This textbook explains the concepts and most important advances of modern physics without resort to higher mathematics. Avoids the traditional division between classical and modern physics and endeavours to present all material so as to develop quantum mechanical concepts.

The textbook is intended for secondary schools and as a teaching aid for physics teachers in general and technical secondary schools. Will be found useful by correspondence students studying 'A' level and first year physics.

*Contents. Vol. II.* Vibrations and Waves. Quantum Physics of Atoms, Molecules and Solids. Physics of the Nucleus and Elementary Particles.

Prof. BORIS YAVORSKY, Doctor of Physical and Mathematical Sciences, is in the department of theoretical physics at the Lenin State Pedagogical Institute in Moscow. He has been lecturing in higher educational institutions of the Soviet Union for the last 35 years. Prof. Yavorsky has written about 300 published works: books and articles in various scientific journals and collected papers, among them the *Complete Course in Physics* (in three volumes) and some other textbooks and aids for institute teachers.

Assoc. Prof. ARKADY PINSKY, Candidate of Pedagogical Sciences, is a senior scientist at the Scientific Research Institute for Teaching Practice of the USSR Academy of Pedagogical Sciences. He has specialized in the methods of physics teaching at school and university levels. In this field he has published over 50 works, including several books on methods of teaching physics.

Б. М. Яворский, А. А. Пинский

# ОСНОВЫ ФИЗИКИ

## Том 2

## Колебания и волны; основы квантовой физики атомов, молекул и твердых тел; физика ядра и элементарных частиц

Издательство «Наука»

B. M. YAVORSKY and A. A. PINSKY

# FUNDAMENTALS of PHYSICS

VOLUME

II

VIBRATIONS AND WAVES.

QUANTUM PHYSICS OF ATOMS, MOLECULES

AND SOLIDS.

PHYSICS OF THE NUCLEUS

AND ELEMENTARY PARTICLES

Translated from the Russian
by
BORIS KUZNETSOV

MIR PUBLISHERS MOSCOW

First published 1975
Revised from the 1974 Russian edition

*На английском языке*

# CONTENTS

## Part six
## VIBRATIONS AND WAVES

## Part seven

## BASIC QUANTUM PHYSICS OF ATOMS, MOLECULES AND SOLIDS

## Part eight

## THE BASIC PHYSICS OF THE NUCLEUS AND ELEMENTARY PARTICLES

# PART SIX VIBRATIONS AND WAVES

## Chapter 49
## HARMONIC VIBRATIONS

### 49.1. THE HARMONIC OSCILLATOR

1. In Sec. 8.4 we examined the motion of a material point, or particle of mass $m$ due to the action of an elastic force $F = -kx$. Assuming the displacement of the particle at the initial time to be $x_0 = 1$ and the initial velocity to be $v_0 = 0$, we found, by a numerical method, a law of motion of the form

$$x = \cos \omega t$$

where $\omega = \sqrt{k/m}$. The instantaneous velocity of the particle was shown to be $v = -\omega \sin \omega t$.

It may be shown that the reverse holds, too. Namely, if a particle is moving so that

$$s = A \cos (\omega t + \varphi), \tag{49.1}$$

then its instantaneous velocity and acceleration will be

$$v = -\omega A \sin (\omega t + \varphi) \tag{49.2}$$

$$a = -\omega^2 A \cos (\omega t + \varphi) = -\omega^2 s \tag{49.3}$$

This is because, by definition (see Sec. 1.6), the average velocity $v_{av} = (s_2 - s_1)/(t_2 - t_1)$. Setting $t_1 = t - \Delta t$ and $t_2 = t + \Delta t$ we get $s_1 = A \cos (\omega t_1 + \varphi) = A \cos (\omega t - \omega \Delta t + \varphi)$, and $s_2 = A \cos (\omega t_2 + \varphi) = A \cos (\omega t + \omega \Delta t + \varphi)$. According to a well-known trigonometric identity

$$\cos \alpha - \cos \beta = -2 \sin \frac{\alpha + \beta}{2} \sin \frac{\alpha - \beta}{2}$$

we have

$$s_2 - s_1 = A\cos(\omega t + \omega\cdot\Delta t + \varphi) - A\cos(\omega t - \omega\cdot\Delta t + \varphi)$$
$$= -2A\sin(\omega t + \varphi)\sin(\omega\cdot\Delta t),$$

$$v_{av} = \frac{s_2 - s_1}{t_2 - t_1} = -\frac{2A\sin(\omega t + \varphi)\sin(\omega\cdot\Delta t)}{2\Delta t} = -\omega A\sin(\omega t + \varphi)\frac{\sin(\omega\cdot\Delta t)}{\omega\Delta t}$$

The sine of a small angle does not practically differ from the angle itself, that is, sin $(\omega\cdot\Delta t) \approx \omega\cdot\Delta t$. Passing to the limit in the expression for the average velocity, we obtain equation (49.2). We leave it for the reader to derive equation (49.3) in a similar way.

2. In equation (49.1) the quantity $s$ is called the *displacement*. Displacement is defined as the change in position of a vibrating particle from its equilibrium position at an arbitrary time. Obviously, the maximum displacement of a particle from its equilibrium position is $A$, since the cosine of an angle can never be greater than unity. This quantity is called the *amplitude of the vibration*:

$$S_{\text{M}} = A \qquad (49.4)$$

where the subscript "M" stands for "maximum". In significance the amplitude is a substantially positive quantity.

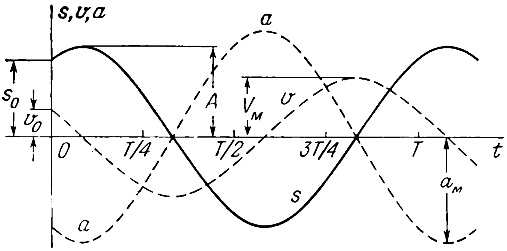


Fig. 49.1

It is an easy matter to see from equation (49.2) that the amplitude of velocity is

$$V_{\text{M}} = \omega A \qquad (49.5)$$

3. The force acting on a body is given by Newton's second law to be

$$F = ma = -m\omega^2 s \qquad (49.6)$$

As is seen, this force is similar to an elastic force in that it is proportional to the displacement and is opposite in sign (see Sec. 5.3). This is the reason why it is called a *quasi-elastic* force (from the Latin "quasi" for "as if it were").

4. The variable quantity $\omega t + \varphi$ is the argument of the cosine and is called the *phase of vibration*; the parameter $\varphi$ is called the *reference* or *initial phase*, or *epoch angle*. In conjunction with amplitude the reference phase defines the position and the velocity of a vibrating particle at the initial instant of time. As a proof, at $t = 0$ the initial displacement $s_0$ and the initial velocity $v_0$, according to equations (49.1) and (49.2), are

$$s_0 = A\cos\varphi, \quad v_0 = -\omega A\sin\varphi \qquad (49.7)$$

Hence, the amplitude and phase of the vibration are decided by initial conditions (see Ch. 8) such that

$$A = \sqrt{s_0^2 + (v_0/\omega)^2}, \quad \tan\varphi = -v_0/\omega s_0 \tag{49.8}$$

5. If the motion of a particle is described by a sinusoidal function of time, equation (49.1), the particle is said to be vibrating harmonically. A particle vibrating harmonically is called a *harmonic oscillator* (from the Latin "oscillum" for "swing").

Plots relating the displacement, velocity and acceleration of a harmonic oscillator are shown in Fig. 49.1. It will be noticed that the phase difference between velocity and displacement is $\pi/2$, and between velocity and acceleration is $\pi$. We leave it for the reader to construct displacement plots at $\varphi = 0$, $\varphi = \pi/2$, and $\varphi = -3\pi/2$.

## 49.2. FREQUENCY AND PERIOD OF VIBRATION

1. The parameter $\omega$ entering the expression for displacement and all the succeeding expressions is called the *radian* or *angular frequency*. The quantity

$$\nu = \omega/2\pi \tag{49.9}$$

is called simply the *frequency*. In electrical and radio engineering, it is denoted $f$ instead of $\nu$. To get the physical meaning of these quantities, we shall express them in terms of the period of vibration.

2. The *period* $T$ is defined as an interval of time required for one complete oscillation to repeat itself, that is, for a vibrating particle to pass through the same positions and in the same direction. By definition,

$$s(t + nT) = s(t) \tag{49.10}$$

where $n$ is an arbitrary integer. It implies that in an arbitrary number of periods the particle will be moving precisely as it is at the present instant.

Substituting (49.1) in (49.10) gives

$$A\cos[\omega(t + nT) + \varphi] = A\cos[\omega t + \varphi]$$

But the cosines of two arguments are equal if these arguments differ by $2n\pi$ (where $2\pi$ is the period of the cosine and sine, and $n$ is an integer). Therefore,

$$\omega t + n\omega T + \varphi = \omega t + \varphi + 2n\pi$$

3. Hence, the period and the frequency are related as follows

$$\omega = 2\pi/T, \quad \nu = 1/T \tag{49.11}$$

Thus, the radian frequency $\omega$ gives the number of complete vibrations over $2\pi$ seconds, and the frequency $\nu$ gives the number of vibrations completed in one second.

The unit of frequency is the *Hertz* (Hz). The frequency $\nu = 1$ Hz, if the period is $T = 1$ s. Like angular velocity, radian or angular frequency is expressed in radians per second.

4. From a comparison of the expression for an elastic force, $F = -ks$, with expression (49.6), we have

$$-ks = -m\omega^2 s$$

whence

$$\omega = \sqrt{k/m} \tag{49.12}$$

In other words, if a harmonic vibration is set up by the action of an elastic force, the frequency of the vibration is independent of the initial conditions and is only decided by the elasticity and mass of the system, that is, by the properties of the oscillator itself. This is why this frequency is called the *natural radian frequency* of an oscillator, designated $\omega_0$.

Then the natural period will be

$$T_0 = 2\pi/\omega_0 = 2\pi\sqrt{m/k} \tag{49.13}$$

### 49.3. ENERGY OF A HARMONIC OSCILLATOR

1. Kinetic energy (see Sec. 16.2) is given by

$$K = mv^2/2 = \frac{1}{2} m\omega^2 A^2 \sin^2(\omega t + \varphi) \tag{49.14}$$

Potential energy (see Sec. 18.6) is given by

$$U = kx^2/2 = \frac{1}{2} kA^2 \cos^2(\omega t + \varphi) \tag{49.15}$$

Noting that $k = m\omega^2$, (49.12), we have

$$U = \frac{1}{2} m\omega^2 A^2 \cos^2(\omega t + \varphi) \tag{49.16}$$

As will be recalled, $\cos^2\alpha + \sin^2\alpha = 1$. Adding together (49.14) and (49.16) gives an expression for the total mechanical energy of an oscillator

$$W = K + U = \frac{1}{2} m\omega^2 A^2 \tag{49.17}$$

Plots of the potential, kinetic and total energy of an oscillator appear in Fig. 49.2. As is seen, the period of kinetic or potential energy is half as great as the period of vibration. This also follows from the relationship $2\cos^2\omega t = 1 + \cos 2\omega t$.

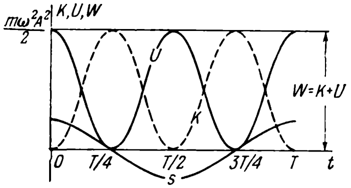


Fig. 49.2

2. A harmonic oscillator is a conservative system (see Sec. 19.1). Its total energy remains unchanged during oscillations; what happens is that potential energy is converted to kinetic and back, while their total amount remains unchanged. It may be shown (see Fig. 49.2)

that the average kinetic energy is equal to the average potential energy and to half the total energy

$$\overline{K}=\overline{U}=W/2=m\omega^2A^2/4 \qquad (49.18)$$

Here, the bar over a symbol designates an average.

## 49.4. RECORDS OF VIBRATORY MOTION

1. A simple record of vibratory motion is shown in Fig. 49.3. This record, usually on paper, is called an *oscillogram* (from the Latin "oscillum" for "swing", and the Greek "gramma" for "record").

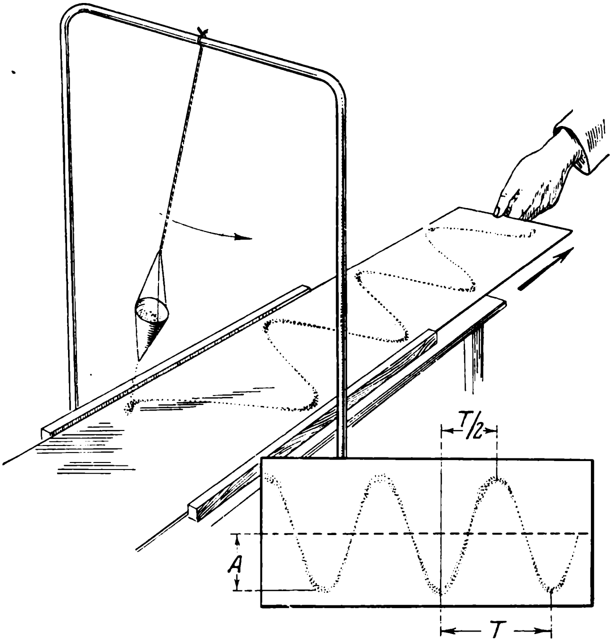


Fig. 49.3

In many cases, it is convenient to convert the vibration under investigation into electric signals which are easier to record. For this purpose, use may be made of electromagnetic induction. If a wire coil attached to a vibrating body is placed in a magnetic field, the oscillations will cause the coil to cut across the magnetic lines of force, and an electric current will be induced in the coil (Secs. 43.1-43.3); the oscillations of the current will precisely follow the motions of the oscillator. The oscillations of the current induced in the loop can readily be displayed on a cathode-ray oscilloscope or a moving-coil oscillograph.

2. The heart of an oscilloscope is a cathode-ray tube (Sec. 47.4). The electric signal representing the vibration under investigation is applied to the input of the vertical-deflection (Y-) amplifier. After amplification, the signal is applied to the vertical-deflection (Y-) plates which cause the electron beam to write a vertical trace on the screen (Fig. 49.4).

In order to obtain an oscillogram, the electron beam should be caused to move horizontally from left to right in addition to moving up and down. This is done by a built-in oscillator called a *time-base generator* which generates a sawtooth voltage (Fig. 49.5).

