), so that their refractive index (31.5.1) $n = \sqrt{\varepsilon_r} = \sqrt{1 + \chi_e}$, where ε_r and χ_e are the relative permittivity (dielectric constant) and the dielectric (or electric) susceptibility of the medium (18.2.4). Therefore, the dispersion of light can be regarded as the dependence of ε_r and χ_e on the frequency of the variable electromagnetic field of the light, causing electronic polarization of the medium (18.2.2). If each atom (or molecule) of the medium contains one optical electron (34.1.1), the polarization vector of the medium (18.2.3) $\mathbf{P}_e = -en_0\mathbf{r}$, where $-e$ is the charge of the electron, r is its displacement from the equilibrium position, and n_0 is the concentration of atoms (or molecules) of the medium. On the other hand (18.2.4), $\mathbf{P}_e = \varepsilon_0\chi_e\mathbf{E}$, where ε_0 is the electric constant (14.2.7), and E is the strength of the electric field of the light.

34.5.2 An optical electron executes forced vibrations under the effect of the following forces:
(a) restoring quasi-elastic force (40.3.5) $\mathbf{F}_{rest} = -m\omega_0^2\mathbf{r}$, where m and ω_0 are the mass of the electron and the angular frequency of its free undamped vibrations;
(b) force of resistance $\mathbf{F}_{resi} = -2\beta m\, dr/dt$, where β is the damping factor of free vibrations of the electron;
(c) driving force $\mathbf{F} = -e\mathbf{E}$, exerted on the electron by a variable field of strength E.
The equation of forced vibrations is

$$\frac{d^2\mathbf{r}}{dt^2} + 2\beta\frac{d\mathbf{r}}{dt} + \omega_0^2\mathbf{r} = -\frac{e\mathbf{F}}{m}.$$

In the case of linearly polarized monochromatic light with the angular frequency ω, the field strength $\mathbf{E} = \mathbf{E}_0\cos\omega t$, where $\mathbf{E}_0 = $ const is the amplitude vector. If, in addition, the medium does not absorb light, then $\beta = 0$ and the steady-state forced vibration of the optical electron is executed according to the relation

$$\mathbf{r} = -\frac{e\mathbf{E}}{m\,(\omega_0^2 - \omega^2)}.$$

Here the polarization vector of the medium is

$$\mathbf{P_e} = \frac{n_0 e^2 \mathbf{E}}{m\,(\omega_0^2 - \omega^2)} \quad \text{and} \quad \chi_e = \frac{n_0 e^2}{\varepsilon_0 m\,(\omega_0^2 - \omega^2)}\,.$$

The dependence of the refractive index of the medium on ω is of the form

$$n^2 = 1 + \frac{n_0 e^2}{\varepsilon_0 m\,(\omega_0^2 - \omega^2)}\,.$$

34.5.3 At values of ω close to ω_0, the absorption of light in the medium can no longer be neglected, assuming that $\beta = 0$. In an absorbing medium (i.e. at $\beta \neq 0$), vibrations of the optical electron and of vector $\mathbf{P_e}$ are out of phase with respect to the oscillations of the field strength $\mathbf{E}$ (29.2.2):

$$\mathbf{r} = \mathbf{A}\cos(\omega t + \varphi_0),$$

where

$$\mathbf{A} = -\frac{e\mathbf{E}_0}{m\,\sqrt{(\omega_0^2 - \omega^2)^2 + 4\beta^2\omega^2}} \quad \text{and} \quad \tan\varphi_0 = -\frac{2\beta\omega}{\omega_0^2 - \omega^2}\,.$$

Accordingly

$$\mathbf{P_e} = \frac{n_0 e^2 \mathbf{E}_0 \cos(\omega t + \varphi_0)}{m\,\sqrt{(\omega_0^2 - \omega^2)^2 + 4\beta^2\omega^2}}\,.$$

Introduced to describe the properties of a light-absorbing medium, along with the complex refractive index (34.2.2) $\tilde{n} = n\,(1 - i\varkappa)$, are the *complex dielectric susceptibility* $\tilde{\chi}_e$ and the *complex dielectric constant* $\tilde{\varepsilon}_r$:

$$\tilde{\chi}_e = \frac{\tilde{P}_e}{\varepsilon_0 \tilde{E}} \quad \text{and} \quad \tilde{\varepsilon}_r = 1 + \tilde{\chi}_e, \quad \text{whereas} \quad \tilde{n}^2 = 1 + \tilde{\chi}_e.$$

Here $\widetilde{P}_e$ and $\widetilde{E}$ are complex values of the polarization and strength of the field:

$$\widetilde{P}_e = \frac{n_0 e^2 E_0 e^{i(\omega t + \varphi_0)}}{m \sqrt{(\omega_0^2 - \omega^2)^2 + 4\beta^2 \omega^2}} \quad \text{and} \quad \widetilde{E} = E_0 e^{i\omega t} ,$$

so that

$$\widetilde{n^2} = n^2 (1 - i\varkappa)^2 = 1 + \frac{n_0 e^2}{\varepsilon_0 m \sqrt{(\omega_0^2 - \omega^2)^2 + 4\beta^2 \omega}} e^{i\varphi_0} ,$$

$$n^2 (1 - \varkappa^2) = 1 + \frac{n_0 e^2 \cos \varphi_0}{\varepsilon_0 m \sqrt{(\omega_0^2 - \omega^2)^2 + 4\beta^2 \omega^2}}$$

$$= 1 + \frac{n_0 e^2 (\omega_0^2 - \omega^2)}{\varepsilon_0 m [(\omega_0^2 - \omega^2)^2 + 4\beta^2 \omega^2]} ,$$

$$2n^2 \varkappa = - \frac{n_0 e^2 \sin \varphi_0}{\varepsilon_0 m \sqrt{(\omega_0^2 - \omega^2)^2 + 4\beta^2 \omega^2}} = \frac{2 n_0 e^2 \beta \omega}{\varepsilon_0 m [(\omega_0^2 - \omega^2)^2 + 4\beta^2 \omega^2]} .$$

34.5.4 In the classical electron theory of light dispersion in gases, each molecule of the gas is dealt with as a system of q linear oscillators. If ω_{0j} and β_j are the natural angular frequency and damping factor of the jth oscillator, then

$$n^2 (1 - \varkappa^2) = 1 + \frac{n_0 e^2}{\varepsilon_0 m} \sum_{j=1}^{q} \frac{(\omega_{0j}^2 - \omega^2) f_j}{(\omega_{0j}^2 - \omega^2)^2 + 4\beta_j^2 \omega^2}$$

and

$$n^2 \varkappa = \frac{n_0 e^2 \omega}{\varepsilon_0 m} \sum_{j=1}^{q} \frac{\beta_j f_j}{(\omega_{0j}^2 - \omega^2)^2 + 4\beta_j^2 \omega^2} .$$

The dimensionless coefficient f_j characterizes the contribution of the jth oscillator to the dispersion and absorption of light. It

is called the *oscillator strength*. In the classical theory of disper-
sion, the values of ω_{0j} and f_j are assumed to be known from
experimental investigations.
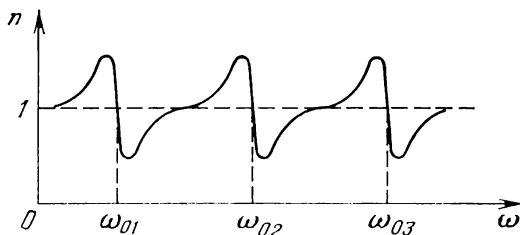
Fig. 34.2

For gases, $\varkappa \ll 1$ and n differs only slightly from unity, so that
$n^2 - 1 = (n + 1)(n - 1) \approx 2(n - 1)$. Therefore, the depen-
dence of n on ω is of the form

$$n = 1 + \frac{n_0 e^2}{2\varepsilon_0 m} \sum_{j=1}^{q} \frac{(\omega_{0j}^2 - \omega^2) f_j}{(\omega_{0j}^2 - \omega^2)^2 + 4\beta_j^2 \omega^2} \cdot$$

The curve of this equation is shown in Fig. 34.2.
Anomalous dispersion is observed close to each of the frequen-
cies ω_{0j}.

34.6 Vavilov-Cherenkov Radiation

34.6.1 *Vavilov-Cherenkov radiation* (or *effect*) is the radiation of
light, differing from luminescence (39.7.1), emitted in the mo-
tion of charged particles in a substance at velocities V greater
than the phase velocity v of light in this substance. The condi-
tion for the existence of this radiation is: $c/n < V < c$, where c
is the velocity of light in free space (or in vacuum), and $n > 1$
is the refractive index of the substance.

In the process of Vavilov-Cherenkov radiation, the energy and velocity of the free emitting particle is reduced, i.e. the particle is retarded. But, in contrast to braking radiation (bremsstrahlung) of slowly moving charged particles (31.3.4), *resulting from their change in velocity*, the reduction in the velocity of particles in Vavilov-Cherenkov radiation is itself a *consequence of* this *radiation*. In other words, if the loss of energy of the particle, consumed in emitting Vavilov-Cherenkov radiation, could be replenished in some way and the particle would continue travelling in the substance at a "faster-than-light" velocity ($V > v$), Vavilov-Cherenkov radiation would still be observed, whereas there would be no braking radiation in this case.

34.6.2 A charged particle gives rise to short-duration polarization (18.2.2) of the substance in the vicinity of the points through which it travels. Hence, the molecules of the medium, lying on the path of the particle, become temporarily acting coherent sources (30.5.1) of elementary electromagnetic waves that interfere upon superposition.

If $V < v = c/n$, the elementary waves cancel one another. Assuming that the charged particle travels at the velocity **V** (where $V < v$) along the OX axis (Fig. 34.3) and is at points A and B at the instants of time t and $t + \Delta t$, respectively. The distance between A and B is $l = V\Delta t$. The path difference of elementary waves, radiated from points A and B in the arbitrary direction **n**, making the angle α with vector **V**, is

$$\Delta = \overline{DF} = (v - V\cos\alpha)\,\Delta t = l\left(\frac{v}{V} - \cos\alpha\right).$$

It is possible, for each value of the wavelength λ, to find a value $l = l_{\alpha\lambda}$ at which $\Delta = \lambda/2$, so that the elementary waves cancel one another. Thus

$$l_{\alpha\lambda} = \frac{\lambda}{2\left(\dfrac{v}{V} - \cos\alpha\right)}.$$

When $l = l_{\alpha\lambda}$, radiation, in the direction **n** from any point M of portion AB of the path of the charged particle, is cancelled upon interference with radiation in the same direction from a like point N on the adjacent portion $\overline{BC} = \overline{AB} = l_{\alpha\lambda}$, the distance between points M and N being $\overline{MN} = l_{\alpha\lambda}$. Therefore,

upon motion of a charged particle at uniform velocity in a straight line in a substance at a "slower-than-light" velocity, the particle does not emit radiation.

34.6.3 If a particle travels in a substance at a "faster-than-light" velocity $V > v = c/n$, the value of $l_{\alpha\lambda}$, complying with the condition for the cancelling of elementary waves,

$$l_{\alpha\lambda} = \frac{\lambda}{2 \left| \dfrac{v}{V} - \cos\alpha \right|}$$

can be found for all values of angle α, with the exception of

$$\theta = \text{arc cos } \frac{v}{V} = \text{arc cos } \frac{c}{nV}.$$

For the direction $\theta = \alpha$, the path difference of the elementary waves emitted by any two points A and B of the path of the charged particle (Fig. 34.3) equals zero. Thus

$$\Delta = \overline{DF} = (v - V \cos\theta)\,\Delta t = 0.$$

Consequently, elementary waves, propagating in the direction $\alpha = \theta$, reinforce one another upon interference, forming a resul-
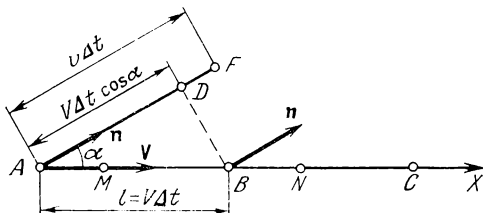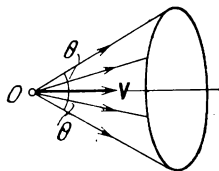
Fig. 34.3   Fig. 34.4

tant radiation in this direction. This is Vavilov-Cherenkov radiation. Light, emitted at each small portion of the path of a charged particle, is propagated along the elements of a cone, whose vertex O (Fig. 34.4) is located on this portion and whose

axis coincides with the path of the particle. The elements make the angle $\theta = \text{arc cos } (c/nV)$ with the axis. The light is polarized so that vector $\mathbf{E}$ is normal to the surface of the cone, and vector $\mathbf{H}$ is tangent to it.

CHAPTER 35   POLARIZATION OF LIGHT

35.1 Polarization of Light in Reflection and Refraction at the Interface Between Two Dielectric Media

35.1.1 Light emitted by ordinary (nonlaser) sources is made up of a set of a great many plane polarized wave trains, or packets (32.1.3), whose electric vectors $\mathbf{E}$ vary along all possible directions perpendicular to the ray (30.2.1). Light is said to be *natural*, or *unpolarized*, if not one of the above-indicated directions of vibration is predominant. In natural light, the resultant strength $\mathbf{E}$ oscillates at each point of the field in a direction that rapidly and randomly varies in a plane perpendicular to the ray.

Light is said to be *partially polarized* if in it there is a predominant direction of oscillation of vector $\mathbf{E}$. Partially polarized light can be regarded as a collection ("mixture") of natural and linearly polarized light, simultaneously propagating in a single direction.

35.1.2 The *polarization of light* is the separation of linearly polarized light out of natural or partially polarized light. Special devices, called *polarizers*, are used for this purpose. Their principle of operation is based on the polarization of light upon being reflected or refracted by the interface between two dielectric media, as well as on the phenomenon of birefringence (35.2.1) and dichroism (35.2.10). The same devices can be utilized as *analyzers*, i.e. devices to determine the kind and degree of polarization of light.

Assume that linearly polarized light, whose electric vector $\mathbf{E}_p$ is along the line p-p and oscillates with the amplitude A_p, falls on the analyzer perpendicularly to the plane of Fig. 35.1. Assume further that the electric vector $\mathbf{E}_a$ of the light passed by the

analyzer is along line a-a, which makes the angle α with line p-p. The incident light can be represented in the form of two waves, linearly polarized in mutually perpendicular planes (31.1.8). The wave whose electric vector E_1 oscillates along the direction perpendicular to a-a, with the amplitude $A_1 = A_p \sin \alpha$, cannot pass through the analyzer. But the other wave, whose electric vector E_2 oscillates along the direction of a-a, with the amplitude $A_2 = A_p \cos \alpha$ passes completely through the analyzer. Consequently, the amplitude of the light emerging from the analyzer is

$$A_a = A_2 = A_p \cos \alpha.$$

The intensities I_a and I_p of the linearly polarized light incident on and emerging from the analyzer, respectively, are related by the *Malus law*:

$$I_a = I_p \cos^2 \alpha.$$
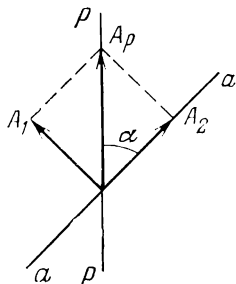
Fig. 35.1

The *principal plane of a polarizer* (or *analyzer*) is the plane of polarization [plane of oscillation, according to previous terminology (31.1.7)] of light passed through the polarizer (or analyzer).

35.1.3 In investigating the laws of light polarization as a result of reflection and refraction of natural light, the light can be conveniently dealt with as a combination of linearly polarized waves, of equal intensities, of two types: s- and p-waves (31.5.4). The reflection coefficient (31.5.6) of the s-waves (R_s) is always greater than that (R_p) of the p-waves. Therefore, in contrast to incident natural light, reflected and transmitted (refracted) light is partially polarized. Oscillations of the electric field vector E of the s-type (perpendicular to the plane of incidence) predominate in reflected light, whereas oscillations of the p-type (in the plane of incidence) predominate in transmitted light.

Brewster's law states that reflected light is completely linearly polarized at the angle of incidence $i = i_{Br}$, complying with the condition $\tan i_{Br} = n_{21}$, where n_{21} is the relative refractive index of a medium reflecting light.

The angle i_{Br} is called *Brewster's angle*. When $i = i_{Br}$, the reflected and refracted rays are perpendicular to each other and

the reflection coefficient of the p-wave is $R_p=0$ (31.5.6). Hence, only s-type waves are reflected. But their reflection coefficient is considerably less than unity (about 0.15 for glass). Thus, the transmitted light is only partially polarized.

35.1.4 Polarization of transmitted light can be enhanced by subjecting the light to a number of consecutive reflections and refractions. This is done in a stack consisting of several identical and parallel plates of some transparent dielectric (for instance, glass), set at the Brewster angle to the incident light beam. If the number of plates in the stack is sufficiently large, then the light passing through the stack is found to be practically linearly polarized (p-type). In the absence of light absorption in the stack, the intensities I_s and I_p of the reflected and transmitted linearly polarized light are the same and equal to one-half of the intensity I_0 of the incident natural light. Thus

$$I_s = I_p = \frac{1}{2} I_0.$$

35.1.5 According to the conceptions of classical electron theory, the formation of the reflected wave is due to the secondary waves emitted by the molecules, the oscillators of the light reflecting medium (34.1.3). Oscillators (vibrating electric dipoles), whose axis is perpendicular to the plane of incidence, correspond to a wave of the s-type. These oscillators are shown in Fig. 35.2 by black dots on the refracted and reflected rays. As is evident from the polar directional diagram (Fig. 31.4), such oscillators should radiate intensively in all directions lying in the plane of incidence, i.e. they should participate in the formation of both the reflected and the refracted s-waves.
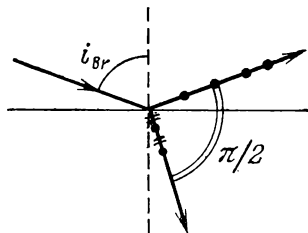
Fig. 35.2

Oscillators, whose axis lies in the plane of incidence and is perpendicular to the refracted ray (shown in Fig. 35.2 by transverse dashes), correspond to a wave of the p-type. Oscillators do not radiate along their axis (Fig. 31.4), and at $i = i_{Br}$ the reflected ray is perpendicular to the refracted one, thereby being

parallel to the axes of these oscillators. Hence, at $i = i_{Br}$ these oscillators do not radiate in the direction of the reflected ray, making no contribution to the reflected wave. Consequently, the reflected light is completely linearly polarized (s-type waves).

35.2 Birefringence (Double Refraction)

35.2.1 Most crystals are optically anisotropic (30.1.6). Their relative dielectric permittivity (dielectric constant) and refractive index depend on the direction of the electric vector E of the light wave. Observed in optically anisotropic crystals is the phenomenon of *double refraction*, or *birefringence*, ;which consists in the splitting of a light ray, falling on a surface of the crystal, into two refracted rays. The birefringence of light in a crystal of Iceland spar ($CaCO_3$) is illustrated in Fig. 35.3.

35.2.2 The *optic axis of a crystal* is the direction in an optically anisotropic crystal along which light propagates without under-
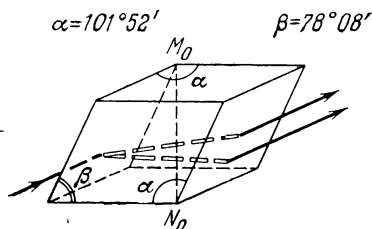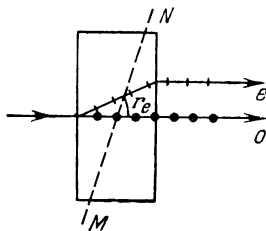
Fig. 35.3   Fig. 35.4

going birefringence. It is necessary to underline that the optic axis of a crystal is no single special straight line similar, for instance, to the axis of symmetry of a body. It only characterizes a certain *direction* in the crystal and can be passed through any point of the crystal.

Depending upon their type of symmetry, optically anisotropic crystals can be either *uniaxial* or *biaxial*, i.e. they can have either one or two optic axes. An example of a uniaxial crystal is

Iceland spar, whose optic axis coincides in direction with the diagonal M_0N_0 of the crystal (Fig. 35.3).

The *principal plane* (or *principal section*) *of a uniaxial crystal* for a certain ray is the plane passing through this ray and intersecting the optic axis.

35.2.3 In a uniaxial crystal, one of the rays formed in birefringence (double refraction) obeys the ordinary laws of light refraction (31.5.3). It lies in the plane of incidence and complies with Snell's law. It is therefore called the *ordinary*, or *O ray*. The second ray is called the *extraordinary* or *E ray* because it generally does not lie in the plane of incidence and does not obey Snell's law. For example, even in the case of normal incidence of light on a plate cut from a uniaxial crystal, the extraordinary ray is refracted (Fig. 35.4). Its refractive angle r_e depends on how the surface of the plate is oriented with respect to the optic axis of the crystal. It may be equal to zero only in two cases: (a) if the surface of the plate is perpendicular to the optic axis (light propagates in the plate along the optic axis and is thereby not subject to birefringence); (b) if the surface of the plate is parallel to the optic axis (light propagates in the plate perpendicular to the optic axis).

In a biaxial crystal both refracted rays behave like extraordinary ones.

35.2.4 Birefringence indicates that a light wave incident on an optically anisotropic crystal produces two waves, propagating through the crystal, generally in different directions. In a uniaxial crystal, they are called the *ordinary* and *extraordinary waves*. The ordinary and extraordinary rays indicate the directions of the Umov-Poynting vectors (31.2.3) of the corresponding waves in the crystal, i.e. the direction of energy transfer by these waves.

The ordinary and extraordinary waves are linearly polarized (31.1.7).* In the ordinary wave, vector **E** is perpendicular to the principal plane of the crystal for the ordinary ray. The electric vector **E** of the extraordinary wave lies in the principal plane of the crystal for the extraordinary ray. The directions of vectors **E** in the ordinary and extraordinary waves are conditionally shown in Fig. 35.4 by black dots on the ordinary

* Frequently, mention is made of linear polarization of the ordinary and extraordinary rays, actually meaning the polarization of the waves corresponding to them.

32*

ray and transverse dashes on the extraordinary ray (it is as-
sumed that both rays, and optic axis MN intersecting them, lie
in the plane of the figure).

35.2.5 The *ray velocity of the wave*, or simply *ray velocity*, in an
optically anisotropic crystal is the velocity **v** of energy transfer
by the wave. In a uniaxial crystal, the *velocity* $\mathbf{v}_o$ *of the ordinary
ray* is numerically equal in all directions: $v_o = c/n_o$, where c is
the velocity of light in free space (or in vacuum), and $n_o = \text{const}$
is the *refractive index of the crystal for the ordinary ray*. According-
ly, the *velocity* $\mathbf{v}_e$ *of the extraordinary ray* is numerically equal
to $v_e = c/n_e$, where n_e is the *refractive index of the crystal for
the extraordinary ray*. The values of n_e and v_e depend on the
direction of the extraordinary ray with respect to the optic axis
of the crystal. For a ray propagating along the optic axis,
$n_e = n_o$ and $v_e = v_o$. The value of n_e differs to the greatest
extent from that of n_o for a direction perpendicular to the optic
axis: $n_e = n_{e0}$.

35.2.6 The *ray surface of a wave in a crystal* is the locus of the
heads of vectors **v** of the ray velocity of the wave that extend
from a certain point O in the crystal in all possible directions.
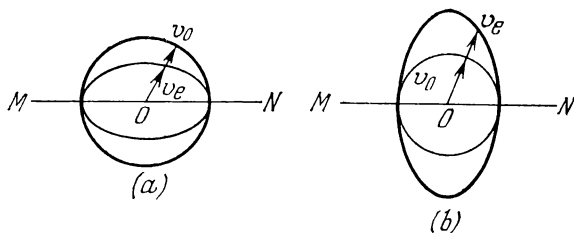
Fig. 35.5

In a uniaxial crystal, the ray surface of the ordinary wave has
the shape of a sphere, whereas that of the extraordinary wave is
an ellipsoid of revolution about the optic axis MN passing
through point O. The ellipsoid and sphere contact each other
at the points of intersection with optic axis MN. When $n_e \geqslant n_o$,
the ellipsoid is inscribed in the sphere (Fig. 35.5*a*), whereas if
$n_e \leqslant n_o$ the ellipsoid is circumscribed about the sphere

(Fig. 35.5*b*). In the former case, the uniaxial crystal is said to be *optically positive*; in the latter, *optically negative*.

35.2.7 Birefringence in a uniaxial crystal can be explained and the directions of the ordinary and extraordinary rays can be found by using the Huygens graphical method. Assume that a plane unpolarized light wave is incident on the plane surface *ab* of a uniaxial optically negative crystal (or a plate cut out of this crystal) at the angle *i* (Fig. 35.6). The optic axis *MN* of the crystal, passing through point *A* of surface *ab*, lies in the plane of the figure and makes the angle γ with plane *ab*.
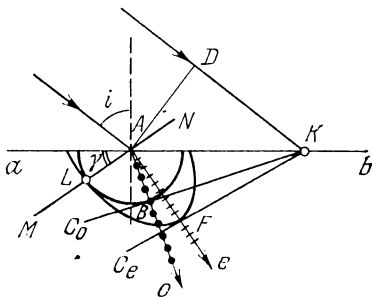
Fig. 35.6

At the instant *t* of time being considered, the front *AD* of the incident wave has reached point *A* on the surface of the crystal, and this point becomes the source of two linearly polarized elementary secondary waves, ordinary and extraordinary, in the crystal. At the instant of time $t + \Delta t$, where Δt is the time required for the incident light to travel the distance $\overline{DK}$, the disturbance, propagating from point *A* in the form of the ordinary elementary wave, reaches the points of a sphere of radius $v_o \Delta t$ having its centre at point *A*. The disturbance, propagating from point *A* in the form of the extraordinary elementary wave, reaches points at the same time on the surface of the ellipsoid of revolution that contacts the sphere of radius $v_0 \Delta t$ at point *L* of its intersection with optic axis *MN*. This ellipsoid is geometrically similar to the ray surface of the extraordinary wave in the crystal (35.2.5).

Planes KC_o and KC_e, perpendicular to the plane of the figure
and tangent, respectively, to the sphere and the ellipsoid,
indicate, according to the Huygens principle (33.1.1), the
positions, at the instant of time $t + \Delta t$, of the fronts of the
ordinary and extraordinary waves that actually propagate in
a uniaxial crystal. Straight lines from point A to the points
of tangency B and F indicate the directions of the ordinary and
extraordinary rays. Both rays lie in the plane of incidence, but
the extraordinary ray is not orthogonal to the wave surface KC_e.
The ordinary and extraordinary waves are linearly polarized
in mutually perpendicular planes. The directions of the electric
vectors $\mathbf{E}_o$ and $\mathbf{E}_e$ in the ordinary and extraordinary waves are
shown in Fig. 35.6 by black dots and transverse dashes located
on the corresponding rays.

Note. If the optic axis MN of the crystal does not lie in the plane
of incidence of the light, then, in general, the extraordinary
ray also does not lie in the plane of incidence. Consequently,
the angle between the planes of polarization of the ordinary
and extraordinary waves differs slightly from a right angle.

35.2.8. The construction for the ordinary and extraordinary
rays in the case of normal incidence of light on the surface of
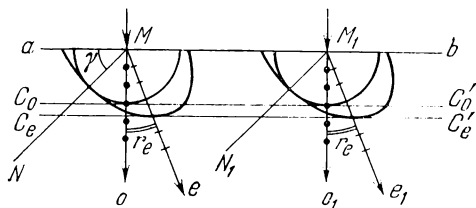
Fig. 35.7

an optically negative uniaxial crystal is shown in Fig. 35.7.
Here ab is the position of the incident wavefront at the instant t
of time, and C_oC_o' and C_eC_e' are the positions of the fronts of the
ordinary and extraordinary waves in the crystal at the instant
of time $t + \Delta t$.

It is assumed that the optic axis MN lies in the plane of inci-
dence and makes the angle γ, differing from 0 and $\pi/2$, with

the refracting surface ab. It is evident from Fig. 35.7 that the ordinary ray is an extension of the incident one and that the extraordinary ray is refracted by the angle $r_e \neq 0$.

Considered in Fig. 35.8 is a case in which light is normally incident on the plane surface ab of an optically negative uni-

axial crystal, whose optic axis MN is parallel to ab. Here, as is evident from the construction, the extraordinary ray is not refracted at the surface ab and coincides in direction with the ordinary and incident rays. But the velocities of the ordinary and extraordinary rays in the crystal in this direction differ and are respectively equal to (35.2.5): $v_0 = c/n_0$ and $v_e =$ $= c/n_{e0}$. Therefore, the two rays (waves) travel the same distance d in the crystal with an optical-path difference (32.3.2):

Fig. 35.8

$$\Delta s = d \, (n_o - n_{e0}).$$

35.2.9. The path of the rays in a *polarizing prism* are shown in Fig. 35.9. This prism is cut out of a crystal of Iceland spar
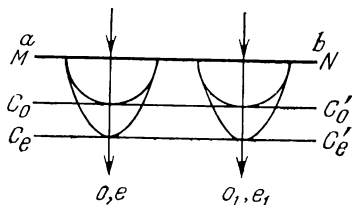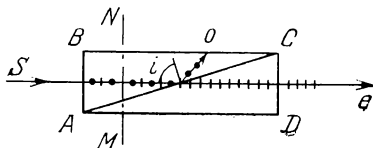
Fig. 35.9

so that its sides AB and CD are parallel to its optic axis MN. Then the prism is cut apart along its diagonal plane AC and glued together along these diagonal surfaces by a thin layer of an optically isotropic transparent substance called Canada balsam. The crystal of Iceland spar is uniaxial and optically nega-

tive; its refractive indices (35.2.5) are: $n_o = 1.658$ and $n_{e0} = 1.486$. The refractive index of Canada balsam is $n_{Cb} = 1.550$, i.e. Canada balsam is a medium less optically dense than the material of the prism for the ordinary ray and more optically dense than this material for the extraordinary ray. Light falls on the prism normal to its side AB (ray S in Fig. 35.9). The ordinary and extraordinary rays propagate in the prism without being refracted until they reach the layer AC of Canada balsam. The dimensions of the prism are selected so that the angle of incidence i of the ordinary ray on surface AC is greater than the critical angle for total internal reflection (31.5.8). Therefore, the ordinary wave is completely reflected by the layer of Canada balsam (ray o in Fig. 35.9). The extraordinary wave passes freely through the layer of Canada balsam and the second half of the polarizing prism. Thus, the polarizing prism can be used as a polarizer (35.1.2).

35.2.10. All birefringent crystals absorb light to one or another degree. This absorption is anisotropic: the absorption factor (34.2.1) depends upon the orientation of the electric vector of the light wave and on the direction of light propagation in the crystal, as well as on the wavelength. This phenomenon is called *dichroism*, or *pleochroism*, because it is manifested in the different colouring of the crystals in different directions. An example of a strongly dichroic crystal is tourmaline, a uniaxial crystal in which the ordinary ray is absorbed much more intensely than the extraordinary ray. Even more pronounced dichroism is displayed by crystals of herapathite, which is used to make thin films, converting natural light into linearly polarized light and called *polaroids*.

35.3 Interference of Polarized Light

35.3.1. Trains of waves with all possible orientations with respect to the ray of their planes of polarization, included in the composition of natural light, are incoherent because they correspond to the radiation of various *independent* atoms of the light source. All these trains participate in the formation of the ordinary and extraordinary waves that propagate in a uniaxial crystal when natural light falls on it. But the contribution of each separate train to these two waves is, in general, not the same. It is greater to the wave whose plane of polarization makes

the smaller angle α with the plane of polarization of the wave train. In other words, the ordinary and extraordinary waves are generated mainly by *different* wave trains included in the composition of natural light. Consequently, the ordinary and extraordinary waves, propagated in a uniaxial crystal on which natural light falls, are not coherent.

35.3.2. The ordinary and extraordinary waves, propagated in a uniaxial crystal upon the incidence of linearly polarized light [obtained from natural light by means, for example, of a polarizing prism (35.2.9) or some other polarizer] are coherent with each other. This is due to the fact that the planes of polarization of all the wave trains included in the composition of the incident light have the same orientation.

Assume that a parallel beam of light, after passing through polarizer P (Fig. 35.10), is normally incident on the plane surface ab
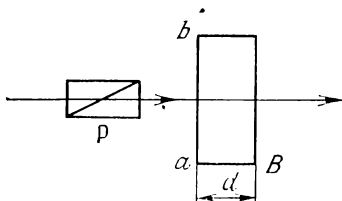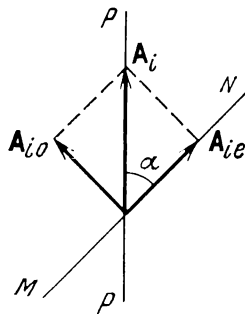
Fig. 35.10   Fig. 35.11

of plane-parallel plate B, cut out of a uniaxial crystal parallel to its optic axis MN (axis MN is parallel to plane ab). Shown in Fig. 35.11 is vector A_i of the amplitude of the ith wave train. This vector is oriented along line p-p, corresponding to the direction of oscillation of the electric vector in the light emerging from the polarizer. The contributions of the ith wave train to the ordinary and extraordinary waves are characterized by the amplitudes $A_{io} = A_i \sin \alpha$ and $A_{ie} = A_i \cos \alpha$, whose ratio

$(A_{io}/A_{ie}) = \tan \alpha$ is the same for all trains. In particular, if $\alpha = \pi/4$, then $A_{io} = A_{ie}$ because pairwise coherent trains, polarized in mutually perpendicular planes, have the same intensity.

35.3.3. At the entrance to crystal plate B (35.3.2), the electric vectors $\mathbf{E}_o$ and $\mathbf{E}_e$ of the ordinary and extraordinary waves oscillate in phase and their vector sum is equal to electric vector of linearly polarized monochromatic incident light: $\mathbf{E}_p = = \mathbf{E}_o + \mathbf{E}_e$. In the plate the ordinary and extraordinary waves propagate at different velocities (35.2.8). Therefore, at the exit from the plate (of thickness d), the mutually perpendicular electric vectors $\mathbf{E}_o'$ and $\mathbf{E}_e'$ of the ordinary and extraordinary waves oscillate with a phase difference of

$$\Delta \varphi = \frac{2\pi \Delta s}{\lambda_0} = \frac{2\pi d}{\lambda_0} (n_o - n_{e0}),$$

where Δs is the optical-path difference of these waves (35.2.8), and λ_0 is the wavelength of light in vacuum. Consequently, as a result of its passing through the plate, the light, in the general case, becomes elliptically polarized (31.1.7): the head of vector $\mathbf{E}' = \mathbf{E}_o' + \mathbf{E}_e'$ describes an ellipse lying in a plane perpendicular to the beam. If α is the angle between the direction of oscillation of vector $\mathbf{E}_p$ and the optic axis MN of the plate, then amplitudes A_o and A_e of vectors $\mathbf{E}_o$ and $\mathbf{E}_e$ equal: $A_o = A_p \sin \alpha$ and $A_e = A_p \cos \alpha$, where A_p is the amplitude of vector $\mathbf{E}_p$. In the absence of light absorption in the plate, the amplitudes of vectors $\mathbf{E}_o'$ and $\mathbf{E}_e'$ are also equal to A_o and A_e.

35.3.4. In accordance with the thickness d of the plate, several special cases are possible.

(a) A *quarter-wave plate* is one whose thickness satisfies the relation: $d(n_o - n_{e0}) = \pm (m + 1/4) \lambda_0$, where $m = 0, 1, 2, ...;$ the plus sign corresponding to an optically negative crystal, and the minus sign to an optically positive one (35.2.6). At the exit from such a plate the oscillations of vectors $\mathbf{E}_o'$ and $\mathbf{E}_e'$ have a phase difference of $\pi/2$. If, in addition, $\alpha = \pi/4$, the light emerging from the plate is circularly polarized (31.1.7).

(b) For a *half-wave plate*: $d(n_o - n_{e0}) = \pm (m + 1/2) \lambda_0$. At the exit from such a plate, the oscillations of vectors $\mathbf{E}_o$ and $\mathbf{E}_e'$ have a phase difference of π. Light emerging from the plate remains linearly polarized. But the directions of oscilla-

tion of vectors $\mathbf{E}_p$ and $\mathbf{E}'$ of the incident and transmitted light are symmetrical with respect to the principal plane of the plate (Fig. 35.12).

(c) For a *full-wave plate*: $d\,(n_0 - n_{e0}) = \pm m\lambda_0$. As a result of passing through the plate, the light remains linearly polarized in the same plane as the incident light.

35.3.5. Coherent waves emerging from crystal plate B (Fig. 35.10) cannot display interference because they are polarized in mutually perpendicular planes. Therefore, one more polarizing prism, the analyzer A (Fig. 35.13) is mounted beyond plate B. The analyzer separates out of the incident coherent waves, the components polarized in one plane. It thus provides the conditions necessary for obtaining the interference of these waves.

Fig. 35.12

The results of this interference depend upon the phase difference $\Delta\varphi$ acquired by the ordinary and extraordinary waves in the plate, on the ratio of the amplitudes of these waves and on the angle β between the principal planes of the analyzer and polarizer (35.1.2).

For example, if the angle between the principal plane of the polarizer and optic axis MN of the plate $\alpha = \pi/4$, the amplitudes and intensities of the ordinary and extraordinary waves are the same. Assume in this case that monochromatic light of the wavelength λ_0 in vacuum is incident on the plate. The two following limiting cases may occur:

$$\Delta\varphi = \frac{2\pi d}{\lambda_0}\,(n_0 - n_{e0}) = \begin{cases} \pm 2m\pi \\ \pm (2m+1)\,\pi \end{cases}$$

(where $m = 0,\ 1,\ 2,\ \ldots$).

In the first case, corresponding to a full-wave plate, the light falling on the analyzer is linearly polarized in the principal plane of the polarizer. Therefore, at $\beta = 0$ (the analyzer is mounted parallel to the polarizer), the intensity I_a of the light passing through the analyzer is maximal, whereas at $\beta = \pi/2$
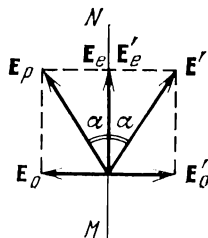
(the analyzer and polarizer are crossed), $I_a = 0$, i.e. at $\beta = 0$ an interference maximum is observed, whereas at $\beta = \pi/2$ a minimum is observed.

In the second case, corresponding to a half-wave plate, the light falling on the analyzer is linearly polarized in a plane making the angle $2\alpha = \pi/2$ with the principal plane of the polarizer. Therefore, at $\beta = 0$ an interference minimum is observed, whereas at $\beta = \pi/2$ a maximum is observed.

If linearly polarized *white light* is incident on plate B (Fig. 35.13), then in observing the light through the analyzer, the plate is
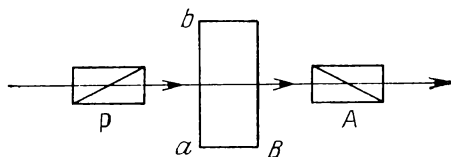
Fig. 35.13

seen to be coloured. Upon rotating the analyzer about the beam, i.e. in varying angle β, the colouring changes. This is due to the fact that the value of the phase difference $\Delta\varphi$, determining the result of interference, depends upon the wavelength of the light. A plate, whose thickness d is not the same at various places, is seen in white light to be oddly coloured, with each coloured interference line (*isochromate*) passing through points of equal thickness d. A similar pattern is observed in a plate, which is of equal thickness at all points, but the value of the difference $(n_0 - n_{e0})$ varies. In this case, each isochromate passes through points of the plate corresponding to the same value of $(n_0 - n_{e0})$.

35.4 Artificial Optical Anisotropy

35.4.1 An optically isotropic transparent body becomes anisotropic if it is subjected to mechanical deformation. This phenomenon is sometimes called *photoelasticity*. Upon linear (uniaxial) tension or compression of an isotropic body along

the OX axis, it acquires the optical properties of a uniaxial crystal (35.2.2), whose optic axis is parallel to OX. The difference between the refractive indices for the ordinary (n_o) and extraordinary (n_{e0}) rays, in the direction perpendicular to axis OX, is proportional to the normal stress σ (40.3.3). Thus

$$n_o - n_{e0} = k\sigma,$$

where k is the proportionality factor depending upon the properties of the body.

35.4.2 The *electrooptical Kerr effect* is the appearance of optical anisotropy in a transparent isotropic solid, liquid or gaseous dielectric upon being placed in an external electric field. By the action of the uniform electric field, the dielectric is polarized and acquires the optical properties of a uniaxial crystal whose optic axis coincides in direction with the field strength vector **E**. The difference between the refractive indices of a polarized dielectric for extraordinary and ordinary rays of monochromatic light, propagating perpendicular to the direction of vector **E**, satisfies the *Kerr law*:

$$n_{e0} - n_o = B\lambda_0 E^2,$$

where λ_0 is the wavelength of the light in vacuum, and B is the *Kerr constant*. The value of B depends upon the nature of the substance, wavelength λ_0 and the temperature, decreasing, as a rule, as the temperature is raised. The sign of the difference $(n_{e0} - n_o)$ is independent of the direction of the field. For the majority of substances $B > 0$, so that they are similar with respect to their optical properties in a uniform electric field to optically positive uniaxial crystals (35.2.6).

35.4.3 The *Cotton-Mouton effect* is the appearance of optical anisotropy in certain isotropic substances (liquids, glasses and colloids) when they are placed in a strong external magnetic field. In a uniform magnetic field, the substance acquires the optical properties of a uniaxial crystal, whose optic axis coincides in direction with the magnetic field strength vector **H**. The difference between the refractive indices of the substance for extraordinary and ordinary rays of monochromatic light, propagating in the direction perpendicular to vector **H**, is proportional to H^2. Thus

$$n_{e0} - n_o = C\lambda_0 H^2,$$

where C is the *Cotton-Mouton constant*, and λ_0 is the wavelength of light in vacuum. The value of constant C depends upon the nature of the substance, wavelength λ_0 and the temperature.

35.5 Rotation of the Plane of Polarization

35.5.1 As linearly polarized light passes through certain substances that are said to be *optically active*, the plane of polarization of the light (31.1.7) rotates about the direction of the ray. Optical activity is displayed by certain crystals (quartz, cinnabar, etc.), pure liquids and solutions (turpentine, a solution of sugar in water, etc.). All substances that are optically active in the liquid state, possess the same property in the crystalline state as well. But some substances that are optically active in the crystalline state, lose this activity in the liquid state. Consequently, optical activity can be due either to the structure of the molecules of the substance themselves or to the arrangement of the particles in the crystal lattice.

35.5.2 In optically active crystals and pure liquids, the angle φ of rotation of the plane of polarization of light is proportional to the thickness l of the layer of the substance through which light passes. Thus $\varphi = \alpha l$. The proportionality factor α is called the *rotatory power*, or *specific rotation*. The rotatory power depends upon the nature of the substance, the temperature and the wavelength λ_0 of the light in vacuum. The dependence of α on λ_0 is called the *rotary*, or *rotatory*, *dispersion*. Far from the band of light absorption by the substance, rotary dispersion obeys *Biot's law*: $\alpha \propto \lambda_0^{-2}$.

35.5.3 The majority of optically active crystals exist in two versions. As light passes through a crystal of one version, said to be *dextrorotatory* or *right-handed*, the plane of polarization is rotated to the right, i.e. clockwise (for an observer looking toward the oncoming beam). In the passing of light through a crystal of the other version, said to be *levorotatory* or *left-handed*, the plane of polarization is rotated to the left (counterclockwise). The values of the rotatory power for the two versions of the same optically active crystal differ only in sign.

35.5.4 The angle of rotation of the plane of polarization of light in travelling a path of length l in an optically active solution is

$$\varphi = [\alpha] \, cl = [\alpha] \, D \, Kl.$$

Here c is the volume-mass concentration of the optically active substance in the solution (in kg/m³), D is the density of the solution, and $K = c/D$ is the weight-part concentration, i.e. the ratio of the mass of the optically active substance to the mass of the whole solution. The proportionality factor $[\alpha]$ is called the *rotatory power*, or *specific rotation, of the solution*. The value of $[\alpha]$ depends on the nature of the optically active substance and of the solvent, the wavelength of the light and the temperature.

35.5.5 When subjected to the action of an external magnetic field, an optically inactive medium acquires the capacity to rotate the plane of polarization of light propagating along the direction of the field. This phenomenon is known as the *Faraday effect*, or *magnetic rotation of the plane of polarization of light*. The angle φ of rotation of the plane of polarization is proportional to the path length of the light in the substance and to the strength H of the magnetic field. Thus $\varphi = VHl$. The proportionality factor V is called the *Verdet constant*. It depends upon the nature of the substance and the wavelength of the light.

The direction of magnetic rotation of the plane of polarization (for an observer looking along the magnetic field) is the same whether the light propagates along the direction of vector **H** or in the opposite direction. In this feature the Faraday effect differs from the rotation of the plane of polarization of light in natural optically active media.

.

CHAPTER 36   THERMAL RADIATION

36.1 Thermal Radiation. Kirchhoff's Law

36.1.1 All bodies emit electromagnetic waves to one or another degree. Highly heated bodies, for instance, glow, whereas at ordinary temperatures they are the source of only invisible infrared radiation.

Electromagnetic radiation, emitted by a substance and produced at the expense of its internal energy (9.1.2) is called *thermal*, or *heat, radiation*. It depends only on the temperature and the optical properties of the radiating bodies. If the energy consump-

tion of a body on thermal radiation is not replenished by sup-
plying more heat to the body, its temperature gradually drops
and the thermal radiation is reduced.

Radiative heat exchange (heat exchange through radiation) is the
spontaneous process of transferring energy in the form of heat
from a more heated body to a less heated one. It is accomplished
by thermal radiation and the absorption of electromagnetic
waves by these bodies.

36.1.2 Thermal radiation is the only kind that can be in a state
of thermodynamic equilibrium (8.3.3) with matter. In equilib-
rium, the energy consumption on thermal radiation is compensat-
ed for by the absorption of the
same amount of energy by the
body owing to the radiation fal-
ling on it. *Radiative (radiation)
equilibrium* is set up in an adia-
batically closed system (i.e. one
that does not exchange heat with
the external medium), all bodies
of which are at the same temper-
ature.
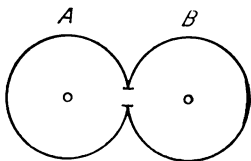
Fig. 36.1

It follows from the second law
of thermodynamics (11.3.2) that
radiative equilibrium is independent of the material of
the bodies that make up the closed thermodynamically equilib-
rium system. The volume density of the energy of radiation
equilibrium and its frequency distribution are universal func-
tions of the temperature. As a matter of fact, if this were no-
so, it would be possible to take two adiabatically closed systemst
A and B (Fig. 36.1), at the same temperature $T_A = T_B = T$,
and accomplish radiative heat exchange between them by mak-
ing a small hole in the heatproof wall separating them. If the
volume density of energy E_A^d of system A is greater than that
E_B^d of system B, i.e. $E_A^d > E_B^d$, then, owing to heat exchange
between the systems, the radiant energy in system A and its
temperature should decrease and the radiant energy in system B
and its temperature should increase until the values of the vol-
ume energy densities in systems A and B become equal: $E_A^{d'} =$
$= E_B^{d'}$. But, in this case, $T_B' > T_A'$, so that the process being
considered contradicts the second statement of the second law

of thermodynamics (11.3.2). Consequently, E_A^d cannot be great-er than E_B^d. In exactly the same way, E_B^d cannot be greater than E_A^d, i.e. $E_A^d = E_B^d = E^d$ (T) is a universal function of the temperature.