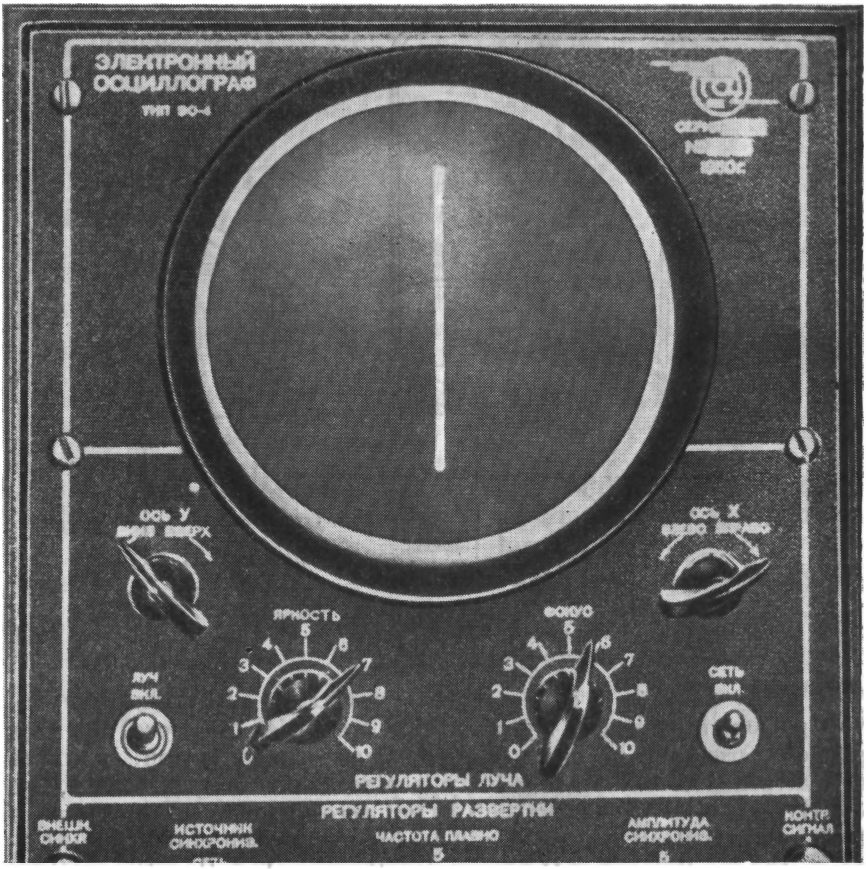


Fig. 49.4

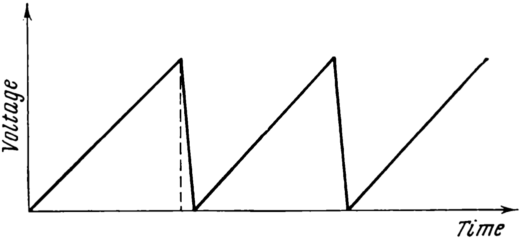


Fig. 49.5

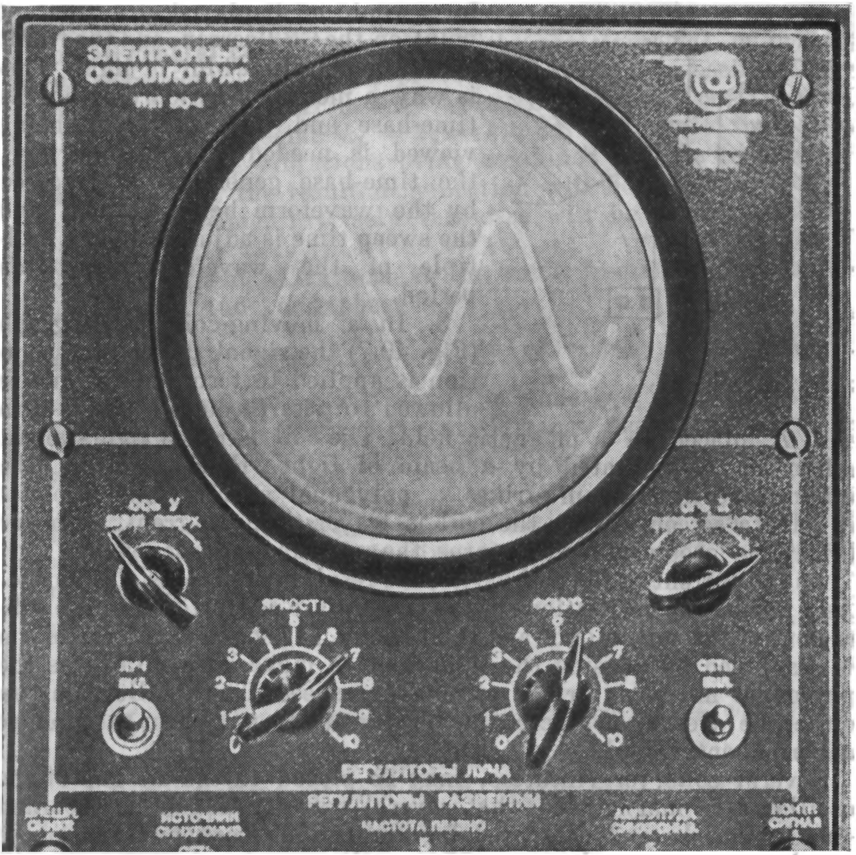


Fig. 49.6

This voltage is applied to the horizontal-deflection (X-) plates and sweeps the electron beam across the face of the CRT screen from left to right at a selected constant speed. At the end of its sweep across the screen, the spot is required to return to the left-hand side (this is called *fly-back*) before it starts its next sweep.

The joint action of the vertical and horizontal deflection plates causes the electron beam to trace out the shape of the waveform, or signal, on the CRT screen (Fig. 49.6).

A waveform must obviously start at the same instant of time during each sweep in the picture, if it is to be exactly superimposed on all earlier pictures. If it did not, each successive picture of the waveform would be displayed at a different point on the CRT screen, and the result would be simply a blur of light. This is why a means of synchronizing the time-base and the waveform being viewed is needed. For this purpose, the time-base generator is triggered by the waveform being viewed, and the sweep time is adjusted to be a multiple of the waveform repetition period.

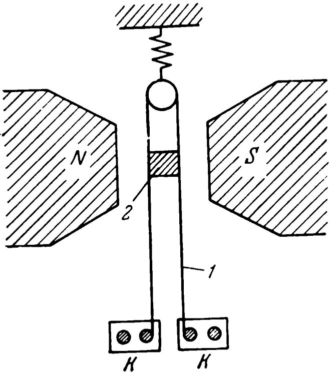


Fig. 49.7

3. In a moving-coil oscillograph (Fig. 49.7) the signal under investigation is applied to terminals $K$ and is allowed to pass through a single-turn coil *1*, placed in a magnetic field. The coil carries a light-weight mirror, *2*, illuminated by a beam of light. Reflected from the mirror, the light beam strikes a polygonal mirror which projects it onto a screen. When the coil is traversed by current, a torque (41.17) is produced which turns the coil and mirror system.

As the current in the coil varies, the resultant torque is also varied. As a result, the oscillations of the current are converted into the vibrations of the coil, mirror and light beam. The shape of the waveform is displayed by causing the light spot to move horizontally as the polygonal mirror is driven at constant speed. The oscillogram can either be viewed on a screen or photographed on a film to produce a permanent record.

Several coils, or vibrators, may be placed in the air gap of the oscillograph magnet, so that several waveforms can be recorded simultaneously and compared for amplitude, frequency and phase at a later time. A multi-beam CRT oscilloscope is a far more complicated piece of equipment to develop and engineer. On the other hand, a CRT oscilloscope has a better frequency response because the electron beam is practically free from time lag. This is why, a CRT oscilloscope may be operated at frequencies from a few tenths of a hertz to tens of megahertz. A moving-coil oscillograph shows a satisfactory performance at frequencies not over one kilohertz.

## 49.5. COMBINING VIBRATIONS HAVING THE SAME FREQUENCY

1. Let two forces $F_1 = -k_1 s_1$ and $F_2 = -k_2 s_2$ be acting on a body. If acted upon by only one of the two forces, the body would be set in vibrations described by the following equations

$$s_1 = A_1 \cos(\omega_1 t + \varphi_1) \quad \text{and} \quad s_2 = A_2 \cos(\omega_2 t + \varphi_2)$$

Now, let us see the body move when the two forces are applied simultaneously.

In the general case, the result will be *non-sinusoidal*, or *non-harmonic* vibration. This can be proved by plotting an oscillogram of a complete vibration. It is only when the elasticity coefficients of the systems are the same ($k_1 = k_2 = k$) and, as a result, the natural frequencies of the vibrations being combined are the same ($\omega_1 = \omega_2 = \omega$) that the resultant vibration will be a harmonic one at the same frequency. Precisely this case will be examined in this section.

2. So, we have vibrations at the same frequency, which only differ in amplitude and phase

$$s_1 = A_1 \cos(\omega t + \varphi_1), \quad s_2 = A_2 \cos(\omega t + \varphi_2) \tag{49.19}$$

The resultant vibration will have the same frequency but a different amplitude, $A$, and a different reference phase, $\varphi$:

$$s = A \cos(\omega t + \varphi) \tag{49.20}$$

To find this amplitude and phase, we recall that when vibrations occur along the same straight line, the displacements may be combined algebraically

$$s = s_1 + s_2$$

or

$$A \cos(\omega t + \varphi) = A_1 \cos(\omega t + \varphi_1) + A_2 \cos(\omega t + \varphi_2)$$

The above expression is an identical equality, that is, it is satisfied at any time. Setting $\omega t = 0$ (or $\pi$, or $2\pi$, etc.) we get

$$A \cos\varphi = A_1 \cos\varphi_1 + A_2 \cos\varphi_2$$

Setting $\omega t = \pi/2$ (or $3\,\pi/2$, or $5\,\pi/2$, etc.), we get

$$A \sin\varphi = A_1 \sin\varphi_1 + A_2 \sin\varphi_2$$

3. The sought quantities $A$ and $\varphi$ can be found from the last two equalities. Dividing the second by the first gives

$$\tan \varphi = (A_1 \sin \varphi_1 + A_2 \sin \varphi_2)/(A_1 \cos \varphi_1 + A_2 \cos \varphi_2) \qquad (49.21)$$

Squaring and adding together the two equalities and noting that $\cos^2 \varphi + \sin^2 \varphi = 1$, we have

$$A^2 = A_1^2 + A_2^2 + 2A_1A_2 (\cos \varphi_1 \cos \varphi_2 + \sin \varphi_1 \sin \varphi_2)$$

However, the expression in brackets is the cosine of the difference of two arguments: $\cos (\varphi_2 - \varphi_1) = \cos \varphi_1 \cos \varphi_2 + \sin \varphi_1 \sin \varphi_2$. Thus, the square of the amplitude is

$$A^2 = A_1^2 + A_2^2 + 2A_1A_2 \cos (\varphi_2 - \varphi_1) \qquad (49.22)$$

## 49.6. VECTOR DIAGRAMS

1. The amplitude and reference phase of the resultant vibration can be found by equations (49.21) and (49.22). As an alternative, this may be done, using a plot such as shown in Fig. 49.8. To begin with, we draw a horizontal axis and construct a vector $\mathbf{A}_1$ making an angle $\varphi_1$ with the axis. Then, from the tip of the vector $\mathbf{A}_1$, a second vector, $\mathbf{A}_2$, is drawn to make an angle $\varphi_2$ with the axis. Now the magnitude of the vector $\mathbf{A}$ which starts from the beginning of the vector $\mathbf{A}_1$ and terminates at the tip of the vector $\mathbf{A}_2$ gives the sought amplitude of the resultant vibration while the angle $\varphi$ that the vector $\mathbf{A}$ makes with the axis gives the sought reference phase.

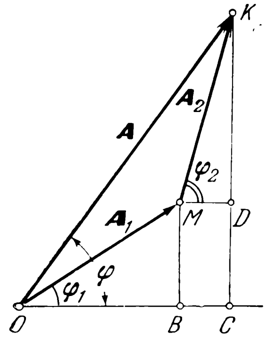


Fig. 49.8

As a proof, refer to the drawing of Fig. 49.8. Here $OB = A_1 \cos \varphi_1$, $BC = MD = A_2 \cos \varphi_2$, $DC = MB = A_1 \sin \varphi_1$ and $KD = A_2 \sin \varphi_2$. Hence, $OC = A_1 \cos \varphi_1 + A_2 \cos \varphi_2$, $KC = A_1 \sin \varphi_1 + A_2 \sin \varphi_2$. But $\tan \varphi = KC/OC$, and, by the Pythagorean theorem $(OK)^2 = (OC)^2 + (KC)^2$. Substituting the numerical values of these terms and making the necessary computations, we obtain the sought expressions, that is, (49.21) and (49.22).

2. The graphical construction illustrated above is called a *vector diagram*. Vector diagrams are convenient, especially in cases where one has to add together several vibrations for which an analytical solution is complicated.

As an example, let us find the amplitude resulting from adding together $N$ vibrations of the same amplitude and frequency whose phases make up an arithmetic progression:

$$\left.\begin{aligned} s_1 &= A\cos(\omega t+\varphi) \\ s_2 &= A\cos(\omega t+\varphi+\alpha) \\ s_3 &= A\cos(\omega t+\varphi+2\alpha) \\ &\cdots\cdots\cdots\cdots\cdots \\ s_N &= A\cos[\omega t+\varphi+(N-1)\alpha] \end{aligned}\right\}$$

A vector dagram for $N = 5$ is shown in Fig. 49.9. Since it is a regular open polygon, it can be inscribed into a circle of radius $R$. It is seen from the drawing that the amplitude of the resultant vibration is

$$B = 2R\sin(\beta/2)$$

From the triangle $MOK$, we have

$$R = \frac{A}{2\sin(\alpha/2)}$$

The angle $\beta = 2\pi - N\alpha$, and so

$$\sin(\beta/2) = \sin(\pi - N\alpha/2) = \sin(N\alpha/2)$$

Substituting it in the expression for the resultant amplitude, we finally get

$$B = A\frac{\sin(N\alpha/2)}{\sin(\alpha/2)} \qquad (49.23)$$

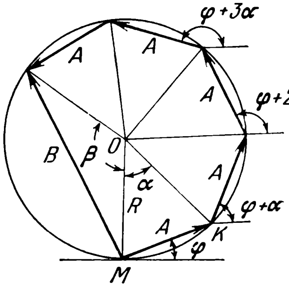


Fig. 49.9

This expression will come in useful at a later time (Sec. 57.6). It should be noted that equation (49.23) is rather difficult to derive analytically, while with vector diagrams the problem has been solved by simple geometrical construction.

# Chapter 50

# HARMONIC ANALYSIS

## 50.1. COMBINING VIBRATIONS AT CLOSELY SPACED FREQUENCIES

1. Let us find the resultant of two harmonic vibrations slightly differing in frequency, $\omega_1 = \omega - \Delta\omega$ and $\omega_2 = \omega + \Delta\omega$, such that $\Delta\omega \ll \omega$. For simplicity, we assume that the vibrations being combined have the same amplitude

$$s_1 = A\cos\omega_1 t, \quad s_2 = A\cos\omega_2 t \qquad (50.1)$$

Then the resultant vibration will be

$$s = s_1 + s_2 = 2A \cos\frac{(\omega_2 - \omega_1)\,t}{2} \cos\frac{(\omega_2 + \omega_1)\,t}{2}$$

$$= 2A \cos(\Delta\omega \cdot t) \cos \omega t \qquad (50.2)$$

Graphically it is shown in Fig. 50.1.

2. As is seen, the resultant vibration is not harmonic. However, provided its harmonic components differ only slightly in frequency, it may be treated as a nearly sinusoidal one of a period such that

$$T_0 = 2\pi/\omega \qquad (50.3)$$

and with a slowly time-varying amplitude

$$B = |\, 2A \cos(\Delta\omega \cdot t) \qquad (50.4)$$

In Fig. 50.1, this time-varying amplitude is shown by the dotted line. In effect, this is not an amplitude in the precise meaning of the word, because, by definition, the amplitude is a constant factor of the cosine. The term $T_0 = 2\pi/\omega$ may be called the period only arbitrarily, because it is in effect the time interval between two consecutive zero values of the function, rather than the period in the strict sense of the word (Sec. 49.2).

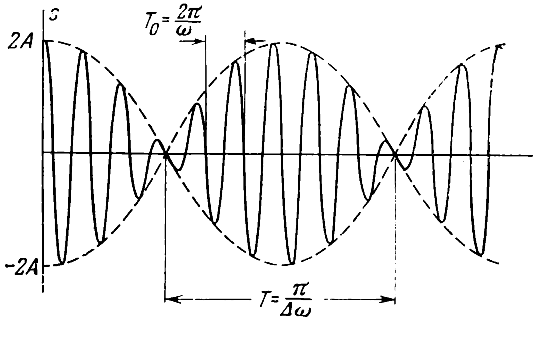


Fig. 50.1

3. The periodic variations in vibration amplitude produced by combining two vibrations slightly differing in frequency as described above are called *beats*. The period of beats is

$$T = \pi/\Delta\omega = 2\pi/(\omega_2 - \omega_1) \qquad (50.5)$$

The beat frequency is the difference of the two component frequencies:

$$\nu = 1/T = (\omega_2 - \omega_1)/2\pi = \nu_2 - \nu_1 \qquad (50\ 6)$$

## 50.2. MODULATED VIBRATIONS

1. Let us find the resultant of three harmonic vibrations, namely

$$s_1 = A\cos\omega t,\quad s_2 = a\cos(\omega + \Omega)t,\quad s_3 = a\cos(\omega - \Omega)t \tag{50.7}$$

Noting that $\cos(\omega + \Omega)t + \cos(\omega - \Omega)t = 2\cos\omega t\cos\Omega t$, elementary manipulation gives

$$s = s_1 + s_2 + s_3 = A\left(1 + \frac{2a}{A}\cos\Omega t\right)\cos\omega t \tag{50.8}$$

Setting $\Omega \ll \omega$ and $k = 2a/A < 1$, the resultant vibration will appear graphically as shown in Fig. 50.2.

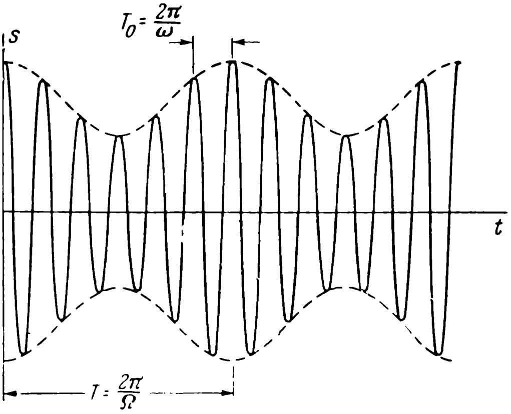


Fig. 50.2

2. From the drawing, it is seen that in this case, too, the resultant vibration may be treated as a nearly sinusoidal one with a time-varying amplitude given by

$$B = A(1 + k\cos\Omega t) \tag{50.9}$$

and with an equivalent period

$$T_0 = 2\pi/\omega \tag{50.10}$$

Changes in the amplitude occur at a period given by

$$T = 2\pi/\Omega \tag{50.11}$$

Since, under the conditions of the problem, $\Omega \ll \omega$, then $T \gg T_0$.

3. The diagram of Fig. 50.2 shows *modulated* vibrations. Generally, the term "modulated" applies to nearly sinusoidal vibrations occurring to a high frequency, $\omega$, whose amplitude varies slowly

with a period $T = 2\pi/\Omega$. The high frequency, $\omega$, is called the *carrier frequency* (or simply the *carrier*), the low frequency, $\Omega$, is called the *modulating* or *modulation frequency*, and the factor $k$ is the *depth of modulation*.

Modulated vibrations, or oscillations, are employed in radio for the transmission of sound or image by means of electromagnetic waves (Sec. 60.2). In these applications, however, more complex waveforms rather than a sinusoid are used for modulation.

## 50.3. COMBINING VIBRATIONS AT MULTIPLE FREQUENCIES

1. Now we shall see what happens when two or more harmonic vibrations at multiple frequencies are combined. As an example, consider two vibrations of radian frequencies $\omega_1 = \omega$ and $\omega_2 = 3\omega$ and of amplitudes $A_1 = A$ and $A_2 = A/2$:

$$s_1 = A \sin \omega t, \quad s_2 = \frac{A}{2} \sin 3\omega t \tag{50.12}$$

The vibration occurring at the lowest frequency is called the *fundamental*; the vibrations at multiple frequencies are called the *harmonics* of the fundamental (in acoustics, overtones).

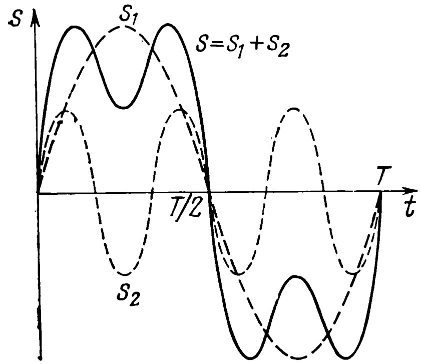


Fig. 50.3

2. To begin with, we shall combine the vibrations graphically. This is done by constructing plots of the component vibrations, measuring the displacements $s_1$ and $s_2$ at each instant, and adding them together, using the common rule for combining displacements, that is, taking their signs into account. As is seen from Fig. 50.3, combining harmonic vibrations at multiple frequencies produces a periodic non-sinusoidal vibration. Its period is the same as that of the fundamental vibration.

Instead of one frequency, there is a whole set of frequencies associated with a complex vibration; therefore, the concept of frequency is meaningful only as regards harmonic vibrations.

3. The character of a non-sinusoidal vibration is fully revealed by its waveform curve which is in turn determined by the number of component harmonics and the relationships bewteen their amplitudes, frequencies, and phases. By adding together a proper number of

simple harmonics, one can construct practically any periodic waveform. As an example, Fig. 50.4 shows waveform curves of vibrations having the same fundamental frequency but different harmonics:

$$\left.\begin{aligned}&\text{(a)}\ s_1 = 2\sin\omega t + 1.5\sin 2\omega t\\ &\text{(b)}\ s_2 = 2\sin\omega t + 3\sin 2\omega t +\\ &\qquad + 1.5\sin 3\omega t\end{aligned}\right\} \tag{50.13}$$

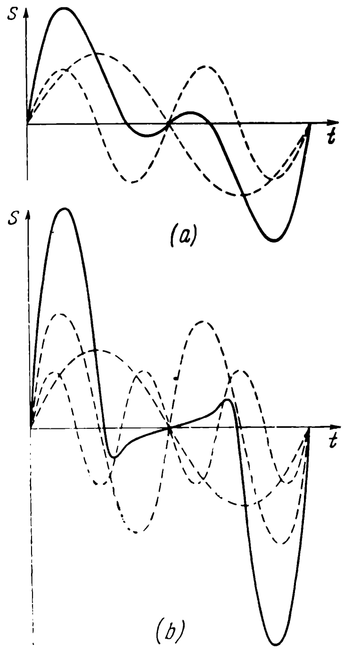


Fig. 50.4

## 50.4. FOURIER SERIES. SPECTRUM

1. In the preceding sections, we have seen that a non-sinusoidal vibration can be constructed by superimposing harmonic vibrations whose frequencies are integer multiples of the fundamental frequency. Thus, the question arises if the opposite is also true. Given a periodic time function of arbitrary shape, will it be possible to decompose this non-sinusoidal waveform into its harmonic components?

The answer is "yes", namely any periodic, single-valued, continuous function of time $f(t)$ with a period $T$ can be expressed as an infinite series of harmonic terms:

$$f(t) = a_0 + a_1\cos(\omega t + \varphi_1) + a_2\cos(2\omega t + \varphi_2) + + a_3\cos(3\omega t + \varphi_3) + \ldots \tag{50.14}$$

Here $\omega = 2\pi/T$, while the amplitudes and phases may be computed by certain rules given in courses on higher mathematics. This discovery is attributed to J.L. Fourier, a French physicist, who in 1822 presented this mathematical thesis in the *Theorie-Analytique de la Chaleur*. Accordingly, expression (50.14) is called the *expansion* of the function $f(t)$ *into a Fourier series*, or simply the *Fourier series*.

As a rule, the amplitudes decrease rather rapidly with inscreasing order of the harmonic, and the Fourier series of a function may be limited to only a few early terms.

2. In any problems of physics, only the amplitudes of harmonics are important, while their phases are immaterial, although they affect the shape of the complex vibration. This is, for example, true of cases where we may be interested not so much in the harmonics, as in their energies which, according to (49.17), are dependent solely

on amplitude and frequency, and independent of phase. Then we shall seek to determine the component frequencies and their respective amplitudes. The analysis of a periodic non-sinusoidal vibration into its harmonic components (ignoring the phases) is called *spectrum analysis*. The graphical representation of each harmonic amplitude as a function of its frequency is called the *spectrum* of a periodic non-sinusoidal vibration.

3. Fig. 50.5 shows the spectra of the vibrations whose Fourier series are represented by equations (50.13) and which are graphically shown in Fig. 50.4. The letters "a" and "b" in the plots and equations refer to the same vibrations. We leave it as an exercise for the reader to construct the spectra of a modulated vibration and beats.

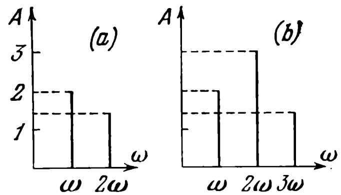


Fig. 50.5

It should be noted that knowledge of the spectrum of a non-sinusoidal vibration is not usually enough for its shape to be determined. Yet, there are cases where it will suffice. Leaving out oscilloscopes which sense the instantaneous magnitude of vibration, the displays used in the study of vibratory systems have an appreciable time lag and only register changes in the average energy over a time interval exceeding one period of oscillation. This is where knowledge of the spectrum is enough for determining the energy contribution of each harmonic component and the average energy of the complex vibration. This is why spectrum analysis figures prominently in the study of vibrations and vibratory systems.

## Chapter 51
## FREE VIBRATIONS

### 51.1. THE SPRING PENDULUM

1. When an elastic system is displaced from its equilibrium position and left to the action of internal forces, it undergoes *free* vibrations. As our first example, consider a spring-mass system, the spring pendulum, made up of a rigid body of mass $m$ and spring of elasticity $k$, joined together as shown in Fig. 51.1. A record of free vibration can be made by advancing a paper strip at a constant speed past a stylus attached to the weight. The record will show that the free oscillations of the spring-mass system gradually damp out, that is, the displacement of the weight from its equilibrium position gradually decreases with time (Fig. 51.2). This is *damped* vibrations (or oscillations).

Damped oscillations are not sinusoidal, and the concepts of amplitude and period in their proper sense do not apply. With light damping, however, when the damping force is a small fraction of the elastic force, we may treat damped vibrations as nearly sinusoidal with a progressively diminishing amplitude and an equivalent period $T_0$. In Fig. 51.2, the amplitude is shown by the dotted line. Our experiment will show that the successive values of amplitude taken at equal time intervals will make up a geometric progression. The period of damped vibration is the time interval between two consecutive maximum displacements of the weight from its equilibrium position in the same direction.

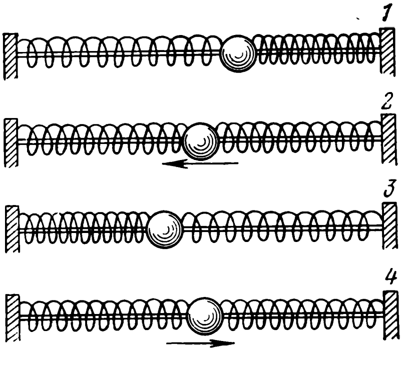


Fig. 51.1

2. It has been found that, although friction affects the manner in which the amplitude decreases, it has practically no effect on the period of damped vibrations. Therefore, the time period $T_0$ may be found from Eq. (49.13) which gives the natural period of a harmonic oscillator (provided the damping is light).

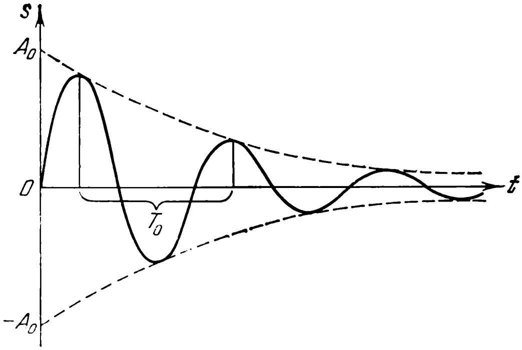


Fig. 51.2

Thus, with light damping, the period and frequency of free vibrations are practically the same as the natural period and frequency of a harmonic oscillator.

## 51.2. DAMPING. $Q$-FACTOR

1. In our further discussion we shall have to differentiate between light and heavy damping. In quantitative terms, this can conveniently be done by using the concept of the $Q$-factor of a vibratory (or oscillatory) system.

In a spring pendulum, the oscillations are damped by friction forces which dissipate energy from the system and reduce the amplitude of motion with time (see Sec. 19.2). Obviously, small friction forces result in a lighter damping.

But, then, how can friction be evaluated? Since a spring oscillates mainly due to its elasticity, it appears reasonable to relate friction force to elastic force. As the two forces are variables, their peak values are to be compared. The peak value of an elastic force is $F_{\mathrm{M}}^{el} = kA$, where $A$ is the amplitude and $k$ is the elasticity coefficient. As long as the pendulum bob oscillates at relatively low velocity, the friction force will be proportional to velocity (see Sec. 11.8), that is

$$F_{\mathrm{M}}^{fr} = hV_{\mathrm{M}} = h\omega_0 A$$

where $\omega_0$ is the natural radian frequency, $V_{\mathrm{M}} = \omega_0 A$ is the peak velocity, (49.6), and $h$ is the coefficient of friction.

2. The ratio of the elastic force to the friction force of a system is called the *Q-factor* of that system. The equation describing the $Q$-factor may be written as follows:

$$Q = F_{\mathrm{M}}^{el}/F_{\mathrm{M}}^{fr} = k/h\omega_0 \tag{51.1}$$

Noting that $k = m\omega_0^2$ (49.12), we get

$$Q = m\omega_0/h = \sqrt{mk}/h \tag{51.2}$$

Thus, as the friction force decreases in comparison with the elastic force, the amount of damping decreases too. In other words, with an increase in the $Q$-factor the amount of damping that free oscillations suffer decreases progressively and they approach harmonic vibrations.

3. Now we shall prove that the $Q$-factor of an oscillatory system is a measure of the relative dissipation of energy.

For this purpose, we shall compare the energy of an oscillator, as given by equation (49.17), with the energy lost through friction over a quarter of a period. In that time, the pendulum bob will have travelled a distance equal to one amplitude and the friction force will have done work given by

$$\Delta W_{heat} = A_{fr} = F_{av}^{fr} A \approx AhV_{\mathrm{M}}/2 = h\omega_0 A^2/2$$

This is the energy lost over a quarter of a period. The ratio of the initial energy to the energy lost gives the $Q$-factor:

$$W/\Delta W_{heat} = m\omega_0^2 A^2 2/2h\omega_0 A^2 = m\omega_0/h = Q \tag{51.3}$$

On the basis of equation (51.3), the $Q$-factor of an oscillatory system may be defined as the ratio of its total energy to the energy lost over a quarter of a period due to energy dissipation.

4. Now we shall find the time, $\tau$, necessary for oscillations to be damped practically completely. It is equal to the energy of the system, $W$, divided by the average power loss, $P_{av} = \Delta W/\Delta t$. Setting $\Delta t = T_0/4$, and the energy loss, according to equation (51.3), $\Delta W = W/Q$, we have

$$P_{av} = 4W/QT_0 = (4W/Q)\,(\omega_0/2\pi) \approx W\omega_0/Q$$

Hence,

$$\tau = W/P_{av} \approx Q/\omega_0 \tag{51.4}$$

### 51.3. THE SIMPLE PENDULUM

1. A simple pendulum is a mechanical oscillator which consists of a weight suspended from a fixed point by an inextensible member of length $l$. Practically, the size of the weight is a small fraction of the suspension length, while its mass is many times that of the suspension (Fig. 51.3).

With small displacements, that is, when the angle $\theta$ does not exceed a few degrees, $\sin\theta \approx \theta$ and the arc $A$ is practically equal to the chord $a$. Then, in a first approximation, the weight may be assumed to be in a rectilinear motion, and the resultant motion is simple harmonic. A simple pendulum has a very high $Q$-factor, because of which its oscillations suffer negligible damping. This can be proved by making a record of these oscillations (see Fig. 49.3, where the pendulum is made in the form of a funnel with a narrow neck, filled with fine sand; the oscillations are recorded on a sheet of paper smeared with glue or moistened with water).

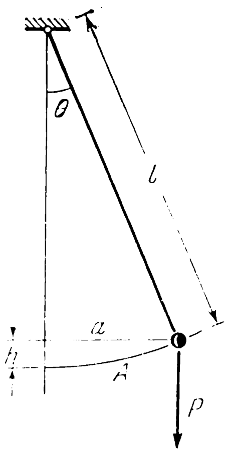


Fig. 51.3

2. The natural frequency and period of a simple pendulum can be found from the condition of energy balance. The maximum potential energy is

$$U_M = mgh = mgl\,(1 - \cos\theta) = 2\,mgl\,\sin^2(\theta/2)$$

For small displacements,

$$U_M = mgl\theta^2/2$$

The maximum kinetic energy is

$$K_M = mV_M^2/2 = mA^2\omega_0^2/2$$

where $\omega_0$ is the natural radian frequency and $A$ is the amplitude From Fig. 51.3 it is seen that $A = l\theta$. Hence,

$$K_M = ml^2\theta^2\omega_0^2/2$$

If damping is neglected, then, according to the law of conservation of energy, the average (and maximum) kinetic and potential energies should be the same. Equating the two expressions and cancelling out like terms, we get

$$\omega_0 = \sqrt{g/l} \tag{51.5}$$

3. The period of a simple pendulum is given by

$$T_0 = 2\pi\sqrt{l/g} \tag{51.6}$$

As is seen, the period is independent of either the mass of the pendulum bob or the amplitude of oscillations (provided the displacements are small).

Equation (51.6) may be used to determine the acceleration of gravity at any point on the Earth, since the length and period of the pendulum can be measured to a fairly high degree of accuracy.

## 51.4. THE PHYSICAL PENDULUM

1. A physical, or compound, pendulum is a rigid body with no restriction on size, shape or composition, capable of undergoing oscillations in a single plane about a suspension point at a distance $l$ from the centre of gravity (Fig. 51.4). If such a pendulum is moved away from the position of equilibrium and left to the effects of the force of gravity, it will undergo free oscillations with a relatively small amount of damping. The frequency and period of a physical pendulum can be found by reasoning along the same lines as for a simple pendulum.

Much as the peak velocity can be expressed in terms of the peak displacement (see equation (49.5)) so the peak angular velocity may be expressed as $\Omega_M = \omega_0\theta$. The maximum kinetic energy (see

Sec. 22.2) is

$$K_M = \frac{1}{2} J\Omega_M^2 = \frac{1}{2} J\omega_0^2\theta^2$$

Since the maximum potential and kinetic energies are equal, we have

$$\frac{1}{2} mgl\theta^2 = \frac{1}{2} J\omega_0^2 \theta^2$$

whence

$$\omega_0 = \sqrt{mgl/J}$$

The period of free oscillations for a physical pendulum is

$$T_0 = 2\pi/\omega_0 = 2\pi\sqrt{J/mgl} \qquad (51.7)$$

2. It should be noted that expression (51.6) is a particular case of equation (51.7). The point is that the moment of inertia of a simple pendulum is $J = ml^2$ (see Sec. 22.2). Substituting this expression for the moment of inertia in equation (51.7) yields an expression for the period of a simple pendulum.