36.1.3 Serving as the spectral characteristic of radiative equi-librium is the *spectral density of the volume energy denstiy of this radiation*. Thus

$$\rho\,(\nu,\ T) = \frac{dE^d}{d\nu}\ ,$$

where dE^d is the energy of radiative equilibrium with frequen-cies from ν to $\nu + d\nu$ enclosed within unit volume of the radia-tion field. The volume energy density of this field is

$$E^d = \int_0^\infty \rho\,(\nu,\ T)\,d\nu.$$

Radiation equilibrium is *isotropic*, i.e. it is not polarized and all directions of its propagation are equally probable. The ener-gy dE of radiation equilibrium in vacuum at frequencies from ν to $\nu + d\nu$ incident in unit time on unit area of the surface of each of the bodies of a thermodynamically equilibrium system is equal to

$$dE = \frac{c}{4}\,\rho\,(\nu,\ T)\,d\nu,$$

where c is the velocity of light in free space (or in a vacuum).

36.1.4 The *radiant exitance** (*total emissive power*) of a body is the physical quantity M_e, numerically equal to the energy of electromagnetic waves of all possible frequencies (or wavelengths) from 0 to ∞** radiated in unit time from unit area of a body's surface.

---

* Formerly called the radiant emittance.
** It is practically sufficient to be restricted to the range of frequencies and wavelengths of optical radiation (31.4.3).

The *emissive power* (*spectral density of radiant exitance*) of a body is the physical quantity numerically equal to the ratio of the energy dE radiated in unit time from unit area of the surface of the body by electromagnetic waves in a narrow range of frequencies from ν to $\nu + d\nu$ (or wavelengths in vacuum from λ to $\lambda + d\lambda$), to the width of this range:

$$r_\nu = \frac{dE}{d\nu} \quad \text{or} \quad r_\lambda = \frac{dE}{d\lambda} ; \quad r_\lambda = \frac{c}{\lambda^2} r_\nu ,$$

where c is the velocity of light in vacuum. The values of r_ν (or r_λ) depend upon the frequency (or wavelength), temperature, chemical composition of the body and the condition of its surface.

The radiant exitance of a body is related to r_ν and r_λ by the equations

$$M_e = \int\limits_0^\infty r_\nu \, d\nu = \int\limits_0^\infty r_\lambda \, d\lambda .$$

36.1.5 The *absorptance* (*monochromatic absorption coefficient*) *of a body* is the dimensionless quantity α_ν which indicates what fraction of the energy of electromagnetic waves of frequencies from ν to $\nu + d\nu$, incident on the surface of the body, is absorbed by it. Thus

$$\alpha_\nu = \frac{dE_{\text{abs}}}{dE_{\text{inc}}} \leqslant 1 .$$

The value of α_ν depends upon the frequency, temperature, chemical composition of the body and the condition of its surface.

A *blackbody* (denoted by a superscript asterisk) is one that completely absorbs all incident radiation, regardless of its direction, spectral composition and polarization, without reflecting or transmitting any: $\alpha_\nu^* \equiv 1$. Though a blackbody is a theoretical ideal, it can be nearly realized in practice by an almost complete enclosure with a small opening (Fig. 36.2). Light entering the enclosure through hole O undergoes repeated reflection from the walls. At this the energy of the incident light is practically completely absorbed by the walls of the enclosure regardless

of the material of which they are made. The emissive power of a blackbody is denoted in the following by r_ν^* (or r_λ^*) and its radiant exitance by M_c^*.

A *greybody* is one whose absorptance is less than unity and is independent of the frequency (or wavelength) of the light, the directions of its propagation and polarization:

$$\frac{\partial \alpha_\nu^{gr}}{\partial \nu} \equiv 0 \quad \text{and} \quad \alpha_\nu^{gr} = \alpha^{gr}.$$

36.1.6 According to the *principle of detailed balance* any microscopic process in an equilibrium system should proceed at the same rate as the reverse process. This principle of statistical physics enables a relation to be found between the emissive power r_ν and the absorptance α_ν of any opaque body. Assume that the body is part of a thermodynamically equilibrium system at the temperature T. The energy radiated per unit time from unit area of the surface of the body being considered in the frequency range from ν to $\nu + d\nu$ is $dE_{rad} = r_\nu d\nu$. During the same time, the same portion of the surface of the body absorbs a part of the energy of the radiation equilibrium (36.1.3) incident on this surface. It is equal to $dE_{abs} = \alpha_\nu (c/4) \rho (\nu, T) d\nu$. Since, according to the principle of detailed balance, $dE_{rad} = dE_{abs}$,
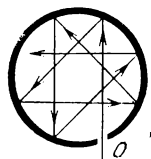
Fig. 36.2

$$\frac{r_\nu}{\alpha_\nu} = r_\nu^* = \frac{c}{4} \rho (\nu, T).$$

This equation expresses *Kirchhoff's law* (for radiation), which states that the ratio of the emissive power of a body to its absorptance does not depend upon the nature of the body and is equal to the emissive power r_ν^* of a blackbody at the same temperature and frequency values.

The dependence of r_ν^* on ν and T is called *Kirchhoff's function*:

$$r_\nu^* = f (\nu, T) = \frac{c}{4} \rho (\nu, T).$$

33*

36.1.7 It follows from Kirchhoff's law that the radiant exitance (36.1.4) of a body is equal to

$$M_e = \int_0^\infty \alpha_\nu r_\nu^* \, d\nu.$$

In particular, the radiant exitance of a greybody is

$$M_e^{gr} = \alpha^{gr} M_e^*, \quad \text{where} \quad M_e^* = \int_0^\infty r_\nu^* \, d\nu,$$

is the radiant exitance of a blackbody at the same temperature. For a nongrey body

$$M_e = \beta M_e^*,$$

where β is the *total degree of blackness of the body*. It depends on the material of the body, the condition of its surface and the temperature. For all bodies except a blackbody, $\beta < 1$.

36.1.8 Radiation equilibrium at the temperature T is identical to thermal radiation of a blackbody at the same temperature. Therefore, it is often called *blackbody radiation*. The relation between the radiant exitance of a blackbody and the volume density of the energy of blackbody radiation is of the form

$$M_e^* = \frac{c}{4} E^d = \frac{c}{4} \int_0^\infty \rho\,(\nu, \ T)\,d\nu.$$

36.2 Stefan-Boltzmann and Wien Laws

36.2.1 The *Stefan-Boltzmann law* states that the radiant exitance of a blackbody is proportional to the fourth power of its absolute temperature. Thus

$$M_e^* = \sigma T^4,$$

where $\sigma = 5.67 \times 10^{-8}$ W-m^{-2}-K^{-4} is the *Stefan-Boltzmann constant*. This law can be derived theoretically by considering

the radiative equilibrium in an enclosed cavity by thermodynamic methods.

36.2.2 The dependence of the emissive power r_ν^* of a blackbody on the frequency ν for several constant values of the temperature is shown in Fig. 36.3. In the region of low frequencies, $r_\nu^* \propto$ $\propto \nu^2 T$, whereas for the region of high frequencies (the right-hand branch of the curves far away from the maxima) $r_\nu^* \propto$ $\propto \nu^3 \exp\left[-a_1 \nu/T\right]$, where a_1 is a constant factor.

The radiant energy of a blackbody is nonuniformly distributed in its spectrum. A blackbody displays almost no radiation in
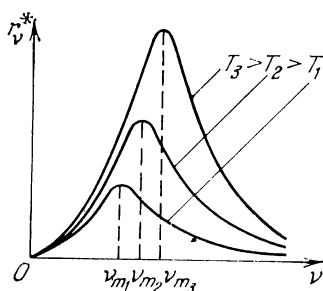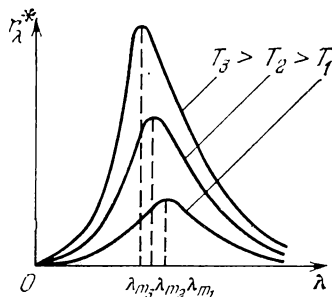
Fig. 36.3   .   Fig. 36.4

the regions of very high and very low frequencies. As the temperature is raised, the maximum of r_ν^* is shifted toward the higher frequencies in accordance with the law: $\nu_m = b_1 T$, where ν_m is the frequency corresponding to maximum r_ν^* at the temperature T, and b_1 is a constant factor.

The dependence of the emissive power of a blackbody $r_\lambda^* =$ $= c r_\nu^*/\lambda^2$ (36.1.4) on the wavelength λ is shown in Fig. 36.4. As the temperature of the blackbody is raised, the maximum of r_λ^* is shifted toward the shorter wavelengths in accordance with the *Wien displacement law*:

$$\lambda_m = \frac{b}{T} ,$$

where $b = 2.9 \times 10^{-3}$ m-K is the *Wien constant*.

36.2.3 All attempts that were made failed to theoretically substantiate the experimentally found form of Kirchhoff's function, $r_\nu^* = f(\nu, T)$, shown in Fig. 36.3, within the framework of classical physics. Using thermodynamic methods, the *Wien formula* was found:

$$r_\nu^* = \nu^3 \varphi \left(\frac{\nu}{T} \right) ,$$

where $\varphi(\nu/T)$ is an unknown function of ν/T.
Rayleigh and Jeans applied the laws of electrodynamics and equipartition of energy in an equilibrium system (10.6.4) to obtain the *Rayleigh-Jeans law*:

$$r_\nu^* = \frac{2\pi\nu^2}{c^2} kT ,$$

where k is Boltzmann's constant (8.4.5).
The Rayleigh-Jeans law agreed with experimental data only in the region of low frequencies. Moreover, it leads to the absurd conclusion that at any temperature the radiant exitance M_e^* of a blackbody and the volume density of energy E^d of radiation equilibrium are infinitely large. This conclusion, arrived at by classical physics in the problem of explaining the spectral distribution of radiation equilibrium, has been aptly named the "ultraviolet catastrophe".

36.3 Planck's Formula

36.3.1 The volume density of energy of radiation equilibrium (blackbody radiation) in a closed cavity, as well as the energy distribution of this radiation in frequency are independent of the material of the cavity walls and are fully determined by the temperature. Therefore, an infinite system of harmonic oscillators with all possible natural frequencies can be taken as a theoretical model of a blackbody. Each of such oscillators corresponds to a monochromatic component of blackbody radiation. Let $\langle E_\nu \rangle$ be the average energy of an oscillator with the natural

frequency v. Then, according to calculations, the radiant exi-
tance of a blackbody is

$$r_v^* = \frac{2\pi v^2}{c^2} \langle E_v \rangle.$$

If kT is taken for $\langle E_v \rangle$, as follows from the classical law of the
equipartition of energy (10.6.4)*, then the preceding expression
for r_v^* coincides with the Rayleigh-Jeans law (36.2.3).

36.3.2 Planck was able to find the correct expression for the
average energy $\langle E_v \rangle$ of an oscillator and for the Kirchhoff
function by introducing the *quantum hypothesis*, which was en-
tirely alien to classical physics. It is assumed in classical phys-
ics that the energy of any system can vary in a *continuous man-
ner*, i.e. that it can take any arbitrarily close consecutive val-
ues. According to Planck's quantum hypothesis, the energy
E_v of an oscillator can have only definite *discrete (quantized)
values*, equal to a whole number of elementary portions, called
energy quanta E_{v0}. Thus

$$E_v = nE_{v0}, \quad \text{where} \quad n = 0, 1, 2, \ldots.$$

If it is assumed that the oscillator distribution in possible dis-
crete energy states is described by the Boltzmann distribution
law (10.4.2), the probability p_n of finding the oscillator in a
state with the energy nE_{v_0} at the temperature T is equal to

$$p_n = C \exp\left[-nE_{v_0}/kT\right].$$

Here k is the Boltzmann constant (8.4.5) and C is a constant
factor determined from the normalization condition:

$$\sum_{n=0}^{\infty} p_n = 1, \quad \text{i.e.} \quad C = \frac{1}{\displaystyle\sum_{n=0}^{\infty} \exp\left[-nE_{v0}/kT\right]}.$$

* One oscillatory degree of freedom of an oscillator accounts for
twice as much energy, on an average, as one degree of transla-
tional or rotary motion, because the oscillator has not only
kinetic, but potential energy as well. On an average, this poten-
tial energy is equal to the kinetic energy.

The average value of the oscillator energy is

$$\langle E_\nu \rangle = \sum_{n=0}^{\infty} p_n n E_{\nu 0} = E_{\nu 0} \frac{\sum_{n=0}^{\infty} n \exp\left[-n E_{\nu 0}/kT\right]}{\sum_{n=0}^{\infty} \exp\left[-n E_{\nu 0}/kT\right]}$$

from which

$$\langle E_\nu \rangle = - E_{\nu 0} \frac{\dfrac{d}{d\xi} \sum_{n=0}^{\infty} \exp\left(-n\xi\right)}{\sum_{n=0}^{\infty} \exp\left(-n\xi\right)}$$

$$= - E_{\nu 0} \frac{d}{d\xi} \ln \sum_{n=0}^{\infty} \exp\left(-n\xi\right),$$

where $\xi = E_{\nu 0}/kT$. Since

$$\sum_{n=0}^{\infty} \exp\left(-n\xi\right) = \frac{1}{1 - \exp\left(-\xi\right)}$$

and

$$-\ln \sum_{n=0}^{\infty} \exp\left(-n\xi\right) = \ln\left(1 - e^{-\xi}\right),$$

then

$$\langle E_\nu \rangle = \frac{E_{\nu 0}}{e^{E_{\nu 0}/kT} - 1} \quad \text{and} \quad r_\nu^* = \frac{2\pi\nu^2}{c^2} \frac{E_{\nu 0}}{e^{E_{\nu 0}/kT} - 1}.$$

Comparing this expression for r_ν^* with the Wien formula (36.2.3), it follows that a *quantum of energy* is equal to

$$E_{\nu_0} = h\nu,$$

where h is a universal constant, called *Planck's constant* (Appendix II).

Planck's radiation formula for the emissive power r_ν^* of a blackbody is

$$r_\nu^* = \frac{2\pi\nu^2}{c^2} \frac{h\nu}{e^{h\nu/kT} - 1} \ .$$

Accordingly, the *spectral density of radiation equilibrium* (36.1.3) is equal to

$$\rho(\nu,\ T) = \frac{8\pi\nu^2}{c^3} \frac{h\nu}{e^{h\nu/kT} - 1} \ .$$

36.3.3 At low frequencies $(h\nu \ll kT)$, $\exp[h\nu/kT - 1] \approx$ $\approx h\nu/kT$ and the Planck radiation formula coincides with the Rayleigh-Jeans law (36.2.3).

Following from Planck's radiation formula is the Stefan-Boltzmann law (36.2.1):

$$M_e^* = \frac{2\pi h}{c^2} \int_0^\infty \frac{\nu^3\, d\nu}{e^{h\nu/kT} - 1}$$

$$= \frac{2\pi k^4 T^4}{c^2 h^3} \int_0^\infty \frac{\eta^3\, d\eta}{e^\eta - 1} = \frac{2\pi^5 k^4}{15 c^2 h^3}\ T^4 \ .$$

The Planck constant is related to the Stefan-Boltzmann constant σ (36.2.1) as follows:

$$h = \sqrt[3]{\frac{2\pi^5 k^4}{15 c^2 \sigma}} = 6.62 \times 10^{-34}\ \text{J-s} .$$

36.3.4 The Planck formula for the emissive power r_λ^* of a blackbody is

$$r_\lambda^* = \frac{2\pi c^2 h}{\lambda^5} \frac{1}{e^{hc/\lambda kT} - 1} \ .$$

The wavelength λ_m corresponding to maximum r_λ^* is determined from the transcendental equation

$$xe^x - 5e^x + 5 = 0, \quad \text{where} \quad x = hc/\lambda_m kT.$$

The root of this equation $x = 4.965$, and λ_m satisfies the Wien displacement law (36.2.2):

$$\lambda_m T = b, \quad \text{where} \quad b = \frac{hc}{4.965\,k} = 2.9 \times 10^{-3}\ \text{m-K}.$$

36.4 Optical Pyrometry

36.4.1 *Optical pyrometry* is the name given to miscellaneous techniques for measuring high temperatures based on the laws of thermal radiation. The instruments used for this purpose are called *pyrometers.*

Radiation pyrometers record the integrated radiation of the heated body being investigated; optical pyrometers record the radiation of a heated body in one or two narrow portions of the spectrum.

36.4.2 The *radiant flux* Φ_e is the average power of optical radiation (31.4.3) during a time considerably greater than a period of oscillation of the electromagnetic field of the light.

The *irradiance* E_e (*radiant flux density*) of a surface is the ratio of the flux of incident radiation on this surface to its area. Thus $E_e = d\Phi_e/dS$, where $d\Phi_e$ is the radiant flux incident on a portion of the surface of area dS.

The *radiant intensity* I_e of a radiative body is the radiant flux emitted by the body within a unit solid angle. Thus $I_e = d\Phi_e/d\Omega$, where $d\Phi_e$ is the radiant flux within the solid angle $d\Omega$.

36.4.3 The *radiance* L_e of the portion dS of a radiative surface in a given direction is the ratio of the radiant intensity dI_e of area dS in the direction being considered to the area of the projection of dS on a plane perpendicular to the direction. Thus

$$L_e = \frac{dI_e}{dS \cos \varphi},$$

where φ is the angle between the specified direction and the normal to the surface of area dS.

The *spectral density of the radiance* is the ratio of the radiance dL_e, corresponding to a narrow band of the optical spectrum, to the width of this band. Thus

$$b_\nu = \frac{dL_e}{d\nu} \quad \text{and} \quad b_\lambda = \frac{dL_e}{d\lambda} .$$

36.4.4 A source of optical radiation is said *to obey Lambert's law*, or *to be a cosine radiator*, if its radiance L_e, as well as its spectral densities b_ν and b_λ are the same in all directions, i.e. are independent of angle φ. A blackbody could be called a cosine radiator.

The radiance of a cosine radiator and its spectral densities are related to its radiant exitance and its spectral densities (36.1.4) as follows:

$$L_e = \frac{M_e}{\pi}, \quad b_\nu = \frac{r_\nu}{\pi} \quad \text{and} \quad b_\lambda = \frac{r_\lambda}{\pi} .$$

36.4.5 In optical pyrometry, distinction is made between radiation, brightness and colour temperatures of a body.

The *radiation temperature* T_r of a body is the temperature of a blackbody at which its radiance L_c^* is equal to the radiance L_e of the given body. If the body being investigated is a cosine radiator, whose total degree of blackness is β (36.1.7), then, from the condition that $L_e(T) = L_e^*(T_r)$, where T is the true temperature of the body, it follows that

$$\beta M_e^*(T) = M_e^*(T_r) \quad \text{and} \quad T = \frac{T_r}{\sqrt[4]{\beta}} \geqslant T_r .$$

36.4.6 The *brightness (luminance) temperature* T_b of a body is the temperature of a blackbody at which its spectral density b_λ^* of radiance for a definite wavelength λ_0 is equal to the spectral density b_λ of radiance for the given body for the same wavelength: $b_\lambda(\lambda_0, T) = b_\lambda^*(\lambda_0, T_b)$. Usually $\lambda_0 = 660$ nm (red light).

For a cosine radiator, whose absorptance (36.1.5) for light of the wavelength λ_0 at the temperature T of the body is equal to

α_λ $(\lambda_0,\ T)$, it follows from Kirchhoff's law (36.1.6) and from Planck's formula (36.3.4) that

$$\alpha_\lambda\,(\lambda_0,\ T)\ r_\lambda^*\,(\lambda_0,\ T) = r_\lambda^*\,(\lambda_0,\ T_{\mathrm{b}})$$

and

$$\alpha_\lambda\,(\lambda_0,\ T)\left[\exp\left(\frac{a_2}{\lambda_0 T_{\mathrm{b}}}\right) - 1\right] = \exp\left(\frac{a_2}{\lambda_0 T}\right) - 1\ ,$$

where $a_2 = hc/k$. Since $\alpha_\lambda\,(\lambda_0,\ T) \leqslant 1$, then $T \geqslant T_{\mathrm{b}}$.

36.4.7 The *colour temperature* T_{C} of a body is the temperature of a blackbody at which the relative distributions of the spectral density of the radiance b_λ^* of the blackbody and that b_λ of the body being considered are close to one another to the maximum degree in the visible region of the spectrum. Thus

$$\frac{b_\lambda\,(\lambda_1,\ T)}{b_\lambda\,(\lambda_2,\ T)} = \frac{b_\lambda^*\,(\lambda_1,\ T_{\mathrm{c}})}{b_\lambda^*\,(\lambda_2,\ T_{\mathrm{c}})}\ .$$

For a cosine radiator

$$\frac{r_\lambda\,(\lambda_1,\ T)}{r_\lambda\,(\lambda_2,\ T)} = \frac{r_\lambda^*\,(\lambda_1,\ T_{\mathrm{c}})}{r_\lambda^*\,(\lambda_2,\ T_{\mathrm{c}})}\ .$$

As a rule, $\lambda_1 = 655$ nm (red light) and $\lambda_2 = 470$ nm (green light). The colour temperature of a greybody (36.1.5) coincides with its true temperature and can be found by Wien's displacement law (36.2.2).

CHAPTER 37   FUNDAMENTALS OF QUANTUM OPTICS

37.1 External Photoelectric Effect (Photoemissive Effect)

37.1.1 *Quantum optics* is the branch of optics that studies phenomena in which the quantum properties of light are manifested. Such phenomena include thermal radiation (36.1.1), the

photoelectric effect, the Compton effect (37.3.1) and photo-
chemical processes.

The *photoelectric effect in gases* consists in the ionization of the
atoms and molecules of the gas due to the action of light, and
is usually called *photoionization*.

In condensed bodies (liquids and solids) distinction is made
between the external and internal photoelectric effects.

The *external photoelectric effect* (*photoemission*) is the emission
of electrons by a substance subject to the action of light. Elec-
trons ejected from the substance in the external photoelectric
effect are called *photoelectrons*, and the electric current formed
by them in their ordered motion in the external electric field
is called the *photoelectric current*.

The *internal photoelectric effect* is the redistribution of electrons
among the energy states in solid and liquid semiconductors and
dielectrics due to the action of light. It is manifested in the
change in the concentration of current carriers in a medium
(20.1.4) and leads to photoconduction or the barrier-layer pho-
toelectric effect. *Photoconduction* is an increase in the electrical
conductivity of a substance due to light falling on it. The *bar-
rier-layer photoelectric effect* is the generating of an emf (photo-
emf) by the action of light in a system consisting of a semicon-
ductor and metal in čontact or two different kinds of semicon-
ductors [for example at a *p-n* junction (41.11.7)].

37.1.2 Shown in Fig. 37.1 is an installation for investigating
the external photoelectric effect in metals. Light enters through
window *D* and falls on the surface of cathode *K*, which is inside
a vacuum tube and is called the *photoelectric cathode*. The shape
of the curve, representing the dependence of the photoelectric
current *I* in the tube on the potential difference *U* between the
anode *A* and cathode *K* upon constant irradiance E_e of the cath-
ode (36.4.2) by monochromatic light, is shown in Fig. 37.2.

The existence of a photoelectric current at negative values of *U*
from 0 to $-U_0$ indicates that the photoelectrons have a certain
initial velocity when emitted from the cathode, and the corre-
sponding kinetic energy. The maximum initial velocity v_{max} of
the photoelectrons is related to the *retarding voltage* (*retarding
potential*) U_0 by the equation

$$\frac{mv_{max}^2}{2} = eU_0,$$

where e and m are the absolute values of the charge and the mass of the electron.

The photoelectric current increases with U only to a definite limiting value I_s called the *saturation photoelectric current*. At this value, all the electrons escaping from the cathode, due to
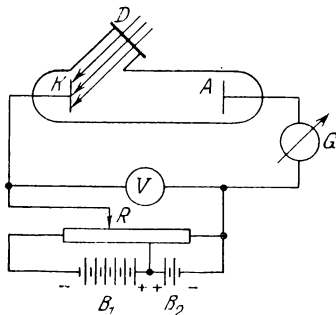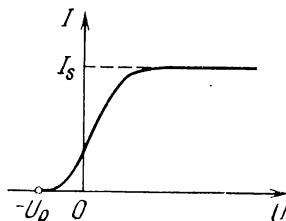
Fig. 37.1   Fig. 37.2

the effect of light, reach the anode. If n_{sec} is the number of photoelectrons emitted from the cathode per second, then

$$I_s = en_{sec}.$$

37.1.3 The external photoelectric effect obeys the following laws.

1. The *Stoletov law* states that at a constant spectral composition of the light incident upon the photoelectric cathode, the saturation photoelectric current is proportional to the irradiance (36.4.2) of the cathode:

$$I_s \propto E_e \quad \text{and} \quad n_{sec} \propto E_e.$$

2. For a given photoelectric cathode, the initial velocity of the photoelectrons depends upon the frequency of the light, but not on its intensity.

3. A *photoelectric threshold* exists for each photoelectric cathode, i.e. the minimum frequency ν_0 of light at which the external photoelectric effect is still feasible. This frequency ν_0 depends

upon the material of the photoelectric cathode and the condition of its surface.

The second and third laws cannot be interpreted on the basis of the classical electromagnetic theory of light. According to this theory the removal of conduction electrons from the metal is a result of their "swaying" in the electromagnetic field of the light wave so that they are "loosened". This effect should increase with the light intensity and with the irradiance of the photoelectric cathode, proportional to the intensity.

37.1.4 Only the quantum theory of light enabled the laws of the external photoelectric effect to be properly explained. In a further development of Planck's ideas on the quantization of the energy of atoms (oscillators) (36.3.2), Einstein proposed the hypothesis that light is not only radiated, but also propagates in space and is absorbed by matter in the form of separate discrete quanta of electromagnetic radiation, which were named *photons*. All photons of monochromatic light of the frequency ν have the same energy $E_{ph} = h\nu$, where h is Planck's constant, and travel in space at the velocity c of light in vacuum (or in free space). In the absorption of light each absorbed photon transmits all of its energy to a particle of the absorbing substance. In the external photoelectric effect, for example, a conduction electron of the metal, in absorbing a photon, receives all of its energy $h\nu$. The work function Φ (41.11.1) is the minimum energy that it is necessary to impart to the electron for it to escape from a metal. Hence, *Einstein's photoelectric emission equation*, expressing the law of the conservation of energy in the photoelectric effect, is of the form

$$h\nu = \Phi + \frac{mv_{max}^2}{2} \, .$$

37.1.5 The second law of the photoelectric effect follows directly from Einstein's equation:

$$\frac{mv_{max}^2}{2} = eU_0 = h\nu - \Phi \, .$$

Thus, v_{max} and U_0 depend only upon the frequency of the light and the work function in emitting an electron from the photoelectric cathode.

The maximum initial kinetic energy of the photoelectrons depends upon the light frequency according to a linear relation. It becomes zero at the frequency ν_0, corresponding to the photoelectric threshold. Thus

$$\nu_0 = \frac{\Phi}{h} \, .$$

Consequently, the photoelectric threshold depends only on the electron work function of the metal.

37.1.6 The photoelectric effect is inertialess, i.e. the emission of photoelectrons begins immediately, as soon as light with the frequency $\nu \geqslant \nu_0$ falls on the photoelectric cathode. This feature of the external photoelectric effect is still another confirmation of the quantum nature of interaction between light and matter. According to classical wave conceptions, quite considerable time is required for an electromagnetic wave of given intensity to impart sufficient energy to the electron for its emission from the metal.

37.1.7 At very high intensities of light, achieved by means of lasers (39.8.8), a *multiple-photon*, or *nonlinear*, *photoelectric effect* is observed. In this case, the electron can simultaneously receive the energy of N photons rather than that of one photon. Then the equation of the law of conservation of energy in the external photoelectric effect under the action of light of frequency ν is of the form

$$Nh\nu = \Phi + \frac{mv_{\max}^2}{2} \, .$$

The photoelectric threshold of the N-photon photoelectric effect is

$$(\nu_0)_N = \Phi/Nh.$$

37.2 Mass and Momentum of the Photon. Light Pressure

37.2.1 The *mass of the photon* (37.1.4) having the energy $E_{\mathrm{ph}} = h\nu$, where ν is the frequency of the light and h is Planck's constant, can be found from the mass-energy relation (5.7.2).

Thus

$$m_{\text{ph}} = \frac{h\nu}{c^2} ,$$

where c is the velocity of light in free space (or in a vacuum). The photon always travels with the velocity c, and its rest mass (5.6.1) equals zero.

37.2.2 The *momentum* $\mathbf{p}_{\text{ph}}$ *of the photon* is numerically equal to

$$p_{\text{ph}} = m_{\text{ph}}c = \frac{h\nu}{c} = \frac{h}{\lambda} ,$$

where λ is the wavelength of the light in vacuum. Since the wavenumber is $k = 2\pi/\lambda$,

$$p_{\text{ph}} = \frac{h}{2\pi} k = \hbar k \quad \text{and} \quad \mathbf{p}_{\text{ph}} = \hbar\mathbf{k},$$

where $\hbar = h/2\pi$, and $\mathbf{k}$ is the wave vector (30.2.7).

37.2.3 Light exerts a pressure on the bodies that reflect or absorb it. In quantum optics, light pressure is interpreted as the result of the transfer of momentum to bodies by photons as the light is reflected or absorbed. The *light pressure* p on the plane surface ab of a body (Fig. 37.3) is equal to the numerical value of the normal component of the total momentum transferred by photons to the body on unit area of the surface being considered during unit time. Assume that monochromatic light of frequency ν is incident upon surface ab at the angle i (Fig. 37.3) and that n_{sec} is the

Fig. 37.3

number of photons incident per second upon unit area of surface ab. If R is the coefficient of reflection (31.5.6) of light from the surface being considered, then, of the n_{sec} photons, Rn_{sec} are reflected and $(1 - R) n_{\text{sec}}$ are absorbed. The photons being reflected transfer a total momentum to the body, directed normal to the surface ab and numerically equal to

$$Rn_{\text{sec}} \frac{2h\nu}{c} \cos i .$$

34—0186
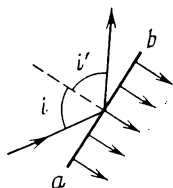
The photons being absorbed transfer a total momentum to the body, whose component normal to surface ab is numerically equal to

$$(1 - R)\, n_{\text{sec}}\, \frac{h\nu}{c} \cos i.$$

Thus the light pressure equals

$$p = R n_{\text{sec}}\, \frac{2h\nu}{c} \cos i + (1 - R)\, n_{\text{sec}}\, \frac{h\nu}{c} \cos i$$

$$= (1 + R)\, n_{\text{sec}}\, \frac{h\nu}{c} \cos i.$$

If n_0 is the photon concentration of the incident light, then $n_{\text{sec}} = n_0 c \cos i$ and $n_0 h\nu = \langle E^{\text{d}} \rangle$ is the average value of the volume energy density of the light. Hence

$$p = n_0 h\nu\, (1 + R) \cos^2 i = \langle E^{\text{d}} \rangle\, (1 + R) \cos^2 i.$$

Thus, the pressure of light can be equally well explained by either wave theory (31.2.6) or quantum theory.

37.3 Compton Effect

37.3.1 The *Compton effect* is the change in the wavelength of X rays when they are scattered by a substance containing light-weight atoms. The wavelength λ' of the radiation, scattered at the angle θ to the direction of propagation of the primary mono-chromatic radiation of wavelength λ, is longer than λ by the amount $\Delta\lambda$, which depends only on angle θ. Thus

$$\Delta\lambda = \lambda' - \lambda = 2\lambda_{\text{C}} \sin^2 \frac{\theta}{2}.$$

The constant $\lambda_{\text{C}} = 2.43 \times 10^{-12}$ m is called the *Compton wavelength of the electron*. The Compton effect cannot be explained on the basis of the classical wave theory of light.

37.3.2 According to quantum theory, the Compton effect is the result of elastic collision of the X-ray photon with a free or effectively free electron (in light atoms, the binding energy of an electron with the atom is substantially less than the energy of an X-ray photon). In this collision the photon transfers a

part of its energy and a part of its momentum to the electron in accordance with the laws of the conservation of energy and momentum. If the electron was initially at rest, it follows from the law of conservation of energy that

$$hv + m_0c^2 = hv' + mc^2, \qquad \text{(a)}$$

where $v = c/\lambda$ and $v' = c/\lambda'$ are the frequencies of the incident and scattered radiation, m_0c^2 is the rest energy of the electron (5.7.3), and mc^2 is the total energy of the electron (5.7.2) after the collision (energy of the *Compton recoil electron*).

Fig. 37.4

It follows from the law of conservation of momentum that

$$\mathbf{p_{ph}} = m\mathbf{v} + \mathbf{p'_{ph}},$$

or, in accordance with Fig. 37.4,

$$(mv)^2 = \left(\frac{hv}{c}\right)^2 + \left(\frac{hv'}{c}\right)^2 - 2\frac{h^2vv'}{c^2}\cos\theta, \qquad \text{(b)}$$

where $\mathbf{v}$ is the velocity of the Compton recoil electron.
The mass of the recoil electron is related to its velocity $\mathbf{v}$ by the equation (5.6.1)

$$m = \frac{m_0}{\sqrt{1-(v/c)^2}}. \qquad \text{(c)}$$

Combining equations (a), (b) and (c), we obtain

$$m_0c^2\,(v-v') = 2hvv'\sin^2\frac{\theta}{2}$$

or

$$\Delta\lambda = \lambda' - \lambda = \frac{2h}{m_0c}\,\sin^2\frac{\theta}{2}\,.$$
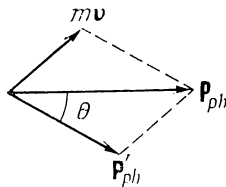
34*

Thus, the *Compton wavelength of the electron* is

$$\lambda_C^- = \frac{h}{m_0 c} \; .$$

37.3.3 The kinetic energy of the recoil electron (5.7.1) is

$$E_k = (m - m_0)\, c^2 = h\, (\nu - \nu') = h \nu \nu' \left(\frac{1}{\nu'} - \frac{1}{\nu} \right) = \frac{hc}{\lambda \lambda'} (\lambda' - \lambda)$$

or

$$E_k = h \nu \, \frac{\Delta \lambda}{\lambda + \Delta \lambda} = h \nu \; \frac{2a \sin^2 \dfrac{\theta}{2}}{1 + 2a \sin^2 \dfrac{\theta}{2}} \; ,$$

where $a = \lambda_C / \lambda$, and $h\nu$ is the energy of the incident photon. The energy E_k is maximal at $\theta = \pi$. Thus

$$E_{k\,max} = \frac{2ah\nu}{1 + 2a} \; .$$

37.3.4 When an electron is tightly bound to its atom, a photon scattered by the electron transfers energy and momentum, not to the electron, but to the atom as a whole. The mass of the atom is many times greater than that of the electron. Hence, only a small part of the photon's energy is transferred to it, so that the wavelength λ' of the scattered radiation does not differ practically from the wavelength λ of the incident radiation. The percentage of electrons that are tightly bound in their atoms increases with the atomic number of the element and, consequently, with the atomic mass. Therefore, the heavier the atoms of the scattering substance, the greater the relative intensity of the unshifted component ($\lambda' = \lambda$) in the scattered radiation. **37.3.5** In contrast to the scattering of photons, accomplished by both free and bound electrons, only *bound* electrons can absorb photons. For example, in the external photoelectric effect, a photon is absorbed by a bound electron, which expends a part of the energy it receives on the work function (i.e. in doing work to escape from the substance), the work function being a measure of how strongly the electron was bound in the substance. The absorption of a photon by a free electron is impossible, because this process would contradict the laws of conservation of energy and momentum. This can be readily shown by consider-

ing the absorption of a photon by a stationary free electron. It follows from the laws of conservation of energy and momentum that the following two relations must be simultaneously complied with:

$$(m - m_0)\, c^2 = h\nu \text{ and } mv = \frac{h\nu}{c}\,, \text{ where } m = \frac{m_0}{\sqrt{1 - (v/c)^2}}\,.$$

But these equations are simultaneous only at $\nu = 0$.

37.4 Wave-Particle Duality of the Properties of Light

37.4.1 Such phenomena as interference (30.5.2) and diffraction of light (33.2.1) are convincing proof of the wave nature of light. At the same time, the laws of thermal radiation equilibrium (36.1.2), the photoelectric effect (37.1.1) and the Compton effect (37.3.1) can be successfully interpreted only on the basis of quantum concepts of light as being a flux of discrete photons (37.1.4). But the wave and quantum (particle) methods of describing light mutually complement, rather than contradict, each other, because light simultaneously possesses both wave- and particlelike properties. Light is a dialectic unity of these opposite properties.

37.4.2 The principal equations relating the wavelike properties of light (frequency ν and wavelength λ in vacuum) to its particlelike properties (the energy E_{ph} and momentum p_{ph} of a photon) are:

$$E_{\mathrm{ph}} = h\nu \text{ and } p_{\mathrm{ph}} = \frac{h}{\lambda}\,.$$

The wavelike properties of light play the decisive role in the laws of its propagation, interference, diffraction, and polarization; whereas the particlelike properties play no less important a role in the interaction of light with matter. The longer the wavelength of light, the less the momentum and energy of the photon, and the more difficult it is to detect the quantum properties of light. For example, the external photoelectric effect occurs only at energies of a photon greater than or equal to the work function in releasing an electron from a substance (37.1.5). The shorter the wavelength of electromagnetic radia-

tion, the higher the energy and momentum of the photons, and the more difficult it is to detect the wavelike properties of this radiation. For example, X rays are diffracted only by an extremely "thin refraction grating, the crystal lattice of a solid (33.3.4).

37.4.3 Quantum optics employs the statistical approach in dealing with the laws of light propagation. According to this approach, the diffraction of monochromatic light by some obstacle (for instance, by a diffraction grating) consists in the redistribution of the photons in space caused by this obstacle. The probability that the photons will hit various points on a screen mounted behind the obstacle is not the same. This explains why a diffraction pattern appears on the screen. The irradiance E_e (36.4.2) of some small portion dS of the screen's surface is proportional to the number dn_{sec} of photons incident upon this portion (dS) per second. Consequently, E_e is proportional to the probability of photons hitting unit area of the screen's surface at the point being considered. On the other hand, according to wave conceptions, E_e is proportional to the square of the amplitude A of the light at the same point on the screen. Hence, the square of the amplitude of the light wave at some point in space is a measure of the probability of a photon being at this point.

37.4.4 Experiments in light diffraction indicate that in changing the intensity of a light flux incident on an obstacle, the appearance of the diffraction pattern, i.e. the relation between the irradiances at various points of the screen, remains unchanged. This indicates that wavelike properties are inherent in each separate photon, and not only in an assembly of a great number of simultaneously travelling photons. As a photon passes through an optical system it is impossible to indicate the particular point of the screen it will hit. It is feasible only to speak of the probability dE^d that the photon will hit some small portion dS of the surface of the screen.

PART SIX ATOMIC
AND MOLECULAR PHYSICS

CHAPTER 38 ELEMENTS OF QUANTUM MECHANICS

38.1 Wave-Particle Dualism of the Properties
of Particles of Matter

38.1.1 *Quantum mechanics** deals with the physics of atoms,
molecules and their associations, crystals in particular, as well
as atomic nuclei and elementary particles. Objects of the micro-
cosm studied by quantum mechanics have linear dimensions
of the order of 10^{-6} to 10^{-13} cm. If the particles travel at veloci-
ties $v \ll c$, where c is the velocity of light in free space, nonrel-
ativistic quantum mechanics is applied; at $v \leqslant c$, relativistic
quantum mechanics is resorted to.**
38.1.2 Quantum mechanics is based on Planck's concepts con-
cerning the discrete nature of changes in the energy of atoms
(36.3.2), Einstein's ideas on photons (37.1.4), and data on the
quantum nature of certain physical quantities (momentum and
energy, for instance) that characterize, under definite condi-
tions, the states of particles of the microcosm.
38.1.3 The basic concept of quantum mechanics is that the wave-
particle dualism of properties, established for light (37.4.1),
is of a universal nature. It should manifest itself for any parti-
cles having a momentum p. All particles with the finite momen-
tum p possess wave properties and their motion is accompanied
by a certain wave process.

---

* Also called *wave mechanics* in its early stages of develop-
ment.
** Information on relativistic quantum mechanics is beyond
the scope of the present handbook. Wherever applied, the term
quantum mechanics refers to nonrelativistic quantum mechanics.

38.1.4 The de Broglie formula establishes the dependence of the wavelength associated with a moving particle of matter on the momentum p of the particle. Thus

$$\lambda = \frac{h}{p} = \frac{h}{mv} \, ,$$

where m is the mass of the particle, v is its velocity, and h is Planck's constant (Appendix II). The waves mentioned here are called *de Broglie*, or *matter*, *waves*.
Another form of the de Broglie formula is

$$\mathbf{p} = \frac{h}{2\pi} \, \mathbf{k} = \hbar \mathbf{k},$$

where $\mathbf{k} = 2\pi \mathbf{n}/\lambda$ is the wave vector, whose magnitude $k = 2\pi/\lambda$ is the wavenumber, i.e. the number of wavelengths per 2π units of length, $\mathbf{n}$ is a unit vector in the direction of wave propagation, and $\hbar = h/2\pi = 1.05 \times 10^{-34}$ J-s.
38.1.5 The de Broglie wavelength for a particle of mass m, having the kinetic energy E_k (3.2.1), is

$$\lambda = \frac{h}{\sqrt{2mE_k}} \, .$$

In particular, for an electron accelerated in an electric field with a potential difference of ΔV volts (16.2.7),

$$\lambda = \frac{12.25}{\sqrt{\Delta V}} \, \text{Å}.$$

38.1.6 The de Broglie formula is confirmed by experiments in the scattering of electrons and other particles by crystals and in the passage of particles through matter. An indication of the wave process in all such experiments is the diffraction pattern observed, showing the distribution of the electrons (or other particles) in the particle detecting devices.
No wave properties are manifested by macroscopic bodies. The de Broglie wavelengths for such bodies are so short that it proves impossible to detect any wave properties.
38.1.7 The phase velocity (30.2.10) of de Broglie waves is

$$v_{\text{ph}} = \frac{\omega}{k} = \frac{c^2}{v} = \frac{c^2}{h} \, m\lambda,$$

where m is the mass of the particle, v is its velocity, and λ is the de Broglie wavelength. Since $c > v$, the phase velocity of de Broglie waves is greater than the velocity of light in free space. The dependence of the phase velocity of de Broglie waves on the wavelength indicates that these waves undergo dispersion (30.3.8). The group velocity (30.4.3) of de Broglie waves is equal to the particle velocity v. Thus

$$u = \frac{d\omega}{dk} = v.$$

Table 38.1 compares the particle and wave properties of particles of mass m, travelling at the velocity v.

TABLE 38.1

Particlelike properties	Wavelike properties
Velocity v	De Broglie wavelength $\lambda = \dfrac{h}{p} = \dfrac{h}{mv}$
Momentum $p = = mv$	De Broglie wave frequency $\nu = \dfrac{E}{h} = \dfrac{mc^2}{h}$
Energy $E = mc^2$.	Group velocity of de Broglie waves $u = v$ Phase velocity of de Broglie waves $v_{\mathrm{ph}} = = \dfrac{c^2}{v}$

38.1.8 Besides the de Broglie formula, it is assumed in quantum mechanics that the following relation exists between the total energy E of a particle and the frequency ν of its de Broglie wave:

$$E = h\nu = \hbar\omega,$$

where $\omega = 2\pi\nu$ is the angular frequency (28.1.2) and $\hbar = = h/2\pi$ (38.1.4).
38.1.9 De Broglie waves are of a specific nature having no analogies among the waves dealt with in classical physics. The square

of the magnitude   of the de   Broglie wave's   amplitude at a
given point is a measure of the probability that the particle
is to be found at this point (this is the *probability*, or *statistical*,
meaning of de Broglie waves). The diffraction patterns, observed
in the experiments mentioned in Sect. 38.1.6, are an indication
of statistical laws. According to these laws the particles, reach
definite places in the receivers, these being the places where
the intensity (30.3.6) of the de Broglie wave is found to be
highest. No particles are observed at the places where, accord-
ing to the statistical interpretation, the square of the magnitude
of the "probability wave" amplitude becomes zero.

38.2 Schrödinger Wave Equation

38.2.1 The position of a particle in space at a given instant
of time is determined in quantum mechanics by specifying its
Schrödinger wave function (also called the *psi function* or the
probability amplitude) ψ $(x,\ y,\ z,\ t)$. The probability dw that
the particle is within the element dV of volume is proportional
to $|\psi|^2$ and to the element dV of volume. Thus

$$dw = |\psi|^2\, dV,$$

where $|\psi|^2$ is the square of the magnitude of the psi function:
$|\psi|^2 = \psi\psi^*$. Here ψ^* is a function that is the complex conju-
gate of ψ. The quantity $|\psi|^2$ is the *probability density* $|\psi|^2 =$
$= dw/dV = \rho$ and specifies the probability of the particle oc-
cupying the given point in space. The intensity of the de Brog-
lie wave is determined by the quantity $|\psi|^2$.
38.2.2 Following from the definition of the psi function is the
normalization condition for probabilities:

$$\int\int\int |\psi|^2\, dV = 1,$$

in which the triple integral with respect to volume is calcu-
lated along the coordinates x, y and z, from $-\infty$ to $+\infty$, i.e.
throughout the whole infinite space. The normalization condi-
tion indicates that it is a certain event that the particle is some-
where in space and the probability of this event should be uni-
ty.
38.2.3 The wave function ψ $(x,\ y,\ z,\ t)$ is the main character-
istic of the state of microobjects (atoms, molecules and ele-

mentary particles). This function is made use of to calculate the average value of some physical quantity L that characterizes the object which is in the state described by the wave function ψ. Thus

$$\langle L \rangle = \int \int \int L \, |\psi|^2 \, dx \, dy \, dz,$$

where $\langle L \rangle$ is the average value of L and integration is carried out as indicated in Sect. 38.2.2.

38.2.4 The fundamental differential equation of quantum mechanics*, concerning the wave function $\psi \, (x, \, y, \, z, \, t)$ is called the *Schrödinger time-dependent wave equation*. It determines the psi function for microparticles in a field of force with the potential energy $U \, (x, \, y, \, z, \, t)$ (3.3.1) at the velocity $v \ll c$, where c is the velocity of light in free space. The Schrödinger equation is of the form

$$i\hbar \, \frac{\partial \psi}{\partial t} = -\frac{\hbar^2}{2m} \, \Delta \psi + U \, (x, \, y, \, z, \, t) \, \psi,$$

where Δ is the Laplace operator, m is the mass of the particle, $\hbar = h/2\pi$, h is Planck's constant, and $i = \sqrt{-1}$.

The Schrödinger equation is supplemented by the following conditions that are imposed on the psi function:

(a) the psi function should be finite, single-valued and continuous;

(b) the partial derivatives $\partial \psi/\partial x$, $\partial \psi/\partial y$ and $\partial \psi/\partial z$ should be continuous;

(c) the function $|\psi|^2$ should be integrable, i.e. the integral $\int\limits_{-\infty}^{+\infty} \int \int |\psi|^2 \, dV$ should be finite. In the simplest cases, this condition is reduced to the normalization of probabilities (38.2.2).