3. Very often, the period of a physical pendulum is found by the equation

$$T_0 = 2\pi\sqrt{L/g} \qquad (51.8)$$

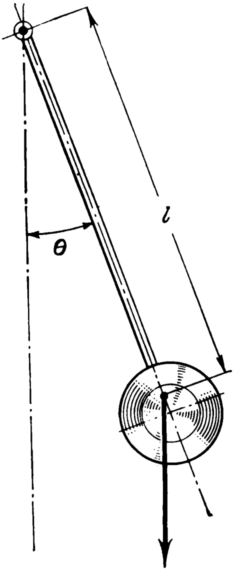


Fig. 51.4

which looks like the expression for the period of a simple pendulum. In this equation the quantity

$$L = J/ml \qquad (51.9)$$

is called the *length of an equivalent simple pendulum*. As its name implies, this length is equal to that of a simple pendulum which has the same period as the physical pendulum in question.

## 51.5. THE OSCILLATORY CIRCUIT

1. An *oscillatory circuit* is an electric circuit containing a coil of inductance $L$, a capacitor of capacitance $C$, and a resistor of resistance $R$, so arranged or connected that a voltage impulse will produce a periodically reversing current around the circuit (Fig. 51.5). The resistance $R$ also absorbs the resistance associated with the coil. Experiments show that the voltage impulse necessary to cause free oscillations can be obtained by charging the capacitor and by closing the circuit.

In a study of these free electrical oscillations, it is convenient to use the set-up of Fig. 51.6. When the switch is turned left, the capacitor is connected to a power source and is charged. When the switch is turned right, the capacitor is connected across the coil, and oscillations are brought about in the circuit, as revealed by the pattern displayed on the CRT screen. Naturally, in a practical test set-up, a manual switch is replaced with an electronic circuit performing the same function. The oscilloscope connected in the circuit of Fig. 51.6 senses the potential difference across the resistor which, according to Ohm's law equation (39.19), is proportional to the current in the circuit. If the oscilloscope be connected directly across the capacitor, it will sense the potential difference across the capacitor and, according to equation (37.20), the capacitor charge.

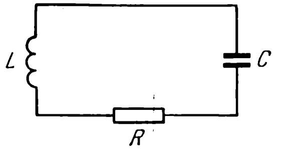


Fig. 51.5

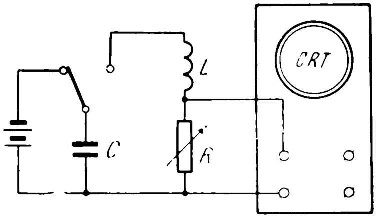


Fig. 51.6

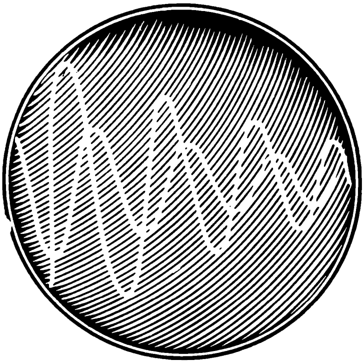


Fig. 51.7

2. Now let us use a double-beam oscilloscope connected so that one input accepts the current oscillations, and the other the voltage (charge) oscillations. The respective waveforms are displayed on the oscilloscope as shown in Fig. 51.7. As should be expected, free oscillations in an oscillatory circuit are subject to damping. By varying the resistance of the resistor, it may be proved that damping increases with an increase in resistance.

The most important finding of the experiment is, however, the fact that there is a phase shift between the current and charge oscillations. Referring to the oscillograms, it is seen that at the instants when the current is zero, the capacitor charge is a maximum and vice versa.

3. The relationship between current and charge, equation (39.14), $i = \lim_{\Delta t \to 0} (\Delta q/\Delta t)$, is analogous to that between velocity and displacement, $v = \lim_{\Delta t \to 0} (\Delta s/\Delta t)$. Then, if variations in the capacitor charge are described by

$$q = q_M \cos \omega_0 t \qquad (51.10)$$

variations in the current round the circuit (see equations (49.1) and (49.2)) will be given by

$$i = -q_M \omega_0 \sin \omega_0 t = I_M \cos (\omega_0 t + \pi/2) \qquad (51.11)$$

Precisely this phase shift is displayed on the CRT screen. Because of this phase shift, the peak values of charge and current are related as

$$I_M = \omega_0 q_M \qquad (51.12)$$

which is analogous to equation (49.5) connecting the peak values of displacement and velocity.

## 51.6. ENERGY, NATURAL FREQUENCY AND $Q$-FACTOR OF AN OSCILLATORY CIRCUIT

1. According to (37.23), the energy of the electric field in a capacitor is $W^e = q^2/2C$, and the energy of the magnetic field around a coil according to (43.19) is $W^m = Li^2/2$. Since the current is a maximum when the charge is zero, and vice versa, the same relationship holds for the energies of the electric and magnetic fields. Hence, there is a continual exchange of energy between the electric field in the capacitor of an oscillatory circuit and the magnetic field around its coil. This process is analogous to the exchange of potential and kinetic energy in a spring pendulum.

2. The maximum energy of the field in a capacitor is

$$W^e_M = q^2_M/2C \qquad (51.13)$$

Similarly, for the magnetic field around a coil

$$W^m_M = LI^2_M/2 = Lq^2_M\omega^2_0/2 \qquad (51.14)$$

Equating the two quantities and cancelling out like terms gives an expression for the natural radian frequency of oscillations in an oscillatory circuit:

$$\omega_0 = \sqrt{1/LC} \qquad (51.15)$$

and an expression for the natural period

$$T_0 = 2\pi\sqrt{LC} \qquad (51.16)$$

3. So far we have neglected the dissipation of energy. Meanwhile, oscillations in an oscillatory circuit are damped out precisely because some of the energy is irreversibly converted to Joule heat. To find the $Q$-factor of an oscillatory circuit, it is essential to relate its heat losses to its total energy. Alternatively, using an analogy with a spring pendulum (see Sec. 51.2), the $Q$-factor can be found as the ratio of the amplitude of the capacitor voltage to the amplitude of the voltage drop across the resistance:

$$Q = v_{\mathrm{M}}^{C}/v_{\mathrm{M}}^{R} = q_{\mathrm{M}}/CRI_{\mathrm{M}} = 1/\omega_0 CR = \sqrt{L/C}/R \qquad (51.17)$$

A low-$Q$ circuit cannot oscillate at all. This is because if heat losses are of the same order of magnitude as the energy stored by the capacitor, all of the energy during discharge will be converted to Joule heat, the energy of the magnetic field will be zero, and the capacitor will not recharge. Thus, if the $Q$-factor of a circuit is close to unity, the energy lost as heat will be of the same order of magnitude as the energy stored by the circuit, and the capacitor will discharge through the coil and resistor, as if there were no coil at all (see Sec. 39.8).

## 51.7. A UNIFIED APPROACH TO VIBRATIONS

1. From a comparison of what happens in a spring pendulum and in an oscillatory circuit, one will immediately witness a striking analogy between the events which appear to have nothing in common at first sight. Indeed, there is little in common between the motion of the pendulum bob under the action of a deformed spring and the motion of electrons due to the recharge of a capacitor. Of course, from this point of view, these are entirely different events which ought to be dealt with in the respective divisions of physics, the spring pendulum in mechanics and the oscillatory circuit in electromagnetism. If, however, emphasis is placed on *why* and *how* rather than on *what* oscillates, it will immediately become clear that the oscillatory processes in both systems have the same physical aspect and that they can be described by the same concepts and the same equations. This discovery forcefully bears out the advisability of using a unified approach to the study of vibrations widely differing in physical nature, and the method of analogies.

The idea of using analogies in the study of vibrations is not new. Already Huygens and Lomonosov used the analogy between sound and light waves. A unified approach to mechanical and electromagnetic vibrations was employed by Rayleigh, Stoletov, Lebedev and other investigators. But it was not until Academician Mandelshtam and his disciples that it became a working tool systematically employed in both theory and experiment. The unified approach has proved a very fruitful method yielding solutions to a number of complex problems in the theory of vibrations. Among them are the theory of nonlinear vibrations, notably self-sustained vibrations (Andronov *et al.*), the theory of Raman scattering (Mandelshtam and Landsberg, Sec. 79.1), parametric resonance and parametric oscillators (Mandelshtam and Papalexi) to name but a few.

2. With a unified approach, the findings obtained in the study of one type of vibrations may be generalized to another kind. Therefore, in our further discussion we shall only take up processes in one particular oscillatory system and apply the results to other vibrations by analogy.

This also applies to experiments. It has been found that in many cases it takes less effort and time to measure electrical rather than mechanical quantities. For example, the instantaneous velocity of a vibrating body is extremely difficult to measure, while it is an easy matter to measure the instantaneous values of current with a CRT oscilloscope. By the same token, it is easier to adjust a variable e.m.f. in a circuit than a variable force in a mechanical system. Last but no least, an electric circuit composed of capacitors, coils and resistors is simpler to assemble for an experiment than an elaborate mechanical system involving weights and springs with variable friction forces. This is why, instead of mechanical oscillatory systems, investigators often prefer to deal with their electrical analogies in conjunction with an oscilloscope. Some of the analogies are listed in Table 51.1.

*Table 51.1*

| Mechanical quantity | Electrical analogy |
|---|---|
| $m$ = mass | $L$ = inductance |
| $k$ = compliance | $1/C$ = elastance |
| $h$ = frictional resistance | $R$ = resistance |
| $F$ = force | $\mathscr{E}$ = emf |
| $s$ = displacement | $q$ = charge |
| $v$ = velocity | $i$ = current |

# Chapter 52
# SELF-SUSTAINED VIBRATIONS

## 52.1. SELF-SUSTAINED OSCILLATORY SYSTEMS

1. In a spring pendulum, oscillations are damped by friction. They would not, however, be damped if energy losses were continually made up for. As an example, consider the mechanism by which con-

tinuous oscillations occur in the set-up of Fig. 52.1. A flexible plate made fast to a weight periodically touches a contact, thereby completing the supply circuit of an electromagnet for a certian time interval. During this time interval, the electromagnet attracts the weight, thereby increasing its kinetic energy. As a result, the loss in energy per cycle due to friction is made up for by the work done by the force of attraction exerted on the weight by the electromagnet; operation of the electromagnet is controlled by the vibrating weight through the breaker point. This set-up is typical of a very wide class of oscillatory systems maintained in continuous oscillations by an energy source which has no oscillatory properties. They are called *self-sustained oscillatory systems.*

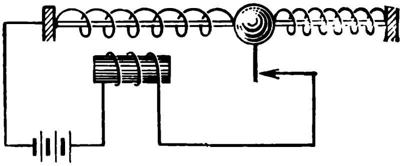


Fig. 52.1

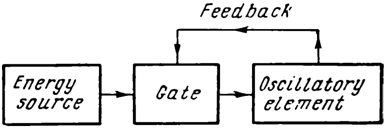


Fig. 52.2

Any self-sustained oscillatory system consists essentially of four elements as follows (Fig. 52.2):

(a) an *oscillatory element* (in the set-up of Fig. 52.1, the weight-and-spring assembly);

(b) an *energy source* that makes up for the loss in energy (in our example, the current source);

(c) a *gate* which admits energy to the oscillatory element in the right amounts and at the right instants (in the set-up of Fig. 52.1, the breaker point);

(d) a *feedback element*, an indispensable part of all self-sustained oscillatory systems. The feedback element controls the gate, utilizing the properties of the oscillatory system itself (in our set-up, the electromagnet which attracts the weight and breaks the contact).

2. Self-sustained oscillations (or vibrations) are a frequent occurrence both in nature and technology. For example, they are utilized in electrical bells, buzzers, steam engines, internal-combustion engines, pavement-breakers, etc. Self-sustained vibrations are produced in the strings of a violin by the bow, in the air columns of wind instruments, the reeds of an accordion, the vocal cords in speaking or singing.

It is important to note that in many self-sustained oscillatory systems feedback is not immediately obvious, and the systems themselves cannot always be broken down into their basic elements.

3. As regards operation of a self-sustained oscillatory system, a factor of crucial importance is the choice of phase for the feedback sig-

nal. It is important that while a force is acting on the system, the force and velocity be in the same direction. Then the energy source will do a positive work on the oscillatory system, that is, impart energy to it. Should the force and velocity be in opposite directions, the work will be negative, and the source will withdraw energy from the oscillatory system, thereby increasing the amount of damping. In the former case, the feedback in the system is said to be positive, while in the latter, negative. Positive feedback is essential to excite oscillations, while negative feedback may be used to suppress them where they are undesirable.

In the set-up of Fig. 52.1, the feedback is positive, owing to which continuous oscillations are excited and maintained. If the contact be placed on the left of the plate, the weight will be attracted to the electromagnet and will come to a stop. Even if the system be caused to build up oscillations initially, they will be damped rapidly, at a much faster rate than in the absence of an electromagnet. In this way, negative feedback discontinues self-sustained oscillations and suppresses free oscillations.

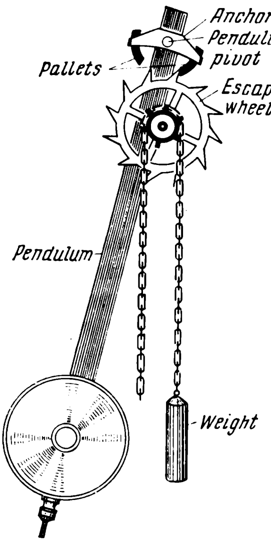


Fig. 52.3

52.2. THE CLOCK

1. Any timepiece is a self-sustained oscillatory system. Fig. 52.3 shows the mechanism of a pendulum clock. The escape wheel is made fast to a toothed drum over which a chain carrying a weight is passed. At the end of the pendulum there is an anchor with two pallets of ruby or other hard material, bent to an arc of a circle with the centre on the axis of the pendulum. In wrist-watches, the weight will be replaced by a main spring, and the pendulum by a balance wheel which rotates and contra-rotates against a hair spring about its pivot.

In this case the oscillatory element is the pendulum or the balance wheel, respectively. The energy source is the pulled-up weight or the wound mainspring. The gate is the anchor which allows the escape wheel to advance one tooth every half-period. Feedback is provided by the interaction of the anchor and the escape wheel.

2. At the instant when the pendulum passes through the position of equilibrium and has a maximum velocity, a tooth of the escape wheel touches momentarily the end of the respective pallet. By scraping against the pallet, the tooth pushes the pendulum, thereby im-

parting it an impulse of energy. As a result, the weight goes down one link of the chain. In this way, the potential energy of the weight (or of the mainspring) is gradually transmitted to the pendulum to make up for friction losses.

### 52.3. THE HARMONIC VALVE OSCILLATOR

1. Let us set up the circuit of Fig. 52.4 and connect it to the input of a CRT oscilloscope. On closing the switch in the circuit, continuous oscillations will set in, as revealed on the CRT screen.

This circuit is a *harmonic valve oscillator*, an electric self-sustained oscillatory system in which the anode battery acts as energy source, the resonant, or tuned, circuit in the anode lead as an oscillatory element, the grid of the triode valve which controls the anode current (Sec. 47.3) as a gate, and the coil connected between the cathode and grid of the triode and inductively coupled to the tuned-circuit coil as a feedback element applying the feedback signal to the grid.

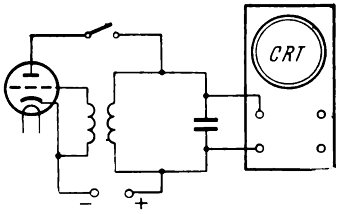


Fig. 52.4

2. It may so happen that no oscillations are produced in an oscillatory circuit after the switch is closed, but they immediately set in, if the leads of the feedback coil are interchanged. This is an indication that previously the phase of the feedback signal was wrong.

### 52.4. BUILD-UP OF SELF-SUSTAINED OSCILLATIONS

1. Let us start a slow sweep on the oscilloscope the same instant as the switch is closed in the resonant circuit of the valve oscillator shown in Fig. 52.4. The oscilloscope should display the pattern shown in Fig. 52.5. As is seen, the oscillations are building up, that is, after the energy source is turned on the amplitude of oscillations continuously rises from zero to a certain steady-state value. This build-up of oscillations is an important feature of any self-sustained oscillatory system.