38.2.5 In the case when the psi function is time-independent, i.e. $\psi = \psi \, (x, \, y, \, z)$, it satisfies the Schrödinger steady-state

---

* See the second footnote to Sect. 38.1.1.

equation:

$$\Delta\psi + \frac{2m}{\hbar^2}(E-U)\psi = 0,$$

where E is the total energy of the particle. The remaining notation is given in Sect. 38.2.4. The psi functions satisfying the Schrödinger equation for the given form of $U = U(x, y, z)$ are called *eigenfunctions*. They exist only at definite values of E, called the *eigenvalues of the energy*. The whole set of eigenvalues of E forms the *energy spectrum of the particle*. If U is a monotonic function and $U \to 0$ at infinity, then in the region $E < 0$, the eigenvalues of the energy form a *discrete spectrum*. The most important problem of quantum mechanics is to find the eigenvalues and eigenfunctions of particles (or systems of particles).

38.2.6 Schrödinger's time-dependent equation has the solution:

$$\psi(x, y, z, t) = \psi(x, y, z)\, e^{-iEt/\hbar}.$$

The state of the particle at the given instant of time is described by a periodic function of time with the angular frequency $\omega = E/\hbar$, determined by the total energy E of the particle. This corresponds to the relation between the total energy E of the particle and the de Broglie wave frequency (38.1.8).

If a particle is in a definite energy state with the energy $E = $ const, the probability dw of finding it in the element dV of volume is independent of time. Thus

$$dw = |\psi|^2\, dV = \psi\psi^*\, dV.$$

This is said to be the *steady state* of the particle. An atom in the steady state has constant energy and does not radiate electromagnetic waves (39.1.7).

38.3 Motion of a Free Particle

38.3.1 In free motion of a particle ($U = 0$), its total energy E coincides with its kinetic energy. If the OX axis is taken along vector $\mathbf{v}$ of the particle's velocity ($\mathbf{v} = $ const), Schrödinger's steady-state equation (38.2.4) has the following solution:

$$\psi = Ae^{-(i/\hbar)\sqrt{2mE}\,x} + Be^{(i/\hbar)\sqrt{2mE}\,x},$$

where m is the mass of the particle, $\hbar = h/2\pi$, h is Planck's constant, and A and B are certain constants. The Schrödinger time-dependent equation (38.2.4) has, in this case, the solution

$$\psi\,(x,\ y,\ z,\ t) = A e^{-i\left(\frac{E}{\hbar}t - \frac{\sqrt{2mE}\,x}{\hbar}\right)} + B e^{-i\left(\frac{E}{\hbar}t + \frac{\sqrt{2mE}\,x}{\hbar}\right)},$$

which is the superposition of two plane monochromatic waves (31.1.6) of equal frequency $\omega = E/\hbar$, propagating, one in the positive direction of the OX axis with the amplitude A, and the other in the opposite direction with the amplitude B.

38.3.2 In quantum mechanics a free particle is described by a plane monochromatic de Broglie wave with the wavenumber k (38.1.4):

$$k = \frac{1}{\hbar}\,\sqrt{2mE}\,.$$

The probability of finding the particle at any point in space is constant. For waves propagating along the positive direction of the OX axis,

$$|\,\psi\,|^2 = \psi\psi^* = |\,A\,|^2.$$

38.4 A Particle in a One-Dimensional Infinitely Deep Potential Well

38.4.1 A *potential well* is a region of space in which the potential energy U of a particle is less than a certain value $U_{\max}$. In particular, at $U = U\,(x)$ and $U_{\max} = \infty$, there is a one-dimensional potential well of infinite depth. If the potential energy of the particle outside and inside the potential well has the following values (Fig. 38.1):
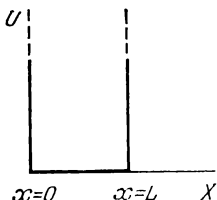
$$U = 0 \quad \text{at} \quad 0 \leqslant x \leqslant L,$$
$$U = \infty \quad \text{at} \quad x \leqslant 0 \quad \text{and} \quad x \geqslant L,$$

the well has a "flat bottom".

The motion of collectivized electrons in a metal is dealt with in classical electronic theory as motion in a potential well,

the potential energy of the electron outside the metal equalling zero, and that inside the metal being negative and numerically equal to the work function of the metal (energy required to remove an electron from the metal) (41.11.1). The potential well illustrated in Fig. 38.1 is simpler than the real potential well of electrons in a metal.

38.4.2 The Schrödinger steady-state equation (38.2.5) for particles in a potential well, considered in the preceding subsection, is of the form

$$\frac{d^2\psi}{dx^2} + \frac{2m}{\hbar^2} E\,\psi = 0$$

Fig. 38.1

at the boundary conditions $\psi\,(0) = \psi\,(L) = 0$, which indicate that $\psi = 0$ and $|\psi|^2 = 0$ outside the region $0 \leqslant x \leqslant L$, i.e. that the probability of finding the particle outside the potential well is zero.

The solution of the Schrödinger equation is

$$\psi\,(x) = A\,\cos kx + B\,\sin kx,$$

where A and B are constants, and $k = \sqrt{2mE}/\hbar$ is the wavenumber (38.1.4). It follows from the boundary conditions that $A = 0$, $B \neq 0$ and $\sin kL = 0$, i.e. the wavenumber can assume a set of discrete values that comply with the requirement: $k_n L = n\pi$, where $n = 1, 2, 3, \ldots$.

The last equation means that

$$\frac{2\pi}{\lambda_n} = \frac{\pi n}{L} \quad \text{or} \quad \lambda_n = \frac{2L}{n}.$$

A whole number of de Broglie half-waves should fit into the length of the potential well.

38.4.3 Physical quantities that can have only definite discrete values are said to be *quantized* (*quantized physical quantities*). The eigenvalues of the energy E_n of a particle in a one-dimensional infinitely deep potential well,

$$E_n = \frac{n^2\pi^2\hbar^2}{2mL^2} \; (n = 1, \ 2, \ \ldots),$$

comprise a discrete set of energy values that are quantized (see also Sect. 38.2.5).

The quantized values of E_n are called *energy levels*, and the numbers n, determining the energy levels of a particle in a potential well, are called *quantum numbers*.

38.4.4 At large quantum numbers $(n \gg 1)$, the energy levels of a particle in a potential well become relatively closer to one another. The relation $\Delta E/E_n \approx 2/n \ll 1$, where $\Delta E = E_{n+1} - E_n = (2n + 1)\, n^2 \hbar^2/2mL^2$. The inequality $\Delta E \ll E_n$ at $n \gg 1$ indicates that the quantization of energy at large quantum numbers yields results close to those obtained in classical physics: the energy levels become *quasi-continuous (quasi-continuous energy levels* at $n \gg 1$).

Bohr's correspondence principle states that at high quantum numbers the conclusions and results obtained in quantum mechanics should correspond to those of classical mechanics.

A more general statement of the correspondence principle is that between any theory developed from a classical one and the initial classical theory there must be a regular relationship: in definite limiting cases, the new theory must go over into the old one. For example, the formulas of kinematics and dynamics of the special theory of relativity become the formulas of Newton's mechanics at velocities where $(v/c)^2 \ll 1$ (5.3.4). Geometrical optics is a limiting case of wave optics if the wavelength can be neglected $(\lambda \to 0)$.

38.5 Linear Harmonic Oscillator

38.5.1 A *linear (one-dimensional) harmonic oscillator* is a particle of mass m, vibrating with the natural angular frequency ω_0 (28.1.2) along a certain axis OX due to the action of quasi-elastic force F, which is proportional to the displacement x of the particle from its equilibrium position. Thus $F = -kx$. Here k is a factor of the quasi-elastic force, related to m and ω_0 by the equation: $k = m\omega_0^2$ (28.2.3). The potential energy U of the harmonic oscillator (28.2.3) is

$$U(x) = \frac{kx^2}{2}.$$

38.5.2 In classical physics the amplitude (28.1.3) of small vibration of a harmonic oscillator is determined by the margin of its total energy E (Fig. 38.2). At points B and A, with the coordinates $\pm a$, the total energy E is equal to the potential energy: $E = \begin{cases} U(a) \\ U(-a) \end{cases}$, where a is the amplitude of vibration of the clas-
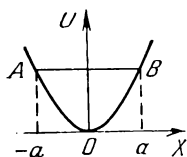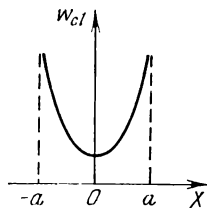
Fig. 38.2   Fig. 38.3

sical harmonic oscillator. Such an oscillator cannot get outside the limits of the region $(-a, +a)$.

38.5.3 The probability $w_{cl}(x)\,dx$ of finding the oscillator on the line segment from x to $x + dx$ in accordance with classical mechanics,

$$w_{cl}(x)\,dx = \frac{1}{\pi a}\,\frac{dx}{\left(1 - \dfrac{x^2}{a^2}\right)^{1/2}},$$

is represented by the curve in Fig. 38.3.

38.5.4 The vibration of a linear harmonic oscillator is investigated in quantum mechanics by means of the Schrödinger steady-state equation (38.2.5):

$$\frac{d^2\psi}{dx^2} + \frac{2m}{\hbar^2}\left(E - \frac{m\omega_0^2 x^2}{2}\right)\psi = 0,$$

The solutions of this equation, complying with the conditions in Sect. 38.1.4, are the eigenfunctions (wave functions) of a linear harmonic oscillator:

$$\psi_n(x) = \frac{1}{\sqrt{x_0}}\,e^{-\xi^2/2}\,H_n(\xi), \quad \xi = \frac{x}{x_0},$$

where $x_0 = \sqrt{h/2\pi m\omega_0}$, $H_n(\xi)$ is a Chebyshev-Hermite polynomial of the nth order:

$$H_n(\xi) = \frac{(-1)^n}{\sqrt{2^n n! \sqrt{\pi}}} \, e^{\xi^2} \frac{d^n e^{-\xi^2}}{d\xi^n}.$$

The eigenfunctions for $n = 0$, 1, 2 are:

$$\psi_0(x) = \frac{1}{\sqrt{x_0 \sqrt{\pi}}} \, e^{-x^2/2x_0^2},$$

$$\psi_1(x) = \frac{1}{\sqrt{2x_0 \sqrt{\pi}}} \, \frac{2x}{x_0} \, e^{-x^2/2x_0^2};$$

$$\psi_2(x) = \frac{1}{\sqrt{8x_0 \sqrt{\pi}}} \left(4\frac{x^2}{x_0^2} - 2 \right) e^{-x^2/2x_0^2}.$$

The *node of a wave function* is its value equal to zero. The number of nodes of the function ψ_n is equal to the quantum number n (38.4.3).

38.5.5 The eigenvalues of the energy E_n of a linear harmonic oscillator,

$$E_n = \left(n + \frac{1}{2} \right) h\nu_0 = \left(n + \frac{1}{2} \right) \hbar\omega_0 \quad (n = 0,\ 1,\ 2,\ \ldots),$$

where $\nu_0 = \omega_0/2\pi$, and ω_0 is the natural angular frequency (38.5.1), are the set of equally spaced energy levels shown in Fig. 38.4. At $n \gg 1$ (where $n + 1/2 \approx n$), the energy levels of the oscillator coincide with the values of quantized energy of the oscillator $E'_n = n\hbar\omega_0$, postulated by Planck in the theory of blackbody radiation (36.1.5).

38.5.6 The minimum energy that can be had by a harmonic oscillator is called the *zero-point energy* E_0:

$$E_0 = \frac{h\nu_0}{2} = \frac{\hbar\omega_0}{2} \quad (\text{at } n = 0).$$

In classical physics and in Planck's theory it was assumed that $E_0 = 0$ (at $n = 0$). This means that the oscillator does not vibrate and is in a state of equilibrium. Atoms, as oscillators,

should not, according to classical physics, execute any vibra-
tions at the temperature of absolute zero $(T = 0)$. It has been
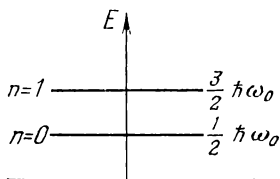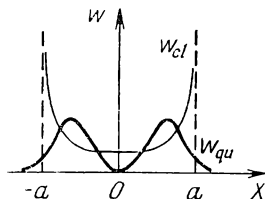
Fig. 38.4   Fig. 38.5

proved in modern quantum mechanics, that the zero-point ener-
gy of a harmonic oscillator cannot be reduced however much it
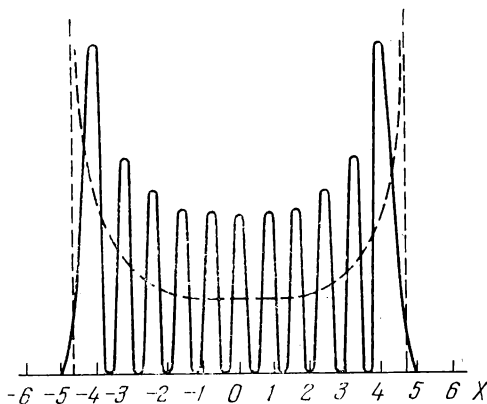
Fig. 38.6

is cooled, even down to absolute zero (11.8.4). Zero-point energy
of the oscillator corresponds to its *zero vibration*. In quantum

mechanics, zero-point energy is a typical feature of any system of particles. At temperatures close to absolute zero, a substance is in a condensed state and its atoms (or molecules or ions) are dealt with as vibrating oscillators. Zero-point energy is the least amount of energy that a quantum oscillator should have in its lowest energy state (at $n = 0$) in order to comply with the indeterminacy relation (38.6.2).

Fig. 38.7

38.5.7 The probability of finding a quantum linear harmonic oscillator on the OX axis in the region from x to $x + dx$ is

$$w_{qu}(x)\, dx = |\,\psi_n\,(x)\,|^2\, dx.$$

Compared in Fig. 38.5 are the quantum probability density at $n = 1$ and the classical probability density $w_{cl}(x)$. The existence of nonzero values of $w_{qu}(x)$ beyond the limits of the classically allowed region $|\,x\,| \geqslant a$ is explained by the feasibility of the passing of particles possessing wavelike properties through a potential barrier (38.7.2).

38.5.8 As the number n is increased, the probability distribution curve $|\,\psi_n\,|^2$, illustrated in Fig. 38.6 for $n = 10$, becomes more and more like the classical probability distribution curve (Fig. 38.3). This agrees with Bohr's correspondence principle (38.4.4).

38.5.9 Making use of the eigenfunctions of the energy of a linear harmonic oscillator, the average energy $\langle E \rangle$ of such an oscillator can be calculated:

$$\langle E \rangle = \frac{\hbar \omega_0}{2} + \frac{\hbar \omega_0}{\exp{(\hbar \omega_0 / kT)} - 1} = \frac{\hbar \omega_0}{2} + \langle E_1 \rangle.$$

Accurate to within zero-point energy, $\langle E \rangle$ is expressed by the term $\langle E_1 \rangle$, which was obtained by Planck in developing his theory of the thermal radiation of a blackbody. Graphically, this term is represented in Fig. 38.7, at $T = \text{const}$, as a function of the frequency. It is evident that the greatest contribution to $\langle E_1 \rangle$ is made by vibrations with low frequencies, corresponding to long wavelengths.
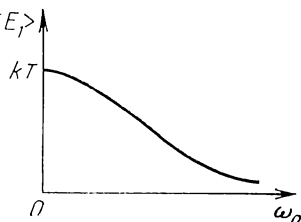
At a high temperature $(kT \gg \hbar\omega_0)$

$$\langle E_1 \rangle = \frac{\hbar\omega_0}{1 + \dfrac{\hbar\omega_0}{kT} - 1} = kT .$$

This result coincides with that obtained from the principle of equipartition of energy (10.6.4).

38.6 Heisenberg Indeterminacy Principle

38.6.1 The wavelike properties of microparticles (38.1.4) introduce limitations into the possibility of applying the concepts of coordinates and momentum in their classical form to such particles.

Limitations also exist in classical physics in the application of certain concepts to definite items. It is of no meaning, for instance, to apply the idea of temperature to a single molecule; the concept of point localization (location at a certain point) is inapplicable in determining the position of a wave in space, etc. But in classical mechanics, a definite value of the particle's coordinate corresponds to exact values of its velocity and momentum. In quantum mechanics there exist fundamental restrictions in the possibility of simultaneously determining the exact coordinate of a particle and its exact momentum. These restrictions are associated with the wave-particle dualism of the properties of microparticles (38.1.3).

38.6.2 The *Heisenberg indeterminacy relations* are the inequalities

$$\Delta x \times \Delta p_x \geqslant \frac{\hbar}{2} , \quad \Delta y \times \Delta p_y \geqslant \frac{\hbar}{2} , \quad \Delta z \times \Delta p_z \geqslant \frac{\hbar}{2} .$$

Here Δx, Δy and Δz are the ranges of the coordinates within which a particle, described by a de Broglie wave (38.1.4), may be localized if the projections of its momentum along the coordinate axes are within the ranges Δp_x, Δp_y and Δp_z. The Heisenberg relations indicate that the coordinates x, y, and z of the particle and the projections p_x, p_y and p_z of its momentum on the corresponding axes cannot simultaneously have values exactly equal to x and p_x, y and p_y, and z and p_z. These physical

quantities can have values given to an accuracy determined by the Heisenberg relations. The more accurately the position of the particle is determined, i.e. the smaller the values of Δx, Δy and Δz, the less accurately the values are determined for the projections of its momentum (i.e. the greater are Δp_x, Δp_y and Δp_z). If the position of the particle along the OX axis is exactly determined and $\Delta x = 0$, then $\Delta p_x = \infty$ and the value of the projection p_x of the momentum on this axis becomes entirely indeterminate.

38.6.3 The indeterminacy relations put definite restrictions in quantum mechanics on the feasibility of describing the motion of a particle along a certain path.

In classical theory each point of the path of a particle has definite coordinates x, y, and z and the particle has a definite momentum $\mathbf{p}$ with projections p_x, p_y and p_z along the coordinate axes. This is realized in quantum mechanics only in cases when the particle travels in a macroscopic region of space (for instance, the particle may leave a trace on a photographic plate or on the screen of an oscillograph). If, for example, the position of an electron is fixed with an accuracy determined by the linear dimensions of a grain of photographic emulsion, subject to the effect of the electron, then $\Delta x \approx 10^{-6}$ m. This corresponds to an indeterminacy in the momentum equal to $\Delta p_x \geqslant \hbar/2\Delta x \approx$ $\approx 10^{-28}$ kg-m/s and in the velocity equal to $\Delta v_x = \Delta p_x/m \approx$ $\approx 10^2$ m/s. This indeterminacy at electron velocities of the order of 10^6 to 10^7 m/s enables us to assume that the electron is travelling along a definite path with an exact velocity specified at each point of this path.

38.6.4 If the particle is travelling in a microscopic region of space, the indeterminacy relations substantially affect the nature of its motion. For example, the position of an electron travelling in an atom can be determined with an accuracy to the dimensions of the atom, i.e. $\Delta x \approx 10^{-10}$ m. But the indeterminacy of the electron's velocity Δv_x, in this case, is found to be of the same order of magnitude as the velocity itself: $\Delta v_x \approx$ $\approx 10^6$ m/s $\approx v$. The path of an electron in an atom with the exact velocity specified at each point does not have the meaning it has in classical mechanics (1.2.3). This by no means indicates that the indeterminacy principle places fundamental limitations on our knowledge of the microcosm. It only indicates the limited applicability of the concepts of classical physics to the region of the microcosm.

38.6.5 No restrictions are made by the indeterminacy principle with respect to the feasibility of applying the conception of the coordinate and momentum of a macroscopic body in the classical sense. The wave properties of such bodies are not manifested in any way (38.1.6) and, consequently, the indeterminacy relations are not applied to them.

38.6.6 The indeterminacy relations for energy E and time t are

$$\Delta E \times \Delta t \geqslant \hbar,$$

where ΔE is the indeterminacy of the energy of a particle, which is in a state with the energy E during the length of time Δt. The longer a particle is in the given state, the more accurately its energy in this state can be determined.

38.6.7 *Measurement* is the process of interaction of an instrument with the item being investigated as a result of which certain information is obtained on the properties of the item. This process proceeds in space and time and is an objective process. The interaction between the measuring instrument and the item being measured essentially differs for microscopic and macroscopic items. In the latter case, the measuring process is described with more or less accuracy by the laws of classical physics, and the instrument has no effect on the item being measured that cannot be accurately taken into account in the terms and concepts of classical physics or reduced to as small a value as desired. In quantum mechanics, owing to the objectively existing duality of the properties of microscopic items (38.1.3), the measuring process is inevitably associated with substantial effects of the instrument on the course of the phenomenon being investigated. For example, to determine the position of the electron it must be "illuminated" by light of the highest possible frequency. As a result of the collision between the photon and the electron, the momentum p_x of the latter is changed by an amount determined by the indeterminacy relation: $\Delta p_x \geqslant \hbar/2\Delta x$ (38.6.2). This effect on the item being measured cannot be considered small or inessential because the state of the item is changed. This change is such that as a result of measurement, definite classical characteristics of a particle, for instance, its momentum, are found to be specified only within the limits restricted by the indeterminacy relations.

38.7 Tunnel Effect

38.7.1 Potential wells in which particles are located may be of much more complicated shape than those dealt with in Sect. 38.4. If the potential energy of the particle is of the form $U = U(r)$, as shown in Fig. 38.8, the particle moves in a field of forces that
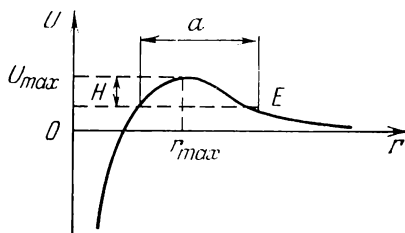
Fig. 38.8

can be represented in the form of a *potential barrier*. This means that the forces act ón the particle in a certain limited region of space. Outside of this region the particles move as free ones. The height H of the barrier and its width a for a given particle depend upon the energy E of the particle. In order for a particle to get out of a 'potential well or to enter it from outside, it is necessary, according to classical mechanics, to impart energy to the particle equal to or greater than $| U_{max} - E | = H$.

38.7.2 The *tunnel effect* is the passing ("leaking") of particles through potential barriers. The tunnel effect is a quantum-mechanical one and is due to the wavelike properties of the particles (38.1.3).

The *penetrability (penetrability factor)* D of a potential barrier is the quantity

$$D = \frac{I_{tran}}{I_{inc}},$$

where I_{tran} is the intensity (30.3.6) of de Broglie waves passing through the potential barrier, and I_{inc} is the intensity of the waves incident on the barrier,

The penetrability of a rectangular potential barrier of the height U_0 and width L (Fig. 38.9) is

$$D = D_0 \ \exp \left[-\frac{2}{\hbar} \sqrt{2m \ (U_0 - E)} L \right],$$

where m is the mass of the particle, and E is its energy. For a potential barrier of complex shape

$$D = D_0 \ \exp \left\{ -\frac{2}{\hbar} \int_{x_1}^{x_2} \sqrt{2m \ [U(x) - E]} \, dx \right\},$$

where x_1 and x_2 are the coordinates of the points at the beginning and end of the potential barrier $U \ (x)$ for a given value E of
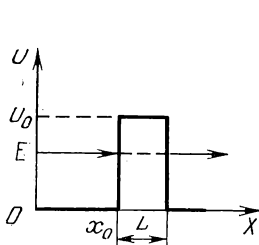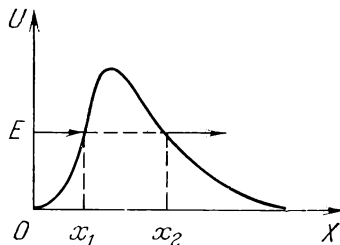
Fig. 38.9   Fig. 38.10

the total energy (Fig. 38.10). In these formulas D_0 is a constant factor close to unity.

38.7.3 The tunnel effect plays an important role when the penetrability is not too low. This occurs in the cases when the linear dimensions of the potential barrier are commensurable with atomic dimensions. For example, at $U_0 - E = 10$ eV for an electron ($m \approx 10^{-30}$ kg) and $L = 10^{-10}$ m, we obtain $D \approx 1/e$. Under the same conditions, excepting that $L = 10^{-2}$ m, $D \approx \approx \exp (-10^8)$. The penetrability of the barrier decreases with

an increase in the mass of the particle and in the difference $U_0 - E$.

38.7.4 The *paradox of the tunnel effect* is that the passing of a particle through the potential barrier enables it to be found in a region that is forbidden from the classical point of view. In this region the potential energy of the particle exceeds its total energy; its kinetic energy is negative and, consequently, the velocity (momentum) of the particle becomes an imaginary quantity.

Actually no such paradox exists. The tunnel effect is a purely quantum-mechanical phenomenon and here an unexpected difficulty arises, from the classical point of view, in representing the total energy E in the form of the sum of the kinetic and potential energies: $E = p^2/2m + U(x)$. Owing to the fact that the indeterminacy relation (38.6.2) excludes the possibility of simultaneously specifying definite values, to any degree of accuracy, for the coordinate x and momentum p of a particle, it is found to be inadmissible in quantum mechanics to represent the total energy as the sum of two exactly definite parts: the kinetic $p^2/2m$ and potential $U(x)$ energies. If the coordinate x of the particle is fixed in the region Δx and its potential energy $U(x)$ is determined with sufficient accuracy, then the indeterminacy Δp introduced into the value of the particle's momentum ($\Delta p \approx \hbar/\Delta x$) excludes the possibility of specifying the exact value of the kinetic energy $p^2/2m$ of the particle. Consequently, the previous value of the particle's total energy E cannot be specified either. It is found, in this case, that the change in the kinetic energy ΔE_k, caused by fixing the particle's coordinate, exceeds the difference between the barrier height U_0 (Fig. 38.9) and the total energy E of the particle:

$$\Delta E_k > U_0 - E.$$

Thus, ΔE_k exceeds the energy that the particle, inside the potential well, lacks to surmount the barrier "in the classical way" to get out of the well.

The role played by the tunnel effect in solid-state physics and in the phenomenon of radioactive decay is dealt with in Sects. 41.8.3 and 42.5.2.

CHAPTER 39   STRUCTURE OF ATOMS AND MOLECULES AND THEIR OPTICAL PROPERTIES

39.1 The Hydrogen Atom and Hydrogen-Like Ions

39.1.1 An *atom* is the smallest particle of matter possessing all the chemical properties of the given chemical element. An atom consists of a positively charged nucleus and electrons moving about in the electric field of the nucleus. The nuclear charge Ze (42.1.2) is equal in magnitude to the total charge of all the electrons of the atom. An *ion* is the electrically charged particle formed when an atom or molecule loses or gains electrons.

39.1.2 The simplest atomic system is the hydrogen atom, consisting of a single proton in the nucleus and a single electron moving in the Coulomb field of the nucleus. *Hydrogen-like ions (isoelectronic with hydrogen)* are such as He^+, Li^{++} and Be^{++}, which also have a nucleus with the charge Ze and a single electron.

39.1.3 Among the optical properties of atoms, the most important is their emission spectrum. The frequency ν of the lines in the discrete line spectrum of hydrogen is described by the Balmer-Rydberg formula:

$$\nu = cR' \left(\frac{1}{n^2} - \frac{1}{m^2} \right) = R \left(\frac{1}{n^2} - \frac{1}{m^2} \right),$$

where

$$R = \frac{me^4}{8\varepsilon_0^2 h^3} \text{ (in SI units),}$$

$$R = \frac{2\pi^2 me^4}{h^3} \text{ (in Gaussian units).}$$

Here c is the velocity of light in free space, m is the mass of the electron, e is the electron charge, h is Planck's constant, and ε_0 is the electric constant in SI units (14.2.7).
Quantities R and $R' = R/c$ are called the *Rydberg constant* in s^{-1} and in cm^{-1} or m^{-1}, respectively: $R = 3.2931193 \times 10^{15} \text{ s}^{-1}$ and $R' = 1.0973731 \times 10^7 \text{ m}^{-1}$.

The whole numbers n and m are called the *principal quantum numbers*, and $m = n + 1$, $n + 2$, etc. A group of lines having the same n value is called a *series*. The hydrogen spectrum has the following series: *Lyman series* at $n = 1$, *Balmer series* at $n = 2$, *Paschen series* at $n = 3$, *Brackett series* for $n = 4$, *Pfund series* for $n = 5$ and the *Humphreys series* for $n = 6$.

For hydrogen-like ions (39.1.1) the Balmer-Rydberg formula is

$$\nu = Z^2 R \left(\frac{1}{n^2} - \frac{1}{m^2} \right),$$

where Z is the atomic number of the element in the Mendeleev periodic table (39.3.5).

39.1.4 Each value of the quantum number n in the Balmer-Rydberg formula corresponds to the *series limit* with the maximum frequency (at $m = \infty$), called the *term* T_n. Thus

$$T_n = \frac{R}{n^2} \text{ (for hydrogen)}, \quad T_n = \frac{Z^2 R}{n^2} \text{ (for hydrogen-like ions)}.$$

39.1.5 The energy of a hydrogen-like ion in a state with the principal quantum number n is:

$$E_n = -\frac{Z^2 R h}{n^2}, \quad T_n = \frac{|E_n|}{h}.$$

The magnitude E_n is called the *binding energy* of the electron in the atom. The minimum value E_1 (at $n = 1$) corresponds to the *ground*, or *normal*, state of the atom (39.1.9). All energy values at $n > 1$ characterize *excited states of the atom*. The most vital distinguishing feature of excited states is the finite lifetime τ of the electron in these states: $\tau \approx 10^{-8}$ s. In the ground state of the atom, isolated from all external effects, τ is unlimited. The maximum energy value $E_{\max} = 0$ as $n \to \infty$ corresponds to *ionization* of the atom or *ion*, i.e. to the removal of an electron from it.

In magnitude, the ionization energy is equal to the binding energy of an electron in the atom (or ion).

The *ionization potential* (16.2.6) of an atom of hydrogen or of a hydrogen-like ion in the state with the principal quantum number n is equal to $V = Z^2 R h / e n^2$, where e is the magnitude of the charge of the electron. A diagram of the energy levels in the hydrogen atom and the arrangement of its series of spectral lines are given in Fig. 39.1.
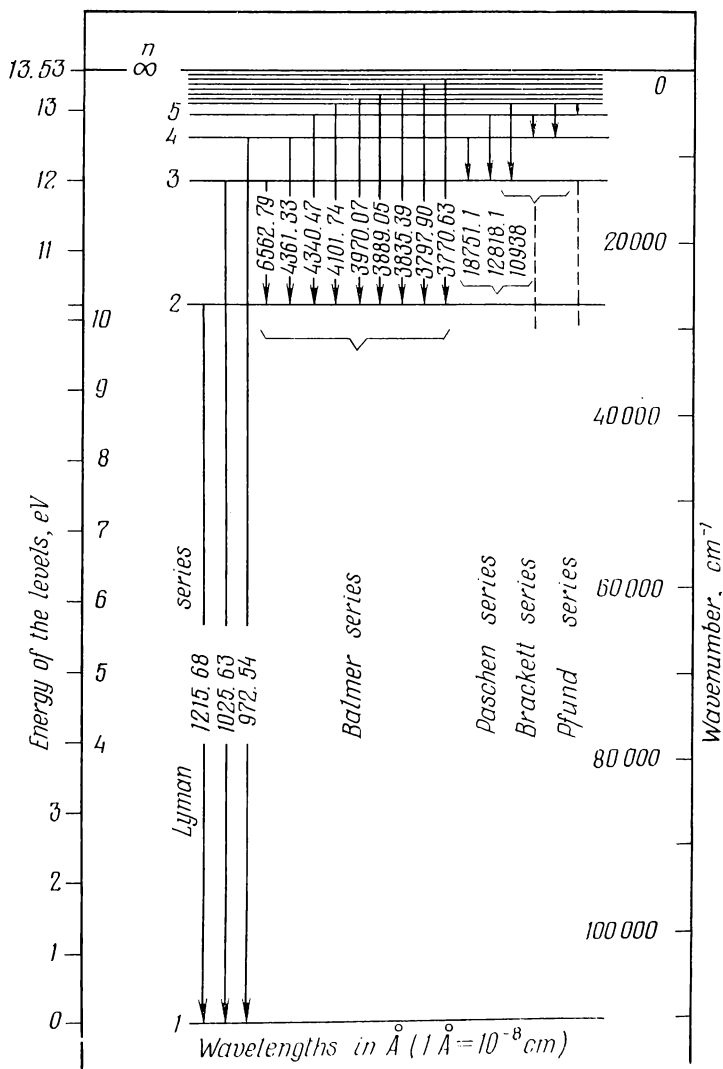
Fig. 39,1

39.1.6 The spectrum and the energy levels of the hydrogen atom were first interpreted by means of Bohr's postulates. *Bohr's first postulate (steady states postulate)* states that there are a set of stationary, or steady, states of the atom in which it does not radiate electromagnetic waves. The steady states correspond to steady orbits (39.1.8) along which electrons travel with acceleration, but no light is radiated (cf. Sect. 31.3.4). Bohr's first postulate was explained by quantum mechanics (38.2.6).

Bohr's second postulate (orbit quantization rule) states that in the steady state of the atom, the electron, travelling along a circular orbit, has quantized values of its angular momentum that comply with the condition

$$L_k = mvr = k\hbar \quad (k = 1, 2, 3, \ldots).$$

Here m is the mass of the electron, v is its velocity, r is the radius of the kth orbit, and $\hbar = h/2\pi$. The whole number k is equal to the number of de Broglie wavelengths for the electron (38.1.4) per length of the circular orbit: $2\pi r/\lambda = 2\pi rmv/h = k$.

The quantization of the angular momentum in quantum mechanics is discussed in Sect. 39.1.8.

Bohr's third postulate (the Bohr frequency condition) states that in going from one steady state orbit to another, the atom either emits or absorbs one photon. A photon is emitted when the atom passes from a state with more energy to one with less energy. Upon transition in the opposite direction, a photon is absorbed. The energy $h\nu$ of the photon is equal to the difference in energy in the two states. Thus

$$E_n - E_m = h\nu.$$

When $E_n > E_m$ a photon is emitted; when $E_n < E_m$ a photon is absorbed.

In quantum mechanics, Bohr's frequency condition follows from the theory of quantum transitions from one energy state to another. Information on this theory is beyond the scope of the present handbook.

39.1.7 Schrödinger's steady-state equation (38.2.5) for the motion of an electron in the Coulomb field of a nucleus with the charge Ze (42.1.2) is of the form

$$\Delta\psi + \frac{2m}{\hbar^2}(E - U)\psi = 0,$$

where $U\ (r) = -Ze^2/4\pi\varepsilon_0 r$ is the potential energy of an electron located at the distance r from the nucleus, ε_0 is the electric constant in SI units (14.2.7), and E is the total energy of the electron in the atom, which it is necessary to determine under the assumption that the wave functions ψ comply with the conditions indicated in Sect. 38.2.4.

39.1.8 A solution of the Schrödinger steady-state equation for an electron in a centrally symmetrical Coulomb field of the nucleus yields the following:

(a) the angular momentum of the electron in the atom is quantized according to the formula:

$$L_l = \sqrt{l\,(l+1)}\,\hbar,$$

where the *orbital quantum number* l, determining the angular momentum, varies within the limits: $l = 0,\ 1,\ \ldots,\ (n-1)$, where n is the principal quantum number (39.1.3);

(b) at $E < 0$, when the electron is "bound" in the atom, its motion is periodic, and the value of the total energy E is quantized. The eigenvalues E_n (38.2.5) are determined by the formula:

$$E_n = -\frac{Z^2 m e^4}{8h^2\varepsilon_0^2}\,\frac{1}{n^2} = -\frac{Z^2 Rh}{n^2} \quad \text{(in SI units)},$$

$$E_n = -\frac{2\pi^2 Z^2 m e^4}{h^2} = -\frac{Z^2 Rh}{n^2} \quad \text{(in Gaussian units)}.$$

Here R is the Rydberg constant (39.1.3). The solution of the Schrödinger equation for an electron in a hydrogen-like ion leads to energy levels of the Balmer-Rydberg type (39.1.4).

39.1.9 In accordance with the orbital quantum number, the following letters have been assigned to designate the state of the electron in atoms:

s-state at $l = 0$, p-state at $l = 1$,
d-state at $l = 2$, f-state at $l = 3$, etc.

The s-state of the electron in the hydrogen atom at $n = 1$ is called the ground state (see also Sect. 39.1.5). This state is spherically symmetrical. The wave function of this state depends only upon the distance r of the electron from the nucleus [$\psi = \psi\,(r)$] and is of the form

$$\psi = \psi\,(r) = Ce^{-r/a_0},$$

where C is a constant determined from the normalization condition for probabilities (38.2.2), and a_0 is the *first Bohr radius*:

$$a_0 = \frac{4\pi\varepsilon_0\hbar^2}{me^2} \text{ (in SI units)},$$

$$a_0 = \frac{\hbar^2}{me^2} \text{ (in Gaussian units)}.$$

In quantum mechanics, the electron orbits in the atom are considered to be the locus of points at which the probability of finding the electron is a maximum. In particular, for the *s*-state of the hydrogen atom, such an orbit is the first circular Bohr orbit of a radius equal to a_0.

39.2 Space Quantization

39.2.1 *Space quantization* is the existence of definite discrete orientations in space of the angular momentum vector $\mathbf{L}_l$ of the electron. This phenomenon has been proven in quantum mechanics. The only allowed orientations of $\mathbf{L}_l$ are those in which the projection L_{lz} of vector $\mathbf{L}_l$ along the direction Z of the external magnetic field assumes values that are multiples of $\hbar$. Thus

$$L_{lz} = m\hbar,$$

where m is a whole number called the *magnetic quantum number* and assumes the values $m = 0, \pm 1, \pm 2, \ldots, \pm l$, and l is the orbital quantum number (39.1.8). Given in Fig. 39.2 is the space quantization of vectors $\mathbf{L}_l$ for electrons in the *p*- and *d*-states (39.1.9).

39.2.2 Since the angular momentum vector of the electron is proportional to the orbital magnetic moment of the electron (26.1.2), it follows from space quantization that the vector $\mathbf{p}_m$ of the orbital magnetic moment of the electron (26.1.2) in the external magnetic field cannot assume arbitrary orientations.

39.2.3 It was experimentally established that space quantization is observed in atoms having a single outer-shell (valence) electron in the s-state (where $l = 0$) (39.1.9). In this state the atoms lack angular momentum ($L_0 = 0$; see Sect. 39.1.8). Space quantization, detected in such experiments, pertained to electron spin, and confirmed the existence of two possible spin orientations in the external magnetic field (26.1.3). The magnitude of the spin angular momentum L_{ls} of the electron is determined by the formula (39.1.8):
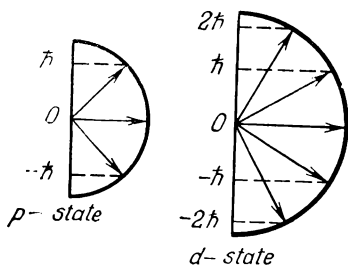
$$L_{ls} = \sqrt{s(s+1)}\,\hbar,$$

where s is the *spin quantum number*, equal to $s = 1/2$. Therefore, the numerical value of electron spin equals (26.1.3):

$$L_{ls} = \sqrt{\frac{1}{2}\left(\frac{1}{2}+1\right)}\,\hbar = \frac{\sqrt{3}}{2}\,\hbar.$$

Space quantization of spin means that the projection L_{lsz} of the spin vector $\mathbf{L}_{ls}$ on the direction of the external magnetic field is determined by the formula

$$L_{lsz} = m_s \hbar,$$

where m_s is the *spin magnetic quantum number*, differing from the spin quantum number s in that it can assume two values: not only $+1/2$, but $-1/2$ as well*.

Fig. 39.2

p— state

d— state

---

* Frequently, no distinction is made between the two spin quantum numbers s and m_s, referring, in this case, to m_s to which two values, $\pm 1/2$, are assigned.

39.3 Pauli Exclusion Principle. Mendeleev's Periodic Table

39.3.1 The simplest statement of the *Pauli exclusion principle* is: in any atom there cannot be two electrons occupying two identical steady states, which are specified by a set of four quantum numbers: the principal n, orbital l, magnetic m and spin magnetic m_s quantum numbers.

Also obeying the Pauli principle, in addition to electrons, are other particles having half-integral spin (in $\hbar$ units). In any system of fermions (41.2.4) no two particles can exist in identical quantum states (41.1.4).

For the electrons in an atom, the Pauli principle is written as follows:

$$Z_1\ (n,\ l,\ m,\ m_s) = 0\ \text{or}\ 1,$$

where $Z_1\ (n,\ l,\ m,\ m_s)$ is the number of electrons in the state specified by the given set of quantum numbers.

39.3.2 The maximum number $Z_2\ (n,\ l,\ m)$ of electrons occupying states specified by a set of three quantum numbers, n, l and m, and differing only in the orientations of the spins of the electrons, is

$$Z_2\ (n,\ l,\ m) = 2,$$

because the spin magnetic quantum number m_s can assume only two values: $1/2$ and $-1/2$.

39.3.3 The maximum number $Z_3\ (n,\ l)$ of electrons occupying states specified by two quantum numbers, n and l, is

$$Z_3\ (n,\ l) = 2\ (2l + 1).$$

This is based on the fact that for a given l value vector $\mathbf{L}_l$ can assume $(2l + 1)$ different orientations in space.

39.3.4 The maximum number of $Z\ (n)$ electrons occupying states specified by the value n of the principal quantum number is

$$Z\ (n) = \sum_{l=0}^{l=n-1} Z_3\ (n,\ l) = 2n^2.$$

Table 39.1 lists the maximum number of electrons occupying states characterized by the given values of the principal n and orbital l quantum numbers.

36—0186

TABLE 39.1

n	Shell	Number of electrons occupying the states					Maximum number of electrons
		$s\ (l = 0)$	$p\ (l = 1)$	$d\ (l = 2)$	$f\ (l = 3)$	$g\ (l = 4)$	
1	K	2	—	—	—	—	2
2	L	2	6	—	—	—	8
3	M	2	6	10	—	—	18
4	N	2	6	10	14	—	32
5	O	2	6	10	14	18	50

39.3.5 The systematics for filling the electron states in atoms and the periodicity with which changes occur in the properties of chemical elements enabled them all to be arranged into *Mendeleev's periodic system of elements*. Modern theory of the periodic system is based on the following principles:
(a) The atomic number Z of a chemical element is equal to the total number of electrons in an atom of the given element.
(b) The state of the electrons in an atom is determined by a set of four quantum numbers: n, l, m and m_s. The distribution of the electrons according to the energy states should comply with the principle of minimum potential energy: as the number of electrons is increased, each consecutive electron should occupy the energy state with the least possible energy.
(c) The filling of the energy states in the atom by electrons should proceed in accordance with the Pauli principle (39.3.1).
39.3.6 An *electron shell* is a collection of all the electron states in an atom, which have the same value of the principal quantum number n (39.1.3). Distinction is made between the following electron shells (Table 39.1): K at $n = 1$, L at $n = 2$, M at $n = 3$, N at $n = 4$, etc.
Inside the electron shell, the electrons are distributed among the *subshells*, each of which corresponds to a certain value of the orbital quantum number l (39.1.8).
39.3.7 The order for the filling of the electron energy states in an atom in the shells and, within one shell, in the subshells, should correspond to the sequence of the energy levels with the given n

and l values, and comply with the Pauli principle (39.3.5b). For light-weight atoms, this order consists of first filling the shell with the lower value of n, and only then filling the next shell with electrons. Inside a given shell, first the state with $l = 0$ is filled, then the states with larger and larger l values up to $l = n - 1$.

Such an ideal periodic system of elements would have a structure and number of elements in each period (lengths of the periods) corresponding to Table 39.1.

39.3.8 Deviations from the order indicated in Sect. 39.3.7 begin with potassium ($Z = 19$) and are explained as follows. The interaction between electrons in the atom leads, at sufficiently large principal quantum numbers n, to a situation in which states with larger n and lesser l values can have less energy, i.e. be more favourable from the energy viewpoint than states with less n, but with higher l values. As a result we have elements with incompletely filled preceding shells, but in which subsequent shells are being filled. Chemical elements, in which the filling of preceding shells is completed after the partial filling of subsequent shells, are called *transition elements*.

39.3.9 The *outer (valence) electrons* of an atom are those in the shell with the highest n value that belong to the s- and p-subshells, i.e. they have $l = 0$ or $l = 1$ (39.1.9). These electrons determine the chemical and optical properties of the atoms (see also Sect. 20.3.1). The total number of electrons in the s- and p-subshells is equal to 8 (Table 39.1). Most chemical reactions are based on giving up or attaching outer (valence) electrons. It proves expedient, theoretically, for an atom, with the $(s + p)$-subshell less than half-filled to give up valence electrons, and for one with a more than half-filled $(s + p)$-subshell to attach more electrons from other atoms in the process of chemical reactions.

39.4 Chemical Bonds and Molecular Structure

39.4.1 A *molecule* is the smallest particle of a given substance, possessing its principal chemical properties. Molecules consist of either the same or different atoms (39.1.1), joined together by *interatomic chemical bonds*. Chemical bonds are based on various interactions of the outer (valence) electrons of the atoms (39.3.9). This is indicated by the changes in the optical spectrum, which explains the behaviour of these electrons when molecules are formed of atoms.

36*

A definite amount of work must be done to dissociate molecules into their constituent atoms. On the contrary, the formation of molecules is accompanied by the evolution of energy. This shows that there are forces which bind atoms into molecules. The energy evolved in forming molecules is a measure of the forces of interaction that join atoms into molecules.

39.4.2 At large distances r between atoms, distances that considerably exceed the linear dimensions d of the atoms, forces of mutual attraction exist between them. At distances r comparable to d, the forces become ones of mutual repulsion, which do not permit the electrons of a given atom to penetrate too deeply into the electron shells of an adjacent atom. As in the case of interaction between molecules (12.1.4), the repulsive forces between atoms are more short-ranged, than those of attraction (Fig. 12.1). The signs of the forces of attraction and repulsion are discussed in Sect. 12.1.4.
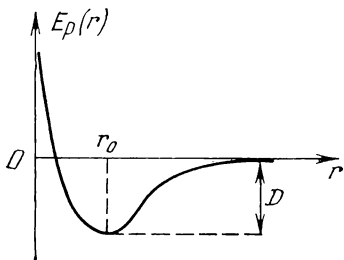
Fig. 39.3

39.4.3 At a certain distance r_0 between the atoms of a diatomic molecule, the oppositely [directed forces of attraction and repulsion counterbalance each other and their vector sum equals zero. The distance r_0 corresponds to the minimum mutual (potential) energy $E_p(r)$ of the atoms of a diatomic molecule (cf. Sect. 12.1.6) (Fig. 39.3). The equilibrium interatomic distance r_0 in a molecule is called the *bond length*, or *distance*. The quantity D is called the *dissociation*, or *bond, energy*. It is numerically equal to the work that must be done to dissociate a molecule into its constituent atoms and to separate them to infinite distances from one another. The dissociation energy is equal to the energy evolved in forming a molecule, but is of opposite sign. Dissociation energy is negative, whereas the energy evolved in forming molecules is positive.

39.4.4 *Heteropolar (ionic)* molecules are ones formed as the result of the conversion of interacting atoms into oppositely electrically charged and mutually attracting ions (39.1.1). This kind of bond between the atoms of a molecule is called a *heteropolar* (or *ionic* or *electrovalent*) *bond*. The formation of heteropolar molecules is

based on the behaviour of atoms with differently filled outer $(s + p)$-subshells, discussed in Sect. 39.3.9. Typical heteropolar (ionic) molecules are those of the alkali-halide salts, formed by ions from the elements of the I and VII groups of the periodic system (39.3.5). These include: NaCl (Na^+Cl^-), CsI (Cs^+I^-) and others. Metals of the first group have low ionization potentials V (39.1.5), whereas atoms of group VII are characterized by a high *electron affinity*, which is the amount of energy released when an electron is attached to a metalloid atom. The transition of an electron from a metal atom to a metalloid atom leads to the formation of ions, each of which has a stable eight-electron $(s + p)$-subshell (39.3.9). As a result, a stable heteropolar (ionic) molecule is formed.

39.4.5 *Homopolar* (*atomic*) molecules are ones formed by the mutual attraction of neutral atoms. The chemical bond of the atoms of a homopolar molecule is said to be *covalent*. A covalent bond has the property of being saturated (saturation of a covalent bond), which is manifested as the definite valency of the atoms. An atom of hydrogen can be bound to only one other atom, and an atom of carbon to no more that four other atoms. The simplest molecule with a covalent bond is the hydrogen molecule H_2; it consists of two electrons and two nuclei—protons.

The basis of the quantum-mechanical explanation of the covalent bond in the H_2 molecule is the indistinguishability of identical particles, e.g. the electrons in a hydrogen molecule (see also Sect. 41.1.2). The electrons in the H_2 molecule, each of which "belongs" to a definite nucleus, can be interchanged and this still leads to the same system: an H_2 molecule consisting of two electrons and two nuclei. The indistinguishability of electrons leads to the occurrence of a special *quantum-mechanical exchange interaction* between two identical electrons. This interaction consists in the electron of each of the atoms in the hydrogen molecule spending a certain part of its time at the nucleus of the other atom. This constitutes the bond between the two atoms forming the molecule. It follows from quantum-mechanical calculations that when two hydrogen atoms approach each other to a distance commensurable with the Bohr radius (39.1.9), and under the condition that the spins of the electrons in the atoms are antiparallel (26.1.3), the two atoms attract each other and a stable H_2 molecule is formed. With parallel spins of the electrons, the two atoms of hydrogen repel each other, and no molecule of hydrogen is formed.

39.5 Optical Properties of Molecules. Molecular Spectra

39.5.1 Owing to their specific appearance, the spectra of molecules, or *molecular spectra*, are called *band spectra*. They consist of a set of more or less wide bands formed of closely spaced spectral lines. Bands are observed in molecular spectra in the infrared, visible and ultraviolet ranges of the scale of electromagnetic waves, (31.4.3). Closely spaced bands form a *group of bands*. Several groups of bands are observed for the simplest diatomic molecules. Continuous broad emission (or absorption) bands are observed in the visible and ultraviolet regions of the spectra of polyatomic complex molecules.

39.5.2 As in the spectra of atoms (39.1.3), a separate spectral line in a molecular spectrum appears as a result of a change in the energy of a molecule. The total energy E^m of a molecule can be presented as the sum of the following parts, assumed independent in a first approximation:

$$E^m = E_t^m + E_e^m + E_v^m + E_r^m + E_n^m,$$

where E_t^m is the energy of translational motion of the centre of mass (2.3.3) of the molecule, E_e^m is the energy of motion of the electrons in the molecule's atoms, E_v^m is the energy of vibrational motion about their equilibrium positions of the nuclei in the molecule's atoms, E_r^m is the energy of rotary motion of the molecule as a whole, and E_n^m is the energy of the nuclei in the molecule's atoms. The energy E_t^m is not quantized (38.4.3) and its change cannot lead to the production of a molecular spectrum; the effect of E_n^m on the molecular spectrum can be neglected in a first approximation. The energy E'^m of the molecule, whose change governs the molecular spectrum, constitutes the sum of three components:

$$E'^m = E_e^m + E_v^m + E_r^m.$$

39.5.3 According to the Bohr frequency condition (39.1.6), the frequency ν of the photon emitted by the molecule when its energy

state changes is equal to

$$\nu = \frac{\Delta E'^{\mathrm{m}}}{h} = \frac{\Delta E_{\mathrm{e}}^{\mathrm{m}}}{h} + \frac{\Delta E_{\mathrm{v}}^{\mathrm{m}}}{h} + \frac{\Delta E_{\mathrm{r}}^{\mathrm{m}}}{h},$$

where $\Delta E_{\mathrm{e}}^{\mathrm{m}}$, $\Delta E_{\mathrm{v}}^{\mathrm{m}}$ and $\Delta E_{\mathrm{r}}^{\mathrm{m}}$ are the changes in the corresponding parts of the energy E'^{m}, which assume discrete quantized values. The production of closely spaced lines, forming bands at various parts of the spectrum, is explained by the fact that

$$\Delta E_{\mathrm{r}}^{\mathrm{m}} \ll \Delta E_{\mathrm{v}}^{\mathrm{m}} \ll \Delta E_{\mathrm{e}}^{\mathrm{m}}.$$

In the far-infrared region of the spectrum (wavelengths of the order of 0.1 to 1 mm), the transition of the molecule from one rotational energy level to another produces spectral lines of a *rotation spectrum*.

39.5.4 In the infrared region of the spectrum (wavelengths from one to several dozen micrometres), transitions of molecules be-

Vibrational energy levels   *Rotational energy levels*
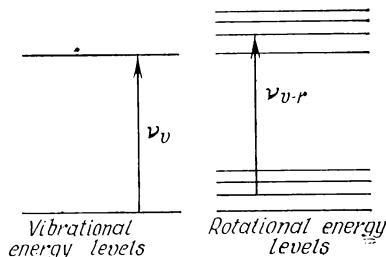
Fig. 39.4

tween vibrational energy levels produce the lines of the *vibrational spectrum* of the molecule. Upon changes in the vibrational energy levels of the molecule, its rotational energy levels are also changed at the same time. Hence, transitions of molecules between vibrational levels are vibrational-rotational transitions, producing a *vibrational-rotational spectrum* with the frequencies ν_{vr} (Fig. 39.4). This spectrum consists of groups of closely spaced lines

determined by the fact that the given vibrational transition is accompanied by rotational transitions.

39.5.5 The visible and ultraviolet regions of the spectrum are produced as a result of transitions of the molecule between the various electron energy levels. Each electron energy level corresponds to various possible vibrations of the nucleus in the molecule,
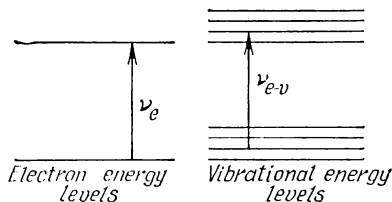
Fig. 39.5

i.e. to a set of vibrational energy levels. Transitions between electron-vibrational levels lead to the production of an *electron-vibrational spectrum* of the molecule, characterized by the frequencies ν_{ev} of the various lines (Fig. 39.5). Since each vibrational energy state corresponds to a system of rotational levels (39.5.4), each electron-vibrational transition corresponds to a certain band. The whole electron-vibrational spectrum of the molecule in the visible and nearby regions consists of a system of several groups of bands.

39.6 Raman Scattering of Light *

39.6.1 *Raman scattering of light*, or the Raman effect, refers to the shifted frequencies ν_s and ν_a observed, along with the frequency ν_0 of the source of radiation, in the spectrum of light scattered by a solid or liquid body. A spectral line with the frequency $\nu_s = \nu_0 - \nu$ is called a *Stokes (red) line*; one with the frequency $\nu_a = \nu_0 + \nu$ is called an *anti-Stokes (violet) line*. The set of lines

---

* Commonly called combination scattering in the Soviet literature.

with the frequencies ν_s and ν_a form the Raman spectrum of the molecule, containing red and violet *satellites with the frequency* ν_0. The intensity of the violet satellites is less than that of the red ones, but increases with the temperature. The intensity of the red satellites is practically independent of the temperature.

39.6.2 The quantum-mechanical explanation of the Raman scattering phenomenon consists in an analysis of the interaction between a photon of incident light having the frequency ν_0 and a molecule at the ground vibrational energy level with the energy E_{1v}^m. If, as a result of the interaction with the photon, the molecule is moved to higher vibrational energy level with the energy $E_{2v}^m > E_{1v}^m$, the energy $\Delta E^m = E_{2v}^m - E_{1v}^m$ required for this purpose is borrowed from the incident light. A photon with the energy $h\nu_0$ is absorbed and, in its place, a photon is produced with lower frequency ν and an energy equal to $h\nu = h\nu_0 - \Delta E^m$. At this the frequency ν appears in the scattered light. It equals

$$\nu_s = \nu_0 - \frac{\Delta E^m}{h} ,$$

and corresponds to the frequency of a red satellite. The whole set of red satellites are produced in transitions of the molecule to various excited vibrational energy states.

39.6.3 If a molecule at the vibrational energy level with the energy E_{2v}^m (39.6.2) moves, due to interaction with a photon having the energy $h\nu_0$, to an energy state with less energy (E_{1v}^m), a violet satellite of the frequency ν_a is produced in the scattered light spectrum. This frequency

$$\nu_a = \nu_0 + \frac{\Delta E^m}{h}$$

where $\Delta E^m = E_{2v}^m - E_{1v}^m$.

The probability of Raman scattering with an increase in frequency is less than the same scattering with a reduction in frequency. This is due to the fact that the number of molecules at the higher energy levels is less than the number at the lower levels. Consequently, the intensity of violet satellites is less than that of the red ones. When the temperature is raised, the molecule popula-

tion of the upper energy levels increases, as does the intensity of the violet satellites. The number of molecules in the ground state changes only slightly upon heating. Therefore, the intensity of the red satellites, remains practically unchanged as the temperature is raised.

39.7 Luminescence. X rays

39.7.1 *Luminescence* is the radiation of light by bodies at a temperature T, which is over and above the thermal radiation (36.1.1) at this temperature. The duration of luminescence substantially exceeds the periods of radiation of atomic systems. Depending upon the method applied to excite the luminescent glow, distinction is made between cathodoluminescence, electroluminescence, photoluminescence and chemiluminescence. Luminescent substances are called *luminophors*.

39.7.2 Luminescent radiation is a nonequilibrium one (cf. Sect. 36.1.2) and is caused by a comparatively small number of *luminescence centres*—atoms, molecules or ions—put into an excited state by the source of luminescence. The return of the excited centre to its ground or less excited state is accompanied by *luminescent radiation*. The duration of this radiation is determined by the duration of the excited state, which depends on the properties of not only the radiating centres, but of the medium surrounding them as well. The duration of a metastable excited state can reach 10^{-4} s, thereby correspondingly increasing the duration of the luminescence.

39.7.3 Luminescence that immediately ceases together with the action of the glow exciting source is called *fluorescence*.

Luminescence that persists long after the action of the glow exciting source has ceased to act is called *phosphorescence*.

The phenomenon of fluorescence is associated with the transitions of atoms, molecules or ions from ordinary excited states with a duration of the order of 10^{-8} s to the ground state. Phosphorescence, which provides for prolonged glow, is due to the transition of luminescence centres from metastable states to the ground state (39.7.2). The subdivision of luminescence into fluorescence and phosphorescence is arbitrary; it is sometimes difficult to establish a time-based boundary between them.

39.7.4 The excitation of luminescence by means of an electron beam is feasible provided that the kinetic energy of a bombarding

electron satisfies the inequality

$$\frac{mv^2}{2} \geqslant E_e - E_g,$$

where E_e and E_g are the total energies of the luminescent particle in the excited and ground states.

39.7.5 *Photoluminescence* is excited by electromagnetic radiation in the visible or ultraviolet ranges and complies with the *Stokes law*: the wavelength of photoluminescence is usually longer than that of the exciting light. The quantum-mechanical basis for the Stokes law is that in the absorption of a photon of exciting light with the energy $h\nu$, a photon with the energy $h\nu_{lum}$, less than $h\nu$, is produced. The surplus energy $h\nu - h\nu_{lum} = E^m$, where E^m is energy expended on various processes except photoluminescence.

As a rule, $E^m > 0$ and $\nu_{lum} < \nu$, i.e. $\lambda_{lum} > \lambda$ in accordance with the Stokes law. Sometimes, *anti-Stokes luminescent radiation* is observed. It complies with a condition that is the opposite of the Stokes law, i.e. $\lambda_{lum} < \lambda$. This occurs when a definite part of the energy of thermal motion of the particles of the luminescent substance is added to the energy $h\nu$ of the photon that excites the luminescence. Thus

$$h\nu_{lum} = h\nu + akT,$$

where a is a factor depending upon the nature of the luminescent substance, k is Boltzmann's constant (8.4.5), and T is the absolute temperature.

39.7.6 *Photoluminescent energy efficiency* is the ratio of the energy of luminescent radiation to the energy from the source absorbed under steady-state conditions by the luminophor.

The *quantum yield of photoluminescence* is the ratio of the number of photons of photoluminescent radiation to the number of photons of the exciting light, the latter having constant energy. *Vavilov's law* states that the photoluminescent energy efficiency increases proportionally to the wavelength λ of the absorbed radiation. Then, after reaching a maximum value in the wavelength range from λ to λ_{max}, drops rapidly to zero upon a further increase in wavelength. With an increase in the wavelength of the exciting light of given energy, there is an increase in the number of photons contained in the given energy of primary radiation. Each of these photons can produce a photon of the energy $h\nu_{lum}$. There-

fore, the photoluminescent energy efficiency increases with the wavelength λ. The sharp drop in energy efficiency at $\lambda = \lambda_{max}$ indicates that photons with the frequency less than $\nu_{min} = c/\lambda_{max}$ cannot excite the particles of the luminophor.

39.7.7 *X rays* are electromagnetic waves with a wavelength from 0.01 to 800 Å that are produced by braking high-speed electrons by a substance. There are two types of short-wave X rays.

At electron energies, not exceeding a certain critical value that depends upon the substance in which the electrons are braked, *white*, or *continuous*, *X rays* are produced. It is radiated by braked electrons and has a continuous spectrum, bounded on the short-wavelength side by a certain boundary λ_{min}, called the *continuous-spectrum boundary*. The boundary wavelength λ_{min} depends upon the kinetic energy E_k of the electrons that produce the bremsstrahlung and is reduced with an increase in the kinetic energy E_k. The existence of λ_{min} is explained by the fact that the maximum energy $h\nu_{max}$ of a photon of X rays, produced at the expense of the energy of the electron, cannot exceed E_k. Thus

$$h\nu_{max} = E_k.$$

Consequently, $\lambda_{min} = c/\nu_{max} = ch/E_k$. Making use of data on the values of E_k and λ_{min}, this formula enables Planck's constant (Appendix II) to be determined. In its time this method was one of the most accurate and reliable ones.

39.7.8 The second type of X rays, *characteristic X radiation* of atoms of matter, has a line spectrum. It is an individual characteristic of a substance and does not change when the substance enters a chemical compound. It follows that, contrary to optical spectra, characteristic X rays are associated with processes in the deep-lying filled inner electron shells of atoms (39.3.6), which undergo no changes in chemical reactions of the atoms. *Line X-ray spectra* consist of lines composing several series. The same types of line series are observed in various elements, differing only in that the like series of lines in the heavier elements are shifted toward the side of shorter wavelengths.

39.7.9 In the order of increasing wavelengths, the series of characteristic X rays are called the K-, L- M-, and N-series. As an electron is removed from one of the inner shells of an atom with the nuclear charge Ze, the vacant space is occupied by an electron from a shell farther from the nucleus and an X-ray photon is emitted. Thus, after removing an electron from the K-shell, the filling of the vacancy by electrons from L-, M- and other shells, produces the K_α-, K_β- and K_γ-lines that form the K-series.

The frequency ν of the lines of characteristic X rays can be found by *Mosley's law*:

$$\sqrt{\bar{\nu}} = a\,(Z - b),$$

where *a* is a constant for the given series of lines (in $s^{-1/2}$), and *b* is the screening constant. This constant means that an electron, accomplishing a transition corresponding to a certain line, is subject, not to the whole charge *Ze* of the nucleus, but to the charge $(Z - b)\,e$, which has been weakened by the screening action of other electrons.

39.8 Stimulated Emission of Radiation. Lasers

39.8.1 An atom located in an electromagnetic field at an excited energy level can, with a certain probability, be transferred by the effect of the field to a lower energy state. It is as if the electromagnetic field "tumbles" the atom down from its excited level to the ground state or only to a less excited state. Such a transition is accompanied by *stimulated (induced) radiation* of the substance, produced by the action of an electromagnetic wave on it.
39.8.2 From the viewpoint of wave optics (32.1.1) the phenomenon of stimulated emission of radiation means that as the electromagnetic wave passes through a substance its intensity (31.2.4) is increased, i.e. *negative absorption of light occurs*. But the frequency (30.2.5) of the wave, direction of its propagation, its phase (30.2.5) and polarization (31.1.7) remain unchanged. Stimulated emission of radiation is *strictly coherent* (30.5.1) to the electromagnetic wave passing through the substance and producing the emission.
39.8.3 From the quantum-mechanical point of view, the coherence mentioned in Sect. 39.8.2 means that the new photon, produced as a result of the stimulated emission, in no way differs from the photon it was produced by. The new photon, produced in stimulated emission, strengthens the light passing through the medium. As is shown in Fig. 39.6*b*, the process of stimulated emission leads to the production of two photons with the energy *hν* in place of one such photon. But, in addition to stimulated emission, absorption of light also occurs. As a result of the absorption of a photon by an atom at the energy level E_1, the photon disappears and

the atom moves to the energy level E_2 (Fig. 39.6a). This process reduces the intensity of the light that passes through the substance.

39.8.4 A medium is said to be active if in it the intensity of transmitted light increases. This indicates that in an active medium the process of stimulated emission predominates over the process of light absorption. Otherwise, when the chief role is played

Before interaction   *After interaction*

(a) Absorption

(b) Stimulated emission
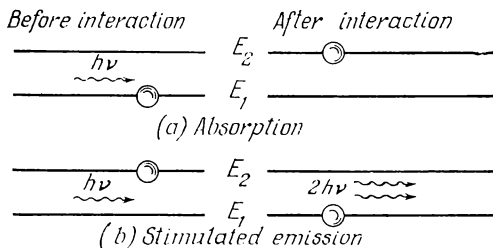
Fig. 39.6

by the absorption of light, the medium weakens the radiation passing through it. An active medium is also called one with negative absorption (39.8.1).

39.8.5 The absorption of light in matter proceeds in accordance with the *Bouguer-Lambert* law: $I = I_0 e^{-\alpha x}$, where $\alpha > 0$ is the absorption factor, x is the thickness of the absorbing layer, I_0 is the intensity (31.2.4) of light at the entrance to the layer (at $x = 0$), and I is the intensity of the light that has passed through the layer of thickness x. The *Bouguer-Lambert-Fabrikant* law is valid for a medium with negative absorption. Thus

$$I = I_0 e^{|\alpha| x},$$

where $|\alpha| > 0$ is a positive quantity corresponding to the strengthening, rather than the weakening, of light passing through an active medium. At this the intensity of the light increases drastically with the thickness of the layer of the medium (Fig. 39.7). In other words, the absorption factor α is a negative quantity for an active medium.

39.8.6 Three types of optical processes, illustrated in Fig. 39.8 can occur between two energy levels with the energies E_1 and E_2. Under conditions in which spontaneous emission can be neglected, i.e. radiation in which the excited atoms (or molecules or ions) spontaneously drop to the ground state, the optical properties of a medium are determined by competition between two processes: absorption and stimulated emission. The number of events of absorption is proportional to the concentration N_1 of particles
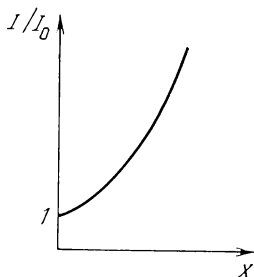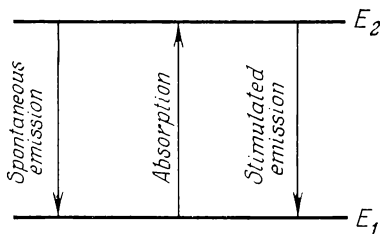
Fig. 39.7   Fig. 39.8

with the energy E_1 at the lower energy level. The number of events of stimulated emission is proportional to the concentration N_2 of particles at the upper energy level. The absorption factor α in the Bouguer-Lambert-Fabrikant law (39.8.5) is proportional to the difference between the numbers of events of absorption and stimulated emission. Thus

$$\alpha = k\,(N_1 - N_2),$$

where $k > 0$ is the proportionality factor.

39.8.7 When the system is in a state of thermodynamic equilibrium (8.3.3), $N_2 < N_1$ and $\alpha > 0$. This means that the number of events of absorption usually exceeds the number of transitions accompanied by stimulated emission of radiation. To obtain a medium with a negative absorption factor it is necessary to provide for a nonequilibrium state in the system, at which $N_2 > > N_1$. Such states are said to be *inverted*. In such a state the

number of events of stimulated emission of radiation exceeds the number of events of absorption.

39.8.8 *Lasers* (from the acronym for "*L*ight *A*mplification by the *S*timulated *E*mission of *R*adiation"), or *generators of coherent light*, are sources of light operating on the principle of stimulated emission of radiation in an active medium with an inverted population of energy levels (39.8.7). If these devices operate in the optical range, they are known as *lasers* proper. Those that operate in the range of ultrashort radiowaves are called *masers* (*M*icrowave *A*mplification by *S*timulated *E*mission of *R*adiation).

39.8.9 The process of obtaining an inverted state (population inversion) in a medium (39.8.7), required for laser operation, is
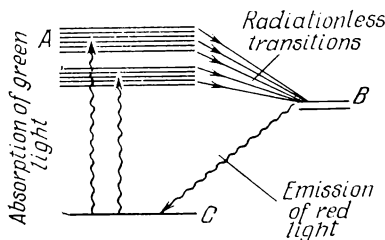
Fig. 39.9

called *pumping the active medium*. In practice, pumping is accomplished by the *three-level procedure*. One of the first lasers operating according to a three-level pumping procedure had a ruby for the active medium (alumina Al_2O_3 with the addition of chromium oxide Cr_2O_3). In the crystal lattice of alumina, a part of the Al atoms are replaced by Cr^{3+} ions, which serve as the active substance accomplishing the transitions with stimulated emission. A schematic diagram of the energy levels of Cr^{3+} is given in Fig. 39.9. The nearest to the ground-state level C are two broad energy bands A (41.8.2) and the double metastable level B. Intensive radiation, applied to the ruby by green light of a powerful pumping flashtube, filled with neon and krypton, transfers the chromium ions to the level of the A band. From here radiationless

transitions occur to the B level. The surplus energy is imparted to the ruby crystal lattice. As a result, a population inversion of the chromium ions at the B and C levels is obtained (39.8.7) and the laser emits red light of the wavelengths 6927 Å and 6943 Å, corresponding to chromium ion transitions from level B to level C.

39.8.10 The avalanche-type increase in intensity in an active medium (39.8.5) indicates that such a medium operates as an
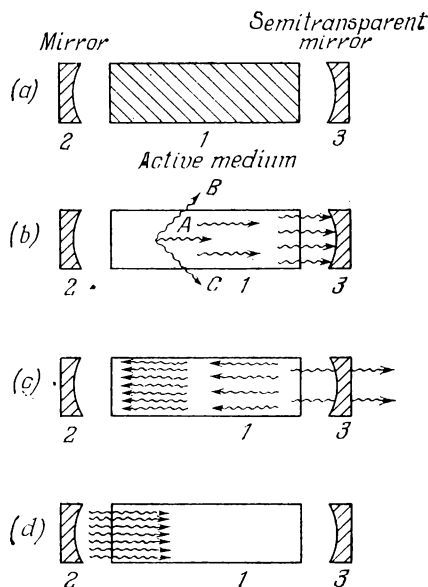
Fig. 39.10

amplifier of electromagnetic waves. The light amplification effect in a laser is increased by passing the light many times through the same layer of the active medium. This takes place by means of the arrangement shown in Fig. 39.10a. Photon A, travelling

37—0186

parallel to the axis of the active medium (a ruby rod) *1*, gives rise to an avalanche of photons moving in the same direction (Fig. 39.10*b*). A part of this avalanche passes through semitransparent mirror *3* and leaves the laser, and a part is reflected and increases again in the active medium (Fig. 39.10*c*). A part of the avalanche of photons that reaches the nontransparent mirror *2* is absorbed by it, but after reflection from this mirror the rest move in the same direction as the initial "starting" photon *A* (Fig. 39.10*d*). The multiply amplified stream of photons, leaving the laser through the semitransparent mirror *3*, produces an extremely intense pencil of light rays, highly directional and with very little angular spread.

Photons *B* and *C* (Fig. 39.10*b*), travelling at an angle to the axis of the active medium, build up an avalanche of photons which, after multiple reflection, disappears from the medium and does not participate in light amplification. The high coherency, sharp directional properties and great intensity of laser radiation are the basis for its ever-increasing application in science and engineering.

PART SEVEN
BASIC SOLID-STATE PHYSICS

CHAPTER 40 STRUCTURE AND CERTAIN PROPERTIES OF SOLIDS

40.1 Structure of Solids

40.1.1 *Solids* are bodies having constant shape and volume. Distinction is made between *crystalline* and *amorphous* solids.

Externally, crystals have a regular geometric shape. They have a periodically repeated arrangement of their constituent particles throughout the whole crystal. This arrangement is called the *crystal*, or *space*, *lattice*. In this sense, a crystal is said to have *long-range order* (cf. Sect. 13.1.2). Crystals are bounded by flat *faces*, orderly arranged with respect to one another, which converge at the *edges* and *vertices*. Large separate crystals having the shape of regular polyhedrons are called *monocrystals*. Their shape is determined by their chemical composition. *Polycrystals* have a fine crystalline structure made up of a great number of small randomly arranged crystals (*crystal grains* or *crystallites*) that have grown together.

Amorphous solids (pitch, glass, etc.) are supercooled liquids and have no clearly defined properties of crystals.

In the following, discussion is confined to crystalline bodies.

40.1.2 Each particle in a crystal lattice is subject to the forces of intermolecular interaction (12.1.4). An equilibrium arrangement of all the particles of a solid at the points of the crystal lattice (8.1.4) corresponds to minimum free energy of the crystal (11.4.5) and its most stable state. The particles at the points of the crystal lattice are located at certain equilibrium distances from one another called the *lattice constants*.

40.1.3 The main types of crystalline solids, distinguished from one another by the nature of the forces of interaction between their particles and the kinds of particles located at the crystal lattice points, are the following.

(a) *Ionic crystals* (NaCl, calcium carbonate and other salts). Located at the lattice points are regularly alternating positive and negative ions, between which are heteropolar bonds (39.4.4).

(b) *Valence (covalent) crystals* (C, Ge, Te, etc.). The neutral atoms at the lattice points have homopolar bonds (39.4.5) between them. Such crystals are found in semiconductors (41.10.1) and in many organic solids.

(c) *Molecular crystals* (Ar, CH_4, paraffin, etc.). The points of the crystal lattice are occupied by molecules that are bound by Van der Waals forces, mainly of the dispersion type (12.1.8).

(d) *Metals* (Na, Cu, Al, etc.). At the lattice points are positive ions that are formed after the outer (valence) electrons (39.3.9) have been stripped from the atoms. These electrons form an electron gas (16.4.1) of collectivized free particles.

The *metallic bond* is a specific type of chemical bond (39.4.1) that is developed between the ions in the crystal lattice and the electron gas. The electrons "constrict" the positive ions (mainly by electrostatic forces) and thereby compensate for repulsion between the ions. At distances between the ions equal to the lattice constant (40.1.2), the metallic crystal acquires a stable state.

40.1.4 The anharmonic nature of the thermal vibration of the particles at the crystal lattice points (8.1.4) is manifested in the fact that the dependence of the potential energy of interaction of the particles on their displacement from the equilibrium position is not a parabolic one. The force acting on a particle is not quasi-elastic (40.3.5). These facts are of prime importance in understanding certain thermal (40.2.4) and electrical (41.5.4) properties of solids.

40.1.5 A typical property of monocrystals (40.1.1) is their anisotropy *(crystalline anisotropy)*, which is the dependence of the physical properties of solids (thermal, elastic, electrical and optical properties) on the direction in the crystal.

40.2 Thermal Expansion of Solids

40.2.1 A solid is subject to *thermal expansion*, of either *linear* or *volume* type, when its temperature is raised. Both kinds of thermal expansion are characterized by the average *coefficients of linear* α_l and *cubic*, or *volume*, α_V expansion for the given temperature range.

40.2.2 If l_0 is the length of a body at the temperature 0 °C, its elongation Δl upon being heated to the temperature t °C is

$$\Delta l = \alpha_l l_0 t,$$

from which $\alpha_l = \Delta l / l_0 t$.

The linear expansion coefficient characterizes the elongation per unit length $\Delta l / l_0$ as the body is heated one degree. For most solids $\alpha_l \approx 10^{-6}$ to 10^{-5} K^{-1} and depends only slightly on the temperature.

40.2.3 When a body is heated from 0 °C to t °C, its volume increases from V_0 to V according to the law

$$V = V_0 (1 + \alpha_V t),$$

from which $\alpha_V = \Delta V / V_0 t$. The cubic expansion coefficient characterizes the volume increase per unit volume $\Delta V / V_0$ as the body is heated by one degree.

In a first approximation, the relation between the coefficients α_V and α_l is of the form

$$\alpha_V \approx 3\alpha_l.$$

40.2.4 The thermal expansion of solids is due to the anharmonicity of the thermal vibration of the particles in the lattice (40.1.4). If r_0 is the equilibrium distance between adjacent particles, then, at some arbitrary instant, the distance between them is $r = r_0 + x$, where x is the mutual displacement of the particles from their equilibrium positions due to thermal vibration. The forces between the particles in the lattice are not quasi-elastic, but depend on the displacement x according to the relation

$$F_x = -\varkappa x + bx^2,$$

where $\varkappa$ is the coefficient of quasi-elastic force (40.3.5), and b is the *coefficient of anharmonicity* of the vibration. The term bx^2 characterizes the deviation of the vibration from its harmonic type due to the different dependences of the forces of attraction and repulsion on the distance.

When a solid is in an equilibrium state the positions of the crystal lattice points should not vary with time and the average force acting on each particle should equal zero:

$$\langle F_x \rangle = 0.$$

If the vibration of the particles was strictly harmonic ($F_x = -\varkappa x$), the average displacement of the particles would be $\langle x \rangle = = -\langle F_x \rangle / \varkappa = 0$, i.e. there would be no thermal expansion. For real anharmonic vibration, it follows from the condition that $\langle F_x \rangle = 0$, that

$$-\varkappa \langle x \rangle + b \langle x^2 \rangle = 0 \quad \text{or} \quad \langle x \rangle = \frac{b}{\varkappa} \langle x^2 \rangle .$$

For thermal vibration with small amplitudes the potential energy E_{p} of the vibrating particles is equal (28.2.3) to:

$$E_{\mathrm{p}} = \frac{\varkappa x^2}{2} .$$

But, according to the law of equipartition of energy (10.4.6), $E_{\mathrm{p}} = kT/2$, where k is Boltzmann's constant (8.4.5) and T is the absolute temperature. Thus

$$\frac{\varkappa \langle x^2 \rangle}{2} = \frac{kT}{2} \quad \text{or} \quad \langle x^2 \rangle = \frac{kT}{\varkappa} .$$

Finally, $\langle x \rangle = bkT/\varkappa^2$. Hence, the average distance between the particles of a solid increases when it is heated, leading to thermal expansion.

40.3 Brief Information on the Elastic Properties of Solids

40.3.1 *Deformation* of a solid is the change in its dimensions and volume. Deformation is usually accompanied by a change in the shape of the body. Sometimes (in triaxial tension or compression) no change in shape occurs. Deformation may be caused by external forces acting on the body, a change in its temperature (40.2.1) and other reasons.

Deformation of a body leads to displacement of its particles from their initial equilibrium positions at the crystal lattice points (8.1.1) to new positions. This displacement is opposed by the forces of interaction between the particles. Internal elastic forces are developed inside the body to counterbalance the external forces causing the deformation.

40.3.2 Deformation is said to be *elastic* (*elastic deformation*) if it disappears as soon as the external forces causing it cease to act. At this the particles of the solid return to their initial positions of equilibrium (40.3.1). Inelastic deformation of a solid leads to irreversible rearrangement of the crystal lattice (40.1.1) and the shape of the body is not restored when the external forces are removed. Such deformation is said to be *plastic* (40.3.8). The transformation of elastic deformation into plastic deformation may occur upon prolonged action of even small external forces applied to the body. The reverse transformation cannot occur.

The *degree of elasticity* is the ratio of the work that can be done by a deformed body when the deforming forces are slowly removed to the work done by external forces in deforming the body.

40.3.3 The physical quantity equal to the elastic force $d\mathbf{F}_{el}$ per unit area dS of the cross section of the body being deformed is called the *stress* σ:

$$\sigma = \frac{d\mathbf{F}_{el}}{dS} .$$

The stress is said to be *normal* if force $d\mathbf{F}_{el}$ is normal to the surface dS and *tangential*, or *shearing*, if the force is tangential to this surface.

40.3.4 A measure of the deformation is the *unit strain*, or *relative deformation*. It is equal to the ratio of the absolute deformation Δx to the initial value of the quantity x, which characterizes the shape or dimensions of a deformed body: $\Delta x/x$.

Hooke's law states that the stress in an elastically deformed body is proportional to its unit strain. Thus

$$\sigma = K_x \frac{\Delta x}{x} ,$$

where K_x is the modulus of elasticity, equal in magnitude to the stress developed in relative deformation equal to unity. The quantity $\alpha_x = 1/K_x$ is called the *coefficient of elasticity*. Hooke's law is valid only for sufficiently small relative deformation. The stress σ_p at which the deformation is no longer proportional to the stress is called the *proportional elastic limit* (point A in Fig. 40.1).

40.3.5 In addition to elastic forces, there are forces of other origin than elastic forces, but which satisfy the equation

$$F_x = -\varkappa x,$$

where F_x is the projection of the force on the direction in which the absolute linear deformation x takes place. They are called *quasi-elastic forces*. The quantity $\varkappa$ is called the *coefficient of the quasi-elastic force*.

40.3.6 The simplest type of deformation is *longitudinal*, or *uni-axial, extension* (or *compression*), which consists in elongating (shortening) a body by the action of the tensile (compressive) force F. Elastic extension (compression) ceases when $F = F_{el}$, where F_{el} is the elastic force (40.3.1). The unit strain (relative deformation) $\Delta x/x = \Delta l/l$, where Δl is the change in length due to the action of force F, and l is the initial length of the body. According to Hooke's law

Fig. 40.1

$$\sigma = \frac{F}{S} = E\,\frac{\Delta l}{l}\,,$$
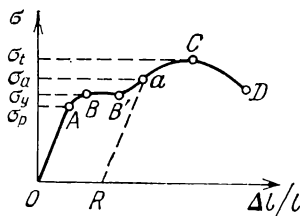
where the modulus of elasticity $K_x = E$ is called *Young's modulus*. Young's modulus is equal to the normal stress σ at which the linear dimension of the body is doubled: $\Delta l = l*$.

40.3.7 The dependence of the normal stress σ on the unit strain $\Delta l/l$ in uniaxial extension is called the *tensile stress-strain diagram* (Fig. 40.1). Beyond the proportional elastic limit (40.3.4), an increase in σ leads to a considerable increase in $\Delta l/l$. At the stress σ_y, corresponding to point B in the stress-strain diagram, the unit strain continues to increase without increasing the stress (horizontal portion BB' in the diagram). The stress σ_y is called the *yield point*. The maximum stress σ_t that the body can withstand before rupture, point C in the diagram, is called the *ultimate strength* or, if the body is in tension, the *tensile strength*. Point D corresponds to the rupture of the body.

40.3.8 If the load on a body, deformed to the stress σ_a (point a in the diagram in Fig. 40.1) is slowly removed, the curve $\sigma = \varphi\,(\Delta l/l)$ is the straight line aR, parallel to the straight portion OA of the diagram. The line segment OR determines the *permanent* et_s, typical of plastic deformation.

---

* Assuming that Hooke's law is valid for such large deformation.

40.3.9 *Shear* is the deformation of a body in which all its flat layers, parallel to a certain *shear plane*, are displaced parallel to one another without distortion or any changes in dimensions (Fig. 40.2). Shear results from the action of tangential force **F**, applied to face BC, parallel to the shear plane. Face AD, parallel to face BC, is fixed. For small shear deformation

$$\gamma \approx \tan \gamma = \frac{\overline{CC'}}{\overline{CD}} \,,$$
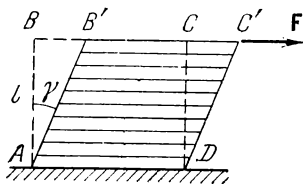
Fig. 40.2

where $\overline{CC'} = \Delta x$ is the absolute shear, γ is the *shear angle*, or *unit shear strain*, expressed in radians. According to Hooke's law 40.3.4), the unit shear strain is proportional to the tangentiall (shearing) stress $\sigma_\tau = F/S$, where S is the area of face BC, i.e.

$$\sigma_\tau = G\gamma.$$

The quantity G is called the *shear modulus*. Its magnitude is equal to the tangential, or shear, stress that produces a unit shear strain equal to unity in the specimen.

40.4 Basic Concepts of Phase Transitions in Solids

40.4.1 The heating of a solid crystalline body leads to an increase in the amplitude of the anharmonic thermal vibration at the points of the crystal lattice (40.2.4) and to an increase in the average interpoint distances in the lattice (thermal expansion, Sect. 40.2.1). Intense heating destroys the crystal lattice and leads to the transition of the substance from its solid to its liquid or vapour phase (first-order phase transition, Sect. 12.4.2).

40.4.2 The transition of a substance from the solid to the gaseous state is called *sublimation*. Transition from the solid to the liquid state is called *melting*, or *fusion*. Melting begins at a definite temperature T_m for the given pressure; it is called the *melting point*. This temperature remains constant during the melting

process. Shown in Fig. 40.3 is the dependence $T(Q)$, where Q is the quantity of heat (9.2.1) delivered to the solid being heated. The isothermal portion BC corresponds to a two-phase system: solid-liquid (cf. Sect. 12.3.2, where the two-phase system consists of a liquid and saturated vapour). In the melting process, the substance is converted from a more ordered crystalline state to a
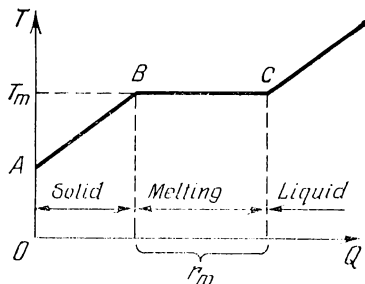
Fig. 40.3

less ordered liquid state. According to the second law of thermodynamics (11.3.2), melting is associated with an increase in the entropy (11.4.3) of the system.

40.4.3 The quantity of heat r_m required to melt unit mass of a solid at the temperature T_m is called the *heat of fusion*, or *of melting*. It is also called the *latent heat*. It follows from the first law of thermodynamics (9.3.1) that

$$r_m = u_l - u_s + p(v_l - v_s),$$

where u_l and u_s are the specific internal energies (9.1.2) of the substance (internal energies per unit volume) in the liquid and solid phases, v_l and v_s are the specific volumes of the liquid and solid phases, and p is the constant pressure at which the first-order transition proceeds (12.4.2).

40.4.4 The melting point T_m depends upon the pressure p. A change of dp in the pressure leads to a change in the melting

point by dT_m. The dependence $T_m = T_m(p)$ is expressed by the Clapeyron-Clausius equation (see also Sect. 13.6.5):

$$\frac{dT_m}{dp} = \frac{T_m(v_l - v_s)}{r_m}.$$

The notation is defined in Sect. 40.4.3. The Clapeyron-Clausius equation is valid for any first-order phase transition.

In the overwhelming majority of cases the melting of a substance leads to an increase in its specific volume, i.e. $v_l > v_s$ and $dT_m/dp > 0$ ($r_m > 0$), i.e. the melting point increases with the pressure. Such substances can be in the solid state at high pressures in the temperature region above the critical temperature (12.3.1).

The specific volume of certain substances (water, bismuth, gallium) increases in melting, so that $v_l < v_s$. Ice, for instance, at 0 °C has a lower density than water. For such substances $dT_m/dp < 0$, i.e. the melting point decreases as the pressure is increased.

40.4.5 In a pT diagram, the dependence of the temperature on the pressure $p = f(T)$ in a first-order phase transition is represented by a curve on which each point corresponds to equilibrium of two existing phases. The vapour generation curve, for example, shows the equilibrium of the two-phase system: liquid-vapour. Since $v_v - v_l > 0$ (where v_v is the specific volume of the vapour),

Fig. 40.4

it follows from the Clapeyron-Clausius equation (40.4.4), that $dp/dT > 0$. The curve $p = f(T)$, separating the regions of the liquid and gaseous states of a substance, ends at critical point C (12.3.1) (Fig. 40.4). Shown in Figs. 40.5 and 40.6 are the equilibrium curves of the two-phase solid-liquid states for the cases when $v_l > v_s$ and $v_l < v_s$, respectively.

40.4.6 In cooling liquids to a certain temperature, called the *freezing point* T_f (or *temperature of solidification*), the substance is solidified (crystallized, or frozen). This is the transition from
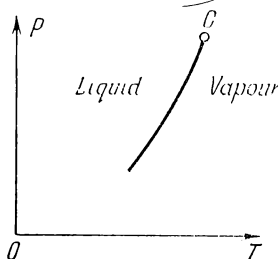
a liquid to a solid crystalline state. In this process a quantity of heat is evolved that is equal to the heat of melting (40.4.3). The freezing point is equal to the melting point and depends upon the presence of impurities in the substance. Impurities lower T_f if they do not form mixed crystals with the substance. These are called *solid solutions*. Sea water, for instance, containing dissolved
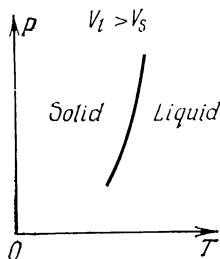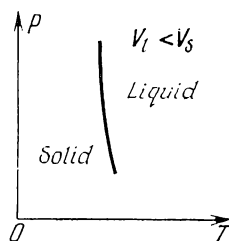
Fig. 40.5   Fig. 40.6

salts, freezes at a lower temperature than distilled water. This serves as the basis for producing cooling mixtures.

40.4.7 The freezing (crystallization) of liquids is associated with a change in the nature of the thermal motion of their particles. The time the particles remain in a temporary settled state (relaxation time, see Sect. 13.2.4) is increased. The forces of mutual attraction between the particles change the thermal motion into random thermal vibration about the points of the crystal lattice. The transition of the substance to a more ordered phase is associated with a reduction in the entropy of the system (11.4.3).

CHAPTER 41   AN OUTLINE OF THE QUANTUM PHYSICS OF SOLIDS

41.1 Basic Concepts of Quantum Statistics

41.1.1 *Quantum statistics* is a statistical method of investigation (8.2.2) that is applied to systems consisting of a large number of particles and obeying the laws of quantum mechanics (38.1.1).

41.1.2 In contrast to the initial propositions of classical statistical physics (10.1.4 and 10.1.5), quantum statistics is based on the *principle of indistinguishability of identical particles*. All identical particles (for example, all the electrons in metals or all the protons in atomic nuclei) are considered to be indistinguishable, in principle, from one another.

41.1.3 Of prime importance in quantum statistics is the problem of distributing the particles with respect to their coordinates and velocities*. It can be stated as the problem of distributing particles amongst the cells of "phase volume", an element of which is equal to $\Delta\Gamma = \Delta x \Delta y \Delta z \Delta p_x \Delta p_y \Delta p_z$ (10.1.5). The wave-particle dualism of the properties of particles (38.1.3) and the Heisenberg indeterminacy relations (38.6.2) lead to a situation in which the element of volume $\Delta\Gamma$ cannot be less than h^3, where h is Planck's constant (Appendix II). In arranging the particles in cells of volume h^3 each, in accordance with the principle of indistinguishability of identical particles (41.1.2), the question posed concerns the number of particles in the given cell, but not which, specifically, of the given sort of particles are in the cell. The state of a system of particles is not changed when identical particles are rearranged within the given cell of the phase space, or interchanged between the cells.

41.1.4 If an element of volume $\Delta\Gamma_i$ is divided into cells with the volume h^3, then the ratio $\Delta\Gamma_i/h^3 = \Delta g_i$ determines the *number of quantum states* contained in the volume $\Delta\Gamma_i$, having energy within the range from E_i to $E_i + \Delta E_i$. Particles, of a number equal to ΔN_i, contained in the volume $\Delta\Gamma_i$, can be distributed in all possible ways between Δg_i states having the energy E_i.

41.2 Bose-Einstein and Fermi-Dirac Distribution Functions

41.2.1 Particles with integer or zero spin (in $\hbar$ units) (26.1.3) are called *bosons* (for instance, photons, phonons and certain nuclei). Systems of such particles are described by *Bose-Einstein quantum statistics*. Bosons do not comply with the Pauli exclusion principle (39.3.1) and they are not subject to restrictions as to the number of particles that may be in a certain quantum state.

---

* In classical statistics, the laws of Maxwell (10.3.1) and Boltzmann (10.4.2) enable a similar problem to be solved.

41.2.2 The *Bose-Einstein distribution function* f_B is the average population of bosons in a state with the given energy, i.e. the average number of particles in one state. Thus

$$f_B = \frac{\Delta N\,(E_i)}{\Delta g_i}\,,$$

where $\Delta N\,(E_i)$ is the number of particles having energies in the range from E_i to $E_i + \Delta E_i$, and Δg_i is the number of quantum states within this energy range (41.1.4). Function f_B is found by considering the thermodynamic probability P (11.5.2) of the distribution of the particles of the system amongst the quantum states. The most probable distribution is found under the condition that the total number N of particles and the total energy E of the system are conserved. Thus

$$\sum_i \Delta N_i = N \quad \text{and} \quad \sum_i \Delta N_i E_i = E.$$

Summation is carried out over all the quantum states of the system.

41.2.3 An application of the method of Lagrangian (undetermined) multipliers to find the conditional extremum enables the following expression to be obtained for the Bose-Einstein distribution function:

$$f_B = \frac{1}{\exp\left(\dfrac{E_i - \mu}{kT}\right) - 1}\,.$$

Here k is Boltzmann's constant (8.4.5), and T is the absolute temperature. The quantity $\mu = (U - TS + pV)/N$ is called the *chemical potential* per particle, where U is the internal energy of the system (9.1.2), S is its entropy (11.4.2), V is the volume of the system, and p is the pressure. The chemical potential is the work done under isobaric-isothermic conditions in increasing the number of particles in the system by one.

41.2.4 Particles with half-integer spin (in $\hbar$ units) (26.1.3) are called *fermions* (electrons, protons, neutrons, etc.). Systems of fermions are described by Fermi-Dirac quantum statistics. Fermions comply with the Pauli exclusion principle (39.3.1) and in a given quantum state of a system of fermions there can be no more than one particle.