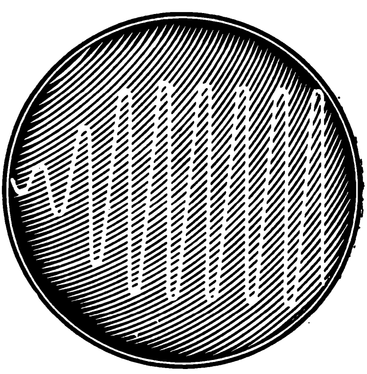


Fig. 52.5

The build-up mechanism of a valve oscillator may be explained as follows. The weak oscillations of current produced in the resonant circuit upon closure of the switch give rise to alternating changes in the magnetic field around the resonant-circuit coil. The reversibly changing magnetic flux induces an emf in the feedback coil (Sec. 43.7), and the grid potential alternately rises above and drops below the cathode potential, bringing about changes in the anode current. With the phase of the feedback signal properly chosen, this causes a further increase in the oscillations of the resonant circuit, increased changes in the grid potential and increased oscillations in the circuit.

2. Why is it then that this mutual build-up of oscillations in the grid and anode circuits does not carry on infinitely? How does a self-sustained oscillatory system attain a steady-state condition which is independent of the initial conditions?

It has been found that in a valve oscillator the build-up ceases because of the nonlinearity in the triode characteristic (Sec. 47.3). It is seen from Fig. 47.8 (see Vol. 1, page 518) that anode current rises with increasing grid potential only within the linear portion of the triode characteristic. It is within this region that the oscillations build up. When the anode current reaches the saturation value, the oscillation build-up ceases. It can be shown that any self-sustained oscillatory system must contain a nonlinear element. No steady-state self-sustained oscillations are possible in a linear system.

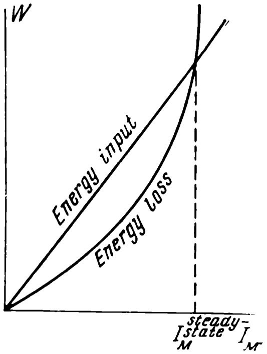


Fig. 52.6

3. The amplitude of steady-state self-sustained oscillations is found from the condition of energy balance which requires that energy input over an oscillatory cycle, $W = k_1 I_{\mathrm{M}} \mathscr{E} T$, should be equal to the energy lost over the same cycle, $\Delta W_{heat} = k_2 I_{\mathrm{M}}^2 R T$, where $k_1$ and $k_2$ are proportionality coefficients. From $W = \Delta W_{heat}$, it follows that

$$I_{\mathrm{M}}^{steady\text{-}state} = k_1 \mathscr{E} / k_2 R = K \mathscr{E} / R$$

The graphs of both functions, namely the linear function of the current amplitude for energy input and the quadratic function for energy losses, are shown in Fig. 52.6. The steady-state amplitude obtains at the point of intersection of the two graphs. As is seen, at current amplitudes below the steady-state value, energy input exceeds energy losses, and the amplitude rises progressively. Conversely, when the amplitude of self-sustained oscillations exceeds the steady-state value, energy losses are greater than energy input from the supply source. Because of this the oscillations begin to die away until their amplitude becomes equal to the steady-state value.

We leave it as an exercise for the reader to analyze the energy balance in a mechanical self-sustained oscillatory system by analogy.

# Chapter 53

# FORCED VIBRATIONS

## 53.1. SINUSOIDAL DRIVING FORCE

1. When a periodic driving force applied externally acts upon an oscillatory system free from self-sustained oscillations, *forced* continuous vibrations occur in the system. Since any complex vibrations may be resolved into sinusoidal components, or harmonics (see Sec. 50.4), we shall limit ourselves mainly to the forced vibrations caused by a *sinusoidal*, or *harmonic*, driving force such that

$$F = F_{\mathrm{M}} \cos \omega t \qquad (53.1)$$

where $F_{\mathrm{M}}$ is the maximum driving force and $\omega$ is its radian frequency.

In addition to the driving force, the oscillatory system (for example, a spring pendulum) is acted upon by elastic and friction forces. According to the main postulate of dynamics (see Sec. 7.1), we have

$$F + F_{el} + F_{fr} = ma \qquad (53.2)$$

or

$$F_{\mathrm{M}} \cos \omega t - ks - hv = ma \qquad (53.3)$$

2. Experience shows that a sinusoidal driving force applied to a spring pendulum will cause its bob to oscillate harmonically at the frequency of the driving force. Similarly, if a source of sinusoidal emf is switched into an oscillatory circuit, forced oscillations of current will be produced in the circuit, whose frequency is the same as that of the driving emf. The latter statement can readily be verified with a dual-beam oscilloscope, feeding the oscillations of emf to one input, and the oscillations of current to the other (Sec. 49.4).

The expression for the forced vibrations has the form

$$s = A \cos (\omega t + \varphi) \qquad (53.4)$$

3. To determine the amplitude $A$ and the reference phase, or epoch, $\varphi$ of forced vibrations, we substitute in (53.3) the values of displacement, velocity and acceleration according to (53.4), (49.2) and (49.3) and get

$$F_{\mathrm{M}} \cos \omega t - kA \cos (\omega t + \varphi) + h\omega A \sin (\omega t + \varphi) = -m\omega^2 A \cos (\omega t + \varphi) \qquad (53.5)$$

Equation (53.5) will be simplified, if the system has a high $Q$, as the term representing friction forces may be discarded. As a result, we obtain

$$F_M \cos \omega t - kA \cos (\omega t + \varphi) = -m\omega^2 A \cos (\omega t + \varphi) \qquad (53.6)$$

This equality must be satisfied identically, which is possible only if $\varphi = 0$, $\varphi = \pi$ or $\varphi = -\pi$.

4. In the former case, equation (53.6) will take the form

$$F_M - kA = -m\omega^2 A \qquad (53.7)$$

Noting that the elasticity coefficient $k = m\omega_0^2$, for the amplitude of displacement we have

$$A = F_M/m\,(\omega_0^2 - \omega^2) \qquad (53.8)$$

Since the amplitude is a substantially positive quantity, expression (53.8) has a meaning when the radian frequency of the driving force, $\omega$, is below the natural radian frequency of the system, $\omega_0$. Then the system will oscillate in phase with the driving force. If, on the other hand, the frequency of the driving force exceeds the natural frequency of the system, the system will oscillate in antiphase with the driving force ($\varphi = -\pi$), in which case the absolute value of expression (53.8) should be taken for the amplitude of displacement.

## 53.2. RESONANCE

1. When the frequency of the driving force, $\nu$, coincides exactly with the natural frequency of the system, $\nu_0$ (or $\omega = \omega_0$), expression (53.8) loses all sense because division by zero is impossible. If, on the other hand, $\omega \rightarrow \omega_0$, then $A \rightarrow \infty$, which has no physical meaning either, because in such a case damping cannot be ignored for fundamental reasons.

The condition when the frequency of the driving force coincides with the natural frequency of the oscillatory system is called *resonance*. In order to find the amplitude and phase at resonance, we substitute $\omega = \omega_0$ in (53.5). Since $k = m\omega_0^2$, equation (53.5) will take the form

$$F_M \cos \omega_0 t + h\omega_0 A_{res} \sin (\omega_0 t + \varphi) = 0 \qquad (53.9)$$

This expression must be satisfied identically, at any instant. This is possible only when $\varphi = -\pi/2$. The displacement at resonance is given by

$$A_{res} = F_M/h\omega_0 = (F_M/m\omega_0^2)\,(m\omega_0/h) = QA_{st} \qquad (53.10)$$

where $A_{st} = F_M/k = F_M/m\omega_0^2$ is the displacement due to a static force.

Fig. 53.1 shows a graph relating the amplitude of displacement to the frequency of the driving force. This graph is called the *resonance curve*. Far away from resonance, the graph is plotted by equation (53.8); at frequencies close to the natural frequency, the amplitude $A$ is very close to the resonant amplitude, $A_{res}$. The sharp resonant peak shown in Fig. 53.1 is characteristic of a high-$Q$ system. The resonance curve for a low-$Q$ system is shown in the same figure by the dotted line.

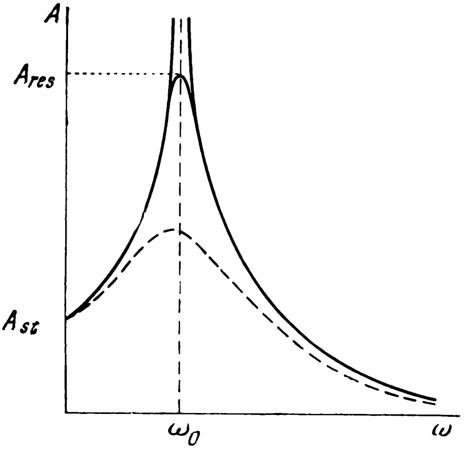


Fig. 53.1

We leave it as an exercise for the reader to derive an equation for the amplitude of velocity, (49.6), at resonant and other frequencies, and to plot a graph of the relationship, that is, the resonance curve for the amplitude of velocity.

2. An unbounded build-up of response amplitude at resonance in a high-$Q$ system may cause its failure. Several cases have been reported where the failure of a structure was caused by operation of low-power engines at a frequency equal to the natural frequency of the structure. For the same reason, unbalanced crankshafts, propeller shafts, air screws, turbine rotors and shafts fail.

Yet, there are ways for using resonance to advantage in mechanics, acoustics and radio. Some of these applications will be discussed later.

## 53.3. RESONANCE AND HARMONIC ANALYSIS

1. Resonance may be utilized in the harmonic analysis of non-sinusoidal oscillatory forces. For this purpose, a system of *resonators* is set up to cover the frequency range of interest, and is subjected to the force under investigation. Strong oscillations will occur only in the resonators whose frequencies coincide with the respective harmonics of the applied force.

As an example, consider the vibrating reed type of frequency meter (Fig. 53.2). In this instrument, there are numerous reeds, or steel strips with solder weights at one end and attached to a common bar at the other. Also attached to this bar is an armature arranged over the pole of an electromagnet energized from the circuit to be measured. When the coil of the electromagnet is energized with alternating current, the armature starts vibrating and drives the bar and the reeds attached to it.

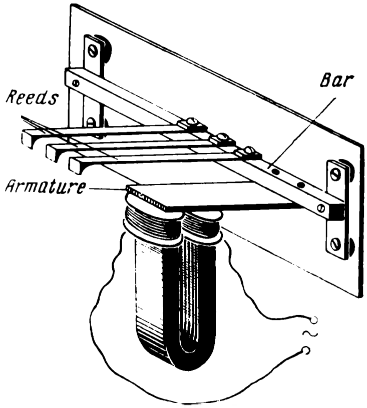


Fig. 53.2

Incidentally, this system is capable of sensing mechanical vibrations as well. In this application, the instrument should be fastened to the mechanical system in which vibrations are to be measured.

2. The reeds are accurately adjusted by means of solder weights and for length so that the frequencies of adjacent strips differ by the same amount, say 0.5 Hz. Then a system of 25 reeds will cover a frequency range of 12 Hz.

When a sinusoidal current energizes the magnet coil, only the reed with a period corresponding to the alternations of the magnetic field will be set in vibration. Fig. 53.3*a* illustrates a case in which one of the reeds resonates at 50 Hz. A different indication is obtained when the electromagnet is energized with a non-sinusoidal current (Fig. 53.3*b*). In this case, three reeds are set in vibration, namely at 47.5, 50 and 52.5 Hz, their amplitudes being different.

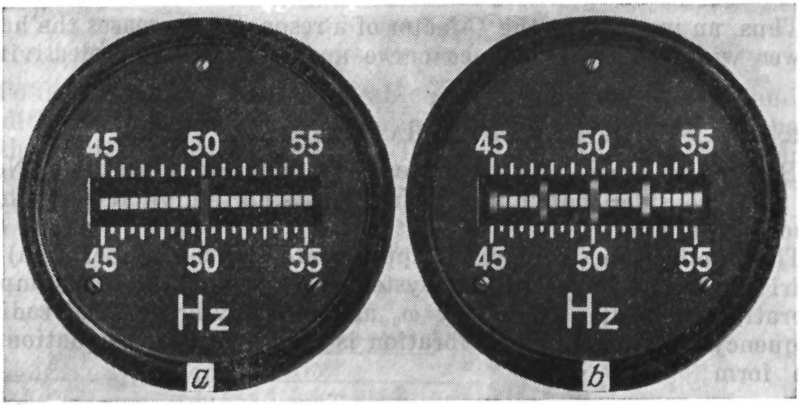


Fig. 53.3

## 53.4. HALF-POWER WIDTH OF THE RESONANCE CURVE. SELECTIVITY

1. A resonator with a sharp resonance curve has a good *selectivity*, that is, of two closely spaced frequencies it will sense only one which coincides with its own natural frequency. In contrast, a resonator with a broad resonance curve will respond to both frequencies almost identically.

Quantitatively, the selectivity of a resonator is described in terms of the *half-power width*, $\Delta\omega$, *of the resonance curve*. The half-power width of the resonance curve is the difference between the resonant frequency $\omega_0$ and the frequency $\omega_1$ at which the energy of forced vibrations in the resonator is one-half that at the natural frequency (with the same amplitude of the driving force). Using the expression for oscillatory energy, (49.17), we have

$$1/2m\omega_1^2A^2 = 1/4m\omega_0^2A_{res} \qquad (53.11)$$

or, noting (53.8) and (53.10)

$$\omega_1^2/(\omega_0^2 - \omega_1^2) = Q/2 \qquad (53.12)$$

For a high-$Q$ resonator, $\omega_1$ and $\omega_0$ are very close to each other, so that $\Delta\omega = |\omega_0 - \omega_1| \ll \omega_0$, and $\omega_1 + \omega_0 \approx 2\omega_0$. Substituting in (53.12) yields finally

$$\Delta\omega \approx \omega_0/Q \qquad (53.13)$$

Thus, an increase in the $Q$-factor of a resonator decreases the half-power width of its resonance curve and increases its selectivity.

## 53.5. BUILD-UP OF FORCED VIBRATIONS

1. So far, we have dealt with steady-state forced vibrations. Now we shall consider their build-up from rest. For simplicity, we shall ignore damping.

The application (through an appropriate switching operation) of a driving force in an oscillatory system gives rise to both free damped vibrations at radian frequency $\omega_0$ and forced vibrations at radian frequency $\omega$. The complex vibration is described by an equation of the form

$$s = A\cos\omega t + B\cos\omega_0 t \qquad (53.14)$$

The velocity and acceleration of the system can be obtained by the same methods as in Sec. 49.1 (compare with (49.2) and (49.3)):

$$v = -A\omega\sin\omega t - B\omega_0\sin\omega_0 t \qquad (53.15)$$

$$a = -A\omega^2\cos\omega t - B\omega_0^2\cos\omega_0 t \qquad (53.16)$$

2. To determine the amplitudes $A$ and $B$, we shall use initial conditions such that at $t = 0$, the displacement is $s_0 = 0$, the velocity $v_0 = 0$, the acceleration $a_0 = F_{\mathrm{M}}/m$. Substituting in (53.14) and (53.16) and carrying out simple manipulations, we obtain

$$B = -A, \quad A = F_{\mathrm{M}}/m\,(\omega_0^2 - \omega^2) \qquad (53.17)$$

Finally, expression (53.14) takes the form

$$s = F_{\mathrm{M}}\,(\cos \omega t - \cos \omega_0 t)/m\,(\omega_0^2 - \omega^2) \qquad (53.18)$$

As is seen, this is a non-sinusoidal vibration. If $\omega$ and $\omega_0$ are close to each other, beats will be produced in the system (see Sec. 50.1).

3. At first sight, it may appear that the above result contradicts experience which shows that a sinusoidal driving force causes the system to oscillate harmonically at the frequency of excitation (see Sec. 53.1). However, this contradiction has appeared because we have ignored the damping of free vibrations. Thus, the results of both sections, that is, Sec. 53.1 and the present one, are correct, but for different instants: the former holds for steady-state vibrations, and the latter applies to vibrations in the making.

In other words, equation (53.14) and its corollary, equation (53.18), may be used for a short interval of time at the beginning of motion, where the damping of free vibrations may be ignored, or more specifically when $t < \tau \approx Q/\omega_0$, (51.4). At $t \gg \tau$, free vibrations are already damped out, and only forced sinusoidal vibrations at the driving force frequency, equation (53.4), exist in the system.