41.2.5 The *Fermi-Dirac distribution function* f_F is determined
in a way similar to that used in Sect. 41.2.2. Thus

$$f_F = \frac{\Delta N\,(E_i)}{\Delta g_i} \,.$$

The notation is defined in Sect. 41.2.2. The solution of the
problem of the most probable distribution of the fermions
amongst the states, under the condition of conserving total
energy E in the system and the total number N of its particles,
leads to the following form of the function f_F:

$$f_F = \frac{1}{\exp\left(\dfrac{E_i - \mu}{kT}\right) + 1} \,,$$

where μ is the chemical potential per particle (41.2.3).
41.2.6 The distribution functions in classical and quantum
statistics, introduced as the average number of particles in one
state (41.2.2), can . be ex-
pressed by a single equation:

$$f = \frac{1}{\exp\left(\dfrac{E_i - \mu}{kT}\right) + \delta} \,.$$

For the Maxwell-Boltzmann
distribution (cf. Sect. 10.4.2)
$\delta = 0$ and $\mu = 0$; for the
Bose-Einstein distribution
$\delta = -1$; and for the Fermi-
Dirac distribution $\delta = +1$.
Plots of these three distri-
bution functions are shown
in Fig. 41.1.

Fig. 41.1

Example. Blackbody radiation (36.1.8) is enclosed in a cavity
of volume V at $T = \text{const}$, the radiation being in a state of
thermodynamic equilibrium with the walls of the cavity. This
radiation can be dealt with as a photon gas complying with
Bose-Einstein statistics, because the spin of a photon equals $\hbar$
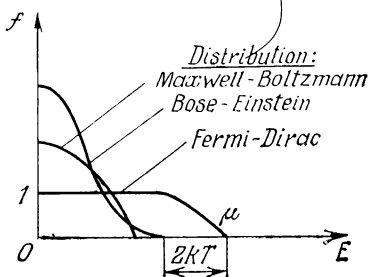
(Table 43.1). The number of quantum states of radiation in volume V (40.1.4) is

$$dg = 2 \, \frac{4\pi p^2 \, dp V}{h^3} = \frac{8\pi E^2 \, dE}{h^3 c} \, V,$$

where p is the momentum of the photon (37.2.2), related to its energy by the equation $p = E/c$, where c is the velocity of light in free space. The factor 2 appears because there are two possible directions in which light can be polarized (31.1.8). The energy of the photon is $E = h\nu$ (37.1.4), where ν is the frequency. According to Sect. 41.2.2, the number of photons with frequencies in the range from ν to $\nu + d\nu$ in the volume V is

$$dN = \frac{dg}{\exp \, (h\nu/kT) - 1} = \frac{8\pi\nu^2 \, d\nu V}{c^3 \, [\exp \, (h\nu/kT) - 1]} \, .$$

Taken into account here is the fact that the chemical potential μ (41.2.3) is equal to zero for a photon gas in which the condition for conserving the total number of particles is not complied with.

The volume density of radiant energy in the frequency range from ν to $\nu + d\nu$ is

$$\omega \, (\nu, \; T) = \frac{h\nu \, dN}{V} = \frac{8\pi\nu^2}{c^3} \, \frac{h\nu \, d\nu}{\exp \, (h\nu/kT) - 1} \, .$$

This result is Planck's radiation formula for the volume energy density of blackbody radiation (36.1.5).

41.3 Degeneracy of Systems of Particles Described by Quantum Statistics

41.3.1 A system of particles (in particular, an ideal gas) is said to be *degenerate* if its properties, described by quantum laws, differ from those of ordinary systems that obey classical laws. Deviation in the behaviour of Bose and Fermi gases from that of classical Maxwellian-Boltzmann gas is called *gas degeneracy*. This becomes appreciable at extremely low temperatures and high densities (41.3.3).

41.3.2 The *degeneracy parameter* A is the quantity

$$A = e^{\mu/kT},$$

where μ is the chemical potential (41.2.3). Under the condition of slight degeneracy ($A \ll 1$), the ones in the denominators of the quantum distribution functions f_B and f_F can be neglected, and these functions become the classical Maxwellian-Boltzmann distribution function f_{M-B}. Thus

$$f_{M-B} = A e^{-E_i/kT}.$$

The degeneracy parameter is found from the condition of normalization of the distribution function: $\sum_i \Delta N_i = N$ (41.2.2),

which indicates the conservation of the total number of particles in the system. The condition for slight degeneracy is of the form

$$A = \frac{n_0 h^3}{(2\pi m kT)^{3/2}} \ll 1,$$

where n_0 is the concentration of particles, m is the mass of the particles, k is Boltzmann's constant, h is Planck's constant, and T is the absolute temperature.

41.3.3 The *degeneracy temperature* T_d is the temperature at which degeneracy becomes appreciable (41.3.1). It is determined from the condition

$$A = \frac{n_0 h^3}{(2\pi m kT_d)^{3/2}} = 1, \quad \text{from which} \quad T_d = \frac{h^2 n_0^{2/3}}{2\pi m k}.$$

The temperature criterion of degeneracy is as follows: when $T \leqslant T_d$ the system of particles is degenerate; when $T \gg T_d$ the system of particles is not degenerate, and its behaviour is described by classical laws.

For example, the degeneracy parameter $A \approx 3 \times 10^{-5} \ll 1$ for hydrogen under standard conditions ($T = 300$ K and $n_0 \approx \approx 3 \times 10^{25}$ m^{-3}). The degeneracy temperature for hydrogen is $T_d \approx 1$ K. For all other gases, heavier than hydrogen, A is even smaller. Under standard conditions gases are not degenerate. Degeneracy, associated with the quantum properties of gases, is manifested to a considerably less degree than the devia-

tions of gases from an ideal gas, which is due to intermolecular interaction. Photon gas is always degenerate ($T_d = \infty$) because the rest mass of a photon equals zero ($m_0 = 0$, Table 43.1). Photon gas is described by Bose-Einstein quantum statistics (41.2.6).

41.3.4 Electrons in metals are an example of a degenerate gas. Under ordinary conditions $n_0 \approx 10^{28}$ to 10^{29} m^{-3}. Since the mass of the electron is small ($m \approx 10^{-30}$ kg), $T_d \approx (16$ to $20) \times$ $\times 10^3$ K. Electron gas does not comply with the classical Maxwell-Boltzmann statistics.

41.3.5 The energy of degenerate electron gas (41.4.5), and other gases described by quantum statistics, is not a linear function of the temperature. Therefore, the simple physical interpretation of the absolute temperature (8.3.4) is not suitable in the region of degenerate gases (41.3.3).

41.4 Degenerate Electron Fermi Gas in Metals

41.4.1 The energy distribution of conduction electrons in metals is described by the Fermi-Dirac distribution function (41.2.5):

$$f_F = \frac{1}{e^{(E-\mu)/kT} + 1}.$$

The number of quantum states of electrons per unit volume of metal, within the energy range from E to $E + dE$ (41.1.4) is

$$dg(E) = \frac{d\Gamma}{h^3} = \frac{4\pi p^2\, dp}{h^3} = \frac{2\pi\,(2m)^{3/2}\, E^{1/2}\, dE}{h^3}.$$

Use is made here of the relation between the momentum p of the electron and its energy:

$$p^2 = 2mE, \quad dp = \left(\frac{m}{2E}\right)^{1/2} dE.$$

The number of electrons $dn_0(E)$ in unit volume of the metal, having energies within the range from E to $E + dE$, equals, according to Sect. 41.2.2:

$$dn_0(E) = 2f_F\, dg(E) = \frac{4\pi\,(2m)^{3/2}}{h^3}\, \frac{E^{1/2}\, dE}{\exp\,[(E-\mu)/kT] + 1}.$$

Factor 2 takes into account the fact that the electrons comply with the Pauli exclusion principle (39.3.1). This equation expresses the energy distribution law for conduction electrons in a metal.

41.4.2 The unusual properties of electron gas have an effect on its behaviour at the absolute zero temperature ($T = 0$ K). If μ_0 is the chemical potential of the gas at $T = 0$ K (41.2.3), then

$$f_F = \frac{1}{e^{(E-\mu_0)/kT} + 1} \, ;$$

at $E < \mu_0$ the difference $E - \mu_0 < 0$ and

$$f_F = \frac{1}{\dfrac{1}{e^{(\mu_0 - E)/kT}} + 1} \, ;$$

at $E - \mu_0 < 0$ and $T \to 0$, $\exp\left[(\mu_0 - E)/kT\right] \to \infty$ and $f_F \to 1$; at $E > \mu_0$, the difference $E - \mu_0 > 0$ and at $T \to 0$, $\exp\left[(E - \mu_0)/kT\right] \to \infty$ and $f_F \to 0$.

The curve of the Fermi distribution function f_F at $T = 0$ K is given in Fig. 41.2. In the energy range from 0 to μ_0, function f_F is equal to unity. In passing through the value $E = \mu_0$, the function drops abruptly to zero. At $T = 0$ K, the electrons of the metal occupy all allowed energy levels with a population equal to unity (41.2.2) up to the level with the energy $E = \mu_0$. All levels with energy exceeding μ_0 are empty; their population equals zero. Thus, μ_0 is the maximum energy that conduction
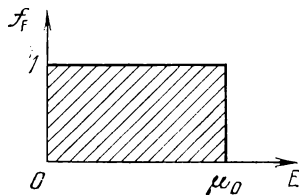
Fig. 41.2

electrons can have at absolute zero temperature. This quantity is called the *Fermi energy*: $\mu_0 = E_F$. If the consecutive numbers of the energy levels are plotted along the axis of abscissas in Fig. 41.2, and the corresponding numbers of conduction electrons per unit volume are the ordinates, the maximum ordinate corresponds to the energy level number $n_0/2$, where n_0 is the electron concentration. The highest level occupied by electrons

38*

is called the *Fermi level*. It corresponds to the Fermi energy E_F, which an electron has at this level.

41.4.3 The energy distribution law for conduction electrons at $T = 0$ K is of the form

$$dn_0\,(E) = \frac{4\pi\,(2m)^{3/2} E^{1/2}\,dE}{h^3}\,.$$

It is represented graphically by the curve in Fig. 41.3.

The total number of conduction electrons in unit volume of a metal is

$$n_0 = \int\limits_0^{E_\mathrm{F}} dn_0\,(E) = \frac{4\pi\,(2m)^{3/2}}{h^3} \int\limits_0^{E_\mathrm{F}} E^{1/2}\,dE = \frac{4\pi\,(2m)^{3/2}}{h^3}\,\frac{2}{3}\,E_F^{3/2}.$$

Hence

$$E_\mathrm{F} = \frac{h^2}{2m} \left(\frac{3n_0}{8\pi} \right)^{2/3}.$$

The average energy $\langle E \rangle$ of the electron at $T = 0$ K is

$$\langle E \rangle = \frac{3}{5}\,E_\mathrm{F} = \frac{3}{5}\,\frac{h^2}{2m} \left(\frac{3n_0}{8\pi} \right)^{2/3}.$$

At $n_0 = 6 \times 10^{28}$ m^{-3} we obtain, after substituting the values of h and m:

$$\langle E \rangle = 9 \times 10^{-19}\ \mathrm{J} = 5.4\ \mathrm{eV}.$$

If we compare this energy with the average energy of a molecule of nondegenerate monatomic gas $\langle E \rangle = 3kT/2$ (10.2.5), we find the energy E_F that the molecule could have at the temperature $T \approx 10^4$ K. In other words, under ordinary conditions, $kT/E_\mathrm{F} \ll 1$.

41.4.4 At a temperature different from zero the Fermi-Dirac distribution function f_F is of the form shown in Fig. 41.4 by the full-line curve. The same figure shows the function f_F at $T = 0$ K in dash lines. The curves differ in the kind of drop the function undergoes near the value $E = \mu$. An abrupt drop

along a vertical line at $E = \mu_0 = E_\mathrm{F}$ for the case when $T = = 0$ K (Fig. 41.2) changes to the smooth curve ABC at $T \neq \neq 0$ K. Of importance is the fact that the states of the electrons, at the energy levels that satisfy the condition $E \ll \mu$, do not change when heated from 0 K to T. Distortion of the function
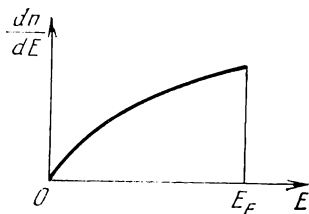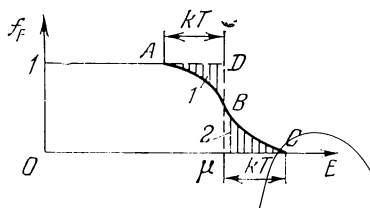
Fig. 41.3   Fig. 41.4

f_F occurs only at its "tail-end", in an energy range of the width $2kT$, near the value $E = \mu$, at which $f_\mathrm{F} = 1/2$.

41.4.5 The chemical potential μ (41.2.3) of an electron gas at the temperature T is

$$\mu = E_\mathrm{F} \left[1 - \frac{\pi^2}{12} \left(\frac{kT}{E_\mathrm{F}} \right)^2 \right].$$

The internal energy U (9.1.2) of a mole of electron gas is

$$U = \frac{3}{5} N_\mathrm{A} E_\mathrm{F} \left[1 + \frac{5}{12} \frac{\pi^2 (kT)^2}{E_\mathrm{F}^2} \right].$$

Here N_A is Avogadro's number (Appendix II), and k is Boltzmann's constant (8.4.5). Quantities μ and U remain practically constant when the temperature is raised.

The molar heat capacity $C_{V\mu}$ (9.5.2) of an electron gas is

$$C_{V\mu} = \frac{\partial U}{\partial T} = \frac{\pi^2}{2} N_\mathrm{A} k \frac{kT}{E_\mathrm{F}}.$$

A comparison with the heat capacity of a nondegenerate monatomic gas (10.7.1) $C_{V\mu}^{\text{class}} = 3N_A k/2$, indicates that

$$\frac{C_{V\mu}}{C_{V\mu}^{\text{class}}} = \frac{\pi^3}{3}\frac{kT}{E_F} .$$

At room temperature $kT/E_F \approx 0.01$ (41.4.3), and therefore $C_{V\mu}/C_{V\mu}^{\text{class}} \approx 0.03$. The heat capacity of degenerate electron gas is negligibly small. This is so because only a slight number of electrons, within the "drop region" of the Fermi-Dirac distribution function f_F (hatched regions 1 and 2 in Fig. 41.4), participate in the process of changing the internal energy of the electron gas when it is heated. This removes one of the chief shortcomings of the classical electron theory of the electrical conduction of metals (20.3.7).

41.5 Quantum Theory of Electrical Conduction in Metals

41.5.1 The theory of the electrical conduction of metals, based on quantum mechanics (38.1.1) and Fermi-Dirac quantum statistics (41.2.5) is called the *quantum theory of electrical conduction in metals*. Using the Fermi-Dirac distribution function (41.2.5) in this theory, Ohm's law for current density (20.3.4) was derived. Thus

$$\mathbf{j} = \varkappa \mathbf{E}.$$

41.5.2 In the quantum theory of electrical conduction, the electrical conductivity $\varkappa$ is calculated by the formula

$$\varkappa = \frac{n_0 e^2 \langle \lambda_F \rangle}{m u_F} ,$$

where n_0 is the number of conduction electrons per unit volume of the metal, $\langle \lambda_F \rangle$ is the mean free path of the electron (20.3.1) having Fermi energy (41.4.2), and u_F is the velocity of thermal motion of such an electron. Notwithstanding the external resemblance between this formula and the formula for $\varkappa$ in the classical electron theory (20.3.4), it has an entirely different

physical content and, in contrast to the classical formula, it completely agrees with the experimental data.

41.5.3 Also explained in the quantum theory of electrical conduction of metals is the dependence of electrical conductivity on temperature ($\varkappa \propto 1/T$), as well as the anomalously large mean free path of the electron in a metal (20.3.7). The ordered motions of electrons in a metal—electric current—is dealt with in quantum theory as the process of propagation of electron de Broglie waves (38.1.4) scattered by anharmonic thermal vibrations of the lattice points in the metal (40.2.4). An analogous phenomenon occurs in optics in the propagation of a light wave through a turbid medium, containing scattering centres (suspended particles in a liquid, colloidal solutions, etc.) (34.3.1). When the distance between the scattering centres are of an order comparable to the wavelength, scattering of the light occurs and the intensity of the light decreases as it propagates in the medium. At distances between the scattering centres substantially less than the wavelength, the medium is optically homogeneous and no light scattering occurs.

41.5.4 An ideal crystal lattice, free of all violations of periodicity and with stationary particles at the lattice points, behaves like an optically homogeneous medium; it does not scatter electron waves and conduction electrons pass through such a lattice without resistance. The scattering of electron waves occurs only when the periodicity of the lattice is violated. These are inhomogeneities that act as scattering centres. Such centres are, for example, density fluctuations in the lattice (11.6.3), resulting from anharmonic thermal vibrations of the positive ions of the metal. With the exception of superconductors (41.6.1), this scattering is what leads to the existence of electrical resistance in pure metals.

41.5.5 The scattering of the electron waves by the thermal vibrations of the lattice points increases with the temperature, whereas the mean free path of the electrons is reduced. The quantity $\langle \lambda_F \rangle$ is calculated by the formula

$$\langle \lambda_F \rangle = \frac{Ed}{\pi n_0 kT} \,,$$

where E is Young's modulus (40.3.6), d is the crystal lattice constant, n_0 is the number of atoms in unit volume, k is Boltzmann's constant (8.4.5), and T is the absolute temperature.

Here the electrical conductivity is found to be inversely proportional to the absolute temperature and independent of n_0 for monovalent metals*. Thus

$$\varkappa = \frac{e^2 E d}{m u_F \pi k T} \cdot$$

This formula agrees with experimental data in the region of room temperatures. The conductivity, for instance, obtained for silver $(E = 10^7$ N/m^2, $d = 3 \times 10^{-10}$ m, $m = 9 \times 10^{-31}$ kg, $kT \approx 4.2 \times 10^{-21}$ J, and $u_F = 1.4$ m/s) is $\varkappa_{theor} \approx 5 \times 10^7$ 1/Ω-m, whereas $\varkappa_{exper} \approx 6.3 \times 10^7$ 1/Ω-m.
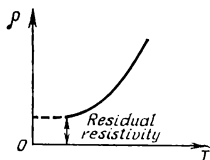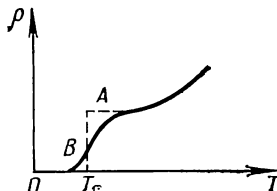
Fig. 41.5   Fig. 41.6

At very low temperatures the formulas given above are not valid. The mean free path under these conditions is inversely proportional to T^5, which agrees with experimental data.

41.5.6 Besides density fluctuations, the cause of electron wave scattering and the electrical resistance of metals is violation in the periodicity of the crystal lattice due to the inclusions of impurity atoms in the lattice. The resistivity ρ of a metal $(\rho = 1/\varkappa)$ consists of two parts:

$$\rho = \rho_T + \rho_{im},$$

where ρ_T is the resistivity due to the scattering of electron waves by density fluctuations, and ρ_{im} is the resistivity due to scattering by impurities. Of importance is the fact that ρ_{im} is independent of the temperature. Since as $T \to 0$, $\rho_T \to 0$ as

---

* The concentration of conduction electrons and ions in the lattice coincide in such metals.

well, $\rho \to \rho_{lm}$ as the temperature is lowered. The quantity ρ_{lm} is called the *residual resistivity*. This resistivity remains a property of the metal when it is cooled even down to absolute zero. The dependence of the electrical resistivity on the temperature is shown in Fig. 41.5. The point of intersection of the extension of the curve (dash line) with the axis of ordinates (to $T = 0$ K) is the residual resistivity.

41.6 Superconductivity

41.6.1 The *phenomenon of superconductivity* consists in the practically complete disappearance of electrical resistivity in certain metals and alloys at a certain temperature T_c, called the *transition temperature to a superconductive state*. Substances having this property are called *superconductors*. Today over 500 pure elements and alloys are known that display superconductivity. The temperature range AB (Fig. 41.6) of the transition to a superconductive state does not exceed thousandths of a degree for pure specimens. It therefore proves expedient to assign a definite value to T_c. The width of the range AB depends on nonhomogeneities in the metal, primarily on the presence of impurities and internal stresses in the specimen. The temperatures T_c known at the present time range from 0.155 K (BiPt) to 23.2 K (Nb$_3$Ge).

The *isotope effect* in superconductors is the fact that the transition temperatures T_c of various isotopes (42.1.3) of the same superconducting metal are inversely proportional to the square roots of their atomic masses.

41.6.2 A sufficiently strong magnetic field at a given temperature destroys the superconductive state of a substance. When a magnetic field is applied to a superconductor, the transition temperature T_c (41.6.1) is lowered.

A magnetic field of the strength H_c which, at the given temperature, causes the transition of a substance from the superconductive to the normal state is called the *critical field*. As the temperature of the superconductor is lowered, the field strength H_c increases. As a first approximation: $H_c = H_0 [1 - (T/T_c)^2]$. The curve in Fig. 41.7 separates the superconductive and normal states of a substance.

The superconductive properties of a conductor disappear when a sufficiently heavy current is passed through them. This is

due to the action of the magnetic field, set up by the current, which destroys the superconductive state.

41.6.3 An external magnetic field, weaker than the critical field (41.6.2) does not penetrate into the interior of a super-conductor. Hence, the magnetic induction B (23.1.2) is always zero in the bulk of the superconductor. Shown in Fig. 41.8 is
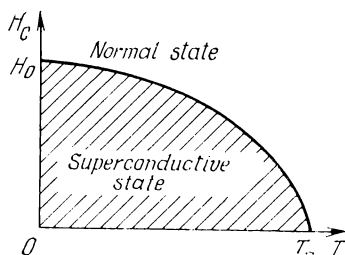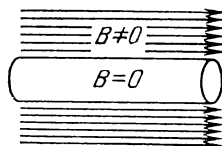
Fig. 41.7   Fig. 41.8

a uniform magnetic field (23.1.3) directed along the axis of a cylindrical superconductor. The superconductor seems to "push" the magnetic field out of the space it itself occupies and is an ideal diamagnetic material with a magnetic susceptibility of $\chi_m = -1$ (26.4.5). Here the relative magnetic permeability $\mu_r = 1 + \chi_m = 0$ and $B = \mu_0\mu_r H = 0$.

41.6.4 The transition of a substance to the superconductive state is accompanied by changes in its thermal properties. In the absence of a magnetic field and at the transition temper-ature T_c (41.6.1), there is an abrupt change in the heat capacity (9.5.1). In the presence of a magnetic field an isothermal transi-tion from the superconductive to the normal state is associated with a jump in the heat conductivity and heat capacity. These phenomena are typical features of a second-order phase transi-tion (12.4.2).

41.6.5 In quantum-mechanical theory, the phenomenon of superconductivity is considered to be the superfluidity (12.4.3) of electrons in a metal with the absence of friction inherent in superfluidity. The conduction electrons move along a super-conductor unimpeded, i.e. without any "friction" between them

and the crystal lattice points. The main feature of superconduc-
tors is that the mutual attraction of the electrons in them leads
to the pairing of electrons. The cause of this attraction is inter-
action between electrons, supplementary to Coulomb repulsion,
accomplished by the action of the crystal lattice and leading
to the pairing. A model of this attraction is illustrated in
Fig. 41.9. Conduction electron
e_1 attracts ion I of the crystal
lattice to itself, displacing the
ion from its equilibrium posi-
tion. This changes the electric
field in the crystal; ion I sets
up an electric field that acts on
the conduction electrons, includ-
ing electron e_2. Thus the inter-
action between electrons e_1 and
e_2 is accomplished by means of
the crystal lattice. The displace-
ment of the ion by an electron
leads to a situation in which

Fig. 41.9

the electron is surrounded by a "cloud" of positive charge, which
exceeds the intrinsic negative charge of the electron. Together
with this "cloud" the electron has a net positive charge and is
thereby attracted to another electron.
41.6.6 In the quantum theory of metals attraction between
electrons (phonon exchange) is associated with the setting up
of elementary excitations (quasi-particles) of the crystal lattice.
An electron, moving in a crystal and interacting with another
electron by means of the lattice, brings the lattice into an excit-
ed state. As the lattice returns to the ground state it emits a
quantum of energy of acoustic frequency, called a *phonon*
(41.7.5), which is absorbed by the other electron. The attraction
between electrons can be conceived of as their exchange of
phonons. This attraction is found to be most effective if the
momenta of the interacting electrons are anti-parallel.
41.6.7 The initiation of a superconductive state in a substance
is associated with the possibility of forming bound pairs of
electrons in the metal. The distance δ between the electrons of
a pair equals
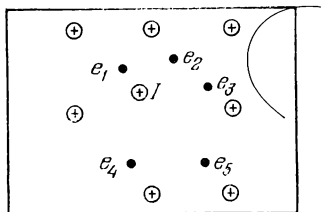
$$\delta = \frac{h u_F}{k T_c} \, ,$$

where h is Planck's constant, u_F is the velocity of an electron at the Fermi level (41.5.2), k is Boltzmann's constant (8.4.5), and T_c is the transition temperature to a superconductive state (41.6.1). An estimation indicates that $\delta \approx 10^{-4}$ cm, i.e. the electrons forming a pair are at a distance from each other of the order of 10^4 crystal lattice constants. The whole electronic system of a superconductor is a bound aggregate extending over huge distances on the atomic scale.

If at arbitrarily low temperatures Coulomb repulsion between electrons predominates over the attraction that forms pairs, the substance (metal or alloy) remains normal with respect to its electrical properties. But if the forces of attraction predominate over those of repulsion at the temperature T_c (41.6.1), the substance goes over to the superconductive state.

41.6.8 The most important feature associated with the pair in the aggregate of electrons in a superconductor is the impossibility of energy exchange between the electrons and the lattice in small portions, smaller than the binding energy of paired electrons. This means that in collisions between the electrons and the crystal lattice points, the energy of the electrons does not change, and the substance behaves as a superconductor with zero resistivity. A quantum-mechanical analysis shows that there is no scattering of electron waves by the thermal vibrations of the lattice or impurities. This is what indicates the absence of electrical resistance.

41.6.9 A definite amount of energy is required to destroy the state of superconductivity. At $T \gg T_c$ (41.6.1), the bound states of the paired electrons are violated, attraction between the electrons ceases and superconductivity no longer exists.

41.6.10 Superconductivity is being used to produce strong magnetic fields because there are no thermal losses when heavy currents, required to set up strong magnetic fields, are passed through superconductors. But, since a magnetic field destroys the state of superconductivity (41.6.2), special *type-2 superconductors* have been developed for obtaining strong magnetic fields. These are certain alloys and superconducting thin films. In such superconductors magnetic fields with a field strength greater than H_c (41.6.2) penetrate the substance in the form of threads piercing the specimen. The substance between the threads is found to be superconductive, and heavy currents can set up ultrastrong magnetic fields. Magnets, based on the application of superconducting solenoids, are being extensively used.

41.7 Heat Capacity of Solids

41.7.1 No distinction is made for solids between the heat capaci-
ties C_V and C_p (9.5.4 and 9.5.7). In nonmetallic solids, the larg-
est contribution to heat capacity is made by the energy of
thermal vibration of the particles at the crystal lattice points.
For metals, a slight contribution is made by the degenerate
electron gas (41.4.5).

41.7.2 The classical theory of the heat capacity of solids is
based on the principle of equipartition of energy (10.6.4).
A homogeneous solid is dealt with as a system of particles,
independent of one another, having three degrees of freedom and
executing thermal vibrations with the same frequency. The
average energy $\langle E \rangle$ per degree of freedom is $\langle E \rangle = kT$ (10.6.5).
The internal energy U (9.1.2) per mole of a solid is

$$U = 3N_A \langle E \rangle = 3N_A kT = 3RT,$$

where N_A is Avogadro's number (Appendix II), k is Boltzmann's
constant (8.4.5), and $R = kN_A$ is the universal gas constant
(8.4.4). The molar heat capacity* (9.5.4) of a solid with an atom-
ic crystal lattice is

$$C_\mu = \frac{\partial U}{\partial T} = 3R = 25 \frac{J}{\text{mol-K}} = 5.97 \frac{\text{cal}}{\text{mol-K}}$$

The *Dulong-Petit law* states that the molar heat capacity of all
chemically simple crystalline solids equals approximately
6 cal/mol-K. According to this law tha molar heat capacity
does not depend on either the temperature or any characteristics
of the crystals. Experiments disprove this statement and indicate
that the heat capacity depends on the temperature, especially
in the region of low temperatures (Fig. 41.10). The reason why
the classical theory of the heat capacity of solids disagrees with
experimental data is in the limited nature of the principle of
equipartition of energy and its inadequacy in the region of low
temperatures. At these temperatures the average energy of the

---

* Frequently, the heat capacity is referred to one gram-atom of
the substance.

vibrating particles must be calculated in accordance with
the laws of quantum mechanics (38.5.9).
41.7.3 In the initial quantum theory of heat capacity of solids,
a crystal was treated as a system of N atoms, each of which is
a quantum harmonic oscillator. All the atoms vibrate indepen-
dently of one another with the same frequency ω. The average
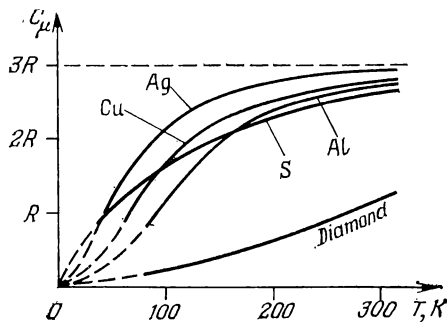
Fig. 41.10

energy $\langle E \rangle$ per degree of freedom of the atom (as a quantum
harmonic oscillator) is equal (38.5.9) to

$$\langle E \rangle = \frac{\hbar\omega}{\exp(\hbar\omega/kT) - 1} = \frac{h\nu}{\exp(h\nu/kT) - 1} \; .$$

The internal energy U (9.1.2) of a mole of a solid is expressed
as follows:

$$U = 3N_A \langle E \rangle = 3N_A \frac{h\nu}{\exp(h\nu/kT) - 1} \; ,$$

from which the molar heat capacity of a solid can be found

$$C_\mu = 3R \left(\frac{h\nu}{kT} \right)^2 \cdot \frac{\exp(h\nu/kT)}{[\exp(h\nu/kT) - 1]^2} \; .$$

If we introduce the *characteristic temperature* $T_\theta = h\nu/kT$, then

$$C_\mu = 3R \left(\frac{T_\theta}{T} \right)^2 \frac{e^{T_\theta/T}}{(e^{T_\theta/T} - 1)^2}.$$

This result qualitatively describes the dependence of the heat capacity of solids on the temperature. Curves of this dependence are plotted in Fig. 41.10.
At high temperatures ($h\nu \ll kT$), as shown in Sect. 38.5.9, $\langle E \rangle = kT$ in accordance with the law of equipartition of energy (10.6.4) and $C_\mu = 3R$.
At low temperatures ($h\nu \gg kT$)

$$e^{h\nu/kT} - 1 \approx e^{h\nu/kT} \text{ and } C_\mu = 3R \left(\frac{T_\theta}{T} \right)^2 e^{-T_\theta/T}.$$

As $T \to 0$, $T_\theta/T \to \infty$ and $\exp(-T_0/T) \to 0$ more rapidly than the increase in $(T_\theta/T)^2$. Therefore, as $T \to 0$, the heat capacity $C_\mu \to 0$. This qualitatively agrees with experimental data. But simple quantum theory does not explain the quantitative behaviour of solids in the vicinity of absolute zero.
41.7.4 The assumption that all the atoms of a solid execute thermal vibrations independently of one another at the same frequency oversimplifies the true picture of particle vibration in a crystal lattice. The interaction between the atoms (or other particles) of a solid is so strong that all N particles of the body form a bound system having $3N$ degrees of freedom, and the vibration of all the atoms may proceed at various frequencies. The extremely complex problem of the frequency distribution of atomic vibration in a solid was, in its time, the basis for the refined theory of heat capacity of solids. A solid has a wide spectrum of vibrational frequencies. There are vibrations with sufficiently low and with higher frequencies. The low frequencies correspond to elastic vibration of the crystal in the acoustic (or ultrasonic) range. Owing to the bonds between the particles of a crystal lattice, elastic acoustic waves propagate in the crystal.
    The physical idea of the theory of the heat capacity of solids, refining the theory discussed in Sect. 41.7.3, is that the main contribution to the energy of thermal vibration of the crystal is made by low-frequency vibration, corresponding to elastic waves with wavelengths, exceeding the crystal lattice constant. This follows, in particular, from Fig. 38.7, which shows that

the maximum contribution to the average energy of a quantum oscillator is made by vibrations with low frequencies (38.5.9). The propagating elastic waves behave as if they were propagating in a continuous medium. The atomic structure of the crystal has no effect on the propagation in it of elastic waves of wavelengths exceeding $\lambda_{min} = v/\nu_{max}$, where v is the velocity of the corresponding elastic wave, and ν_{max} is its frequency. The wavelength λ_{min} should be commensurable with the lattice constant, i.e. $\lambda_{min} \approx d \approx (V/N)^{1/3}$, where N is the number of particles (crystal lattice points) in a crystal of volume V.

41.7.5 Elastic waves in a crystal have quantum properties consisting in the existence of a minimum portion of energy, a quantum of energy of the wave with the given frequency ν. This enables a wave with the frequency ν to be correlated to a *quasi-particle*, the *phonon*, whose propagation at the velocity of sound v corresponds to an acoustic wave.

A phonon has the energy $E = h\nu$, where h is Planck's constant (Appendix II), and the quasi-momentum $p = h\nu/v$. The *quasi-momentum* **p** *of the phonon* is in the direction coinciding with that of the propagation of the acoustic wave. The most essential difference between the quasi-momentum and the momentum is that in the collision of phonons in crystals the quasi-momentum can be transferred in discrete portions to the crystal lattice. In this case it is not conserved*.

In short, the quantization of the acoustic field leads to phonons in the same way as the quantization of the electromagnetic field leads to photons (37.1.4).

41.7.6 Phonon spin equals zero and, therefore, phonons comply with Bose-Einstein statistics (41.2.2). Phonons can be emitted and they can be absorbed, but their number is not maintained constant, hence the chemical potential for phonon gas (41.2.3) equals zero (cf. Sect. 41.2.6).

The energy of a crystal can be dealt with as the energy of a phonon gas and can be calculated as in Sect. 41.2.6. The number dN of phonons with frequencies in the range from ν to $\nu + d\nu$ is

$$dN = \frac{dg}{e^{h\nu/kT} - 1} ,$$

---

* A detailed discussion of the properties of quasi-particles, associated with the role of periodicity of crystal structure, is beyond the scope of this handbook.

where dg is the number of quantum states (number of cells) in volume V of the crystal:

$$dg = 3 \; \frac{4\pi p^2 \, dp V}{h^3} = 3 \; \frac{4\pi v^2 \, dv}{v^3} \, V,$$

where $p = hv/v$ is the quasi-momentum of the phonon, and v is the velocity of sound in the crystal. The factor 3 takes into account the fact that longitudinal and transverse waves, with two mutually perpendicular polarizations, can propagate in a crystal* (31.1.8). Thus

$$dN = \frac{12\pi v^2 \, dv V}{v^3 \, (e^{hv/kT} - 1)} \, .$$

The internal energy U (9.1.2) of a crystal (accurate to zero-point energy) is

$$U = \int\limits_0^{v_{max}} hv \, dN = \frac{12\pi V h}{v^3} \int\limits_0^{v_{max}} \frac{v^3 \, dv}{e^{hv/kT} - 1} \, ,$$

where $v_{max} = v \, (3N/4\pi V)^{1/3}$ is the upper frequency boundary of the phonons that contribute to the energy of thermal vibration of the crystal. An assessment of the order of magnitude of v_{max} is given in Sect. 41.7.4.

Introduced in calculating U is the *Debye characteristic temperature* $T_D = hv_{max}/k$. The following two limiting cases are considered.

(a) *High temperatures* $T \gg T_D$.

Here $e^{hv/kT} - 1 \approx \dfrac{hv}{kT}$ and $U = \dfrac{12\pi V}{v^3} \, kT \int\limits_0^{v_{max}} v^2 \, dv = \dfrac{12\pi V kT}{v^3} \times$

$$\times \frac{v_{max}^3}{3} = 3NkT.$$

* The difference in the velocities of the longitudinal and transverse waves is ignored, the number of quantum states of the phonons being simply tripled.

For one mole of the crystal $N = N_A$, where N_A is Avogadro's number (Appendix II), and the molar heat capacity C_μ complies with the Dulong-Petit law (41.7.2):

$$C_\mu = \frac{\partial U}{\partial T} = 3N_A k = 3R.$$

(b) *Low temperatures* $T \ll T_D$.

In calculating the integral $\int\limits_0^{\nu_{max}} \frac{\nu^3 \, d\nu}{\exp(h\nu/kT) - 1}$ a new variable, $\xi = h\nu/kT$, is introduced and the upper limit is replaced by ∞. Thus

$$U = \frac{12\pi V h}{v^3} \int\limits_0^{\nu_{max}} \frac{\nu^3 \, d\nu}{\exp(h\nu/kT) - 1} = \frac{12\pi V h}{v^3} \left(\frac{kT}{h}\right)^4 \int\limits_0^\infty \frac{\xi^3 \, d\xi}{e^\xi - 1}$$

$$= \frac{4\pi^5 k^4 V}{5h^3 v^3} \, T^4.$$

The molar heat capacity is proportional to the cube of the absolute temperature and obeys the *Debye* T^3 *law*. Thus

$$C_\mu = \frac{\partial U}{\partial T} = \frac{16\pi^5 k^4 V T^3}{5h^3 v^3} = \text{const } T^3.$$

41.8 Band Theory of Solids

41.8.1 Long-range order in crystals (40.1.1) leads to a situation in which an electric field, which is a periodic function of the coordinates, exists in solids. In a metal, for example, where the positive ions are located in strict order at the lattice points, the potential energy of an electron varies along some direction OX as shown in Fig. 41.11. Energy minima correspond to the places where the positive ions are located.

41.8.2 A periodic electric field in a crystal of any type (40.1.3) essentially changes the energy states of the electrons in a solid

as compared to their state in single atoms. In single atoms the electrons are in discrete energy states (38.2.5). In a solid, the energy states of the electrons are determined both by their interaction with the nucleus of their atom and by the electric field of the crystal lattice, i.e. by their interaction with other atoms. As a result of this interaction, the energy levels of the
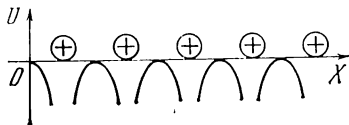
Fig. 41.11

electrons are split. Thus, in a solid containing N interacting atoms, N closely spaced energy levels are formed in place of the discrete energy level typical of a single atom. These N energy levels make up an *energy band*. A *band energy spectrum of the electrons* is formed in crystals.

41.8.3 The formation of a band energy spectrum in a crystal follows from the indeterminacy principle (38.6.6). Owing to the finite lifetime τ of the electron in an excited state ($\tau \approx 10^{-8}$ s) (39.1.5) the width ΔE of the energy level in a single atom is

$$\Delta E \approx \frac{\hbar}{\tau} \approx 10^{-7} \ \text{eV}$$

(*natural* energy level *width*).

The valence electrons (39.3.9) of the atoms in a crystal are more weakly bound to the nuclei than the inner-shell electrons and can pass from one atom to another by means of the tunnel effect (38.7.2). In this way electrons penetrate the potential barrier (38.7.1) separating the atoms of the crystal.

To assess the average lifetime τ of a valence electron in a given atom it is assumed that the penetrability D of the barrier is expressed by the simple formula (38.7.2)

$$D \approx e^{-\frac{2}{\hbar} \ \sqrt{2m(U_0 - E)} \ L} \ ,$$

39*

where $U_0 - E$ is the barrier height, consisting of the ionization energy, and amounts to approximately 10 eV, and L is the barrier width, which is commensurate with the crystal lattice constant: $L \approx d \approx 10^{-10}$ m. The frequency ν of penetration of an electron through the barrier is

$$\nu = \frac{v}{d}\, D \approx \frac{v}{d}\, e^{-\frac{2}{\hbar}\sqrt{2m(U_0 - E)}\, L} ,$$

where v is the velocity of the electron in the atom (potential well) and can be taken equal to $v \approx 10^6$ m/s. The average lifetime τ of an electron in the given atom is

$$\tau = \frac{1}{\nu} \approx \frac{d}{v}\, e^{\frac{2}{\hbar}\sqrt{2m(U_0 - E)}\, L} .$$

After substituting the values of all the quantities, $\tau \approx 10^{-15}$ s is obtained. From the indeterminacy relation

$$\Delta E \approx \frac{\hbar}{\tau} \approx 1 \text{ eV}.$$

Instead of the natural width $\Delta E \approx 10^{-7}$ eV of the electron energy level in a single atom, a band of allowed values of energy is formed in a crystal.

41.8.4 The frequency of the penetration of inner-shell electrons in atoms through a potential barrier and their transitions to other atoms is negligible. This is due to the increase in barrier height: $U_0 - E \approx 10^3$ eV and the increase in barrier width: $L \approx 3 \times 10^{-10}$ m. According to calculations, $\tau \approx 10^{20}$ years. The broadening of the energy levels of the inner-shell electrons is unessential, and these electrons behave in crystals practically in the same way as in single atoms.

41.8.5 If N is the total number of atoms in a solid, then the energy band, formed from the electronic energy level of the valence electron of the atom consists of N closely spaced levels. Adjacent energy levels in a band are at about 10^{-22} eV from each other.

The allowed energy bands are separated by regions that are *forbidden bands* of electron energy values. The width of the for-

bidden bands is commensurable with that of the allowed bands.
With an increase in energy the width of the allowed energy
bands increases and that of the forbidden bands decreases.
A schematic diagram of the energy bands of a solid is shown
in Fig. 41.12.
41.8.6 The allowed energy bands in a solid can be filled in
different ways by electrons. In limiting cases they can be com-
pletely filled or entirely empty (41.9.1). The electrons in a solid
can move from one allowed band to another. For an electron
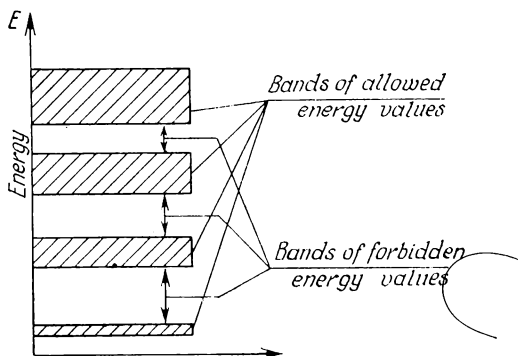
Fig. 41.12

to move from a lower band to the next higher one it is necessary
to expend energy equal to the width of the forbidden band be-
tween the two allowed ones (of the order of several eV).
Very little energy is required for intraband transitions. For
example, an energy of 10^{-4} to 10^{-8} eV is sufficient for this pur-
pose, and it is acquired by an electron in a metal by the action
of an electric field over a free path length at ordinary potential
differences. But this energy is insufficient to move an electron
from one band to another. The action of thermal excitation can
impart various energies to electrons, sufficient both for intra-
band and interband transitions.

41.9 Metals and Dielectrics in the Band Theory

41.9.1 Differences in the electric properties of solids are ex-
plained by the band theory (41.8.2) by the fact that the allowed
energy bands (41.8.2) are differently filled by electrons and by
the different widths of the forbidden bands (41.8.5). These two
factors determine whether a given solid is a conductor of electric
current or is a dielectric. A necessary condition for a solid to be
a conductor is the availability of empty energy levels to which
the electric field can raise the electrons. It should be taken into
consideration that the field can bring about only intraband
transitions of the electrons (41.8.6).
41.9.2 Typical examples of conductors are the metals of the
first group of Mendeleev's periodic table (39.3.5), for instance,
sodium. In a single atom of sodium there are two filled electron
shells, containing 2 and 8 electrons, respectively (39.3.4).
According to the Pauli principle (39.3.1), the eleventh valence
electron of the sodium atom only fills half of the upper energy
level. In a crystal of sodium, the first two filled shells of single
atoms correspond to the energy bands completely filled with
electrons.
The valence electrons of the sodium atoms in a crystal half
fill the levels of the band of allowed energy values (horizontally
hatched in Fig. 41.13). This is called the *conduction band* because
the electrons in it participate in producing the conduction cur-
rent (20.1.2). By the action of the electric field, set up in the
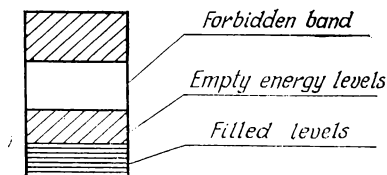
Fig. 41.13

crystal by the source of electrical energy, the valence electrons
acquire more energy and move to higher, unoccupied energy
levels in the conduction band. This initiates their ordered motion
and produces a current in the crystal.

Thus, if the band is not completely occupied by valence electrons, a solid is always a conductor of electric current.

41.9.3 *Hybridization* of allowed energy bands may occur in crystals. A band formed in splitting an upper excited level may overlap a band produced by the splitting of a lower state of valence electrons. This phenomenon is observed in the crystals of elements of the second group in Mendeleev's periodic table (Be, Cd, Mg, Zn and others). This forms a wider *hybrid band* in which valence electrons are located, filling it only partly. Hence, a hybrid band is a conduction band, and such crystals are conductors of electric current.

41.9.4 The energy bands do not overlap in solid dielectrics, and the band that combines the energy levels of the valence electrons of atoms or ions is completely filled with electrons, whereas all higher bands are entirely empty at $T = 0$ K. The band completely filled with electrons is called the *valence band*. The empty bands are conduction bands. An example of a crystalline dielectric is common salt (NaCl). The NaCl molecule has a heteropolar bond (39.4.4), which leads to the formation of Na^+ and Cl^- ions with completely filled electron shells. The NaCl crystal has a valence band of the Cl^- ion and all lower bands are completely filled, whereas the upper Na^+ ion band, located above, is entirely empty. The Cl^- and Na^+ bands are separated by 6 eV, and the source of electrical energy cannot excite the electrons from the completely filled Cl^- band to the empty Na^+ conduction band. This explains the dielectric properties of NaCl.

41.10 Electrical Conduction of Semiconductors

41.10.1 *Semiconductors* include a large class of substances whose resistivity varies in a wide range from 10^{-5} to 10^8 Ω-m, and is reduced very rapidly, according to an exponential law, as the temperature is raised (41.10.2). The most typical and extensively employed semiconductor materials are the chemical elements germanium, silicon and tellurium. In the outer shell of germanium and silicon atoms there are four valence electrons, which form covalent bonds (39.4.5) with the valence electrons of adjacent atoms. There are no "free" valence electrons in chemically pure crystals of these semiconductors. On the basis of the band theory (41.8.2) crystalline semiconductors belong

to the type of solids in which the valence band (41.9.4) is sepa-
rated from the empty conduction band (at $T = 0$ K) by a com-
paratively narrow energy region ΔE_0 (Fig. 41.14), less than that
in dielectric crystals (41.9.4).
For silicon $\Delta E_0 = 1.1$ eV;
for germanium $\Delta E_0 = 0.72$ eV.
41.10.2 The electrical con-
duction of chemically pure
semiconductors is called the
*intrinsic conduction. Electron
conduction (n-type conduction)*
occurs when electrons are
transferred from the va-
lence band to the conduction
band. This requires an expen-
diture of energy, not less than
the width ΔE_0 of the for-
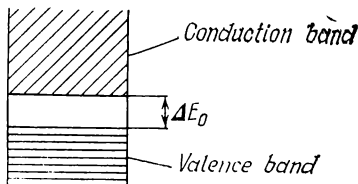
Fig. 41.14

bidden band (Fig. 41.14). The quantity ΔE_0 is called the *activa-
tion energy* of intrinsic conduction. As the temperature of the
semiconductor is raised, an increased number of electrons move
from the valence band to the conduction band and participate
in electrical conduction. The conductivity (20.3.4) of semicon-
ductors increases with the temperature T according to the law

$$\varkappa = \varkappa_0 e^{-\Delta E_0/2kT},$$

where k is Boltzmann's constant (9.4.5). The resistivity (20.3.4)
of semiconductors drastically decreases as the temperature is
raised according to the relation

$$\rho = \rho_0 e^{\Delta E_0/2kT}.$$

This feature of semiconductors essentially differentiates them
from metals (20.3.7).
41.10.3. The transition of an electron from the valence band
of a semiconductor to the conduction band indicates that the
covalent bonds (39.4.5) of the atoms are violated in the semi-
conductor crystal. A certain one of the valence electrons of one
of the atoms in the lattice leaves its place. The abandoned place,
or vacancy, displays surplus positive charge. Such a vacancy
is called a *positive hole*, and is treated as a positive charge
carrier. In the band theory this means that a vacant energy
level appears in the valence band of the crystal. A positive

hole behaves like a positive charge, equal in magnitude to the charge of the electron. Another electron may move to the place (hole) abandoned by the previous electron. This is equivalent to the movement of the positive hole; it appears at a new place, left by the second electron. In the external electric field, the electrons throughout the mass of the semiconductor move in the direction opposite to that of the electric field strength (15.1.2). The positive holes move in the direction of the field strength, i.e. in the direction that a positive charge would travel when subject to an electric field.

The electrical conduction of a semiconductor that is due to the motion of holes is called *hole*, or *p-type*, *conduction*.

41.10.4 The intrinsic conduction of semiconductors is due to the presence of two kinds of current carriers: electrons in the conduction band and holes in the valence band. Each electron excited to the conduction band corresponds to one hole in the valence band. The electron concentration n_e and hole concentration n_h are the same and increase rapidly with the temperature T according to the law

$$n_e = n_h = \text{const } e^{-\Delta E_0/2kT},$$

where ΔE_0 is the activation energy of intrinsic conduction (41.10.2), and k is Boltzmann's constant (9.4.5). This explains the nature of the dependence of the electrical conductivity and resistivity of semiconductors (41.10.2) on the temperature, which strongly differs from the dependence of these quantities in metals (20.3.7).

41.10.5 The electrical conduction of semiconductors due to the presence in them of impurity centres is called *extrinsic*, or *impurity*, *conduction*.

Impurity centres (or simply *impurities*) may be atoms or ions of foreign elements, and various defects and distortion of the crystal lattice (vacant points, displacements due to deformation of the crystal, etc.).

Impurities change the periodic electric field in a solid and affect the motion of the electrons and their energy states. The energy levels of the valence electrons of impurity atoms are not located in the allowed energy bands (41.9.2) of the main, or host, crystal, and *impurity energy levels (localized levels)* are formed in the forbidden band.

41.10.6 Impurities can be a supplementary source of electrons in a solid. For example, in the substitution of one tetravalent atom of germanium by an atom of an impurity having five valence electrons (phosphorus, antimony or arsenic) one of the electrons of the impurity cannot form a covalent bond (39.4.5) with the germanium atoms and is thereby "superfluous". The energy levels of such electrons are below the conduction band of the host crystal (Fig. 41.15). Such levels, filled with electrons,
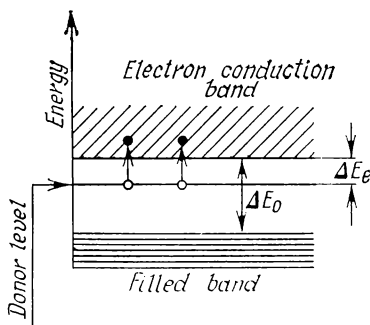
Fig. 41.15

are called *donor levels*, and the atoms of the impurity that supply the "superfluous" electrons are called *donor atoms*. Only a small amount of energy ΔE_e is required to excite electrons from the donor levels to the conduction band. This energy is obtained, for instance, from thermal excitation. For silicon $\Delta E_e =$ $= 0.054$ eV if the impurity is arsenic. When electrons are transferred from the donor levels to the conduction band, *electron impurity*, or *extrinsic*, *conduction* (*n-type impurity conduction*) occurs. Such semiconductors are called *electron impurity semiconductors* (*n-type semiconductors*).

41.10.7 If a tetravalent atom in the lattice of the semiconductor is replaced by a trivalent impurity atom (boron, aluminium or indium) there will be a lack of one electron to form saturated covalent bonds. The lacking electron can be borrowed from the lattice of the neighbouring atom of germanium, in which a posi-

tive hole is thereby formed (41.10.3). The consecutive filling
of the holes formed in the germanium atoms by electrons is
equivalent to the movement of the holes and leads to conduction
in the semiconductor. The impurity energy levels, not occupied
by electrons, are called *trapping*, or *acceptor*, *levels*. In this
case, the impurity atoms are said to be *acceptor atoms*. Acceptor
levels are located above the upper edge of the valence band
(41.9.4) of the host crystal (Fig. 41.16). When silicon crystals
are doped with boron (the addition of impurities to a semicond-
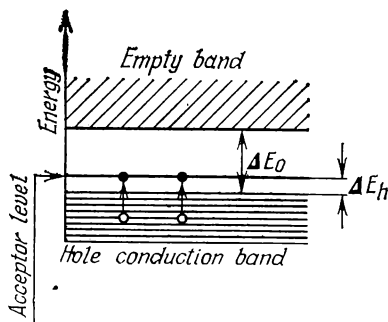
Fig. 41.16

ductor is called doping), the acceptor levels are an amount
$\Delta E_h = 0.08$ eV above the valence band. The transfer of elec-
trons from the filled valence band to the acceptor level produces
positive holes in this band, and the valence band becomes
a band of hole conduction. This leads to *hole impurity conduction*
(*p-type impurity conduction*). Such semiconductors are called
hole impurity semiconductors (*p-type semiconductors*).
41.10.8 When a semiconductor is doped with both donor and
acceptor impurities at the same time, the nature of the conduc-
tion (*n*- or *p*-type) depends upon which impurity produces the
higher concentration of current carriers. The concentration and
energy of the electrons (and holes) in semiconductors depend
very strongly, in contrast to those of metals, on the tempera-
ture, and increase when the temperature is raised.

41.11 Concept of Contact Electrical Phenomena in Metals and Semiconductors

41.11.1 The *work function* Φ in removing an electron from a metal is the work it is necessary to do for the electron to emerge from the metal into vacuum. The work Φ is done against the attractive forces developed by the surplus positive charge in the metal resulting from the removal of the electron. Besides, it is necessary to overcome the repulsion forces of the previously emerged electrons if they have not been removed and form an electron "cloud" at the surface of the conductor. The work function has a value of the order of several eV and depends upon
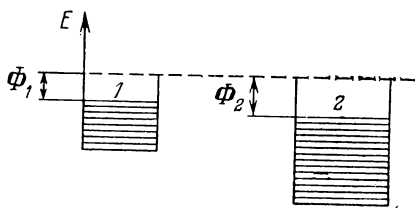
Fig. 41.17

the kind of metal and the condition of its surface; it is changed by the presence of dirt, traces of moisture, etc. In the quantum theory of solids, the work function is reckoned from the upper electron-filled Fermi level (41.4.2) (Fig. 41.17).

41.11.2 Since the fastest electrons fly out of the metal, a lack of negative charge is formed in the metallic conductor and a surplus in the surrounding space. This is evident in a very thin layer (of a thickness equal to several interatomic distances in the metal) on both sides of the surface of the conductor. It is contended that the surface of the metal is a *double electric layer* (*contact layer*) similar to an extremely thin capacitor (17.2.4).

41.11.3 The potential difference ΔV (16.2.7), characterizing the electric field of the double layer, is called the *surface poten-*

tial jump at the metal-vacuum boundary, or the *contact potential difference* between a metal and a vacuum. Thus

$$\Delta V = \frac{\Phi}{e},$$

where Φ is the work function (41.11.1), and e is the magnitude of the electron charge.

There is no electric field outside the double layer, and the potential outside the metal and beyond the double layer equals zero. Inside the metal the potential is positive and equal to ΔV. The potential energy of the conduction electrons (20.3.1) is negative and is equal to $-e\Delta V = -\Phi$. The whole volume of the metal is a potential well (38.4.1) for the conduction electrons; the depth of the well equals the work function Φ.

41.11.4 In the contact of two dissimilar metals, *1* and *2*, having the work functions Φ_1 and Φ_2 electrons pass predominantly from the metal with the lower work function to that with the higher one. The metals acquire unlike charges and, in a state of thermodynamic equilibrium (8.3.3), there is a potential difference ΔV_{12} between the two contacting dissimilar metals. It is called the *internal contact potential difference*.

The potential difference $\Delta V'_{12}$ between two points located close to but outside the surfaces of the first and second contacting metals is called the *external contact potential difference*.

In the contact of the two metals shown in Fig. 41.17 ($\Phi_1 < \Phi_2$), the electrons pass predominantly in the direction from metal *1* to metal *2* and they become charged positively and negatively, respectively. At the same time, a shift occurs in the energy levels of the electrons: upward in metal *2* and downward in metal *1*. In a state of thermodynamic equilibrium, the Fermi levels (41.4.2) in the two metals coincide (Fig. 41.18).

It is evident from Fig. 41.18 that the external contact potential difference $\Delta V'_{12}$ depends upon the work functions Φ_1 and Φ_2. Thus

$$\Delta V'_{12} = -\frac{\Phi_1 - \Phi_2}{e}.$$

The internal contact potential difference depends upon the Fermi levels in the metals. Thus

$$\Delta V_{12} = V_1 - V_2 = \frac{E_{F_1} - E_{F_2}}{e}.$$

The potential changes from V_1 to V_2 within the limits of the double electric layer (41.11.2) of thickness l (Fig. 41.19) that can be approximately assessed by the formula

$$l = \sqrt{\frac{2\varepsilon_0 \Delta V'_{12}}{en_0}} \text{ (in SI units).}$$

Here n_0 is the concentration of charges in the double layer, ε_0 is the electric constant in SI units (14.2.7), and e is the electron charge.

The thickness of the double layer in contacting metals is of the order of the lattice constants in metals, and the resistivity
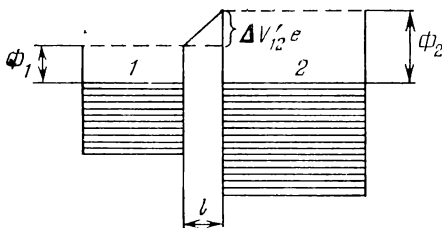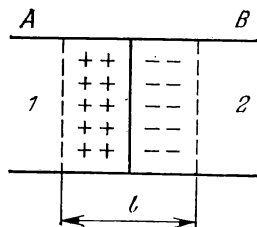
Fig. 41.18   Fig. 41.19

of this layer is the same as that of the rest of the metal. Electric current passes equally well in both directions through the contact layer of two metals, and there is no *rectifying (barrier-layer) effect of the contact,* or *junction,* i.e. *unilateral conduction.*

41.11.5 Contact between a metal and an n-type semiconductor (41.10.6) leads to unilateral conduction of the junction. Assume that $\Phi_1 > \Phi_n$, where Φ_1 and Φ_n are the work functions of the metal and the n-type semiconductor, respectively. The arrangement of the conduction band of the metal (41.9.2), the completely filled band of the semiconductor and its donor levels before contact has been made are shown in Fig. 41.20. After the metal makes contact with the semiconductor, electrons from the donor levels of the n-type semiconductor (41.10.6) move into the metal.

The charge in the contact layer (41.11.2) is positive on the side of the n-type semiconductor and negative on the metal side. Owing to the low electron concentration in the semiconductor in comparison to that of the metal (10^{15} cm^{-3} instead of 10^{22} cm^{-3}), the thickness of the contact layer at the metal-semiconductor boundary is found to be approximately 10^4 times greater than that in the metal.

41.11.6 The small number of current carriers in the contact layer and its large thickness indicate that the resistivity of this layer is substantially greater than that of the rest of the
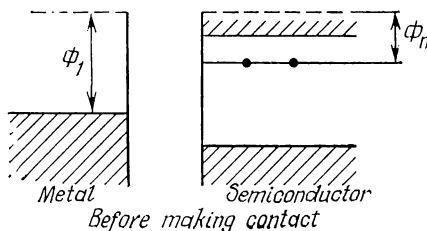
Fig. 41.20

semiconductor. Here the contact layer is called the *depletion*, or *space-charge*, *layer* (formerly called the *barrier*, or *blocking*, *layer*). The depletion layer is the basis for rectification of alternating current at a metal-semiconductor junction. When the metal is connected to the positive terminal of a source of electrical energy, and the semiconductor to the negative terminal, the external electric field is directed from the metal toward the semiconductor. In this case the electrons are pulled from the bulk of the semiconductor into the double layer. The thickness of this layer is reduced and the conduction increases. In this *conducting direction*, electric current passes through the metal-semiconductor junction. When the external electric field is directed from the semiconductor toward the metal, the electrons are forced out of the double layer deep into the semiconductor. This increases the width and resistance of the depletion layer and in this *reverse (nonconducting) direction*, the metal-semiconductor junction blocks the electric current. This unilateral

conduction of the junction means that an alternating current
passing through it is rectified.

41.11.7 The region of contact of two semiconductors having
different, n- and p-type, conduction is called an *electron-hole*
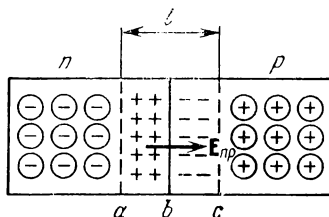
Fig. 41.21
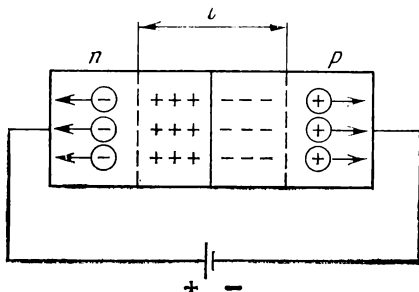
Fig. 41.22
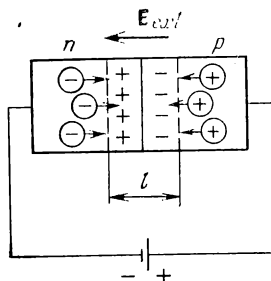
Fig. 41.23

junction (p-n junction). A double electric layer (41.11.2) is
formed at a p-n junction as a result of electron and hole motion
through the interface. Electrons from the n-type semiconductor
pass into the p-type semiconductor, and the holes pass in the
opposite direction. This forms surplus charges of opposite sign
in the ab regions of the n-type semiconductor and the bc regions
of the p-type semiconductor (Fig. 41.21). The double layer of

thickness l sets up a contact electric field with the strength $\mathbf{E}_{np}$ and a certain potential difference at the boundaries of the layer. This field impedes further counter migration of electrons and holes. At a definite thickness of the p-n junction a state of equilibrium is set up, corresponding to the equalizing of the Fermi levels in the two semiconductors. This forms an *equilibrium contact layer*, which is a barrier layer (41.11.6), having higher resistance than that of the remaining regions of the semiconductors.

41.11.8 When a p-n semiconductor junction is connected into the external circuit of a source of electrical energy in the way shown in Fig. 41.22, the external (applied) electric field, strengthening the field of the contact layer, causes electrons to migrate to the n-type semiconductor and holes to the p-type semiconductor at opposite sides of the junction. The thickness of the barrier layer and its resistance increase. This direction of the applied electric field is said to be the *reverse* one. In this direction practically no current passes through the p-n junction. When the polarity of the external applied voltage is reversed (Fig. 41.23), the external electric field with the strength $\mathbf{E}_{ext}$ opposes the field of the contact layer (41.11.7). The counter motion of electrons and holes, forced by the external field to move from the bulk of the semiconductors to the region of the p-n junction, increases the number of mobile current carriers at the junction. This reduces the thickness and resistance of the contact layer, and, in this forward direction, current passes through the p-n junction. The blocking effect of a p-n junction is similar to the rectification of a two-electrode valve, or tube, called a diode. Accordingly, a semiconductor device containing one p-n junction is called a *semiconductor diode. Crystal triodes*, or *transistors*, contain two p-n junctions.

PART EIGHT NUCLEAR PHYSICS AND ELEMENTARY PARTICLES

CHAPTER 42 STRUCTURE AND BASIC PROPERTIES OF ATOMIC NUCLEI

42.1 Main Properties and Structure of the Nucleus

42.1.1 The *nucleus* is the central part of an atom; practically all of the atom's mass and its positive charge are concentrated in the nucleus. All atomic nuclei are made up of elementary particles (43.1.1) called *protons* and *neutrons*, considered to be two charge states of a single particle called the *nucleon*. The proton has a positive electric charge, equal in magnitude to that of the electron (14.1.2). The neutron has no electric charge.

42.1.2 The *charge of the nucleus* is the quantity Ze, where e is the electron charge, and Z is the atomic number of the chemical element in Mendeleev's periodic table (39.3.5) and is equal to the number of protons in the nucleus. Up to date nuclei are known with Z ranging from $Z = 1$ to $Z = 107$. For all nuclei, with the exception of $^1_1$H, $^3_1$He and certain other neutron-deficient nuclei, $N \geqslant Z$, where N is the number of neutrons in the nucleus. For light nuclei $N/Z \approx 1$; for the nuclei of chemical elements at the end of the periodic system, $N/Z \approx 1.6$.

42.1.3 The number of nucleons (42.1.1) in a nucleus, $A = N + Z$, is called the *mass*, or *nucleon*, *number*. Nucleons (proton and neutron) are assigned a mass number equal to unity. The electron has zero mass number.

Nuclei with the same Z value, but different A values are called *isotopes*. Nuclei with the same A value, but different Z values are called *isobars*. The nucleus of chemical element X is denoted by the symbol ^A_ZX, where X is the symbol of the chemical element.

About 300 stable isotopes of chemical elements are found in nature, and there are over a 1000 man-made (radioactive) isotopes.

42.1.4 The size of a nucleus is specified by the *nuclear radius r*, which has only an arbitrary meaning because the nucleus has no clearly defined boundary. The empirical formula for the nuclear radius is $r = r_0 A^{1/3}$, where $r_0 = (1.3$ to $1.7) 10^{-15}$ m can be interpreted as the proportionality of the volume of the nucleus to the number of nucleons it contains.

The density of nuclear matter is of the order of magnitude of 10^{17} kg/m³ and is constant for all nuclei. It substantially exceeds the density of the most dense of ordinary substances.

42.1.5 Nucleons in atomic nuclei are fermions (41.2.4) and their spin equals $\hbar/2$. The nucleus of an atom has intrinsic angular momentum, called *nuclear spin*, equal to $L_N = \sqrt{I(I+1)}\,\hbar$, where I is the *inner (total) spin quantum number* (cf. Sect. 26.1.3), $\hbar = h/2\pi$, and h is Planck's constant (Appendix II).

The number I may assume any whole or half-integer value: 0, 1/2, 1, 3/2, 2, etc. Nuclei with even mass numbers A have integer spin (in $\hbar$ units) and comply with Bose-Einstein statistics (41.2.1). Nuclei with odd mass numbers A have half-integer spin (in $\hbar$ units) and comply with Fermi-Dirac statistics (41.2.4).

42.1.6 Nuclear particles have their intrinsic magnetic moments, which determine the *total nuclear magnetic moment* P_{mN}. The unit of measurement of nuclear magnetic moments is the *nuclear magneton* μ_N analogous to Bohr's magneton (26.1.4). Thus

$$\mu_N = \frac{e\hbar}{2m_p} \quad \text{(in SI units)},$$

$$\mu_N = \frac{e\hbar}{2m_p c} \quad \text{(in Gaussian units)},$$

where e is the magnitude of the electron charge, m_p is the mass of the proton, and c is the electrodynamic constant (Appendix II). The nuclear magneton is a factor of $m_p/m_e = 1836.5$ times less than the Bohr magneton. It follows that the magnetic properties of atoms are determined by those of their electrons.

The following relation exists between nuclear spin L_N, measured in $\hbar$ units, and the magnetic moment of the nucleus, measured in nuclear magnetons:

$$P_{mN} = \gamma_N L_N,$$

40*

where γ_N is the *nuclear gyromagnetic ratio* (cf. Sect. 26.1.4). The magnetic moments of the neutron and proton are associated with their complex internal structure (43.6.2). The neutron has a negative magnetic moment equal to $-(1.91314 \pm \pm 0.00005)$ μ_N. The spin of the neutron and its magnetic moment are opposite in direction. The magnetic moment of the proton is positive and is equal to $(2.79277 + 0.00003)$ μ_N. Its direction coincides with the direction of proton spin.

42.1.7 In the general case, the distribution of electric charge of the proton in the nucleus is asymmetrical. A measure of the deviation of this distribution from spherical symmetry is the *electric quadrupole moment of the nucleus* Q. If the charge density is assumed the same throughout the nucleus, Q is determined only by the shape of the nucleus. For an ellipsoid of revolution

$$Q = \frac{2}{5} \, Ze \, (b^2 - a^2),$$

where b is the semiaxis of the ellipsoid along its spin, and a is the semiaxis in the perpendicular direction. For a nucleus stretched out along its spin, $b > a$ and $Q > 0$. For a nucleus flattened in this direction, $b < a$ and $Q < 0$. For spherical distribution of the charge in the nucleus, $b = a$ and $Q = 0$. This is valid for nuclei with spin equal to 0 or to $(1/2) \, \hbar$.

42.2 Binding Energy of Nuclei. Mass Defect

42.2.1 In nuclei, nucleons are in states that differ essentially from their free states. With the exception of the nucleus of ordinary hydrogen, all nuclei contain at least two nucleons, between which there exists a special *nuclear strong interaction*—attraction—that ensures stability of the nucleus notwithstanding the repulsion of the protons of like charge.

42.2.2 The *binding energy of a nucleon* in a nucleus is the physical quantity equal to the work required to remove the nucleon from the nucleus without imparting kinetic energy to it.

The *nuclear binding energy* is the amount of work required to take the nucleus apart into its component nucleons without imparting any kinetic energy to them. It follows from the law of energy conservation that in forming a nucleus the same amount

of energy is released as is required in splitting the nucleus into its component nucleons. The nuclear binding energy is the difference between the total energy of all the free nucleons that make up the nucleus and their total energy in the nucleus.

42.2.3 In forming a nucleus, its mass is reduced: the mass of the nucleus is less than the total mass of its component nucleons. This reduction in mass in the formation of a nucleus is due to the release of the binding energy. If ΔE_b is the amount of energy released in forming a nucleus, the corresponding mass Δm (5.7.6), equal to

$$\Delta m = \frac{\Delta E_b}{c^2},$$

is called the *mass defect* and characterizes the reduction in the total mass in forming a nucleus from its component nucleons. If a nucleus of mass m_N is formed of Z protons with the mass m_p and of $(A - Z)$ neutrons of mass m_n, then

$$\Delta m = Z m_p + (A - Z) m_n - m_N.$$

Instead of the mass m_N of the nucleus, the quantity Δm can be expressed in terms of the atomic mass m_a. Thus

$$\Delta m = Z m_H + (A - Z) m_n - m_a,$$

where m_H is the mass of the hydrogen atom.

In practice, in calculating Δm, the masses of all the particles and atoms are expressed in atomic mass units (Appendix II). The mass defect serves as a measure of the nuclear binding energy. Thus

$$\Delta E_b = \Delta m c^2 = [Z m_p + (A - Z) m_n - m_N] c^2.$$

One atomic mass unit corresponds to one *atomic unit of energy*, which is equal to 931.5016 MeV.

42.2.4 The *specific binding energy* ΔE_b^{nu} of the nucleus is the binding energy per nucleon: $\Delta E_b^{nu} = \Delta E_b / A$. The quantity ΔE_b^{nu} equals 8 MeV/nucleon on an average. A curve showing the dependence of the specific binding energy on the mass number A and characterizing the different bonding strengths for the nucleons in the nuclei of various chemical elements, is given in Fig. 42.1. The nuclei of elements in the middle of the periodic system (39.3.5) $(28 < A < 138)$, i.e. from $^{28}_{14}\text{Si}$ to

$^{138}_{56}$Ba have the strongest bonds. In these nuclei ΔE_b^{nu} is close to 8.7 MeV/nucleon. As the number of nucleons in the nucleus increases, the specific binding energy decreases. The atomic nuclei of chemical elements located at the end of the periodic system (for instance, the nucleus of uranium) have the binding
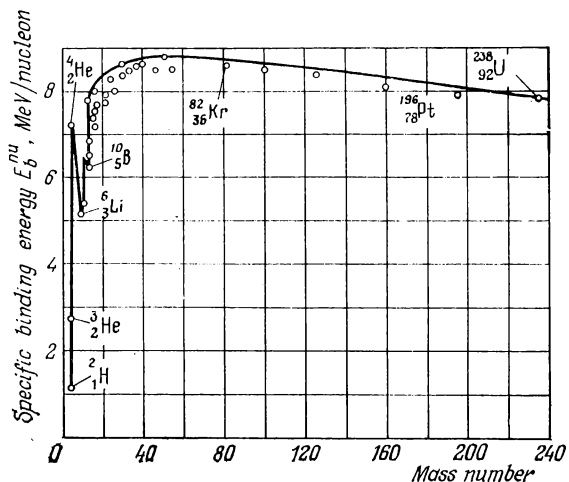
Fig. 42.1

energy $\Delta E_b^{nu} \approx 7.6$ MeV/nucleon. This explains the feasibility of releasing energy by the fission of heavy nuclei (42.9.2). In the region of low mass numbers there are sharp peaks of specific binding energy. Maxima are typical for nuclei with even numbers of protons and neutrons ($^4_2$He, $^{12}_6$C, $^{16}_8$O). Minima are typical for nuclei with odd numbers of protons and neutrons ($_3$Li, $_5$B, $_7$N).
42.2.5 Data on the binding energy of the nucleus and an application of the liquid-drop model of the nucleus* enabled certain laws of nuclear structure to be established.

---

* The liquid-drop model of the nucleus is discussed in the physics course of secondary schools.

The *atomic nuclear stability criterion* is the relation between the numbers of protons and neutrons in a stable nucleus for the given isobars (42.1.3) (i.e. $A = $ const). The condition for minimum energy of the nucleus yields the following relation between Z_{st} and A:

$$Z_{st} = \frac{A}{1.98 + 0.015 A^{2/3}} \; .$$

The stable atomic numbers Z_{st} are the nearest whole numbers to those obtained by this formula.

At low and medium values of A there are approximately equal numbers of protons and neutrons in stable nuclei: $N \approx Z$ or $A \approx 2Z$. Especially stable are nuclei (when $Z = A/2$) for which N or $Z = 2, 8, 20, 28, 50, 82$ or 126. These have been called *magic numbers*. Nuclei in which both N and Z are magic numbers are said to be *double-magic*.

As Z increases, the Coulomb forces of repulsion between the protons increases in proportion to $Z(Z-1) \approx Z^2$ (pairwise proton interaction), and, to compensate for this repulsion by nuclear attraction, the number of neutrons must increase more rapidly than the number of protons.

42.3 Nuclear Forces

42.3.1 Nuclear interaction indicates that special *forces* exist *in nuclei* that do not belong to any of the types of forces known in classical physics (e.g. gravitational and electromagnetic forces).

42.3.2 Nuclear forces are of short-range action. They only act at the extremely short distances, of the order of 10^{-15} m, between nucleons in a nucleus. This distance, ranging from 1.5 to 2.2 $\times$ $\times 10^{-15}$ m, is called the *action radius*, or *range*, *of the nuclear forces*.

42.3.3 Nuclear forces display *charge independence*: the attraction between any two nucleons is the same and is independent of the charge state of the nucleons, regardless of whether they are protons or neutrons (42.1.1). The charge independence of nuclear forces follows from a comparison of the binding energy in *mirror nuclei*. These are nuclei having the same total number of nucleons, but the number of protons in one is equal to the

number of neutrons in the other. The nuclei of helium $^3_2$He and tritium $^3_1$T (superheavy hydrogen) can be taken as an example. The binding energies (42.2.4) in these nuclei are 7.72 MeV and 8.49 MeV, respectively. Assuming charge independence of nuclear forces, the difference in the nuclear binding energies, equal to 0.77 MeV, is associated with the energy of Coulomb repulsion between two protons in the nucleus of $^3_2$He. Assuming this quantity equal to $e^2/4\pi\varepsilon_0 r$ (16.2.6), the average distance r between the protons in the $^3_2$He nucleus is found to equal 1.9 $\times$ 10^{-15} m. This is in agreement with the action radius of nuclear forces (42.3.2).

42.3.4 Nuclear forces have the property of *saturation*. This is manifested in the fact that a nucleon in the nucleus interacts with only a limited number of the closest nucleons among its neighbours. Besides the short-range action of nuclear forces (42.3.1), this saturation also follows from the linear dependence of the binding energy in nuclei on the mass number A (42.1.3). If each nucleon in a nucleus interacted simultaneously with all $(A - 1)$ nucleons in the nucleus, the binding energy of the nucleus would be proportional to the possible number of interacting pairs of nucleons in the nucleus, i.e. the number of combinations of A taken two at a time: $A (A - 1)/2$*. The dependence, in this case, of the binding energy on A would be quadratic, rather than linear, and this is in contradiction with the experimental data. Practically total saturation of nuclear forces is reached for the alpha particle (42.4.2), which is a stable structure.

42.3.5 Nuclear forces depend upon the orientation of the spins of the interacting nucleons. This is confirmed by the different kinds of scattering of neutrons by the molecules of ortho- and parahydrogen. In a molecule of orthohydrogen, the spins of the two protons are parallel to each other; in a molecule of parahydrogen they are antiparallel. If the interaction of a neutron with a proton was independent of the mutual orientation of their spins, the scattering of neutrons by molecules of either orthohydrogen and parahydrogen would be the same, even though in the case of parahydrogen the spin of the neutron is parallel to that of one of the protons, and in the case of orthohydrogen the spin of the neutron does not necessarily coincide with the spins of the protons. Experiments showed that the scattering

---

* Assuming, of course, pairwise nuclear interaction.

of neutrons by parahydrogen is 30 times more than the scattering by orthohydrogen. This is a clear indication of the dependence of nuclear forces on the orientations of the spins of the interacting nucleons.

Nuclear forces are not central forces (3.3.4). They cannot be represented as ones acting along lines joining the centres of the interacting nucleons.

42.3.6. No rigorous theory exists that can comprehensively explain the nature and laws of nuclear forces. One of the simplest theories is the exchange interpretation of nuclear forces, which are also called *exchange forces*, or forces arising in an exchange interaction. This theory is based on the idea that interaction between two particles is accomplished by the exchange of a third particle. The exchange theory of nuclear forces assumes that the interaction between the nucleons of a nucleus is due to the exchange of special particles called pions (43.3.4) and having the rest mass m_π. The time Δt during which nucleons exchange pions (i.e. the time interval from the emission of a pion by one nucleon and its absorption by another), should not, according to the Heisenberg relation (38.6.5), exceed $\Delta t \geqslant \hbar/\Delta E$, where ΔE is the indeterminacy in the value of the nucleon energy. The quantity ΔE is related to the rest energy E_π of the new particle created in the region of the action radius R of the nuclear forces (42.3.2). Assuming that $\Delta E = E_\pi = m_\pi c^2$, $\Delta t \approx \hbar/m_\pi c^2$. If it is assumed that the pion leaves the nucleon at the velocity $v_\pi \approx c$, it travels, during the time Δt, the distance

$$R_\pi = v_\pi \Delta t = c \, \frac{\hbar}{m_\pi c^2} = \frac{\hbar}{m_\pi c} \, .$$

It is quite natural to assume further that the pion travels a distance equal to R during the time Δt, i.e. $R_\pi = R \approx 1.5 \times 10^{-15}$ m (42.3.2). Consequently, the mass of the pion m_π is equal to $m_\pi = \hbar/R_\pi c \approx 250 \, m_e$, where m_e is the rest mass of the electron. This agrees well with experimental data on the masses of pions (see Table 43.1).

42.4 Radioactivity

42.4.1 *Radioactivity* is the transformation of unstable isotopes (42.1.3) of one chemical element into isotopes of another element, accompanied by the emission of certain particles. Radioactivity

TABLE 42.1

Radioactivity	Change in nuclear charge Z	Change in mass number A	Essence of the process
Alpha decay	$Z-2$	$A-4$	Emission of an alpha particle, which is a system of two protons and two neutrons bound together.
Beta decay	$Z\pm1$	A	Mutual conversions in the nucleus of a neutron ($_0^1$n) into a proton ($_1^1$p) and vice versa
β^- decay	$Z+1$	A	$_0^1\text{n} \rightarrow {}_1^1\text{p} + (_{-1}^0\text{e} + {}_0^0\widetilde{\nu}_e)$
β^+ decay	$Z-1$	A	$_1^1\text{p} \rightarrow {}_0^1\text{n} + (_{+1}^0\text{e} + {}_0^0\nu_e)$
Electron capture (e or K capture)	$Z-1$	A	$_1^1\text{p} + {}_{-1}^0\text{e} \rightarrow {}_0^1\text{n} + ({}_0^0\nu_e)$, where $_0^0\nu_e$ and $_0^0\widetilde{\nu}_e$ are the electron neutrino and antineutrino. Particles emitted from the nucleus are enclosed in parentheses.
Spontaneous fission	$Z-(1/2)Z$	$A-(1/2)A$	Fission of the nucleus into two fragments of approximately equal mass and charge.
Proton radioactivity	$Z-1$	$A-1$	Emission of a proton from the nucleus.
Double proton radioactivity	$Z-2$	$A-2$	Simultaneous emission of two protons from the nucleus.

is also the mutual conversion of certain elementary particles into other ones (for instance, the conversion of a neutron into a proton with the creation of an electron and an electron anti-neutrino) (42.6.1).

Natural radioactivity is that observed in the unstable isotopes that exist in nature.

Artificial, or *induced*, *radioactivity* is that acquired by isotopes as a result of nuclear reactions (42.9.1).

42.4.2 The main kinds of radioactivity are listed in Table 42.1.

42.4.3 Usually, all kinds of radioactivity are accompanied by the emission of gamma rays, which is hard, short-wave electro-magnetic radiation. Gamma rays are the main form that reduces the energy of excited products of radioactive transformations. A nucleus subject to radioactive decay is called a *parent* nucleus; the result of decay is called a *daughter*, or *product*, nucleus and, as a rule, is found to be excited. Its transition to the ground state is accompanied by the emission of a γ photon.

42.4.4 Spontaneous decay of atomic nuclei obeys the radioactive decay law:

$$N = N_0 e^{-\lambda t},$$

where N_0 is the number of nuclei in the given volume at the initial instant of time $t = 0$, N is the number of nuclei in the same volume at the instant of time t, and λ is the *decay constant*, which is the probability that the nucleus will decay in one second and is equal to the fraction of the nuclei that decays in unit time.

The law of spontaneous radioactive decay is based on two assumptions: (1) the decay constant is independent of external conditions, and (2) the number of nuclei, which decay during the time dt, is proportional to the number of nuclei present. These assumptions signify that radioactive decay is a statistical process and the decay of any given nucleus is a chance event, having a definite probability. The quantity $1/\lambda$ is the *average life* (or *lifetime*) of a radioactive isotope. As a matter of fact, the total lifetime dN of a nucleus is equal to $t \mid dN \mid = t\lambda N \, dt$. The average life τ of all initially existing nuclei is

$$\tau = \frac{1}{N_0} \int_0^\infty \lambda N t \, dt = \lambda \int_0^\infty t e^{-\lambda t} \, dt = \frac{1}{\lambda}.$$

42.4.5 A characteristic of nuclear stability against decay is the radioactive *half-life* (half-value period) $T_{1/2}$. This is the length of time after which half of the initial number of nuclei of the given radioactive substance undergoes decay. The relation between λ and $T_{1/2}$:

$$T_{1/2} = \frac{\ln 2}{\lambda} = \frac{0.693}{\lambda} = 0.693\tau.$$

42.4.6 Natural radioactivity (42.4.1) is observed for the nuclei of atoms of chemical elements that are located beyond lead in Mendeleev's periodic system (39.3.5). Natural radioactivity of light and medium-weight nuclei is observed only for the nuclei $^{40}_{19}K$, $^{87}_{37}Rb$, $^{115}_{49}In$, $^{138}_{57}La$, $^{147}_{62}Sm$, $^{176}_{71}Lu$ and $^{187}_{75}Re$. The law of the conservation of electric charges (14.1.3) in nuclear radioactive decay is

$$Z_N e = \sum_i Z_i e,$$

where $Z_N e$ is the charge of the parent nucleus (42.4.3), and $Z_i e$ is the charge of the nuclei and particles created as a result of the radioactive decay. This law is also applicable in investigating all nuclear reactions (42.9.1). The rule for the conservation of the mass numbers (42.1.3) in natural radioactive phenomena is

$$A_N = \sum_i A_i,$$

where A_N is the mass number of the parent nucleus, and A_i is the mass number of the nucleus or particles obtained as a result of radioactive decay.

42.4.7 *Soddy's displacement* (*Fajans-Soddy*) *law* for radioactive α and β^- decays:

$$^A_Z X \rightarrow ^{A-4}_{Z-2} Y + ^4_2 He \quad \text{for} \quad \alpha \text{ decay}$$

$$^A_Z X \rightarrow ^A_{Z+1} Y + ^0_{-1} e \quad \text{for} \quad \beta^- \text{ decay}.$$
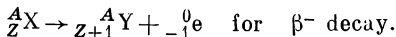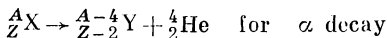
Here $^A_Z X$ is the parent nucleus, Y is the symbol for the daughter nucleus, $^4_2 He$ is a helium nucleus, $^0_{-1} e$ is the symbol for designating an electron, for which $A = 0$ and $Z = -1$. If the daughter nucleus is also radioactive, a radioactive series,

or chain, of transformations is formed. Natural radioactive nuclei constitute *three radioactive families*, called the uranium series ($^{238}_{92}U$), thorium series ($^{232}_{90}Th$) and the actinium series ($^{235}_{89}Ac$). They are named after the longest-lived parent isotope of the series with the maximum half-life (42.4.5). After undergoing a sequence of α and β^- decays, all the families terminate in one of the stable nuclei of lead isotopes: $^{206}_{82}Pb$, $^{208}_{82}Pb$ and $^{207}_{82}Pb$. The neptunium series, beginning with the transuranium element neptunium $^{237}_{93}Np$, is obtained artificially (being a man-made element) and terminates in bismuth $^{209}_{83}Bi$.

42.4.8 If there is a whole series of radioactive decays and, during the time dt, $\lambda_p N_p \, dt$ nuclei decay from the total number N_p of parent nuclei, whereas $\lambda_d N_d \, dt$ daughter nuclei decay during the same time, the total change dN_d in unit time in the number of nuclei of the daughter substance is expressed by the relation

$$\frac{dN_d}{dt} = \lambda_p N_p - \lambda_d N_d.$$

In the case of dynamic equilibrium between the parent and daughter substances $dN_d/dt = 0$, and the conditions of radioactivity equilibrium are complied with. Thus

$$\lambda_p N_p = \lambda_d N_d,$$

from which

$$\frac{N_p}{N_d} = \frac{\lambda_d}{\lambda_p} = \frac{T_p}{T_d},$$

where T_p and T_d are the half-lives (42.3.5) of the parent and daughter nuclei. The product $A = \lambda N$ is called the *activity* of the given radioactive substance and represents the number of decays in unit time. Activity referred to unit mass of the substance is called *specific activity*. Activity is measured by the number of nuclei of a radioactive substance that decays in unit time (s^{-1}). The unit of activity is called the becquerel (Bq).

42.5 Alpha Decay

42.5.1 *Alpha decay* is the emission of alpha particles by the nuclei of certain chemical elements. Alpha decay (42.5.2) is a property of heavy nuclei having mass numbers $A > 200$

and nuclear charges $Ze > 82$. Formed inside such nuclei are separate alpha particles, each consisting of two protons and two neutrons. This is facilitated by the saturation of nuclear forces (42.3.4). The formed alpha particles are subject to more Coulomb force repulsion from the protons of the nucleus than single protons are. At the same time, an alpha particle is subject to less nuclear attraction to the nucleons in the nucleus than single nucleons.

42.5.2 The nucleus is a potential barrier (38.7.1) for alpha particles and its height U is greater than E, the energy of the alpha particle in the nucleus. Alpha decay occurs by the penetration of alpha particles through the potential barrier by means of the tunnel effect (38.7.2). The formula for the penetrability D of the potential barrier indicates that even negligible changes in the energy of an alpha particle in the nucleus lead to a drastic change of the quantity

$$\exp \left\{ -\frac{2}{\hbar} \int_{x_1}^{x_2} \sqrt{2m_\alpha \left[U\left(x\right) - E \right]} \, dx \right\}.$$

This explains the great differences in the half-lives of alpha radiators, from 10^9 years to 10^{-7} s, at a comparatively small increase in the energy of the alpha particles, from 4 to 9 MeV.

42.5.3 The decay constant λ (42.4.4) is related to the penetrability D of the potential barrier for alpha particles. For a simplified model of a potential barrier of length L (38.7.2)

$$\lambda = Dn,$$

where n is the number of collisions of the alpha particle with the barrier wall in unit time, equal to $n = v/2L$, and $v = \sqrt{2E/m_\alpha}$ is the velocity of the alpha particle in the nucleus. The quantity L is commonly taken equal to the radius R of the nucleus ($2R$ is the width of the potential barrier).

The simplified formula for the alpha-radioactive decay constant is

$$\lambda = \frac{1}{2R} \sqrt{\frac{2E}{m_\alpha}} \exp \left[-\frac{2}{\hbar} \sqrt{2m_\alpha \left(U_0 - E\right)} \, R \right].$$

42.5.4 The *Geiger-Nuttall rule*, or *relation*, states that the larger the decay constant λ of a radioactive element, the longer

the range R_α (path length) of the alpha particles it emits in air. Thus

$$ln\,\lambda = A + B \ln R_\alpha,$$

where A and B are empirical constants, having different values for the various radioactive families (42.4.7).

42.5.5 On the basis of data, which is given in Sects. 42.5.3 and 42.5.4, it has been shown that the same alpha-radioactive element has several groups of alpha particles with different ranges (path lengths). Constancy of the ranges is observed within each group. It follows that the alpha particles emitted by the nuclei have a definite energy spectrum and, consequently, the atomic nuclei have *discrete energy levels*.

42.6 Beta Decay

42.6.1 The concept of beta decay unites three kinds of nuclear transformations: *electron* decay (β^-), *positron* decay (β^+) and *electron capture* (*e* or *K* capture). The first two kinds of transformations consist in the emission by the nucleus of an electron (or positron) and an electron antineutrino (or electron neutrino). These processes involve the transformation of one kind of nucleon in the nucleus into another: a neutron into a proton or a proton into a neutron. The transformations occur as follows:

$${}_0^1n \to {}_1^1p + {}_{-1}^0e + {}_0^0\tilde{\nu}_e \quad (\beta^- \text{ decay}),$$

$${}_1^1p \to {}_0^1n + {}_{+1}^0e + {}_0^0\nu_e \quad (\beta^+ \text{ decay}).$$

Here ${}_0^1n$ and ${}_1^1p$ are the symbols for the neutron and proton, ${}_{-1}^0e$ and ${}_{+1}^0e$ are the symbols for the electron and positron, and ${}_0^0\nu_e$ and ${}_0^0\tilde{\nu}_e$ are the symbols for the electron neutrino and antineutrino (42.4.2).

In the case of *e* capture, the proton is transformed into a neutron as follows:

$${}_1^1p + {}_{-1}^0e \to {}_0^1n + {}_0^0\nu_e.$$

This procedure consists in the disappearance of one of the electrons from the *K* shell of the atom, which is closest to the nucleus. The proton, in being transformed into a neutron, seems to capture an electron. This gave rise to the term electron cap-

ture (or e capture). A feature of this kind of beta decay is the emission from the nucleus of only a single particle, the neutrino $^0_0\nu_e$. An example of e capture is the transformation of the radioactive nucleus of beryllium $^7_4$Be into a stable nucleus of lithium $^7_3$Li. In contrast to $\beta^{\pm}$ decay, electron capture is accompanied by characteristic X radiation, belonging to the K line of the corresponding element (39.7.9).

42.6.2 β^- decay occurs in both natural and artifical radioactive nuclei; β^+ decay is typical of only artificial radioactivity, i.e. the initiation of intrinsic radioactive radiation of nuclei due to the action of alpha particles, neutrons or other particles. This violates the conditions of stability of the atomic nucleus (42.2.5). For example, the artificially radioactive isotope of carbon $^{14}_6$C is obtained from the stable nucleus of nitrogen by the action of neutrons and with the emission of protons:

$$^{14}_7\text{N} + ^1_0\text{n} \rightarrow ^{14}_6\text{C} + ^1_1\text{p}$$
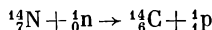
and, after undergoing beta decay, is again transformed into the stable isotope $^{14}_7$N. Thus

$$^{14}_6\text{C} \rightarrow ^{14}_7\text{N} + ^{\ 0}_{-1}\text{e} + ^0_0\tilde{\nu}_e.$$
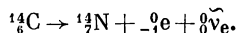
Violation of the stability conditions (42.2.5) by introducing surplus protons into the nucleus leads to artificial β^+ decay. This is evident from the following example:

$$^{10}_5\text{B} + ^4_2\text{He} \rightarrow ^{14}_7\text{N} \rightarrow ^{13}_7\text{N} + ^1_0\text{n},$$

$$^{13}_7\text{N} \rightarrow ^{13}_6\text{C} + ^0_{+1}\text{e} + ^0_0\nu_e,$$
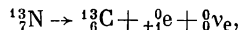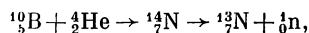
where $^0_{+1}$e is a positron and $^0_0\nu_e$ is an electron neutrino.

42.6.3 Natural β^- decay takes place so that the neutron $^1_0$n is spontaneously transformed into a proton. The rest energy (5.7.3) of the neutron exceeds that of a hydrogen atom (i.e. the proton and electron taken together) by 782 keV. Hence, a transformation of the type (42.6.1)

$$^1_0\text{n} \rightarrow ^1_1\text{p} + ^{\ 0}_{-1}\text{e} + ^0_0\tilde{\nu}_e$$
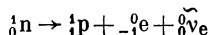
is feasible with respect to energy outside the nucleus as well. The *radioactive decay of free neutrons* has been observed in high-intensity streams of neutrons released in nuclear reactors.

This decay takes place with a half-life (42.4.5) of $(1.01 \pm 0.03) \times 10^3$ s. In heavy nuclei, overloaded with neutrons, such transformations lead to β^- natural radioactivity.

Transformations (42.6.1) of the type $_1^1p \rightarrow {}_0^1n + {}_{+1}^0e + {}_0^0\nu_e$ are only possible in nuclei where the required energy is borrowed from neighbouring particles. This transformation leads to artificial β^+ decay.

42.6.4 The half-lives of beta decays (42.4.5) vary for various sources of $\beta^\pm$ radioactive radiation in an extensive time range from 10^{-2} s to 10^{12} years, which is incommensurably long compared to nuclear time (10^{-22} to 10^{-23} s) (43.2.6). This indicates that beta decay is due to weak interaction (43.2.8).