## 53.6. BUILD-UP OF VIBRATIONS AT RESONANCE

1. To determine the manner in which vibrations are built up under conditions of resonance, in expression (53.18) it is necessary to pass to the limit on letting $\omega \to \omega_0$. Transforming the numerator and denominator of (53.18) and noting that the sine of a small angle is practically equal to the angle in radians, we obtain

$$s = \lim_{\omega \to \omega_0} \frac{2F_{\mathrm{M}} \sin \dfrac{(\omega + \omega_0)\,t}{2} \sin \dfrac{(\omega_0 - \omega)\,t}{2}}{m\,(\omega_0^2 - \omega^2)}$$

$$= \frac{F_{\mathrm{M}} t}{2m\omega_0} \sin \omega_0 t \lim_{\omega \to \omega_0} \frac{\sin \dfrac{(\omega_0 - \omega)\,t}{2}}{(\omega_0 - \omega)\,t/2}$$

The last limit is unity. As a consequence, the build-up of vibrations at resonance is described by

$$s = F_{\mathrm{M}} t \sin \omega_0 t/2m\omega_0 \qquad (53.19)$$

Thus, if a high-$Q$ oscillatory system at rest is acted upon by a driving force whose frequency coincides with the natural frequency of the system (the condition of resonance), its amplitude will rise with time

$$A = F_{\mathrm{M}}t/2m\omega_0 \tag{53.20}$$

2. If the oscillatory system were free from friction, the amplitude of the response would grow without bound. Actually, the response rises in amplitude until the work done by the friction force becomes equal to the work done by the driving force (Fig. 53.4). This condition determines the steady-state resonant amplitude given by equation (53.10).

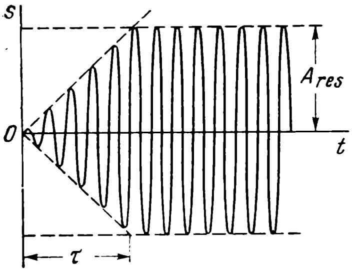


Fig. 53.4

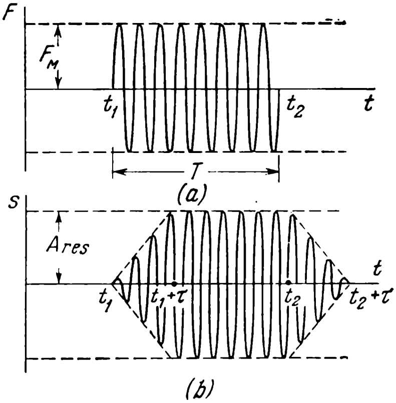


Fig. 53.5

To find the time necessary for the resonant amplitude to attain its steady-state value, we equate expressions (53.10) and (53.20) and get

$$F_{\mathrm{M}}/h\omega_0 = F_M\tau/2m\omega_0 \quad \text{or} \quad \tau = 2m/h = 2Q/\omega_0 \tag{53.21}$$

Thus, the rise time of vibrations at resonance and the damping time of free vibrations, (51.4), are practically equal. It should be noted that the two expressions only define the order of magnitudes.

## 53.7. RESPONSE TO SINUSOIDAL PULSES

1. Consider the response of a resonator to a driving force which has the shape of a sinusoidal pulse (Fig. 53.5*a*). This waveform is also known as an interrupted sine wave or a wave packet. Incidentally, these waveforms are obtained in keying dashes and dots in Morse code. Let the period of the driving force be $T_0 = 2\pi/\omega_0$ where $\omega_0$ is the natural frequency of the resonator. The pulse duration is $T \gg T_0$, the pulse rise and decay times are $\tau = Q/\omega_0$.

As the excitation is suddenly applied at time $t_1$, oscillations in the resonant amplitude begin to build up so that by time $t_1 + \tau$ they attain the resonant amplitude. At time $t_2 = t_1 + T$ the driving force is removed, but the oscillations in the resonator persist until time $t_2 + \tau$, when they are damped out (Fig. 53.5*b*).

2. As is seen, the resonator responds to a sinusoidal pulse as if it were a non-sinusoidal oscillation similar in waveform to beats. If the build-up (rise) time, $\tau$, is a fraction of the pulse duration, $T$, the waveform of the sinusoidal pulse is distorted insignificantly. If, on the other hand, $\tau \geqslant T$, the sinusoidal pulse will be distorted heavily.

This implies that in order to reproduce the original sinusoidal pulse faithfully, the resonator should have a short rise time, that is, a low $Q$-factor. Thus, the requirement for high fidelity (faithful reproduction of a signal) runs counter to the requirement for high selectivity (see Sec. 53.4). Incidentally, this conflict between fidelity and selectivity is typical of the reception of any modulated signals, and not only of sinusoidal pulses.

## 53.8. UNCERTAINTY RELATIONS FOR FREQUENCY AND TIME

1. A resonator with which an oscillatory process is investigated or simply sensed introduces an amount of uncertainty in the measurement of both frequency and time. This is because the limit to the accuracy with which frequency can be measured is set by the half-power width of the resonance curve, and the limit to the accuracy of time measurement is set by the rise time of the resonator. Therefore, instead of a single value of frequency or time, we have a frequency range, $\Delta\omega$, and a time range, $\Delta t$. The uncertainty in frequency can be evaluated by equation (53.13), and that in time by equation (53.21):

$$\Delta\omega \approx \omega_0/Q, \quad \Delta t \approx Q/\omega_0$$

2. As is seen, varying the $Q$-factor of a resonator will minimize the uncertainty in one of the quantities, while the uncertainty in the other will increase in proportion. The product of the uncertainties in the two conjugate variables, frequency and time, is independent of the properties of a resonator:

$$\Delta\omega \cdot \Delta t \approx 1 \tag{53.22}$$

The above expression is the uncertainty relationship for frequency and time. It will come in useful in discussing the elements of quantum mechanics (Sec. 70.2).

# Chapter 54
# ALTERNATING CURRENT

## 54.1. THE SYNCHRONOUS ALTERNATOR

1. A general view of a synchronous alternator is shown in Fig. 54.1. It consists essentially of a stationary part, called the *stator*, and a revolving part, called the *rotor*. The stator is assembled from electrical-sheet steel laminations of low coercive force, that is, with a narrow hysteresis loop, and carries a winding. The rotor which revolves inside the stator is essentially an electromagnet (in small alternators, it is a permanent magnet). As the rotor turns, its magnetic field revolves too, and the magnetic flux linking the stator winding is varying continuously.

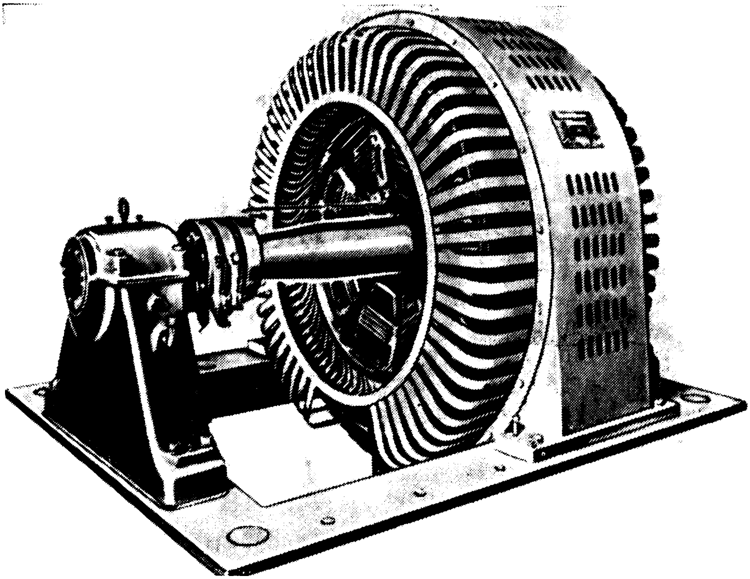


Fig. 54.1

In a certain position, the stator winding is linked by a maximum magnetic flux; in a quarter of a revolution the flux through the winding will be zero; in another quarter of a revolution the flux will again be a maximum in magnitude but opposite in sign because the magnet poles have exchanged position. By giving an appropriate shape to the pole pieces, the magnetic flux may be caused to vary as a cosine function

$$\Phi = \Phi_M \cos \omega t \qquad (54.1)$$

where $\omega$ is the angular velocity of the rotor.

2. Change in the magnetic flux induces in the stator winding an emf which, according to Faraday's law of magnetic induction, (43.10), has the form $\mathscr{E} = -\Delta\Phi/\Delta t$. The relationship between the induced emf and the magnetic flux is precisely the same as that between velocity and displacement, (49.1) and (49.2). Similarly, it is proved that the emf induced in the winding is a sinusoid:

$$\mathscr{E} = \omega\Phi_M \sin \omega t = \mathscr{E}_M \sin \omega t \qquad (54.2)$$

The peak value, or *amplitude*, of the emf, $\mathscr{E}_M$ is equal to the product of the angular velocity of the rotor, $\omega$, and the peak value of the magnetic flux, $\Phi_M$:

$$\mathscr{E}_M = \omega\Phi_M \qquad (54.3)$$

3. The alternator shown in Fig. 54.1 is called *synchronous* because its emf varies in synchronism with the rotation of the rotor, and their frequencies coincide. The mechanical energy required to drive the rotor is supplied by a prime mover, which may be a hydraulic turbine at a hydraulic power station, a steam turbine at a fuel-fired power plant, or a heat engine. As a rule, the alternator rotor is mounted on the same shaft with the rotor of the prime mover.

The Soviet Union builds the world's largest alternators. For example, the Leningrad Metal Works and the "Elektrosila" Works turn out hydraulic turbine-alternator units for power outputs of over 500 MW and steam turbine-alternator units for power outputs of over 800 MW.

4. The sinusoidal emf gives rise to what is called an *alternating current*. In effect, an alternating current represents forced oscillations of current in an electric circuit, and all that we have learned in Chapter 53 in the analysis of forced vibrations in mechanical oscillatory systems is applicable to it. The relationships thus obtained can be translated from "mechanics language" to the "language of electric circuits" by the use of Table 51.1.

At this point, however, a very important remark is in order. In Chapter 53 we limited ourselves to the examination of forced vibrations in high-$Q$ systems. Meanwhile, alternating-current (a.c.) circuits often have a very small or even nearly zero $Q$-factor. Such circuits are combinations of a resistor and a capacitor, a resistor and a coil, a separate resistor, etc. In such circuits free oscillations cannot be excited, while forced oscillations can, which is proved by experiments. In this chapter we shall mainly deal with such circuits.

## 54.2. A.C. CIRCUITS

1. Let an incandescent lamp be connected in series with a resistor $R$, a bank of capacitors of capacitance $C$, and a high inductance coil $L$ (Fig. 54.2). If this circuit be now connected across the terminals of an alternator, the lamp will illuminate, thereby indicating that a current is flowing around the circuit.

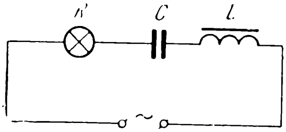


Fig. 54.2

Three alternating electric fields are at work in this circuit. Above all, this is the field due to the external source, the alternator. The emf due to it is given by equation (54.2), $\mathscr{E}=\mathscr{E}_M \sin \omega t$. Then there is the field of self-induction represented by the emf given by equation (43.16), namely $\mathscr{E}_L = -L\,\Delta i/\Delta t$. The third field is due to the charges accumulated on the capacitors, it is represented by a potential difference. According to Ohm's law for a circuit containing different kinds of elements, equation (39.29), we have

$$iR = \mathscr{E} + \mathscr{E}_L + \varphi_2 - \varphi_1 \tag{54.4}$$

2. This expression may be caused to take the following form

$$\mathscr{E} = iR + (\varphi_1 - \varphi_2) - \mathscr{E}_L = u_R + u_C + u_L \tag{54.5}$$

Here, the voltage across the resistance is

$$u_R = iR \tag{54.6}$$

that across the capacitive reactance is

$$u_C = \varphi_1 - \varphi_2 = q/C \tag{54.7}$$

and that across the inductive reactance is

$$u_L = -\mathscr{E}_L = L\,\Delta i/\Delta t \tag{54.8}$$

Thus, we have treated the events that happen in an a.c. circuit in a somewhat different way. Instead of the action of three fields on one resistance, we deal with the action of one external field on three impedances, namely: resistance $R$, capacitive reactance $X_C$, and inductive reactance $X_L$. This approach is especially convenient in the analysis of a.c. circuits.

## 54.3. RESISTANCE

1. Consider a circuit in which the voltage across the reactances is a small fraction of that across the resistor. This is a resistive circuit. Discarding $u_C$ and $u_L$ in (54.5) and substituting for $\mathscr{E}$ its expres-

sion (54.2), we obtain

$$iR = \mathscr{E}_{\mathrm{M}} \sin \omega t \tag{54.9}$$

Thus, the current varies as a sine function

$$i = I_{\mathrm{M}} \sin \omega t \tag{54.10}$$

where the peak value or amplitude of the alternating current, $I_{\mathrm{M}}$, is given by

$$I_{\mathrm{M}} = \mathscr{E}_{\mathrm{M}}/R \tag{54.11}$$

2. As is seen, the sinusoidal emf applied to a resistive circuit gives rise to harmonic oscillations of current at the frequency and phase of the applied emf. The graphs of variations in current and emf for the resistive circuit are shown in Fig. 54.3.

## 54.4. AVERAGE AND ROOT-MEAN-SQUARE VALUES OF CURRENT AND VOLTAGE

1. In a resistive circuit, electric energy is irreversibly converted to the internal energy of the conductor, that is, Joule heat (see Sec. 39.7). The instantaneous heating effect, or instantaneous power, is given by the product of the instantaneous current and the instantaneous emf:

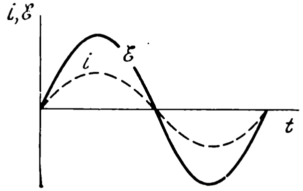


Fig. 54.3

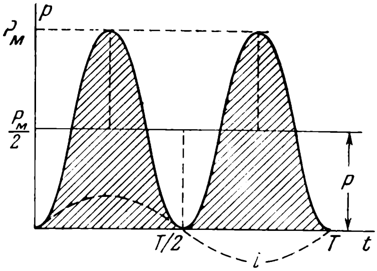


Fig. 54.4

$$p = i\mathscr{E} = I_{\mathrm{M}}^2 R \sin^2 \omega t \tag{54.12}$$

In the graph of Fig. 54.4 this function is represented by the solid line. For comparison, the dotted line represents the current. The *peak value* or *amplitude* of power is

$$P_{\mathrm{M}} = I_{\mathrm{M}}^2 R \tag{54.13}$$

2. Now let us find the *average* (or *mean*) value of a.c. power over a cycle. It is found by dividing the work done by the current over

the cycle (that is the Joule heat liberated in the mean time) by the period. The computations will be assisted by the plot of Fig. 54.4.

From the definition of power (see Sec. 16.5), an elementary work $\Delta A = p\Delta t$; total work is represented by the area under the curve. From the property of the sinusoid it follows that the area shaded under the curve is equal to the area of a right triangle whose base is equal to the period and whose height is equal to half the peak value of power. Then, the work over a cycle is

$$A = 1/2 P_{\mathrm{M}} T = 1/2 I_{\mathrm{M}}^2 RT \qquad (54.14)$$

and the average (mean) power is

$$P = A/T = 1/2 P_{\mathrm{M}} = 1/2 I_{\mathrm{M}}^2 R \qquad (54.15)$$

3. By comparing equation (54.15) with the expression for d.c. power, $P = I^2R$, we get

$$I^2R = 1/2 I_{\mathrm{M}}^2 R$$

Hence,

$$I = I_{\mathrm{M}}/\sqrt{2} \qquad (54.16)$$

Thus, in its thermal (or mechanical) effect an alternating current of amplitude $I_{\mathrm{M}}$ is equivalent to a direct current of magnitude $I = I_{\mathrm{M}}/\sqrt{2}$. This quantity is called the *root-mean-square* (*rms*) or *effective*, value of alternating current. Accordingly, the quantities

$$\mathscr{E} = \mathscr{E}_{\mathrm{M}}/\sqrt{2} \quad \text{and} \quad V = V_{\mathrm{M}}/\sqrt{2}$$

are called the *rms* or *effective* values of emf and voltage.

## 54.5. CAPACITIVE REACTANCE

1. Let us assemble a circuit similar to that shown in Fig. 54.2, omitting the coil ($u_L = 0$), and choose the capacitance of the bank of capacitors and the resistance of the incandescent lamp such that $u_C \gg u_R$. The result will be a capacitive circuit. Experiments show that oscillations of current in a capacitive circuit are harmonic oscillations at the frequency of the driving emf. In contrast to a resistive circuit where the current and the emf are in phase, in a capacitive circuit the current is a cosine function:

$$i = I_{\mathrm{M}} \cos \omega t = I_{\mathrm{M}} \sin (\omega t + \pi/2) \qquad (54.17)$$

In a capacitive circuit, the current is said to lead the applied emf in phase by $\pi/2$. Naturally, it may be said that the applied emf lags behind the current in phase by the same angle (Fig. 54.5).

2. It may be shown that the peak values of current and emf in a capacitive circuit are related as

$$I_M = C\omega\mathscr{E}_M = \mathscr{E}_M/X_C \qquad (54.18)$$

where $X_C$ is the capacitive reactance defined as

$$X_C = 1/C\omega \qquad (54.19)$$

## 54.6. INDUCTIVE REACTANCE

1. On shorting out the bank of capacitors in the circuit of Fig. 54.2, we obtain a circuit containing only an inductive reactance ($u_C = 0$, $u_L \gg u_R$). From the graph of Fig. 54.6 it is seen that the current through an inductive circuit lags on the voltage by an angle $\varphi = -\pi/2$. The instantaneous current through an inductive circuit is given by

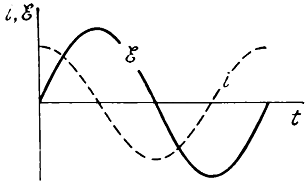


Fig. 54.5

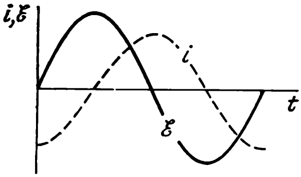


Fig. 54.6

$$i = -I_M \cos \omega t = I_M \sin (\omega t - \pi/2) \qquad (54.20)$$

2. It may be shown that the relationship between the peak values of current and emf in an inductive circuit has the form

$$I_M = \mathscr{E}_M/L\omega = \mathscr{E}_M/X_L \qquad (54.21)$$

where $X_L$ is the inductive reactance defined as

$$X_L = L\omega \qquad (54.22)$$

## 54.7. OHM'S LAW FOR AN A.C. CIRCUIT

1. Let us go back to the a.c. circuit containing a resistance, a capacitance, and an inductance (see Fig. 54.2). The same current flows through the three elements connected in series:

$$i = I_M \sin \omega t$$

The voltages across the three elements have the form

$$\left.\begin{aligned} u_R &= iR = I_M R \sin \omega t \\ u_C &= I_M X_C \sin\left(\omega t - \frac{\pi}{2}\right) = -I_M X_C \cos \omega t \\ u_L &= I_M X_L \sin\left(\omega t + \frac{\pi}{2}\right) = I_M X_L \cos \omega t \end{aligned}\right\} \qquad (54.23)$$

According to equation (54.5), the emf $\mathscr{E} = \mathscr{E}_M \sin(\omega t + \varphi)$ is a sum of voltage oscillations, (54.23). Since they differ in phase, they can conveniently be combined, using a vector diagram (see Sec. 49.6). A vector diagram for voltages is shown in Fig. 54.7.

2. Using the Pythagorean theorem, we have

$$\mathscr{E}_M = \sqrt{I_M^2 R^2 + I_M^2 (X_L - X_C)^2} \tag{54.24}$$

As is seen, the peak values of current and emf are related as

$$I_M = \mathscr{E}_M / Z \tag{54.25}$$

where the quantity

$$Z = \sqrt{R^2 + (X_L - X_C)^2} = \sqrt{R^2 + (L\omega - 1/C\omega)^2} \tag{54.26}$$

is called the *impedance* of an a.c. circuit.

Expression (54.25) is known as *Ohm's law for an a.c. circuit*. Dividing the left-hand and right-hand sides of equation (54.25) by $\sqrt{2}$ shows that this law also holds for the rms values of current and emf.

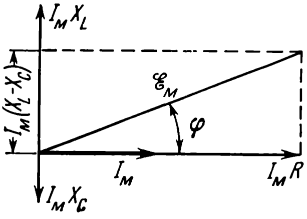


Fig. 54.7

3. The phase shift between the current in and the emf across a circuit may be found, using a vector diagram. From Fig. 54.7 it follows that

$$\cos\varphi = I_M R / \mathscr{E}_M = R/Z. \tag{54.27}$$

4. It follows from equation (54.25) that the peak value of current depends on the frequency of the emf. The amplitude of current is a maximum when the impedance is a minimum, that is, at $L\omega - 1/C\omega = 0$. The respective frequency, $\omega = \sqrt{1/CL} = \omega_0$ is equal to the natural frequency of the circuit. In other words, the amplitude of current is a maximum at resonance.

## 54.8. A.C. POWER

1. In an a.c. circuit containing resistance, capacitance, and inductance, energy is dissipated irreversibly only across the resistance $R$, while the peak value of current is limited by the impedance $Z$. To find the active power, that is, the average power associated with the irreversible dissipation of energy in an a.c. circuit, we shall use expression (54.15). Substituting $I_M = \mathscr{E}_M/Z$ from equation (54.25) gives:

$$P = 1/2\, I_M(\mathscr{E}_M/Z)\, R = (I_M/\sqrt{2})\,(\mathscr{E}_M/\sqrt{2})\,(R/Z)$$

However, the first two terms are the rms values of current and emf (see Sec. 54.4), while the last term $R/Z = \cos\varphi$, (54.27). Thus, the power dissipated in the resistive element, known as the *active* or *true power*, is given by

$$P = I\mathscr{E} \cos \varphi \tag{54.28}$$

2. The quantity $R/Z = \cos \varphi$ is called the *power factor* (abbreviated *PF* or *pf*). It plays an important part in electrical engineering. The point is that if there is a marked phase shift between the current in, and the emf across, a circuit, the power factor will be small, and the load will draw a small active power from the generator. At the same time, the generator should supply a total power given by

$$S = I\mathscr{E} \tag{54.29}$$

The same power should be supplied to the generator by its prime mover. Thus, at a low power factor the load will take only part of the energy supplied by the generator. The balance will be exchanged between the generator and the load and dissipated in transmission lines.

3. It follows from equation (54.28) that transfer of energy from a supply to the associated system will be a maximum at resonance. This is because at resonance the power factor will be unity ($\cos \varphi = 1$), and the peak value of current will be a maximum. Precisely the same result can be obtained for a mechanical oscillatory system.

## 54.9. THE TRANSFORMER

1. It may sometimes happen that a single supply of alternating current is required to energize loads designed for different voltages. For example, in a TV set plugged into the 220-V mains the valve filaments operate on 6.3 V, the anode circuits on 200 V to 500 V, and the picture tube on 15 kV. It is obvious that the TV set should use one or several devices for stepping up or down the mains voltage to the requisite values. Such devices are called transformers.

The arrangement of a transformer is shown in Fig. 54.8. It has a core assembled from electrical-sheet steel laminations of low coercive force. The core carries a primary winding with turns $w_1$ and a secondary winding with turns $w_2$ (or several secondaries).

2. When its secondary circuit is open, a transformer is said to be operating at no-load. If properly designed, the primary winding will then draw a very small no-load current, and the transformer will dissipate a very small no-load power, $P_{no\text{-}load}$, which is practically equal to the power expended to magnetize the core cyclically, that is, the hysteresis losses, also known as iron losses, $P_{iron}$. Thus,

$$P_{no\text{-}load} \approx P_{iron} \tag{54.30}$$

3. Let us determine the voltages across the transformer windings. The no-load current causes the core to be magnetized cyclically, and,

the two transformer windings are linked by an alternating magnetic flux, $\Phi$. According to Faraday's law, equation (43.10), we have

$$\mathscr{E}_1 = -w_1 \Delta\Phi/\Delta t, \quad \mathscr{E}_2 = -w_2 \Delta\Phi/\Delta t$$

Hence, the induced emfs are proportional to the number of turns in the windings

$$\mathscr{E}_1/\mathscr{E}_2 = w_1/w_2 = k \tag{54.31}$$

Here, $k = w_1/w_2$ is the turns or transformation ratio. In step-up transformers, the secondary winding carries a greater number of turns

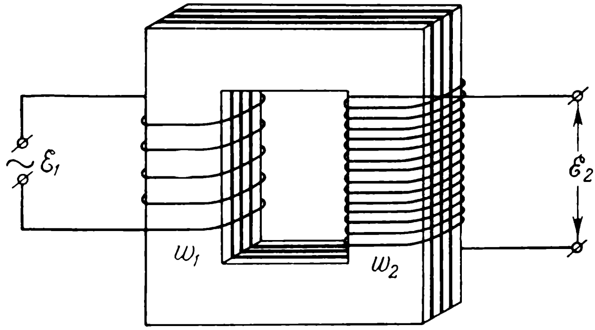


Fig. 54.8

than the primary winding, that is, $w_2 > w_1$. In other words, the voltage across the secondary winding is higher than that across the primary winding.

4. When the transformer secondary is connected across a resistive load, a current begins to flow in the secondary winding, whose rms value is $I_2$; the voltage across the secondary winding will be $V_2$, and the phase shift will be $\varphi_2$. According to Lenz's law, the current through the secondary winding opposes the change of the magnetic flux in the core, because of which the inductive reactance of the primary winding is decreased and the current in the primary winding increases, and the rms current in the primary of a loaded transformer is greater than its no-load current, that is, $I_1 > I_{no\text{-}load}$.

According to the law of conservation of energy,

$$P_2 = P_1 - P_{copper} - P_{iron} \tag{54.32}$$

Here, $P_2 = I_2 V_2 \cos \varphi_2$ is the power output from the secondary winding, $P_1 = I_1 V_1 \cos \varphi_1$ is the power input into the primary winding from the supply mains, $P_{copper} = I_1^2 r_1 + I_2^2 r_2$ represents the copper losses, that is, the power dissipated as heat in the windings of resistance $r_1$ and $r_2$, and finally, $P_{iron}$ represents the iron losses, that is, the power expended to magnetize the core cyclically. Then the efficiency of a transformer will be

$$\eta = P_2/P_1 = P_2/(P_2 + P_{copper} + P_{iron}) \qquad (54.33)$$

5. At loads close to nominal, the efficiency of a transformer is very high, being of the order of 90 to 95%, and the phase shift is zero very nearly. The terminal voltages differ but little from the emf, because the resistance of the transformer windings is relatively low. Under these conditions, equation (54.31), which holds for emfs, is sufficiently true of the terminal voltages, that is, the voltages across the transformer windings:

$$V_1/V_2 \approx w_1/w_2 = k \qquad (54.34)$$

## 54.10. TRANSMISSION OF ELECTRIC POWER OVER DISTANCES

1. When electric power is transmitted from a generator to a load, some losses are inevitable because of conductor heating. Let us see how these losses may be minimized.

Let there be a load operating on a voltage $V$ at a power factor $\cos \varphi$; the load power is $P$. The length of the power transmission line is $l$ and the cross-sectional area of the line conductors is $A$. Then the resistance of a two-wire transmission line will be $R = 2\rho l/A$. The power lost as heat in the conductors will be $\Delta P = I^2R$. Since $I = P/V \cos \varphi$, then

$$\Delta P = 2\rho l P^2/AV^2 \cos^2 \varphi \qquad (54.35)$$

2. It is seen that for a given load power $P$ and a given transmission line length $l$, the losses might be minimized by improving the power factor, or raising the operating voltage. As regards the conductor material, practically nothing can be done since the material is usually specified in advance (copper and/or aluminium). As regards the cross-sectional area of wires, an increase in it would offer no advantage because it would entail additional difficulties in construction work and an increase in the weight of the metal which means a higher cost of the line.

An improvement in the power factor is attractive. For example, raising the power factor from 0.63 to 0.88 will nearly halve the losses. Yet by far the most efficient method of reducing heat losses in a power transmission line is to raise the load voltage. This is why long-haul power transmission lines operate on very high voltages, from a few tens to several hundred kilovolts. For example, the line between the Lenin Hydro near Kuibyshev on the river Volga and Moscow operates on 400 kV, and that between the Volgograd Hydro and Moscow, on 500 kV.

## 54.11. THE REVOLVING MAGNETIC FIELD

1. Let the stator of an electric motor carry four coils connected in series pairwise so that they make up two windings, $AX$ and $BY$ (Fig. 54.9). A current traversing the winding $AX$ will give rise to a magnetic field with the magnetic induction vector directed as shown in the figure; any change in the direction of current flow will bring about a change in the direction of the magnetic induction vector. Then, if the winding $AX$ or $BY$ is traversed by an alternating current, the vector of magnetic induction inside the stator will oscillate harmonically at the frequency of the current.

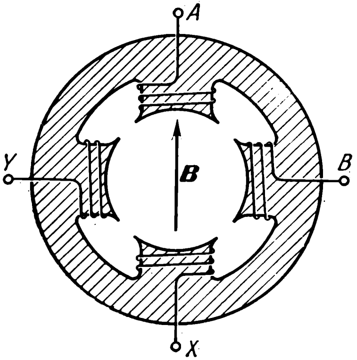


Fig. 54.9

2. If the two windings are connected in parallel, the vector of magnetic induction will oscillate harmonically along the bisector of the angle between the $AX$ and $BY$ directions. An entirely different picture will emerge, however, if the phase of the current in one of the windings is changed by, say, connecting a capacitor in series with that winding. The resultant phase shift will be $\varphi = \pi/2$. Now the magnetic induction in this winding will lag behind that of the other winding by a quarter of a cycle. As a result, the vector of magnetic induction will turn through 90° over a quarter of a cycle, and will complete a revolution over a cycle.

Thus, if the stator of a motor carries two windings traversed by an alternating current such that its phase in one lags behind the phase in the other by $\varphi = \pi/2$, a *revolving magnetic field* will be produced inside the stator.

## 54.12. SYNCHRONOUS AND INDUCTION MOTORS

1. As any other electrical machine, a synchronous generator is reversible. In addition to its operation as a generator, that is, an electrical machine converting mechanical energy into electrical, it can also operate as a motor, that is, a machine converting a.c. energy into mechanical energy. The principle on which a synchronous machine operates as a motor can readily be understood by recalling that a magnetic dipole placed in a magnetic field is acted upon by a torque (see Sec. 41.10).

The rotor of a synchronous machine may be likened to a giant magnetic dipole. Then, if its magnetic moment and the vector of magnetic induction make an angle $\theta$, the torque acting on the rotor will be $M = p_M B \sin \theta$, (41.17). The revolving magnetic field pulls along the rotor, causing it to rotate in synchronism with the magnetic field. For this to happen, however, the rotor should be brought up to the speed of the revolving field (the condition of synchronism).

2. A short-circuited copper or aluminium coil placed in a revolving magnetic field will be caused to rotate in the direction of the revolving field. This can be explained as follows. Let the angular velocity of the coil, $\omega_{coil}$, be somewhat lower than that of the revolving field, $\omega$ (asynchronous rotation). Then the coil will slip relative to the field; the amount of slip is $s = (\omega - \omega_{coil})/\omega$. Thus, the coil rotates relative to the magnetic field with an angular velocity proportional to the slip:

$$\omega_{rel} = \omega - \omega_{coil} = s\omega \qquad (54.36)$$

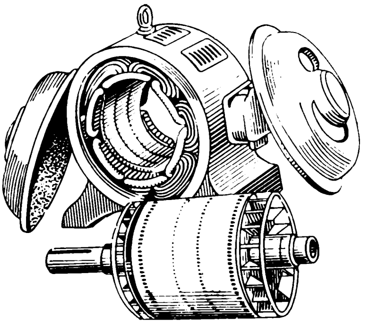


Fig. 54.10

Therefore it is traversed by an induction current proportional to the slip. According to Lenz's law (see Sec. 43.8), the induced current interacts with the field so that the coil is repelled by the field. Since the magnetic field is revolving, the coil is caused to rotate.

The torque acting on the coil is proportional to the induced current and, as a consequence, to the slip. This torque is balanced out by external load. Thus, in this type of machine the coil will always rotate at a slower speed than the field. This is asynchronous (that is, out-of-step) rotation, and motors utilizing this effect are called *asynchronous*, or more frequently, *induction motors*.

3. The stator of an induction motor does not differ from that of a synchronous generator or motor. The rotor core is assembled from iron sheets in which slots are made. The rotor winding consists of copper or cast-aluminium bars dropped in the rotor slots on its outer periphery and solidly connected to conducting end rings on each end, thus forming a squirrel-cage structure. A squirrel-cage induction motor in disassembled form is shown in Fig. 54.10.

Among the advantages of an induction motor are simplicity of design and enhanced reliability, because it has no slip-rings. An induction motor develops a sufficient starting torque and can readily be reversed. This is the reason why induction motors are most commonly used.

On the negative side, it should be noted that induction motors have a very low power factor even under nominal operating conditions ($\cos \varphi \approx 0.86$). The power factor is especially low in operation on underload and at no-load. To improve the power factor, it is important to choose a motor having the right power rating and to avoid sustained operation at no- or reduced load. Sometimes, the phase shift in power transmission lines may be made up for by banks of capacitors or synchronous motors with a high power factor.

# Chapter 55
# ELASTIC WAVES

## 55.1. TRANSVERSE AND LONGITUDINAL WAVES

1. If a vibratory disturbance occurs at any point in an elastic medium, it will be transmitted from one point to another, causing them to vibrate about their equilibrium positions. It has been found that this propagation of disturbances from point to point occurs not only in elastic media, but also in the electromagnetic field (see Chapter 59).