42.6.5 The decisive experimental fact in understanding the mechanism of β^- decay and establishing a suitable theory was
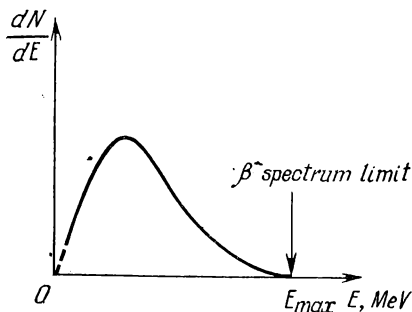
Fig. 42.2

an investigation of the energy spectrum of the emitted electrons. This spectrum was found to be continuous, extending to $E = E_{max}$ (Fig. 42.2). The energy E_{max} is called the *upper energy limit* of the β^- spectrum; it is a characteristic of the source of β^- radioactive radiation. Electrons with an energy exceeding E_{max} cannot be obtained with the given source.

42.6.6 To reconcile the continuity of the electron energy spectrum with the discrete nature of the energy levels of the nucleus (42.5.5), it is necessary to assume that, together with the elec-

tron $_{-1}^{0}$e, another particle, the electron antineutrino* $_{0}^{0}\widetilde{\nu}_e$, is emitted from the nucleus. The total energy lost by the nucleus in a beta decay is equal to E_{max}, but it is variously distributed between the electron and electron antineutrino**. In particular, the limit point of the curve in Fig. 42.2 means that all the energy of the β^- decay is carried away by the electron. Zero energy of the electron on the curve would correspond to all the energy being carried away by the antineutrino $_{0}^{0}\widetilde{\nu}_e$.

For the β^- radioactivity of free neutrons $E_{max} = 782$ keV, which fully corresponds to that discussed in Sect. 42.6.3.

42.6.7 Neither the mass number A nor the spin of the nucleus (42.1.5) changes in a beta decay. The conservation of spin is explained by the fact that the spin of the electron antineutrino is $\hbar/2$ (see Table 43.1), and the combined spins of this particle and of the electron, also equal to $\hbar/2$, are equal to zero when the spins of the two particles are oppositely oriented.

42.7 Gamma Rays

42.7.1 *Gamma rays* constitute hard electromagnetic radiation whose energy is released in nuclear transitions from excited energy states to the ground or less excited states (42.5.5), as well as in nuclear reactions. In the first case, according to the Bohr frequency condition (39.1.6), the energy of a photon of gamma rays is equal to the difference in the energies of the final and initial energy levels of the nucleus:

$$\hbar\nu_{ih} = E_i - E_h = \Delta E_{ih},$$

where ν_{ih} is the frequency of the photon corresponding to the transition of the nucleus from the state with energy E_i to the state with energy E_h. The magnitude of ΔE_{ih} is of the order of 0.1 MeV and considerably exceeds the difference of the energies of the electron levels in the atom (39.1.5). Gamma rays are extremely short-wave electromagnetic radiation of a wavelength not exceeding 10^{-2} nm, i.e. 0.1 Å. The discrete line spectrum of

---

* Antiparticles are discussed in Sect. 43.5.1.
** It is evident that E_{max} determines the difference ΔE_{ih} in the energies of two levels of the nucleus.

gamma rays is confirmation of the aforesaid (see Sect. 42.7.4).

42.7.2 Gamma radiation is not an independent type of radio-activity (42.4.2). It accompanies the processes of α and β decays and causes no changes in the charge or mass number of nuclei. It has been established that gamma rays are emitted by a daughter nucleus (42.4.3) which turns out to be excited at the instant it is formed (42.7.4). The excited nucleus loses its energy during a time ranging from 10^{-13} to 10^{-14} s, which is considerably less than the lifetime of the excited atom (approx. 10^{-8} s).

42.7.3 The origin and properties of gamma rays are confirmed by the laws of *internal conversion* of gamma rays, which consists in photoelectric-effect phenomena (37.1.1) of the electrons of the atom's inner shells as a result of gamma radiation of its nucleus. Electrons formed by such an internal photoelectric effect are said to be *conversion electrons*.

In many cases all the energy of the gamma rays is expended on internal conversion phenomena, and only conversion electrons are observed instead of gamma radiation. The energy eV_n of a conversion photoelectron is related to the energy $h\nu$ of a photon of gamma rays by Einstein's photoelectric emission equation (37.1.4):

$$eV_n = h\nu - \Phi_n,$$

where Φ_n is the work function of an electron from the nth electron shell of the atom, numerically equal to the energies of electrons located at definite energy levels in the atom (39.1.5). These energies are known from data on characteristic X-ray spectra of atoms (39.7.9). A gamma photon of the energy $h\nu$ can remove an electron from any inner shell (K, M, etc.) of the atom (39.3.6). Accordingly, the energies eV_n of the conversion electrons are expressed by

$$eV_L = h\nu - \Phi_L, \quad eV_M = h\nu - \Phi_M, \quad eV_N = h\nu - \Phi_N, \text{ etc.}$$

where Φ_L, Φ_M, Φ_N, etc. are the X ray energy levels of the atom.

42.7.4 The result of internal conversion (42.7.3) is the loss of electrons from the inner shells of the atom and, consequently, the provision of conditions for producing the X ray characteristic spectrum (39.7.9). Internal conversion is accompanied by the emission of characteristic X rays. A measurement of the

energy of the conversed electrons and the data on the quanti-
ties Φ_L, Φ_M, Φ_N, etc. (42.7.3) enabled it to be shown that the
nucleus can emit a definite series of monochromatic gamma rays,
i.e. that gamma rays have a line spectrum. Besides, these data
showed that gamma rays are emitted by a daughter nucleus
rather than a parent one.

42.7.5 Gamma rays have a very strong effect on matter, in
particular, on biological substances. The effect of gamma rays
and other kinds of ionizing radiation is evaluated by the absor-
bed *radiation dose D*, the ratio of the energy of radiation to the
mass of the irradiated matter. The unit of the dose is J/kg, which
is the dose of absorbed radiation at which a mass of 1 kg of irra-
diated matter receives 1 joule of energy of ionizing radiation*.
This unit is called the *gray* (Gy). Also used is the off-system
unit called the rad: 1 rad $= 10^{-2}$ Gy.

The *dose rate N of radiation* is the absorbed dose D referred to
unit time. Thus

$$N = D/t.$$

The unit of dose rate is a watt per kilogram: W/kg $=$ Gy/s.

42.7.6 The energy characteristic of radiation, evaluated on
the basis of the ionization of dry atmospheric air, is the *exposure
dose D_e*. It is measured in coulombs per kilogram (C/kg). One
coulomb per kilogram of exposure dose of X or gamma radiation
is the amount at which the sum of the electric charges of ions
of the same sign, produced by the electrons released by irradia-
ted air of 1 kg mass, fully utilizing the ionizing capacity, is
equal to one coulomb. An off-system unit of exposure dose is the
roentgen R, with 1 R $= 2.58 \times 10^{-4}$ C/kg.

With an exposure dose equal to 1 R, a total charge of ions of
the same charge equal to $(1/3) \times 10^{-9}$ C is obtained in 10^{-6} m³
of dry air at standard atmospheric pressure.

The *exposure dose rate N_e* is the exposure dose per unit time:
$N_e = D_e/t$, measured in amperes per kilogram (A/kg). One
unit of exposure dose rate is the electromagnetic radiation at
which the exposure dose increases by one coulomb per kilogram
in one second. The off-system units of the exposure dose rate

---

* Owing to the specific nature of information on the biological
effect of gamma rays, the pertinent units of physical quantities
are given here instead of in the Appendix.

are: 1 R/s = 2.58×10^{-4} A/kg, 1 R/min = 4.30×10^{-6} A/kg and 1 R/h = 7.17×10^{-8} A/kg.

42.7.7 A radiation dose can also be evaluated on the basis of its biological effect. The unit applied for this purpose is called the *roentgen equivalent man* (rem). This is the absorbed energy of radiation that produces the same biological effect as that produced by one roentgen. Thus

1 rem = 10^{-2} J/kg.

The safe dose rate for humans is considered to be one about 250 times that produced by the background of space and by radioactive radiation from the earth's interior.

42.8 Mössbauer Effect

42.8.1 All excited energy levels of the nucleus have energy values determined to an accuracy of ΔE, which is determined from the indeterminacy relation (38.6.6). Thus

$$\Delta E \approx \frac{\hbar}{\Delta t} ,$$

where Δt is the lifetime of the nucleus in the excited state. Only for the ground state of a stable nucleus is $\Delta t = \infty$ and $\Delta E = 0$, i.e. the energy value of the nucleus is exactly equal to E^*. For example, a nucleus of iridium $^{191}_{77}$Ir goes over from the excited state with the energy $E = 129$ keV to the ground state, emitting a gamma photon, during the time Δt, which can be taken equal to the half-life (42.4.5) $T = 10^{-10}$ s. The energy indeterminacy ΔE is found to equal $\Delta E \approx 5 \times 10^{-6}$ eV. The finite lifetime of the excited energy states of the nucleus leads to a lack of monochromaticity of the gamma radiation that accompanied the transition of the nucleus from the excited to the ground state. This lack of monochromaticity is called the *natural linewidth* of gamma radiation, and the indeterminacy

* This is equally true for the energy levels of electrons in atoms.

ΔE of the energy of an excited state is called the *natural width* Γ *of the energy level of the nucleus**.

42.8.2 *Resonance absorption* of gamma rays by nuclei implies the absorption by the nuclei of gamma photons of such frequency ν that the energy $h\nu$ of a photon is equal to the difference in the energies of one of the excited states and the ground state of the nucleus. The name means that the line of the gamma photon, emitted in the transition of the nucleus from an excited state to its ground state, has the same frequency ν.

When gamma photons are emitted or absorbed by a nucleus, the *recoil of the nucleus* is taken into account. In a transition of a nucleus from an excited state with the energy E to the ground state (of an energy taken equal to zero), the gamma photon acquires the energy E_{ph} equal to

$$h\nu_{em} = E_{ph} = E - E_N < E,$$

where E_N is the nuclear recoil energy.

In exciting the nucleus and its transition from the ground state to the state with the energy E, the gamma photon should have the energy E'_{ph}, equal to $h\nu_{abs} = E'_{ph} = E + E_N > E$.

The frequencies at the maxima of the emission lines (ν_{em}) and absorption lines (ν_{abs}) are shifted with respect to one another by the amount

$\nu_{abs} - \nu_{em} = \Delta\nu$, which is such that $h\Delta\nu = 2E_N$.

The energy E_N of nuclear recoil can be determined from the momentum $\mathbf{p}_{ph}$ of the photon (37.2.2), which, in the processes of emitting and absorbing a gamma photon, should equal the momentum of the nucleus ($\mathbf{p}_{ph} = \mathbf{p}_N$):

$$E_N = \frac{p_N^2}{2M_N} = \frac{p_{ph}^2}{2M_N} = \left(\frac{h\nu}{c}\right)^2 \frac{1}{2M_N},$$

where M_N is the mass of the nucleus. For the nucleus ${}^{191}_{77}\text{Ir}$, with the energy $E = 129$ keV of its excited state, calculations yield $E_N = 0.05$ eV and a shift $\Delta\nu = 2E_N/h$ of the maxima of

---

* This definition also holds for transitions of atoms from an excited to the ground state. But here the lack of monochromaticity is called the *natural spectral linewidth*, and the indeterminacy is the *natural width of energy levels of electrons* in the atom.

emission and absorption lines. At this, $h\Delta\nu = 0.1$ eV, which considerably exceeds the natural width of level Γ (42.8.1).

42.8.3 A drastic reduction in the energy of nuclear recoil in absorbing and emitting gamma rays is evident when these processes are observed in nuclei in the crystal lattice, i.e. in the bound state. Under these conditions, the momentum and energy are imparted to the crystal lattice as a whole, rather than to the single nucleus that emits (or absorbs) a gamma photon. The mass of the crystal is substantially greater than that of a nucleus, and the loss of energy E_N in emission or absorption of gamma rays becomes extremely small. In this case, resonance absorption and emission of a gamma photon of a strictly definite frequency ν is observed and the linewidth is comparable to the natural width (42.8.1). The phenomenon of resonance absorption (or emission) of gamma rays without recoil is called the *Mössbauer effect*.

42.8.4 The Mössbauer effect is made use of in *nuclear spectroscopy* to make precise measurements of the energy levels of atomic nuclei. In the gamma transition in ^{57}Fe nuclei, for instance, with the energy of the transition $E = 14.4$ keV, the change in the energy of the level is determined with an accuracy to $\Gamma/E = 3 \times 10^{-13}$. For the gamma transition in ^{67}Zn with the energy of the transition $E = 93$ keV, the quantity Γ/E was found to equal 5×10^{-16}.

The Mössbauer effect has been used to check the conclusion on shift of spectral line frequencies in a gravitational field (6.2.1). As a photon travels in a gravitational field its energy is changed by the amount $\Delta E = -m(\varphi_2 - \varphi_1)$, where $m = h\nu/c^2$ is the mass of the photon (37.2.1), and φ_1 and φ_2 are the potentials of the gravitational field at points 1 and 2 (6.2.4). The minus sign indicates that the increase in the energy of a photon in a gravitational field is at the expense of its own energy $E = h\nu$:

$$h\Delta\nu = -m\Delta\varphi.$$

The relative change in frequency of a photon as it passes through a gravitational potential difference $\Delta\varphi$ is

$$\frac{\Delta\nu}{\nu} = -\frac{\Delta\varphi}{c^2},$$

Here $\Delta\varphi > 0$ because the potential of zero gravity of the sun increases with the distance from the sun. At the earth's surface

it is greater than at that of the sun. Consequently, $\Delta\nu/\nu < 0$ and all frequencies of spectral lines of the sun and stars, recorded on the earth, are shifted toward the red end of the spectrum. This effect is called the *gravitational red shift*.

The Mössbauer effect enabled the gravitational frequency shift of the gamma photon to be observed in its motion in the earth's gravitational field. In vertical motion from the floor to the ceiling of the laboratory to a height of about 10 m, the relative change in frequency is

$$\left|\frac{\Delta\nu}{\nu}\right| = \frac{\Delta\varphi}{c^2} = \frac{g\,\Delta h}{c^2} \approx 10^{-15},$$

where g is the acceleration due to gravity (7.3.3). To record this frequency shift it is necessary to provide for resonance absorption of gamma photons in such manner that the source and detector of the gamma rays have a relative linewidth equal to or less than 10^{-15}. Then there will be no absorption if the frequency of the gamma photon incident on the nucleus differs from the frequency of the photon that the nucleus can absorb by the amount $\Delta\nu = 10^{-15}\,\nu$.

This experiment was conducted with two identical crystal sources of gamma rays, one of which could be placed at a height of 20 m above the other. When the detector of gamma photons was at the same height as the source of gamma photons, resonance absorption occurred. As the detector was raised 20 m the absorption ceased owing to the gravitational shift in frequency. The Doppler effect (30.6.1) was employed to restore absorption. At a definite velocity at which the detector and source approach each other, the Doppler increase in frequency exactly compensated for its gravitational reduction, and resonance absorption was restored. This experiment was a confirmation under laboratory conditions of the gravitational red shift.

42.9 Nuclear Reactions

42.9.1 *Nuclear reactions* are transformations of atomic nuclei brought about by their interactions with one another or with elementary particles (43.1.1). As a rule, nuclear reactions involve two nuclei and two particles. One pair, a nucleus and a particle, is said to be the *initial*, and the other, the *final* pair.

Nuclear reactions are represented in one of the following ways:

$A + a \rightarrow B + b$ or $A\,(a,\ b)\,B,$
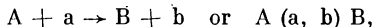
where A and B are the initial and final nuclei, and a and b are the initial and final particles in the reaction. In many cases a nuclear reaction can proceed in more than one way, the preceding reaction being $A + a \rightarrow C + c$, i.e. $A\,(a,\ c)\,C$, or some other version. The possible ways in which a nuclear reaction may proceed are called its *channels*. The initial phase of a reaction is called its *entrance channel*.

42.9.2 A nuclear reaction is characterized by its energy Q (called the *nuclear reaction energy*), which is equal to the difference between the energies of the final and initial pairs in the reaction (42.9.1). When $Q < 0$, the reaction proceeds with the absorption of energy and is said to be *endothermic* (or *endoenergic*); when $Q > 0$, the reaction proceeds with the release of energy and is said to be *exothermic* (or *exoergic*). An endothermic nuclear reaction is found to be feasible at a certain minimum (threshold) kinetic energy E_{th} of the nuclei or particles causing the reaction:

$$E_{th} = \frac{M_A + M_a}{M_A}\,Q,$$

where M_A is the mass of the stationary (target) nucleus, and M_a is the mass of the particle (or nucleus) bombarding the target nucleus.

Nuclear reactions comply with the laws of conservation of energy, momentum, electrical charge and mass number. If the kinetic energy of the particles entering a reaction is sufficient for the creation of a nucleon-antinucleon pair (43.5.6), the mass number may change. Besides, nuclear physics has certain conservation laws not found in other branches of physics (see Sects. 43.4.1 through 43.4.4).

42.9.3 The efficiency of a nuclear reaction is determined by the magnitude of the effective cross section σ of the given reaction (reaction cross section). Quantity σ has the dimension of area and characterizes the "yield" of the reaction per particle irradiating the nucleus. Thus

$$\sigma = \frac{dn_0}{n_0 N_0\,dx}\,.$$

It is assumed in this definition that a plane-parallel stream, containing n_0 particles is incident in unit time on unit area of the cross section of a substance containing N_0 nuclei in unit volume; dn_0 is the number of the particles that undergo the nuclear reaction in a layer of the thickness dx.

42.9.4 Depending upon the nature of the interaction of particle a with the target A, distinction is made between *direct interaction*, in which nuclear reactions proceed in a single stage according to some process (42.9.1) and nuclear reactions that proceed in two stages with the formation of a *compound nucleus*. In the first stage the incident particle is captured by the target nucleus and the particle's energy is not imparted to any particular nucleon, but is shared equally by all the particles of the compound nucleus. Thus none of the particles gains sufficient energy to be ejected from the nucleus*. The compound nucleus is dealt with as an excited statistical system of particles, having random motions similar to those of particles in a drop of liquid. Such rapid redistribution of energy between the particles of the nucleus is possible only upon frequent collisions of the particles, this being typical for the redistribution of energy between the particles in a drop of liquid. As the result of chance deviations from uniform distributions of the excitation energy amongst the particles of the compound nucleus, sufficient energy is concentrated on one of them to be ejected from the nucleus. This second stage of the nuclear reaction occurs after a lapse of time (10^7 to 10^8) τ_N following the first stage, where τ_N is the characteristic nuclear time (43.2.6).

The sequence of a nuclear reaction proceeding with the formation of a compound nucleus is as follows:

$$\begin{smallmatrix}A_1\\Z_1\end{smallmatrix}X + a \rightarrow \begin{smallmatrix}A_2\\Z_2\end{smallmatrix}Y \rightarrow \begin{smallmatrix}A_3\\Z_3\end{smallmatrix}C + b,$$
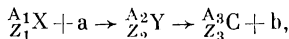
where $\begin{smallmatrix}A_1\\Z_1\end{smallmatrix}X$ is the initial target nucleus, a is the incident particle, $\begin{smallmatrix}A_2\\Z_2\end{smallmatrix}Y$ is the compound nucleus, $\begin{smallmatrix}A_3\\Z_3\end{smallmatrix}C$ is the product nucleus and b is the particle ejected from the nucleus as a result of the reaction. If a $=$ b, particles are being scattered by the nucleus (scattering is elastic or inelastic according to whether the energy of the particles is the same or different before and after scattering).

---

* The "missile" captured by the nucleus, for instance, an alpha particle or deuteron, can consist of several particles.

If $b \neq a$ a nuclear reaction is proceeding in the true sense of the term.

42.9.5 Nuclear reactions are classified according to various of their features: according to the energy of the particles that initiate them, according to the energy of the particles participating in them, and according to the nature of the nuclear transformations that occur in them. Nuclear reactions at low energies (of the order of 1 eV) proceed mainly with the participation of neutrons. Reactions at medium energies (up to several MeV) are induced, in addition to neutrons, by various charged particles (alpha particles, protons, deuterons, nuclei of heavy hydrogen, etc.), as well as gamma photons. Charged particles that cause nuclear reactions may be multiply charged ions of heavy chemical elements, as well as charged particles accelerated in accelerators. Reactions at high energies (hundreds and thousands of MeV) lead to the creation of elementary particles (mesons, hyperons, etc.) that are absent in the free state.

42.9.6 The following are examples of nuclear reactions initiated by alpha particles and deuterons ^2_1D:

(a) historically, the first nuclear reaction for the transformation of nitrogen into oxygen was the following:

$$^{14}_{7}N + ^4_2He \rightarrow (^{18}_9F) \rightarrow ^{17}_8O + ^1_1p,$$
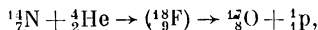
or in shorter notation:

$$^{14}_{7}N\ (\alpha,\ p)\ ^{17}_8O;$$
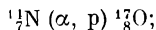
(b) the nuclear reaction in which neutrons were first obtained was:

$$^9_4Be + ^4_2He \rightarrow ^{12}_6C + ^1_0n;$$
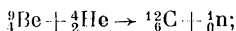
(c) nuclear reactions initiated by deuterons ^2_1D, the nuclei of heavy hydrogen (deuterium), can lead to the fusion of heavy nuclei, tritium ^3_1H, or the light isotope of helium ^3_2He, with the formation of a proton or neutron:

$$^2_1D + ^2_1D \rightarrow ^3_1H + ^1_1p \quad \text{and} \quad ^2_1D + ^2_1D \rightarrow ^3_2He + ^1_0n$$
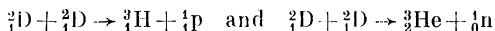
(see also 42.9.13).

42.9.7 The action of neutrons ^1_0n leads to the formation of artificial radioactive isotopes, such as radiocarbon $^{14}_6C$ with a half-life (42.4.5) exceeding 5000 years:

$$^{14}_7N + ^1_0n \rightarrow ^{14}_6C + ^1_1p. \tag{*}$$

Subsequent decay proceeds according to the process

$$^4_6C \rightarrow ^{14}_7N + _{-1}^0e + ^0_0\widetilde{v}_e,$$
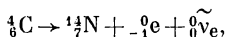
where $_{-1}^0e$ and $^0_0\widetilde{v}_e$ represent the electron and electron antineu-
trino. The long half-life of $^{14}_6C$ is the basis for the radiocarbon
(carbon-14) dating technique used in archaeology. By deter-
mining the relative amount of undecayed radioactive carbon,
which ceased to be accumulated in a living organism when it
died, it is possible, according to reaction (∗), to establish the
time when the organism no longer absorbed the isotope $^{14}_6C$ from
the atmosphere. This isotope is formed in the atmosphere from
nitrogen by the action of cosmic-ray neutrons.

42.9.8 *Slow neutrons* are ones whose de Broglie wavelength
(38.1.4) exceeds the nuclear radius R: $\lambda \gg \hbar/mv$ and, accordingly,
$v \ll \hbar/mR$. Their energy does not exceed 100 keV. Neutrons
with energies up to 0.5 eV are called *thermal neutrons*. As such
neutrons pass through matter they undergo *capture*, the absorp-
tion of the neutrons by the nuclei. If the energy of the neutrons
coincides with that of the compound nucleus (42.9.4), *resonance
absorption* (resonance capture) of the neutrons occurs. This
process is the basis of the technique for obtaining *transuranic
chemical elements*. The transuranic element neptunium $^{239}_{93}Np$
is formed upon resonance capture of neutrons by the most abun-
dant isotope of uranium $^{238}_{92}U$ according to the reaction

$$^{238}_{92}U + ^1_0n \rightarrow ^{239}_{92}U \xrightarrow[\text{23 min}]{\beta^-} {}^{239}_{93}Np,$$
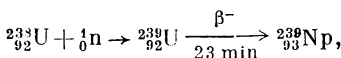
where the half-life of radioactive isotope $^{239}_{92}U$ is indicated,
which is transformed into $^{239}_{93}Np$. Next, the nucleus of isotope
$^{239}_{93}Np$ is transformed into plutonium $^{239}_{94}Pu$:

$$^{239}_{93}Np \xrightarrow[\text{2.3 days}]{\beta^-} {}^{239}_{94}Pu.$$
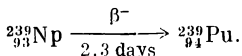
Owing to its efficient fission by the action of thermal neutrons
(42.9.10), plutonium plays a prominent role in the production
of nuclear energy. Plutonium $^{239}_{94}Pu$ is radioactive by alpha par-
ticle emission and has a huge half-life (24 000 years). It is trans-

formed into the stable isotope of uranium $^{235}_{92}$U according to the reaction:

$$^{239}_{94}\text{Pu} \xrightarrow[2.4\times10^4 \text{ years}]{\alpha} {}^{235}_{92}\text{U}.$$

42.9.9 Nuclear reactions using accelerated nuclei of chemical elements enabled advancement in Mendeleev's table up to the element with the number $Z = 107$. The element with $Z = 105$ was named nilsbohrium (Ns). Upon bombarding the isotope $^{242}_{94}$Pu with nuclei of neon $^{22}_{10}$Ne a chemical element with $Z = 104$ (kurchatovium $^{260}_{104}$Ku) is obtained. After capture of the nucleus $^{22}_{10}$Ne by plutonium, the compound nucleus of kurchatovium $^{264}_{104}$Ku is formed. Then, in one out of ten thousand million compound nuclei, the nucleus of the element $^{260}_{104}$Ku was produced after emitting four neutrons. Thus

$$^{242}_{94}\text{Pu} + {}^{22}_{10}\text{Ne} \to {}^{264}_{104}\text{Ku} \to {}^{260}_{104}\text{Ku} + 4{}^1_0\text{n}.$$
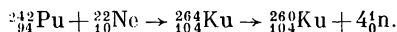
In the most powerful beams of accelerated neon nuclei, one nucleus of kurchatovium is created in several hours.

42.9.10 A heavy compound nucleus, excited in the resonance capture of a neutron, can divide into two approximately equal parts (*fission reaction of heavy nuclei*). The parts that are formed are called *fission fragments*. The instability of heavy nuclei against fission is due to the large number of protons they contain which are subject to Coulomb repulsion from one another.

The fissure of a heavy nucleus into two fragments is accompanied by the evolution of a huge amount of energy. This follows from a comparison of the specific binding energies of nuclei of elements located in the middle and at the end of Mendeleev's periodic system. The energy released in the fission reaction is proportional to the difference in the specific binding energies of an unstable "friable" nucleus and two "tightly packed" stable fission fragments. This difference constitutes 1.1 MeV per nucleon (42.2.4). In the fission of a nucleus of uranium $^{238}_{92}$U, containing 238 nucleons, the evolved energy should be of the order of 200 MeV. In the fission of nuclei contained in 1 gram of uranium $^{235}_{92}$U, the energy produced is 8×10^{10} J or 22 000 kWh. Heavy nuclei are capable of fission if they comply with the condition: $Z^2/A \geqslant 17$, where Z^2/A is called the *fission parameter*. This condition is complied with for all nuclei beginning with

silver $^{108}_{47}\mathrm{Ag}$ for which $Z^2/A \approx 20$. Nuclei for which $(Z^2/A)_{\mathrm{cr}} \geqslant$ $\geqslant 49$ (the *critical fission parameters*) are entirely unstable with respect to fission and cannot exist. For kurchatovium $(Z = 104)$ (42.9.9). $Z^2/A \approx 41$. *Spontaneous nuclear fission* is possible at values $Z^2/A < (Z^2/A)_{\mathrm{cr}}$, which proceeds similar to radioactivity by alpha-particle emission (42.5.1), employing the tunnel effect (38.7.2). The half-life (42.4.5) for spontaneous nuclear fission is from 10^{16} to 10^{17} years.

42.9.11 At the instant they are formed fission fragments have a neutron excess, i.e. the number of neutrons are in excess of the number of protons. Excess neutrons, emitted by the fragments, are called *fission neutrons*. Their number may vary and the nuclear fission process is accompanied by neutron multiplication, characterized by the average number $\langle \nu \rangle$ of emitted neutrons per fission event. For nuclei of plutonium $^{239}_{94}\mathrm{Pu}$ and uranium $^{235}_{92}\mathrm{U}$, in which fission is due to thermal neutrons, $\langle \nu \rangle$ equals 3.0 and 2.5, respectively. Fission neutrons include *prompt* (*secondary*) and *delayed neutrons*. Prompt neutrons are emitted coincident with nuclear fission during a time of the order of 10^{-14} s. Delayed neutrons are emitted by the fission products at some time after fission occurs.

42.9.12 If each of the prompt neutrons emitted in a fission reaction (42.9.11) interacts with neighbouring nuclei of the fissible substance, inducing fission in them, there will be an avalanche-like build-up of acts of nuclear division. This is called a *chain fission reaction*. A condition for the initiation of a nuclear chain reaction is the availability of neutron multiplication.

The *neutron multiplication factor* k is the ratio of the number of neutrons emitted at a given stage of the reaction to the number of such neutrons in the preceding stage. The condition for a self-sustaining reaction is $k \geqslant 1$. The practical feasibility of nuclear fission chain reactions has been proved by the development of *nuclear power engineering*, a branch of engineering in which various types of *nuclear reactors* have been designed and developed. Such reactors are devices in which controlled nuclear chain reactions are carried out.

42.9.13 The reaction of the fusion of nuclei of tritium or helium from nuclei of deuterium, discussed in item (c) of Sect. 42.9.6, is a second method of producing intranuclear energy in addition to the fission of heavy nuclei. The specific binding energies (42.2.4) in three nuclei—$^2_1\mathrm{D}$, $^3_1\mathrm{H}$ and $^3_2\mathrm{He}$—are approximately

in the ratio 1 : 3 : 6. This means that the nuclear reactions discussed in Sect. 42.9.6 are accompanied by the evolution of great amounts of energy: 4.04 MeV in the first reaction and 3.27 MeV in the second. Still more energy, 17.58 MeV, is released in the reaction

$$^2_1D + {}^3_1H \rightarrow {}^4_2He + {}^1_0n.$$
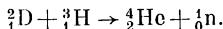
Per particle this energy is equal to 17.6/5 MeV = 3.5 MeV, i.e. approximately four times as much as in the fission of uranium $^{238}_{92}U$ (42.9.10): 200/238 MeV = 0.85 MeV. Even more efficient with respect to the specific evolution of energy is the fusion of nuclei of helium ^4_2He from four protons. This energy amounts to 6.70 MeV per particle.

42.9.14 The fusion reaction of light nuclei, requiring the overcoming of the potential energy of their repulsion, can be efficiently carried out at super-high temperatures of the order of 10^8 or 10^9 K, which exceeds the temperature in the central regions of the sun ($T = 1.3 \times 10^7$ K). Such reactions are called *thermonuclear reactions* and they take place in matter in a plasma state (22.6.1). Thermonuclear reactions are evidently the sources of energy of the stars, compensating for their loss of energy through radiation. The sun radiates 3.8×10^{26} J of energy per second. This corresponds to the evolution of energy per unit mass per second of only 1.88×10^{-4} J/s-kg. This is only one per cent of the specific energy production of a living organism in the process of metabolism.

It is thought that thermonuclear reactions take place on the sun in the form of *thermonuclear chains*, or cycles, in which energy is released by the transformation of hydrogen nuclei into helium nuclei.

One of the versions of the *proton-proton chain*, or deuterium cycle, begins with combination of two protons to form deuterium with the emission of a positron and an electron neutrino:

$$^1_1p + {}^1_1p \rightarrow {}^2_1D + {}^0_{+1}e + {}^0_0\nu_e.$$
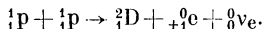
The chain continues according to the reaction

$$^2_1D + {}^1_1p \rightarrow {}^3_2He + \gamma$$
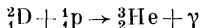
where gamma radiation is emitted, not only as excess energy, but in the combination of the positron with electrons, which

always exist in plasma (43.5.4). A probable continuation of the chain is a reaction with the evolution of energy:

$$_2^3\text{He} + _2^3\text{He} \rightarrow _2^4\text{He} + 2_1^1\text{p},$$
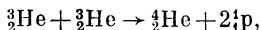
where $_2^4\text{He}$ is the symbol for the alpha particle.

In the *carbon-nitrogen cycle*, the carbon nuclei are "catalysts" in the reaction for the fusion of hydrogen nuclei into a helium nuclei.

At the beginning of the cycle a fast proton penetrates a carbon nucleus:

$$_6^{12}\text{C} + _1^1\text{p} \rightarrow _7^{13}\text{N} + \gamma.$$

The radioactive isotope of nitrogen $_7^{13}\text{N}$, with a half-life of 14 min, decays (42.6.1): $_1^1\text{p} \rightarrow _0^1\text{n} + _{+1}^0\text{e} + _0^0\nu_e$ and a nucleus of a carbon isotope is formed:

$$_7^{13}\text{N} \rightarrow _6^{13}\text{C} + _{+1}^0\text{e} + _0^0\nu_e.$$
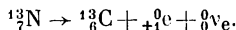
Approximately after each 2.7 million years, a nucleus of $_6^{13}\text{C}$ captures a proton, forming the nucleus of the stable nitrogen isotope $_7^{14}\text{N}$:

$$_6^{13}\text{C} + _1^1\text{p} \rightarrow _7^{14}\text{N} + \gamma.$$
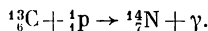
On an average of 32 million years later, the nucleus of $_7^{14}\text{N}$ captures a proton and is transformed into a nucleus of oxygen $_8^{15}\text{O}$:

$$_7^{14}\text{N} + _1^1\text{p} \rightarrow _8^{15}\text{O} + \gamma.$$
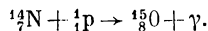
The unstable nucleus of $_8^{15}\text{O}$, with a half-life of 3 min, emits a positron and a neutrino as it is transformed into the nucleus of $_7^{15}\text{N}$:

$$_8^{15}\text{O} \rightarrow _7^{15}\text{N} + _{+1}^0\text{e} + _0^0\nu_e.$$
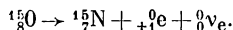
The cycle is completed by a reaction that takes place approximately after 100 thousand years:

$$_7^{15}\text{N} + _1^1\text{p} \rightarrow _6^{12}\text{C} + _2^4\text{He}.$$
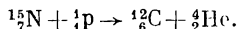
The result of the cycle is the transformation of four protons into a helium nucleus with the emission of two positrons and gamma radiation. The released energy amounts to 26.8 MeV per helium nucleus. This corresponds to an energy output of 700 thousand kWh per gram-atom of helium. This energy is sufficient to com-

pensate for the energy the sun loses in radiation. Various stages of the cycle are separated by tremendous lengths of time when judged according to the time scale on the earth. But this cycle is closed and continuous. Hence, all the stages of the cycle proceed simultaneously on the sun, having begun at various instants of time.

42.9.15 Conditions, close to those prevailing in the interior of the sun, were provided in a *thermonuclear*, or *fusion*, *bomb* (more commonly called the *hydrogen*, or *H*, *bomb*). In this bomb a self-sustaining thermonuclear fusion reaction of an explosive nature is initiated in a mixture of deuterium and tritium. Thus

$$^2_1D + {}^3_1H \rightarrow {}^4_2He + {}^1_0n. \tag{*}$$

The high temperature required for a thermonuclear reaction was obtained by the explosion of an "ordinary" atomic bomb, operating on the principle of a rapid chain reaction in the fission of heavy nuclei.

The theoretical basis for controlled thermonuclear fusion is a reaction of the type

$$^2_1D + {}^3_1H \rightarrow {}^4_2He + {}^1_0n,$$

as well as of the type

$$^2_1D + {}^2_1D \rightarrow {}^3_1H + {}^1_1p \quad \text{or} \quad ^2_1D + {}^2_1D \rightarrow {}^3_2He + {}^1_0n. \tag{**}$$
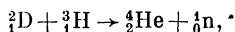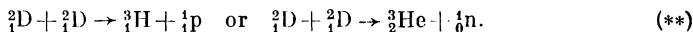
The feasibility of these reactions reduces to the necessity to meet two requirements: the availability of a certain least temperature and a definite restriction of the product $n\tau$:

$n\tau > 10^{16}$ and $T > 10^9$ K for reaction (**),

$n\tau > 10^{14}$ and $T > 10^8$ K for reaction (*).

Here n is the concentration of matter in the plasma, and τ is the time the particles are confined in the plasma. The preceding inequalities indicate that the number of nuclei reacting in thermonuclear fusion should be proportional to the concentration of matter and the time of existence of the nucleus.

The practical realization of controlled thermonuclear fusion reactions is a timely problem of enormous importance and its solution should be reached in the near future.

CHAPTER 43   ELEMENTARY PARTICLES

43.1 Preliminary Information on Elementary Particles

43.1.1 *Elementary*, or *fundamental*, *particles* are ones to which an internal structure that is a simple combination of other stable particles cannot be attributed at the present state of knowledge.

TABLE 43.1

Particles and antiparticles	Symbol	Charge, e		Mass, MeV	Lifetime, s
Photon	γ	0		0	Stable
Leptons *Neutrinos:* electron neutrino and antineutrino	ν_e $\widetilde{\nu}_e$	0	0	0	Stable
mu-meson neutrino and antineutrino	ν_μ $\widetilde{\nu}_\mu$	0	0	0	Stable
Electrons: electron and positron	e^- e^+	-1	$+1$	0.511	Stable
Muons: μ^- meson and μ^+ meson	μ^- μ^+	-1	$+1$	106	2.2×10^{-6}
Mesons: *Pions:* π^+ meson and π^- meson	π^+ π^-	$+1$	-1	140	2.6×10^{-8}
π^0 meson	π^0	0		135	0.8×10^{-16}
Kaons K^+ meson and K^- meson	K^+ K^-	$+1$	-1	494	1.2×10^{-8}

In its interaction with other particles and fields, an elementary
particle behaves as an integral unit. The term "elementary"
particle is essentially of a conditional nature.
43.1.2 In energy regions less than $2m_0c^2$ (where m_0 is the rest
mass of the particle), the structure of particles has no effect on
interaction between them and other particles and fields. In
these cases elementary particles are dealt with as particles in

Spin, $\hbar$	Baryon charge		Strange- ness		Iso- spin	Projection of isospin		Dominant decay mode
1	0							
1/2	0	0	.					
1/2	0	0						
1/2	0	0						
1/2	0	0						$\mu^- \to e^- + \nu_\mu + \widetilde{\nu}_e$
0	0	0	0	0	1	+1	−1	$\pi^+ \to \mu^+ + \nu_\mu$
0		0		0	1		0	$\pi^0 \to 2\gamma$
0	0	0	+1	−1	1/2	+1/2	−1/2	$K^+ \to \mu^+ + \nu_\mu$

TABLE 43.1 (*cont.*)

Particles and antiparticles	Symbol	Charge, e	Mass, MeV	Lifetime, s
K^0 meson, anti-K^0 meson	K^0 $\widetilde{K}^0$	0 0	498	$K_S^0 : 0.86 \times 10^{-10}$ $K_L^0 : 5.38 \times 10^{-8}$
η^0 meson	η^0	0	549	2.4×10^{-19}
Baryons *Nucleons:*				
proton and antiproton	p $\widetilde{p}$	+1 —1	938.2	Stable
neutron and antineutron	n $\widetilde{n}$	0 0	939.6	0.93×10^3
Hyperons:				
Λ^0 hyperon and anti-Λ^0 hyperon	Λ^0 $\widetilde{\Lambda}^0$	0 0	1116	2.5×10^{-10}
Σ^+ hyperon and anti-Σ^+ hyperon	Σ^+ $\widetilde{\Sigma}^-$	+1 —1	1189	0.8×10^{-10}
Σ^- hyperon and anti-Σ^- hyperon	Σ^- $\widetilde{\Sigma}^+$	—1 +1	1197	1.5×10^{-10}
Σ^0 hyperon and anti-Σ^0 hyperon	Σ^0 $\widetilde{\Sigma}^0$	0 0	1192	$< 10^{-11}$
Ξ^- hyperon and anti-Ξ^- hyperon	Ξ^- $\widetilde{\Xi}^+$	—1 +1	1321	1.7×10^{-10}
Ξ^0 hyperon and anti-Ξ^0 hyperon	Ξ^0 $\widetilde{\Xi}^0$	0 0	1315	3×10^{-10}
Ω^- hyperon and anti-Ω^- hyperon	Ω^- $\widetilde{\Omega}^+$	—1 +1	1672	1.3×10^{-10}

Spin, $\hbar$	Baryon charge	Strange-ness	Iso-spin	Projection of isospin	Dominant decay mode
0	0 0	+1 −1	1/2	−1/2 +1/2	$K_S^0 \to 2\pi$ $K_L^0 \overset{\pi\mu\nu}{\underset{\pi e \nu}{\to 3\pi}}$
0	0	0	0	0	$\eta^0 \overset{2\gamma}{\underset{\pi^+\pi^-\pi^0}{\to 3\pi^0}}$
1/2	+1 −1	0 0	1/2	+1/2 −1/2	
1/2	+1 −1	0 0	1/2	−1/2 +1/2	$n \to p + e^- + \widetilde{\nu}_e$
1/2	+1 −1	−1 +1	0	0 0	$\Lambda^0 \to p + \pi^-$
1/2	+1 −1	−1 +1	1	+1 −1	$\Sigma^+ \to N + \pi$
1/2	+1 −1	−1 +1	1	−1 +1	$\Sigma^- \to n + \pi^-$
1/2	+1 −1	−1 +1	1	0 0	$\Sigma^0 \to \Lambda^0 + \gamma$
1/2	+1 −1	−2 +2	1/2	−1/2 +1/2	$\Xi^- \to \Lambda^0 + \pi^-$
1/2	+1 −1	−2 +2	1/2	+1/2 −1/2	$\Xi^0 \to \Lambda^0 + \pi^0$
3/2?	+1 −1	−3 +3	0	0 0	$\Omega^- \overset{}{\underset{\Lambda^0 + K^-}{\to \Xi^- + \pi^0}}$

the sense of material points, but possessing many properties, such as: rest mass, electric charge, spin and a dominant decay mode (Table 43.1). The concept of a point elementary particle meets the requirements of relativity theory. An extended particle of finite dimensions, which is an integral whole, should not be deformable as otherwise independent movements of its parts would be possible. But this would mean that external action on such an elementary particle would be instantaneously transmitted from certain parts to others, and this contradicts relativity theory (5.1.1).

43.1.3. A structureless charged elementary particle has an infinite potential (16.2.6) and infinite potential energy. Up-to-date particle physicists are working out various methods of introducing nonpoint-like interaction. These methods eliminate the "infinity" of point elementary particles, and introduce definite structures for them without contradicting relativity theory.

It follows from the indeterminacy relations (38.6.2) that a study of the structure of elementary particles requires the investigation of processes taking place at the extremely short distances δ that the momentum indeterminacy $\Delta p \geqslant \hbar/\delta$ corresponds to. The shorter the distances δ, associated with the structure of elementary particles, the greater the momentum p of the particles must be, because it cannot be less than Δp. Consequently, the structure of elementary particles can be cleared up only by research involving high energies. For this reason, particle physics, the new branch of physics engaged in the study of elementary particles, is now more commonly called *high-energy physics*.

43.2 Classification of Elementary Particles and Their Interaction

43.2.1 There are several groups of elementary particles which differ in their properties and the nature of the interaction between one another.

The main classification of elementary particles is based on their various rest masses: distinction is made between *leptons* (light particles), *mesons* (intermediate particles) and *baryons* (heavy particles), with the last being further subdivided into *nucleons* and *hyperons*.

43.2.2 As to their electric charge, measured in units of the

elementary charge e (14.1.2), particles may be positive, negative or neutral. The existence of particles with fractional electric charges, called *quarks*, has been predicted.

With respect to their spin (26.1.3), elementary particles are subdivided into ones with integer, half-integer and zero spin (in units of $\hbar$). Depending upon their spin, assemblies of particles comply with different quantum statistics (41.2.1 and 41.2.4).

43.2.3 Elementary particles are either *stable* or *unstable*. This classification characterizes their lifetime and the dominant decay mode. The characteristics of elementary particles are listed in Table 43.1. The rest masses of the particles are given in energy units, as is now customary practice, on the basis of the relation $E = mc^2$ (5.7.2). Since the rest mass of the electron is approximately 0.5 MeV, the mass of a particle in electron mass units can be found by multiplying its mass in MeV by two.

43.2.4 In addition to the elementary particles listed in Table 43.1, there are also a great many particles called *resonances* (*resonons* and *resonance particles*). Their lifetimes are characteristic of strong interaction (43.2.5); they have definite properties, energies and momenta, and definite decay modes. These features enable resonances to be regarded as special particles.

43.2.5 Mainly three of the four known kinds of interaction exist between elementary particles: strong, electromagnetic and weak. Gravitational interaction between elementary particles is usually not considered. Each of the interactions is characterized by its *interaction coupling constant*, determining its comparative magnitude, and its characteristic interaction time (*time constant of interaction*).

43.2.6 *Strong interactions* refer to processes involving baryons, as well as pions and kaons (Table 43.1). Strong interactions result from nuclear forces between nucleons (42.3.1), as well as the processes of creation and decay of mesons and hyperons in nuclear reactions at high energies. Strong interaction is characterized by the *Fermi constant* $\beta = f^2/\hbar c \approx 1$, where the constant f has the dimension of an electric charge. Processes that display strong interaction are said to be fast and have a characteristic *nuclear time* τ_N. In order of magnitude this time equals the Compton wavelength of the nucleon $h/m_p c$ (where m_p is the mass of the nucleon), divided by the velocity of light in free space, i.e. $h/m_p c^2$.

Particles that take part in strong interactions are called *hadrons* (large and heavy particles).

43.2.7 *Electromagnetic interactions* characterize processes that take place between particles that have electric charges. For instance, Coulomb repulsion in nuclei, and the processes of creation and annihilation of electron-positron pairs are associated with electromagnetic interaction. A measure of this interaction is a dimensionless constant called the *fine-structure constant*: $\alpha = e^2/\hbar c = 1/137$. It has been thus named because α determines the magnitude of fine splitting of the atom's energy level into a number of sublevels. This is caused by the effect of the relativistic dependence of mass on velocity (5.6.1) on the electron's energy in the atom. Processes accompanying electromagnetic interaction are called electromagnetic processes and they proceed during a time of the order of 10^{-20} to 10^{-18} s in accordance with the fact that $\alpha/\beta \approx 10^{-2}$ (43.2.6). Electromagnetic interaction is the only one of the interactions of the microcosm that is also manifested in macroscopic phenomena and processes.

TABLE 43.2

Interaction	Relative strength	Interaction time, s
Strong	1	10^{-23} to 10^{-22}
Electromagnetic	1/137	10^{-20} to 10^{-18}
Weak	10^{-14}	10^{-10} to 10^{-8}

43.2.8 *Weak interactions* are processes involving leptons (Table 43.1). Examples of such processes are beta decay of nuclei (42.6.1), and interaction between nuclei and muons, electrons (or positrons), and neutrinos (or antineutrinos).

Weak interaction is characterized by the *weak interaction constant G*, having the dimension of the inverse square of mass. The dimensionless weak interaction constant $g^2 = Gm_p^2 \approx 10^{-5}$, where m_p is the mass of the proton. Weak interaction is the slowest of the interactions manifested in the microcosm and

proceeds during time of the order of 10^{-10} s. A comparison of the various kinds of interaction between elementary particles is given in Table 43.2.

43.3 Certain Information on Various Elementary Particles

43.3.1 Various kinds of elementary particles (Table 43.1) are produced in the interaction of cosmic rays with the nuclei of the nitrogen and oxygen atoms in the atmosphere. Since the development of modern accelerators, the creation, mutual transformations and structure of elementary particles are investigated in the laboratory.

43.3.2 The feasibility of creating new particles in the interaction between protons and nuclei follows from the relation between mass and energy (5.7.2). When an energy of 10^4 GeV is imparted to the proton, approximately 10^4 times more than its rest energy $m_p c^2$ (where m_p is the rest mass of the proton and c is the velocity of light in free space), a collision between the proton and a nucleus leads to splitting of the latter. High kinetic energy is imparted to the fission products and new particles are created, some having a rest mass and others without one.

43.3.3 *Muons* (*mu mesons*) are the electrically charged particles μ^+ and μ^- with a rest mass equal approximately to $200\ m_e$, where m_e is the rest mass of the electron. The main characteristics of muons are listed in Table 43.1. Muons are not stable, and decay according to the modes:

$$\mu^+ \to {}_{+1}^{0}e + {}_{0}^{0}\nu_e + {}_{0}^{0}\tilde{\nu}_\mu \quad \text{and} \quad \mu^- \to {}_{-1}^{0}e + {}_{0}^{0}\tilde{\nu}_e + {}_{0}^{0}\nu_\mu,$$

where ${}_{-1}^{0}e$ and ${}_{+1}^{0}e$ are the electron and positron, ${}_{0}^{0}\nu_e$ and ${}_{0}^{0}\tilde{\nu}_e$ are the electron neutrino and antineutrino, and ${}_{0}^{0}\nu_\mu$ and ${}_{0}^{0}\tilde{\nu}_\mu$ are the muon neutrino and antineutrino.

The *muon neutrino* (or *antineutrino*) is a particle having neither rest mass nor electric charge, and which is emitted in the creation and decay of mesons. These particles differ from the neutrino (or antineutrino) emitted together with electrons in beta-decay processes (42.6.1). Therefore, the two types of neu-

trino (and antineutrino) are designated by different symbols: $_{0}^{0}\nu_{e}$ ($_{0}^{0}\widetilde{\nu}_{e}$) and $_{0}^{0}\nu_{\mu}$ ($_{0}^{0}\widetilde{\nu}_{\mu}$). The spin of the mu mesons equals $\pm\hbar/2$ and the muon decay reaction complies with the conservation laws (42.9.2). The interaction of muons with nuclei is an example of weak interaction (43.2.8), and muons are nuclearly inactive particles.

43.3.4 *Pions (pi mesons)* are neutral (π^{0}) and electrically charged (π^{+} and π^{-}) particles with a rest mass of approximately 300 m_{e}, where m_{e} is the rest mass of the electron. The main characteristics of pions are listed in Table 43.1. Pions, like muons, are unstable and decay according to the modes:

$$\pi^{+} \to \mu^{+} + _{0}^{0}\nu_{\mu}, \quad \pi^{-} \to \mu^{-} + _{0}^{0}\widetilde{\nu}_{\mu} \quad \text{and} \quad \pi^{0} \to \gamma + \gamma.$$

In these decay reactions μ^{+} and μ^{-} are positive and negative muons, $_{0}^{0}\nu_{\mu}$ and $_{0}^{0}\widetilde{\nu}_{\mu}$ are the muon neutrino and antineutrino, and γ is a gamma-ray quantum (photon). Shown schematically
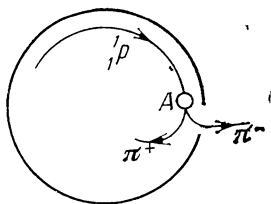
Fig. 43.1   Fig. 43.2

in Fig. 43.1 is the sequence for the (π-μ-e) decays. Pions are spinless particles. This follows from their decay reactions and the requirement that these reactions comply with the conservation laws.

43.3.5 Pions are formed in cosmic rays as the result of the disintegration of the nuclei of atmospheric gas atoms by fast protons and alpha particles. To produce pions the energy of the protons should be about 300 MeV.

The arrangement for obtaining pions in the laboratory is illustrated in Fig. 43.2. Target A of beryllium or carbon is bom-

barded by fast protons $^1_1$p. This knocks pions out of the target
at arbitrary angles. The magnetic field of the accelerator causes
the ejected pions to follow circular paths whose radii are deter-
mined by the velocities of the pions. Pions ejected in the for-
ward direction are divided into two beams, the negative ones
being deflected outside the accelerator chamber and the positive
ones inside. The opposite is true for pions ejected in the back-
ward direction (not shown in Fig. 43.2).

The masses of charged ions are determined by their energies
and momenta by methods employing the deflection of particles
in magnetic and electric fields. The lifetime of pions (Table 43.1)
is established by measuring the time intervals between the
instant of creation of a pion and its instant of decay (43.3.4).
Pions are nuclearly active particles. Their interaction with
nuclei is an example of strong interaction (43.2.6).

Similar methods are applied for investigating other charged
particles. Since neutral particles are not deflected in a magnetic
field, they are investigated by means of the conservation laws
and interaction reactions of the given particle with other parti-
cles. For example, the reaction of a negative pion with a proton
proceeds according to the mode

$$\pi^- + {}^1_1p \rightarrow \pi^0 + {}^1_0n. \quad \cdot$$
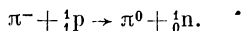
The negative pion is transformed into a neutral one, and the
proton into a neutron. The neutral pion π^0 decays into two gamma-
ray quanta (43.3.4). The mass of the neutral pion (Table 43.1)
is determined from the known masses and energies of the proton,
neutron and negative pion, from the energy of gamma-ray quan-
ta, as well as from data on the momenta of particles in these
transformations.

43.3.6 The principle of applying the law of conservation of
momentum for studying the transformations of elementary par-
ticles can be illustrated by the example of the β^- decay in which
an electron antineutrino $^0_0\tilde{\nu}_e$ is created (42.6.1). If only a single
electron were ejected by the nucleus in a β^- decay, the nucleus
would be subject to recoil in the direction directly opposite to
that of the ejected electron. The momentum of the nucleus
that was stationary before the decay reaction would be equal
in magnitude but opposite in direction to that of the ejected
electron. If, however, in addition to the electron, the nucleus
also ejects an antineutrino, the vector sum of the three momenta,

that of the electron, antineutrino and recoil nucleus, should, according to the law of conservation of momentum (2.7.1), be equal to zero, as before the decay reaction (Fig. 43.3).

43.3.7 *Kaons*, or *K mesons*, are neutral (K^0) and charged (K^+ and K^-) particles with rest masses close to 1000 m_e, where m_e is the rest mass of the electron. As other mesons, kaons are unstable, but, in contrast to the decay reactions of pions and muons (43.3.3 and 43.3.4), the reaction channels of their decays are not unambiguous. As a result of kaon decay, pions, muons and neutrinos (or antineutrino) of both types are formed. Like pions, kaons are spinless. All of these particles (pions and kaons) are united in a single group, the mesons (Table 43.1).
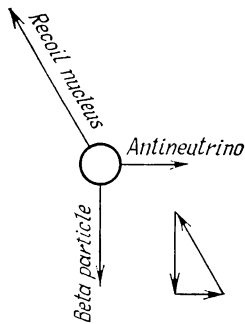
43.3.8 A large group of particles, called *hyperons*, have rest masses greater than that of nucleons, and range approximately from 2183 m_e for the lambda-zero hyperon (Λ^0) to 3280 m_e for the omega-minus hyperon (Ω^-). Hyperons are unstable (see Table 43.1). The lighter hyperons decay into nucleons and pions. The heavier hyperons decay into lighter hyperons, pions and kaons. All hyperons have a spin equal to $\hbar/2$, except the Ω^- hyperon, whose spin is equal to $3\hbar/2$. Nucleons (protons and neutrons) and hyperons are related particles and they belong to a single class of heavy particles called baryons (Table 43.1).

Fig. 43.3

43.4 Certain Conservation Laws in Elementary-Particle Physics

43.4.1 The differences in the rest masses of charged $\pi^\pm$ and the neutral π^0 pions, like the difference in the rest masses of the two states of nucleons—proton and neutron—in the nucleus, are the result of charge independence of nuclear forces (42.3.3) and of the presence of electromagnetic interactions in addition to strong nuclear interactions. With no electromagnetic interactions in the nucleus, charge independence of nuclear forces

would lead to a condition in which the proton and neutron have equal rest mass, and the masses of all types of pions coincide. When the additional electromagnetic interaction is taken into account, the double state of the nucleon (*proton-neutron doublet*) is split into two different states in accordance with their rest masses. The masses of the neutron and proton are found to differ. For the same reason, the masses of the particles that make up the *pion triplet* are also not equal to one another. The masses of the charged $\pi^{\pm}$ pions exceed that of the π^0 pion (Table 43.1).

43.4.2 Introduced to characterize the multiplicity of states of certain kinds of particles are the concepts of *isotopic spin* T (*isospin*) and the *projection* T_ξ *of isotopic spin* on a certain "axis". The introduction of these characteristics of elementary particles makes it possible to formulate special features of the properties and transformation of elementary particles that follow from charge independence in strong interactions. Isotopic spin T characterizes the number of particles in the given "charge multiplet"; this number is equal to $(2T + 1)$. Thus the isotopic spin of the nucleon equals $T = 1/2$, and that of the pion is $T = 1$, in accordance with the number of particles (nucleons and pions) in the doublet and triplet.

The projection T_ξ of isotopic spin of the nucleon is introduced in such a way that it is related to the charge Z (in e units) of a particle that is a member of a charge doublet as follows:

$$Z = T_\xi + 1/2.$$

For the neutron $T_\xi = -1/2$, for the proton $T_\xi = 1/2$.

For the triplet of pions, this relation is of a more complex nature (43.4.3) and leads to the situation in which the projection T_ξ of isotopic spin is equal to the charge of the particle (in e units): $T_\xi = 0$ for the π^0 meson, $T_\xi = 1$ for the π^+ meson and $T_\xi = -1$ for the π^- meson.

The *law of conservation of isotopic spin* states that the total isotopic spin T of all the particles of an isolated system remains constant through all transformations caused by charge-independent strong interactions.

43.4.3 The concept of a *baryon* or *nuclear* (*nucleon*) *charge B* can be introduced for all elementary particles. For baryons (Table 43.1) $B = +1$, for antibaryons $B = -1$, whereas for particles that do not belong to the baryon class $B = 0$.

The *law of conservation of baryon charge* states that in all nuclear

transformations in an isolated system the total baryon charge remains constant.

This law, like the law of conservation of electric charge, is valid both in strong and in electromagnetic interactions. Making use of the baryon charge B, the relation between the electric charge Z and the projection of isotopic spin T_ξ of a particle in the charge multiplets of nucleons and pions can be written as

$$Z = T_\xi + B/2.$$

For nucleons (proton and neutron) $B = 1$ and $Z = T_\xi + 1/2$ (43.4.2). For pions $B = 0$ and $Z = T_\xi$.

Since Z and B are conserved both in strong and electromagnetic interactions, T_ξ is also conserved in these interactions, whereas T is conserved only in strong interactions.

43.4.4 The investigations of kaons and hyperons (Table 43.1) revealed certain of their unusual properties. The creation of these particles is brought about by strong interaction, and the time for their creation corresponds to nuclear time (43.2.6). At the same time, when they decay into nuclearly active pions, kaons have a lifetime typical of weak interaction (Table 43.2). Moreover, it was established that kaons and hyperons are always created in pairs, but not in arbitrary combinations. Finally, a substantial difference was found in the conditions for the creation of kaons and the reactions of their interaction with other particles. These features made it necessary to consider kaons and hyperons to be *strange particles*. A special characteristic of elementary particles was introduced to explain the behaviour of these strange particles. It is called *strangeness* S and is nonzero only for strange particles ($S = \pm 1$, ± 2, . . .) and zero for "ordinary" particles. For strange particles the relation in Sect. 43.4.3 is of the form

$$Z = T_\xi + \frac{B+S}{2}.$$

From the conservation law formulated in Sect. 43.4.3 follows the *law of conservation of strangeness* which states that the total strangeness of an isolated system remains constant in strong and electromagnetic interactions.*

* The information given in this section does not cover all the conservation laws in high-energy physics. But information on the other laws is beyond the scope of this handbook.

43.5 Antiparticles

43.5.1 The majority of elementary particles have corresponding *antiparticles*. Examples of particles and antiparticles are: the electron $_{-1}^{0}e$ and positron $_{+1}^{0}e$, muons μ^{+} and μ^{-}, pions π^{+} and π^{-}, kaons K^{+} and K^{-}, the electron and muon neutrinos, $_{0}^{0}\nu_e$ and $_{0}^{0}\nu_\mu$, and antineutrinos, $_{0}^{0}\widetilde{\nu}_e$ and $_{0}^{0}\widetilde{\nu}_\mu$. The rest masses, spins and lifetimes of particles and their antiparticles are the same. The electric charges of particles and their antiparticles are the same in magnitude, but opposite in sign.

Particles whose properties are completely identical with those of their antiparticles are called *truly neutral particles*. They include the photon and the neutral mesons, π^0 and K^0.

43.5.2 The existence in nature of pairs of charged particles and their antiparticles expresses the *principle of charge conjugation*, which states that each charged elementary particle must have an antiparticle. In accordance with this principle, the proton has an antiparticle, the *antiproton* $_{-1}^{1}\widetilde{p}$ (Table 43.1). The principle of charge conjugation extends to the neutral particles, neutron and neutrino, as well. There is the *antineutron* $_{0}^{1}\widetilde{n}$, as well as the electron antineutrino $_{0}^{0}\widetilde{\nu}_e$ and muon antineutrino $_{0}^{0}\widetilde{\nu}_\mu$.

43.5.3 When a particle combines with its antiparticle, the energy released is not less than twice the rest energy of each of them. The formation of a "particle-antiparticle" pair requires an energy input that exceeds twice the rest energy of each particle. This is due to the necessity to impart momentum and kinetic energy to the pair being produced.

43.5.4 The process of producing an electron-positron pair occurs in the collision of a hard gamma-ray quantum with some kind of charged particle, for instance, with an electron on the shell of an atom or with an atomic nucleus*. This process proceeds according to the reaction $\gamma \rightarrow {}_{-1}^{0}e + {}_{+1}^{0}e$ and is feasible with

---

* The need for a third particle is due to the fact that in producing two particles with the rest mass $m_0 \neq 0$ from a photon, the total momentum of the two particles is less than the momentum $h\nu/c$ of the photon. Hence, a third particle is required to take up part of the momentum of the photon.

a photon energy not less than $2m_ec^2$, i.e. $h\nu > 2m_ec^2 = 1.022$ MeV, where $m_ec^2 = 0.511$ MeV is the rest energy of the electron (or positron). It follows from the reaction for producing a pair by the action of a gamma-ray quantum (photon) that the spin of the photon should equal 0 or 1 (in units of $\hbar$), because the spins of electron and positron are each equal to $\hbar/2$. A great many experimentally established facts and great deal of theoretical investigation lead to the conclusion that the spin of the photon is equal to $\hbar$.

43.5.5 The process of uniting an electron and a positron is called *positron-electron annihilation*. In the great majority of cases, the process results in the production of two gamma-ray quanta:

$$_{-1}^{0}e + _{+1}^{0}e \rightarrow 2\gamma.$$
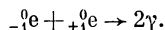
Before combining the particles $_{-1}^{0}e$ and $_{+1}^{0}e$, the total momentum of the two particles, in the coordinate system fixed in the centre of mass (2.3.3) of the electron-positron system, was equal to zero. After annihilation of the pair, two photons are produced whose momenta are in opposite directions, and their total momentum should be equal to zero*. Each photon carries away energy equal to $h\nu = m_ec^2 = 0.511$ MeV.

43.5.6 The minimum energy required to produce a proton-antiproton pair in a coordinate system where one nucleon is at rest is $6m_pc^2 = 5.6$ GeV. This energy is reduced to 4.3 GeV when the $_{1}^{1}p$-$_{1}^{1}\widetilde{p}$ pair is actually produced. Considerably more energy is required to produce a hyperon-antihyperon pair than for a pair of nucleons. This is due to the comparatively large mass of the hyperons. For example, the production of the lightest antihyperon, the antilambda-zero hyperon $(\widetilde{\Lambda^0})$, requires an energy from 5.3 to 5.8 GeV.

43.5.7 A feature of antiparticles is their capacity to combine rapidly with their corresponding particles, for instance, positrons with electrons, antiprotons with protons, and antineutrons with neutrons. The reason is that the matter of which all of nature surrounding us is built up consists of electrons, protons and neutrons. Antiparticles, artificially produced, inevi-

---

* Annihilation with the production of three photons, complying with the laws of conservation of momentum and energy, is also possible.

tably meet their much more numerous particles in substances, combine with them and cease to exist, giving rise to the production of new particles.

43.5.8 The difference between neutral particles and their antiparticles is established by the nature of the interactions of the particles and antiparticles with substances. Examples are the interaction reactions of the antineutrino ${}^{0}_{0}\tilde{v}_{e}$ with protons ${}^{1}_{1}p$ and the neutrino ${}^{0}_{0}v_{e}$ with neutrons:

$$ {}^{0}_{0}\tilde{v}_{e} + {}^{1}_{1}p \rightarrow {}^{1}_{0}n + {}^{0}_{+1}e \quad \text{and} \quad {}^{0}_{0}v_{e} + {}^{1}_{0}n \rightarrow {}^{1}_{1}p + {}^{0}_{-1}e \qquad (*) $$
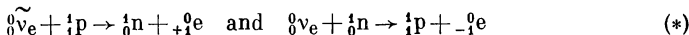
43.5.9 The difference between electron and muon neutrinos is established by investigating the decays of charged pions (43.3.4). If the neutrinos produced in these decays are separated out and their capture by neutrons is accomplished, then instead of the reaction (*) of Sect. 43.5.8, the process will be the following:

$$ {}^{0}_{0}v_{\mu} + {}^{1}_{0}n \rightarrow {}^{1}_{1}p + \mu^{-}, $$
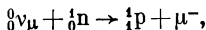
where μ^{-} is the negative muon. This demonstrates the difference between the electron and muon neutrinos (and antineutrinos).

43.6 Structure of the Nucleon

43.6.1 The *structure of the nucleon* (or any other elementary particle) is understood to mean its extent in space and what it consists of. The structure of any elementary particle cannot be treated by itself. It is related to the structure and properties of other particles. Special experiments for studying their structure have been conducted only for nucleons*. Effective methods for investigating the structure of nucleons are elastic collisions of pions with protons and elastic collisions of fast electrons with protons and neutrons. The first method showed that the pion is only slightly deflected by the collision from its initial

---

* The problem of the structure of elementary particles, as well as attempts to establish certain fundamental particles which all others are built up of are the frontiers of high-energy physics. The present handbook contains only general information on the structure of nucleons.

direction, and the proton received only slight recoil, so that the momentum Δp transmitted to the proton is very small. It follows from the indeterminacy relation (38.6.2) that the process of the collision of a pion with a proton takes place in a certain region of space with the linear dimensions $a \geqslant \hbar/\Delta p$, where a characterizes the size of the nucleon.

43.6.2 The emission and absorption of particles and antiparticles proceed continuously in a nucleon. A nucleon is treated as a complex, incessantly changing assembly of many particles.
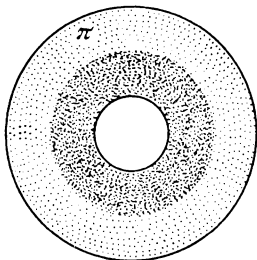
Fig. 43.4

At the central part of the nucleon (bare nucleon) is the nucleon core having a radius of (0.2 to 0.4) $\times$ $\times 10^{-15}$ m. In this region, heavy particles, resonances and nucleon-antinucleon pairs, play some special role that is not yet quite clear. The outer part of the nucleon forms the pion cloud (Fig. 43.4). This concept of nucleon structure enables the difference in masses of the neutron and proton to be cleared up, as well as the difference of the magnetic moment of the proton and neutron from μ_N. These differences are due to energy of electromagnetic interaction between the "core" of the nucleon and the pion cloud.

43.6.3 The scattering of fast electrons having energies up to 550 MeV by protons enabled the investigation of electric charge density distribution in the proton as a function of the distance r from the centre of the "core". It is necessary here to take into account the fact that the charge of the proton is indivisible and is always manifested as a unified whole. Therefore, the distribution of the electric charge in the proton does not imply the possibility of experimentally separating out a definite part of this charge. Shown in Fig. 43.5a is the dependence of the radial density of charge distribution $\rho_q = dq/dr$ on r (measured in fermi units), where dq is the charge contained in a spherical layer with the radii r and $r + dr$. The curve in the figure has a sharply defined maximum at the distance $R_e = 0.77 \times 10^{-15}$ m, which is called the "*electric radius*" *of the proton*. The area under the curve is equal to the charge e of the proton. The results obtained

in similar experiments in scattering fast electrons by neutrons are given in Fig. 43.5*b*.
Scattering proceeds as if the "electric radius" of the neutron was equal to zero. In interaction with fast electrons the neutron
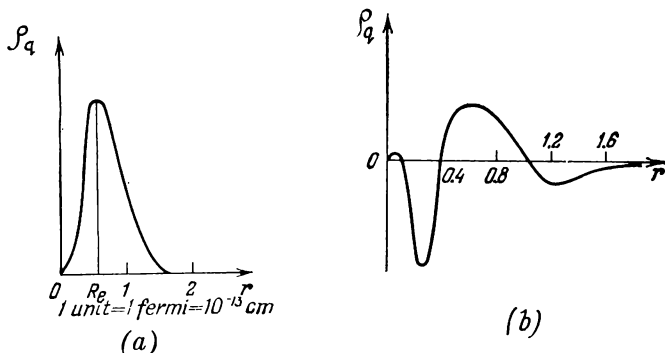
Fig. 43.5

behaves as if its pion cloud coincides in dimensions with its "core". The electric charge is negative in the inner and outer regions of the neutron; in the middle region it is positive. The total charge of the neutron, equal to the area under the curve, is zero.
The structure of elementary particles is being intensively investigated at the present time.

PART NINE APPENDICES

I. SYSTEMS OF UNITS OF PHYSICAL QUANTITIES

I.1 *A unit of a physical quantity* is an arbitrarily selected physical quantity having the same physical meaning as the one being considered. *A system of units* is the whole set of units, pertaining to a certain system of quantities and worked out on the basis of accepted rules. The *base*, or *fundamental*, *units* of a given system are the units of several dissimilar physical quantities, arbitrarily selected in establishing the system. The corresponding physical quantities are called *base quantities*. A system of units is said to be *absolute* if its base physical quantities are length, mass and time.

Derived units are ones established from other units of the given system by means of physical laws that express the relations between the corresponding physical quantities.

I.2 *The dimensional formula of any physical quantity* is an expression characterizing the relations of the units of this physical quantity with the base units of the given system of units. This expression is a monomial in the form of the product of symbols representing the base units raised to the corresponding power (whole or fractional, positive or negative). A physical quantity is said to be *dimensionless* if in its dimensional formula all base units are in zero power. The numerical value of a dimensionless quantity is independent of the choice of the system of units.

I.3 Prefixes and their symbols for decimal multiples and submultiples are listed in Table I.1.

I.4 Preferred in most countries today is the International System (SI), which is based on the application of seven base units: metre, kilogram, second, ampere, kelvin, mole and candela, as well as two supplementary units, the radian and steradian.

The definitions and notation of these units are listed in Tables I.2a and I.2b.

I.5 Derived units of the International System have been worked out according to equations relating the physical quantities, corresponding to the simplest cases of phenomena and bodies. The SI derived units for electric and magnetic quantities have been established in accordance with the rationalized form of electromagnetic field equations (Sects. 14.2.7 and 23.2.2). The most important derived units are listed in Table I.3.

I.6 In addition to SI units, it is permissible in certain cases to use the CGS (for example, in scientific papers in physics). The CGS system makes use (within the scope of mechanics) of three base units: the centimetre (cm) for length, the gram (g) for mass, and the second (s) for time. There are also two supplementary units: the radian (rad) for a plane angle and the steradian (sr) for a solid angle. All of these units either coincide with the corresponding SI units or are decimal fractions of them: 1 cm $= 10^{-2}$ m and 1 g $= 10^{-3}$ kg.

The derived CGS units for mechanical quantities and their relations to SI units are listed in Table I.4.

I.7 Three systems of units for electrical and magnetic quantities, based on the CGS units for mechanical quantities, are used in physics. They are the *absolute electrostatic* (CGS-electrostatic) *system*, the *absolute electromagnetic* (CGS-electromagnetic) *system* and the *absolute Gaussian*, or *symmetrical*, *system of units*.

In the CGS-electrostatic system, the proportionality factor k (14.2.4) in Coulomb's law is considered to be dimensionless and equal to unity. Accordingly, Coulomb's law and all other relations of electrostatics and electrodynamics are not written in the rationalized form.

In the CGS-electromagnetic system, the proportionality factor k in the Biot-Savart-Laplace law (23.2.1) is considered to be dimensionless and equal to unity. Hence, this law and all other relations of electromagnetism are written in the unrationalized form. The quantity c, indicating the number of units of electric charge (or electric current) in the CGS-electrostatic system, is equal to one unit of electric charge (or electric current) in the CGS-electromagnetic system and is called the *electrodynamic constant*. The *electrodynamic constant* is equal to the velocity of light in free space: $c = 2.99792458 \times 10^{10}$ cm/s $\approx 3 \times 10^{10}$ cm/s.

TABLE I.1

Prefixes Used with SI Units

Factor	Prefix		Factor	Prefix	
	Name	Symbol		Name	Symbol
10^{18}	exa	E	10^{-1}	(deci)	d
10^{15}	peta	P	10^{-2}	(centi)	c
10^{12}	tera	T	10^{-3}	milli	m
10^{9}	giga	G	10^{-6}	micro	μ
10^{6}	mega	M	10^{-9}	nano	n
10^{3}	kilo	k	10^{-12}	pico	p
10^{2}	(hecto)	h	10^{-15}	femto	f
10^{1}	(deca)	da	10^{-18}	atto	a

Note. The prefixes enclosed in parentheses may be used only in the names of decimal multiples and fractions of units that are already widely used (for example, hectowatt, decalitre, decimetre, and centimetre).

TABLE I.2a

Base Units of the International System (SI)

Quantity		SI units		
Name	Dimensions	Name of unit	Unit symbol	Definition
Length	L	metre	m	*One metre* is 1 650 763.73 wavelengths in vacuum of the radiation corresponding to the transition between electronic energy levels $2p_{10}$ and $5d_5$ of the krypton-86 atom

TABLE I.2a (*cont.*)

Quantity		SI units		
Name	Dimen-sions	Name of unit	Unit symbol	Definition
Mass	M	kilogram	kg	*One kilogram* is the mass of the platinum standard kept at the International Bureau of Weights and Measures at Sèvres, France
Time	T	second	s	*One second* is the duration of 9 192 631 770 periods of the radiation corresponding to the transition between two hyperfine levels of the ground state of the cesium-133 atom
Electric current	I	ampere	A	*One ampere* is the intensity of constant current which, when maintained in two parallel straight conductors of infinite length and negligible cross section placed one metre apart in a vacuum, would produce between them a force equal to 2×10^{-7} N per metre length
Thermodynamic temperature	Θ	kelvin	K	*One kelvin* is equal to the fraction 1/273.16 of the thermodynamic temperature of the triple point of water

TABLE I.2a (*cont.*)

Quantity		SI units		
Name	Dimen-sions	Name of unit	Unit symbol	Definition
Amount of substance	N	mole	mol	*One mole* is the amount of substance of a system that contains as many elementary units (atoms, molecules, ions, electrons, other particles, or specified groups of particles) as there are atoms in 0.012 kilogram of the pure nuclide carbon-12
Luminuous intensity	J	candela	cd	*One candela* is the luminous intensity, in the perpendicular direction, of a surface of 1/600 000 square metre of a black-body radiator at the temperature of freezing platinum under a pressure of 101 325 newtons per square metre, i. e. 101 325 Pa

Most extensively used in physics is the CGS-Gaussian, or symmetrical, system of units, in which the units for measuring all electrical quantities are the same as in the CGS-electrostatic system and those for measuring magnetic quantities are the same as in the CGS-electromagnetic system. In the Gaussian system, as in the CGS-electrostatic system, the proportionality factor k (14.2.4) in Coulomb's law is dimensionless and equal to unity, whereas the proportionality factor k in the Biot-Savart-Laplace law (23.2.1) is taken equal to $k = 1/c$, where c is the electrodynamic constant (23.2.2).

I.8 Derived CGS units (Gaussian system) for electrical quantities are listed in Table I.5, and for magnetic quantities, in Table I.6.

I.9 Off-system, or arbitrary, units, permitted for use in physics and astronomy, are listed in Table I.7.

I.10 Off-system units permitted for use in conjunction with SI units are listed in Table I.8.

TABLE I.2b

Supplementary Units of the International System (SI)

Quantity		SI units		
Name	Dimensions	Name of unit	Unit symbol	Definition
Plane angle	—	radian	rad	*One radian* is the angle subtended at the centre of a circle by an arc the length of which is equal to the radius
Solid angle	—	steradian	sr	*One steradian* is the solid angle subtended at the centre of a sphere by an area of surface equal to the square of the radius

TABLE I.3
Derived Units

Quantity		Derived SI units		
Name	Dimensions	Name	Symbol	Note
1. Derived Units, Space and Time				
Area	L^2	square metre	m^2	
Volume	L^3	cubic metre	m^3	
Velocity (speed)	LT^{-1}	metre per second	m/s	
Acceleration	LT^{-2}	metre per second squared	m/s^2	
Frequency	T^{-1}	hertz	Hz	
Rotational speed	T^{-1}	second to minus first power	s^{-1}	
Angular velocity	T^{-1}	radian per second	rad/s	
Angular acceleration	T^{-2}	radian per second squared	rad/s^2	
2. Derived Units, Mechanical Quantities				
Density	$L^{-3}M$	kilogram per cubic metre	kg/m^3	
Moment of inertia	L^2M	kilogram-metre squared	$kg\text{-}m^2$	
Momentum	LMT^{-1}	kilogram-metre per second	(kg-m)/s	
Angular momentum	L^2MT^{-1}	kilogram-metre squared per second	$(kg\text{-}m^2)/s$	

TABLE 1.3 *(cont.)*

Quantity			Derived SI units	
Name	Dimensions	Name	Symbol	Note
Force	LMT^{-2}	newton	N	1 N = 1 kg-m/s^2
Moment of force	L^2MT^{-2}	newton-metre	N-m	
Impulse	LMT^{-1}	newton-second	N-s	
Pressure, mechanical stress, modulus of elasticity	$L^{-1}MT^{-2}$	pascal	Pa	1 Pa = 1 N/m^2
Surface tension	MT^{-2}	newton per metre	N/m	
Work, energy	L^2MT^{-2}	joule	J	1 J = 1 N-m
Power	L^2MT^{-3}	watt	W	1 W = 1 J/s
Dynamic viscosity	$L^{-1}MT^{-1}$	pascal-second	Pa-s	
Kinematic viscosity	L^2T^{-1}	metre squared per second	m²/s	

3. Derived Units, Thermal Quantities

Quantity			Derived SI units	
Quantity of heat, internal energy	L^2MT^{-2}	joule	J	1 J = 1 N-m
Specific quantity of heat	L^2T^{-2}	joule per kilogram	J/kg	
Heat capacity and entropy of a system	$L^2MT^{-2}\Theta^{-1}$	joule per kelvin	J/K	
Specific heat capacity	$L^2T^{-2}\Theta^{-1}$	joule per kilogram-kelvin	J/(kg-K)	
Molar heat capacity	$L^2MT^{-2}N^{-1}\Theta^{-1}$	joule per mole-kelvin	J/(mol-K)	
Heat conductivity	$LMT^{-3}\Theta^{-1}$	watt per metre-kelvin	W/(m-K)	

TABLE I.3 (*cont.*)

4. *Derived Units, Electrical and Magnetic Quantities*

| Quantity | | Derived SI units | | |
Name	Dimensions	Name	Symbol	Note
Electric current density	$L^{-2}I$	ampere per square metre	A/m^2	
Electric charge	TI	coulomb	C	1 C = 1 A-s
Electric charge density:				
(a) linear	$L^{-1}TI$	coulomb per metre	C/m	
(b) surface	$L^{-2}TI$	coulomb per square metre	C/m^2	
(c) volume	$L^{-3}TI$	coulomb per cubic metre	C/m^3	
Polarization, electric displacement	$L^{-2}TI$	coulomb per square metre	C/m^2	
Electric dipole moment	LTI	coulomb-metre	$C-m$	
Electric displacement flux	TI	coulomb	C	
Electric potential, voltage, emf	$L^2MT^{-3}I^{-1}$	volt	V	1 V = 1 J/C
Electric field strength	$LMT^{-3}I^{-1}$	volt per metre	V/m	

TABLE I.3 (*cont.*)

| Quantity | | Derived SI units | | |
Name	Dimensions	Name	Symbol	Note
Capacitance	$L^{-2}M^{-1}T^4I^2$	farad	F	$1\ F = 1\ C/V$
Electric constant	$L^{-3}M^{-1}T^4I^2$	farad per metre	F/m	
Electric resistance	$L^2MT^{-3}I^{-2}$	ohm	Ω	$1\ \Omega = 1\ V/A$
Electric resistivity	$L^3MT^{-3}I^{-2}$	ohm-metre	Ω-m	
Electric conductance	$L^{-2}M^{-1}T^3I^2$	siemens	S	$1\ S = 1\ A/V$
Electric conductivity	$L^{-3}M^{-1}T^3I^2$	siemens per metre	S/m	
Magnetic flux	$L^2MT^{-2}I^{-1}$	weber	Wb	$1\ Wb = 1\ T\text{-}m^2 = 1\ V\text{-}s$
Magnetic induction	$MT^{-2}I^{-1}$	tesla	T	$1\ T = 1\ N/(A\text{-}m)$
Magnetomotive force	I	ampere	A	
Magnetic field strength	$L^{-1}I$	ampere per metre	A/m	
Inductance, mutual inductance	$L^2MT^{-2}I^{-2}$	henry	H	$1\ H = 1\ Wb/A$
Magnetic constant	$LMT^{-2}I^{-2}$	henry per metre	H/m	
Magnetic moment of electric current	L^2I	ampere-metre squared	A-m²	
Magnetization	$L^{-1}I$	ampere per metre	A/m	
Magnetic resistance (reluctance)	$L^{-2}M^{-1}T^2I^2$	ampere per weber	A/Wb	

TABLE I.3 (*cont.*)

5. *Derived Units, Light Quantities and Energetic Photometry*

Quantity		Derived SI units		
Name	Dimensions	Name	Symbol	Note
Luminous flux	J	lumen	lm	
Illumination	$L^{-2}J$	lux	lx	1 lx = 1 lm/m²
Luminous emittance	$L^{-2}J$	lumen per square metre	lm/m²	
Luminance	$L^{-2}J$	candela per square metre	cd/m²	
Radiation flux	L^2MT^{-3}	watt	W	
Radiant emittance	MT^{-3}	watt per square metre	W/m²	
Radiance	MT^{-3}	watt per steradian-metre squared	W/(sr·m²)	
Spectral density of radiant emittance				
(a) as function of wavelength	$L^{-1}MT^{-3}$	watt per cubic metre	W/m³	
(b) as function of frequency	MT^{-2}	joule per square metre	J/m²	

TABLE I.4

Derived Units of the CGS System, Mechanical Quantities

Quantity		Derived CGS unit		
Name	Dimen-sions	Name	Symbol	Expressed in SI units
Density	$L^{-3}M$	gram per cu-bic centi-metre	g/cm^3	10^3 kg/m^3
Moment of inertia	L^2M	gram-centi-metre squared	g-cm^2	10^{-7} kg-m^2
Momentum	LMT^{-1}	gram-centi-metre per second	$(g$-cm$)/s$	10^{-5} (kg-m)/s
Angular mo-mentum	L^2MT^{-2}	gram-centi-metre squared per second	$(g$-cm$^2)/s$	10^{-7}(kg/m^2)/s
Force	LMT^{-2}	dyne	dyn	10^{-5} N
Moment of force	L^2MT^{-2}	dyne-centi-metre	dyn-cm	10^{-7} N-m
Impulse	LMT^{-1}	dyne-second	dyn-s	10^{-5} N-s
Pressure, me-chanical stress, modulus of elasticity	$L^{-1}MT^{-2}$	dyne per square cen-timetre	dyn/cm^2	10^{-1} Pa
Surface tension	MT^{-2}	dyne per centimetre	dyn/cm	10^{-3} N/m
Work, energy	L^2MT^{-2}	erg	erg	10^{-7} J
Power	L^2MT^{-3}	erg per sec-ond	erg/s	10^{-7} W
Dynamic vis-cosity	$L^{-1}MT^{-1}$	poise	P	10^{-1} Pa-s
Kinematic viscosity	L^2T^{-1}	stokes	St	10^{-4} m^2/s

TABLE I.5

Derived Units of the CGS System, Electrical Quantities

Quantity		Derived CGS units	
Name	Dimensions	Name	Expressed in SI units
Electric current	$L^{3/2}M^{1/2}T^{-2}$	—	$10/c$ A
Electric current density	$L^{-1/2}M^{1/2}T^{-2}$	—	$10^5/c$ A/m²
Electric charge	$L^{3/2}M^{1/2}T^{-1}$	—	$10/c$ C
Electric charge density:			
(a) linear	$L^{1/2}M^{1/2}T^{-1}$	—	$10^3/c$ C/m
(b) surface	$L^{-1/2}M^{1/2}T^{-1}$	—	$10^5/c$ C/m²
(c) volume	$L^{-3/2}M^{1/2}T^{-1}$	—	$10^7/c$ C/m³
Polarization	$L^{-1/2}M^{1/2}T^{-1}$	—	$10^5/c$ C/m²
Electric dipole moment	$L^{5/2}M^{1/2}T^{-1}$	—	$1/10c$ C-m
Electric displacement flux	$L^{3/2}M^{1/2}T^{-1}$	—	$10/4\pi c$ C
Electric displacement	$L^{-1/2}M^{1/2}T^{-1}$	—	$10^5/4\pi c$ C/m²
Electric potential, voltage, emf	$L^{1/2}M^{1/2}T^{-1}$	—	$10^{-8}c$ V
Electric field strength	$L^{-1/2}M^{1/2}T^{-1}$	—	$10^{-6}c$ V/m
Capacitance	L	centimetre	$10^9/c$ F
Electric resistance	$L^{-1}T$	—	$10^{-9}c^2$ Ω
Electric resistivity	T	—	$10^{-11}c^2$ Ω-m
Electric conductance	LT^{-1}	—	$10^9/c^2$ S

TABLE I.6

Derived Units of the CGS System, Magnetic Quantities

Quantity		Derived CGS units		
Name	Dimensions	Name	Symbol	Expressed in SI units
Magnetic flux	$L^{3/2}M^{1/2}T^{-1}$	maxwell	Mx	10^{-8} Wb
Magnetic induction	$L^{-1/2}M^{1/2}T^{-1}$	gauss	G	10^{-4} T

TABLE I.6 (*cont.*)

Quantity		Derived CGS units		
Name	Dimensions	Name	Sym-bol	Expressed in SI units
Magnetomotive force	$L^{1/2}M^{1/2}T^{-1}$	gilbert	Gb	$10/4\pi$ A
Magnetic field strength	$L^{-1/2}M^{1/2}T^{-1}$	oersted	Oe	$10^3/4\pi$ A/m
Inductance, mutual induc-tance	L	centimetre	cm	10^{-9} H
Magnetic mo-ment of elec-tric current	$L^{5/2}M^{1/2}T^{-1}$	—	—	10^{-3} A-m^2
Magnetization	$L^{-1/2}M^{1/2}T^{-1}$	—	—	10^3 A/m
Magnetic resis-tance (reluc-tance)	L^{-1}	—	—	$10^9/4\pi$ A/Wb

TABLE I.7

Off-System Units Permitted for Use in Physics and Astronomy

Quantity	Off-system units		
	Name	Symbol	Expressed in SI units
Length	astronomical unit	A.U.	1.49600×10^{11} m
	light-year	lt-yr	9.4605×10^{15} m
	parsec	pc	3.0857×10^{16} m
	X-unit	X	1.00206×10^{-13} m
Mass	atomic mass unit (unified)	u	1.66057×10^{-27} kg
Area	barn	b	10^{-28} m^2
Energy	electronvolt	eV	1.60219×10^{-19} J

TABLE I.8

Off-System Units Permitted for Use in Conjunction With SI Units

Quantity	Off-system units		
Name	Name of unit	Symbol	Expressed in SI units
Area (plots of ground)	hectare	ha	10^4 m^2
Volume	litre	l	10^{-3} m^3
Plane angle	degree	...°	1/180 rad
	minute	...′	1/10 800 rad
	second	...″	1/648 000 rad
Time	minute	min	60 s
	hour	h	3600 s
	day	d	86 400 s
	week	w	—
	month	mo	—
	year	yr	—
Mass	tonne	t	10^3 kg
Celsius temperature, temperature difference	Celsius degree	°C	Celsius temperature $t = T - 273.15$, where T is the Kelvin temperature. One Celsius degree is equal to one kelvin

II. FUNDAMENTAL PHYSICAL CONSTANTS

II.1 Listed in Table II.1 are the fundamental constants compiled under the auspices of the CODATA Task Group on Fundamental Constants. They have been officially adopted by CODATA and are taken originally from J. Phys. Chem. Ref. Data, Vol. 2, No. 4, p. 663 (1973) and CODATA Bulletin No. 11 (December 1973).

TABLE II.1
Fundamental Constants

Quantity	Symbol	Numerical value *	Uncert. (ppm) **
Velocity of light in free space	c	299792458 (1.2) m/s	0.004
Magnetic constant (permeability of free space)	μ_0	$12.5663706144 \times 10^{-7}$ H/m	
Electric constant (permittivity of free space)	$\varepsilon_0 = 1/(\mu_0 c^2)$	$8.854187818\ (71) \times 10^{-12}$ F/m	0.008
Fine-structure constant	$\alpha = \mu_0 c e^2/2h$ α^{-1}	0.0072973506 (60) 137.03604 (11)	0.82 0.82
Elementary charge	e	$1.6021892\ (46) \times 10^{-19}$ C	2.9
Planck constant	h $\hbar = h/2\pi$	$6.626176\ (36) \times 10^{-34}$ J/Hz $1.0545887\ (57) \times 10^{-34}$ J/Hz	5.4 5.4
Avogadro's number (Avogadro constant)	N_A	$6.022045\ (31) \times 10^{23}$ mol^{-1}	5.1
Atomic mass unit	$u = 10^{-3}$ kg/mol·N_A	$1.6605655\ (86) \times 10^{-27}$ kg	5.1
Electron rest mass	m_e	$0.9109534\ (47) \times 10^{-30}$ kg $5.4858026\ (21) \times 10^{-4}$ u	5.1 0.38
Proton rest mass	m_p	$1.6726485\ (86) \times 10^{-27}$ kg 1.007276470 (11) u	5.1 0.011

44*

TABLE II.1 (*cont.*)

Quantity	Symbol	Numerical value *	Uncert. (ppm) **
Neutron rest mass	m_n	$1.6749543\ (86) \times 10^{-27}$ kg $1.00866502\ (37)$ u	5.1 0.037
Electron charge to mass ratio	e/m_e	$1.7588047\ (49) \times 10^{11}$ C/kg	2.8
Magnetic flux quantum	$\Phi_0 = h/2e$ h/e	$2.0678506\ (54) \times 10^{-15}$ Wb $4.135701\ (11) \times 10^{-15}$ J/Hz-C	2.6 2.6
Quantum of circulation	$h/2m_e$ h/m_e	$3.6369455\ (60) \times 10^{-4}$ J/Hz-kg $7.273891\ (12) \times 10^{-4}$ J/Hz-kg	1.6 1.6
Faraday constant (number)	$F/N_A e$	$9.648456\ (27) \times 10^{4}$ C/mol	2.8
Rydberg constant	$R_\infty = \dfrac{\mu_0^2 m_e c^3 e^4}{8h^3}$	$1.097373177\ (83) \times 10^{7}$ m^{-1}	0.075
Bohr radius	$a_0 = \alpha/(4\pi R_\infty)$	$0.5291706\ (44) \times 10^{-10}$ m	0.82
Classical electron radius	$r_e = \dfrac{\mu_0 e^2}{4\pi m_e}$	$2.8179380\ (70) \times 10^{-15}$ m	2.5
Bohr magneton	$\mu_B = e\hbar/2m_e$	$9.274078\ (36) \times 10^{-24}$ J/T	3.9
Electron magnetic moment	μ_e μ_e/μ_p	$9.284832\ (36) \times 10^{-24}$ J/T $658.2106880\ (66)$	3.9 0.010

TABLE II.1 (*cont.*)

Quantity	Symbol	Numerical value *	Uncert. (ppm)**
Proton magnetic moment	μ_p	$1.4106171\,(55)\times10^{-26}$ J/T	3.9
	μ_p/μ_B	$1.52103\,2209\,(16)\times10^{-3}$	0.011
	μ_p/μ_N	$2.7928456\,(11)$	0.38
Nuclear magneton	$\mu_N = e\hbar/2m_p$	$5.050824\,(20)\times10^{-27}$ J/T	3.9
Compton wavelength of the electron	$\lambda_{C,\,e} = h/(m_e c)$; $\lambdabar_{C,\,e},\ e/2\pi$	$2.4263089\,(40)\times10^{-12}$ m ; $3.8615905\,(64)\times10^{-13}$ m	1.6 ; 1.6
Compton wavelength of the proton	$\lambda_{C,\,p} = h/(m_p c) = \lambda_{C,\,p}/2\pi$	$1.3214099\,(22)\times10^{-15}$ m ; $2.1030892\,(36)\times10^{-16}$ m	1.7 ; 1.7
Compton wavelength of the neutron	$\lambda_{C,\,n} = h/(m_n c)$; $\lambdabar_{C,\,n},\ u/2\pi$	$1.3195909\,(22)\times10^{-15}$ m ; $2.1001941\,(35)\times10^{-16}$ m	1.7 ; 1.7
Molar volume of ideal gas under *standard conditions* ($T_0 = 273.15$ K, $p_0 = 101\,325$ Pa)	$V_0 = RT_0/p_0$	$0.02241383\,(70)\times$m³/mol	31
Molar gas constant	R	$8.31441\,(26)$ J/mol-K	31
Boltzmann constant	$k = R/N_A$	$1.380662\,(44)\times10^{-23}$ J/K	32
Stefan-Boltzmann constant	$\sigma = \dfrac{\pi^2 k^4}{60\hbar^3 c^2}$	$5.67032\,(71)\times10^{-8}$ W/m²-K⁴	125

TABLE II.1 (cont.)

Quantity	Symbol	Numerical value *	Uncert. (ppm)**
First radiation constant	$c_1 = 2\pi hc^2$	$3.744832\ (20) \times 10^{-16}$ W-m²	5.4
Second radiation constant	$c_2 = hc/k$	0.0143786 (45) m-K	31
Gravitational constant	G	$6.6720\ (41) \times 10^{-11}$ N-m²/kg²	615
Acceleration of free fall (standard value)	g	9.80665 m/s²	
Neutron rest energy	$m_n c^2$	939.5731 (27) MeV	2.8
Proton rest energy	$m_p c^2$	938.2796 (27) MeV	2.8
Electron rest energy	$m_e c^2$	0.5110034 (14) MeV	2.8
Energy corresponding to one atomic mass unit	$u c^2$	931.5016 (26) MeV	2.8

* The numbers in parentheses are the one standard-deviation uncertainties in the last digits of the quoted values.
** The uncertainty (relative error) is given in parts per million (ppm).

INDEX

TO THE READER