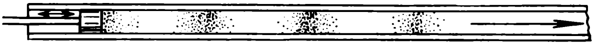


Fig. 55.1

This progressing disturbance in the state of the medium or field is called a *wave*. If it occurs in an elastic medium, it is called an *elastic wave*, and if in the electromagnetic field, an *electromagnetic wave*.

2. Imagine a long tube filled with a gas or liquid, in which a piston is moved to and fro repeatedly (Fig. 55.1). Owing to elastic forces, the harmonic motion of the piston will be transmitted to the gas, and an elastic wave will be propagated along the tube. This elastic wave is a train of periodically alternating regions of compression and rarefaction (or dilatation), that is, what is a region of compression at a given instant becomes that of rarefaction in a half-period, while the region next beyond it undergoes a reverse change, and so on.

In this case the particles of the elastic medium are displaced in the direction of progress of the disturbance, that is, in the direction of propagation of the wave. A wave in which the direction of displacement at each point of the medium is in the direction of wave propagation is called *longitudinal*.

The waves produced on the surface of a liquid are due to either surface tension or gravity, rather than owing to the elasticity of the medium. A distinction of these waves is that the particles are displaced in a vertical direction, while the wave is propagated in a horizontal plane. This type of wave is illustrated in Fig. 55.2. A wave in which the direction of displacement at each point of the medium is at right angles to the direction of wave propagation is called *transverse*.

In solids, both longitudinal and transverse waves may exist. A longitudinal wave may be caused by compressional or dilatational strains, much as it is in gases and liquids (a *compressional* or *dilatational wave*). A transverse wave is produced by shear strain (a *shear wave*). Since gases and liquids do not possess elasticity in shear, no transverse waves can occur in them.

3. The locus of points in a wave having the same phase is called a *wave front* or an *equiphase surface*. If the equiphase surfaces are planes at right angles to the direction of wave propagation, the waves are called *plane*.

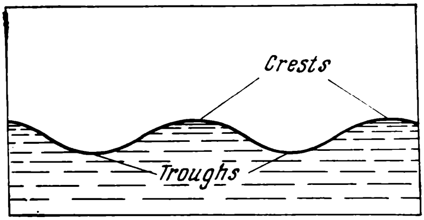


Fig. 55.2

A wave whose equiphase surfaces form a family of concentric spheres is called *spherical*. An example of spherical waves is the waves excited in air around a small isotropic source of sound, say a small bell.

4. The direction of propagation of a wave may be visualized as a line a tangent to which at each point is coincident with the direction of propagation of the wave or, which is the same, the direction of energy transfer, and is called a *ray*. In homogeneous media, rays are straight lines constantly perpendicular to the wave front.

## 55.2. THE VELOCITY OF ELASTIC WAVES

1. Elastic waves of high amplitude are called *shock-waves*, and those of small amplitude are called *sound waves*. Expressions for the velocity of sound waves, (30.13) and (30.18), were obtained in Sec. 30.5. Thus

$$a = \sqrt{\Delta p/\Delta \rho} = \sqrt{\gamma RT/M} \qquad (55.1)$$

For air

$$a = 20\sqrt{T} \qquad (55.2)$$

At $T = 273$ K, $a = 330$ m/s. at $T = 293$ K, $a = 343$ m/s, which agrees well with experiment.

2. The velocity of sound waves in solids and liquids depends on the compressibility (elasticity) and density of the materials. As a proof, we shall use the data of Sec. 31.1. Using equations (31.3) and (31.4), the following expression can be obtained for the velocity of an elastic longitudinal wave ($L$-wave):

$$a_L = \sqrt{\Delta p/\Delta\rho} = \sqrt{1/\beta\rho} = \sqrt{K/\rho} \tag{55.3}$$

From Table 31.1 we have: for water, $a = 1430$ m/s; for copper, $a = 3910$ m/s; for aluminium, $a = 4880$ m/s, etc.

3. To find the velocity of a transverse wave ($S$-wave) in solids, the bulk modulus, $K$, in equation (55.3) should be replaced with shear modulus, $G$:

$$a_S = \sqrt{G/\rho} \tag{55.4}$$

Since the shear modulus is one-half to one-fourth as great as the bulk modulus, the velocity of transverse waves is about half as great as that of longitudinal waves. This has been borne out by experiments. For example, the velocity of a longitudinal wave in granite is $a_L = 5400$ m/s and that of the transverse wave is $a_S = 3300$ m/s. For basalt, the respective figures are $a_L = 6300$ m/s and $a_S = 3700$ m/s.

This difference in velocity of propagation is utilized in the seismic methods of prospecting for minerals. In a blast-hole drilled underground, a charge of explosives is detonated, and the "shot" instant is sensed by a suitable device. The waves reflected from the various rocks are picked up by a set of geophones whose signals are then relayed to a seismic station. At the station, they are amplified and recorded on a paper chart along with time marks to produce permanent records known as seismograms. From these seismograms, one can readily determine the distribution of various rocks. Seismic methods are widely used in prospecting for oil, gas, ores and other mineral deposits.

4. If a longitudinal wave is propagated in a bar rather than in an infinite medium, the bulk modulus, $K$, in equation (55.3) should be replaced with Young's modulus, $E$:

$$a_L = \sqrt{E/\rho} \tag{55.5}$$

For steel, $E = 21.0 \times 10^{10}$ Pa, and $\rho = 7800$ kg/m$^3$, therefore $a_L = 5100$ m/s.

## 55.3. ENERGY AND INTENSITY OF THE WAVE

1. Imagine a volume $V$ in an elastic medium in which a wave is propagated with an amplitude $A$ and frequency $\omega$. According to equation (49.17), the average energy in the volume is $\overline{W} = \frac{1}{2} m\omega^2A^2$. Divi-

ing the average energy by volume gives an expression for the average wave energy density

$$\overline{w}=\overline{W}/V=\frac{1}{2}\rho\omega^2A^2 \tag{55.6}$$

where $\rho$ is the density of the medium.

2. The average rate of flow of energy in the direction of propagation per unit area of the wave front is defined as the *wave intensity*

$$I = \Delta W/S\Delta t = P/S \tag{55.7}$$

where $P$ is the power of the wave. Let $\Delta t \gg T$, where $T$ is the period of oscillation. In the time $\Delta t$, the surface will pass the energy contained within the volume $\Delta V = Su\Delta t$, where $u$ is the wave velocity, and the energy $\Delta\overline{W} = \overline{w}\Delta V = \overline{w}Su\Delta t$. Substituting it in (55.7) and cancelling out like terms gives

$$I=\overline{w}u=\frac{1}{2}\rho u\omega^2S^2 \tag{55.8}$$

3. The product of the density of a medium by the velocity of sound in that medium

$$z = \rho u \tag{55.9}$$

is called the *acoustic impedance*. It characterizes the wave properties of a medium (see Sec. 56.6).

### 55.4. ATTENUATION OF WAVES

1. Elastic waves are always absorbed by the medium in which they are propagated, the amount of absorption being dependent on several factors. Let us derive a law for the absorption of plane waves (parallel rays). For light, this law was discovered and explained by Bouguer in 1729.

Let a plane wave pass through a transmitting layer of thickness $x$. The wave intensity changes from $I_0$ to $I < I_0$. The ratio of the transmitted-wave intensity to the incident-wave intensity gives the transmittance of the layer, designated $D$:

$$D = I/I_0 \tag{55.10}$$

Assume that the transmittance of the layer depends solely on its thickness and is independent of the incident-wave intensity:

$$D = f(x) \tag{55.11}$$

Let a wave pass through two plates of thicknesses $x_1$ and $x_2$, tightly fitting to each other so that the reflection at the interface between them may be neglected (Fig. 55.3). On emerging from the first plate, the wave intensity will be $I_1 = I_0f(x_1)$. On emerging from the second plate, the wave intensity will be

$$I_2 = I_1 f(x_2) = I_0 f(x_1) f(x_2) \tag{55.12}$$

Since this system of two plates may be treated as a single plate of thickness $x = x_1 + x_2$, then

$$I_2 = I_0 f(x_1 + x_2) \tag{55.13}$$

From a comparison of the two equations we obtain a functional equation of the form

$$f(x_1) f(x_2) = f(x_1 + x_2) \tag{55.14}$$

It is an easy matter to show that this functional equation is satisfied by an exponential function of the form $f(x) = a^{\alpha x}$. We choose

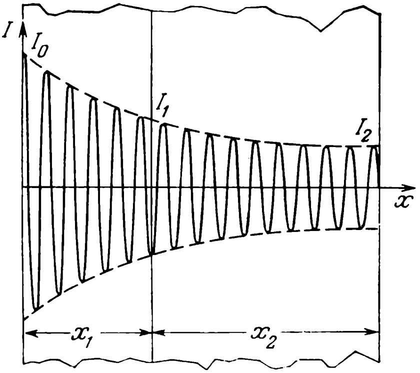


Fig. 55.3

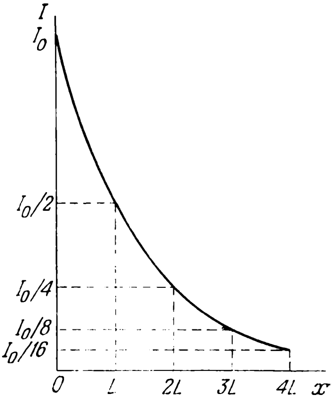


Fig. 55.4

the base to be the number $a = 2$. Noting that $f(x)$ is a decreasing function, we see that the factor $\alpha$ in the exponent should be a negative number, that is, $\alpha = -1/L$. Then the sought function will take the form

$$f(x) = 2^{-x/L} \tag{55.15}$$

2. The Bouguer law of absorption for plane waves can then be written as follows:

$$I = I_0 \cdot 2^{-x/L} \tag{55.16}$$

or differently

$$I = I_0 e^{-\mu x} \tag{55.17}$$

The quantity $L$ is called the *half-thickness*, that is, the thickness of the medium that reduces the intensity of the incident wave to one-half its initial value. Thus, setting $x = L$, the wave intensity transmitted will be $I = I_0 \cdot 2^{-1} = I_0/2$. The quantity $\mu = 0.69/L$ is called the *linear* or *mass absorption coefficient*. In graphic form, the Bouguer law is illustrated in Fig. 55.4.

The assumption that the transmittance of a medium is independent of the incident-wave intensity is pivotal to the derivation of the absorption law. If the transmittance does depend on the wave intensity, equation (55.16) will no longer apply. Incidentally, this is so with shock waves.

# Chapter 56

# WAVE EQUATION

## 56.1. WAVELENGTH

1. Using the set-up of Fig. 56.1, let us apply a sinusoidal voltage from an audio generator to a dynamic speaker and, at the same time, to one input of a double-beam oscilloscope. The waveform of the applied signal is displayed on the CRT screen. The applied voltage causes the diaphragm (or cone) in the speaker to vibrate and set up an elastic wave in the air. This plane wave reaches the microphone and causes its diaphragm to vibrate. In this way, the sound wave is

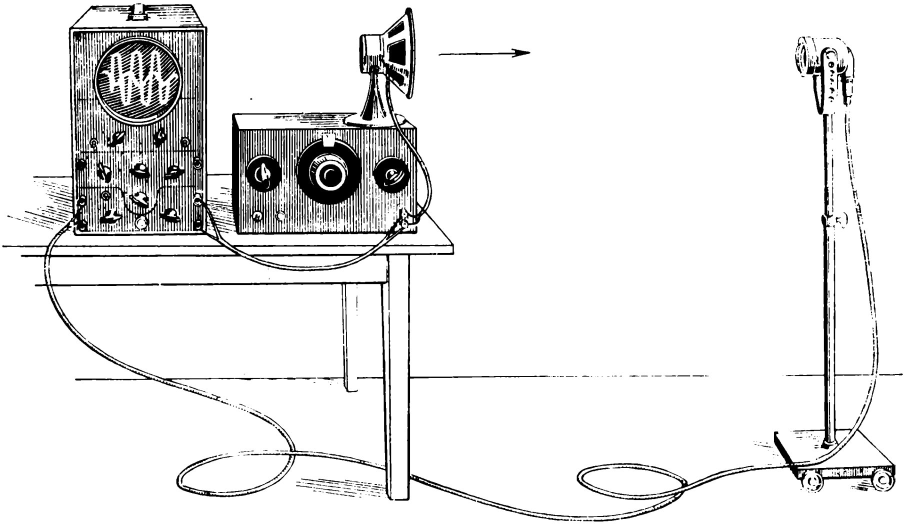


Fig. 56.1

converted into an electric wave which is applied to the second input of the oscilloscope. The screen displays a second waveform representing the wave front at the microphone. On moving the microphone to and away from the speaker, we can investigate vibrations in each wave front.

This experiment shows that the frequencies of both wave fronts are the same, and this is displayed on the oscilloscope, but the sound wave front reaching the microphone lags behind the speaker cone vibrations in phase. Assume that the ambient temperature is $t \approx \approx 17°\text{C}$, the velocity of sound is $a \approx 340$ m/s and that the generator is oscillating at a frequency of $\nu = 500$ Hz. Then, if the distance between the speaker and microphone is $x_1 = 17$ cm, the phase difference between the speaker cone vibrations and the sound wave at the microphone will be $\Delta\varphi_1 = \pi/2$; at $x_2 = 34$ cm, the phase difference will be $\Delta\varphi_2 = \pi$, that is, the phase reversal takes place. Finally, at $x_3 = 68$ cm, the phase difference will be $\Delta\varphi_3 = 2\pi$, that is, the sound wave front and the speaker cone vibrations are in phase. As the microphone is moved farther away from the speaker, the events will be repeated with a regularity. For example, at $x_4 = 136$ cm, the phase shift will be $\Delta\varphi_4 = 4\pi$, that is, the speaker cone vibrations and the sound wave front will again be in phase.

2. The findings of the experiment described above are explained by the fact that waves have a finite velocity of propagation. Then, if the microphone is a distance $x$ from the speaker the sound wave front originating at the speaker will reach the microphone with a delay

$$\Delta t = x/u \tag{56.1}$$

where $u$ is the velocity of a sinusoidal wave. This delay is maintained all the time as a lagging phase shift, $\Delta\varphi$. It can be found, noting that over a period $T$ the phase changes by $2\pi$:

$$\Delta\varphi/\Delta t = 2\pi/T$$

Substituting the expression for $\Delta t$ from equation (56.1) yields

$$\Delta\varphi = 2\pi x/uT = 2\pi x/\lambda \tag{56.2}$$

8. The quantity

$$\lambda = uT = u/\nu = 2\pi u/\omega \tag{56.3}$$

is called the *wavelength*. Setting $x = \lambda$ in equation (56.2), we get $\Delta\varphi = 2\pi$. Thus, if the distance between any two oscillating points (or any two wave fronts) is equal to the wavelength, these points will be in the same state of vibration.

Thus, ***the wavelength is the distance between two successive points in a wave that are in the same state of vibration*** (or ***with a phase difference*** *of* $\Delta\varphi = 2\pi$). Otherwise phrased, the *wavelength is the "space period" of a wave*, that is, the least translation distance that leaves the wave invariant, which is analogous to the time period $T$.

## 56.2. EQUATION OF A PLANE WAVE

1. Let the vibrations of the speaker cone be described by an equation of the form

$$s_0 = A \cos (\omega t + \varphi) \qquad (56.4)$$

Then, neglecting the wave attenuation, the vibrations of the wave front $x$ distant from the speaker will be described by an equation of the form

$$s = A \cos [\omega (t - \Delta t) + \varphi]$$

Substituting the expression for $\Delta t$ from equation (56.1), we get

$$s = A \cos [\omega (t - x/u) + \varphi] \qquad (56.5)$$

This is the equation of a plane sinusoidal wave propagated along the $x$-axis.

2. The reciprocal of the wavelength

$$k = \omega/u \qquad (56.6)$$

is called the *wave number**. From comparison with (56.3), we get

$$k = 2\pi/\lambda \qquad (56.7)$$

The wave number gives the number of wavelengths contained in the distance equal to $2\pi$ metres. It is similar to the radian frequency $\omega$ which gives the number of periods or cycles contained in the time interval equal to $2\pi$ seconds (see Sec. 49.2).

3. Opening the brackets in (56.5) and noting (56.6), we obtain an equation describing a plane sinusoidal wave in a more symmetrical form

$$s = A \cos (\omega t - kx + \varphi) \qquad (56.8)$$

From this equation we can derive:

(a) An expression describing the displacement of a point ($x = x_0 =$ constant) at different instants, the result is a simple harmonic motion

$$s = A \cos (\omega t + \alpha), \quad \text{where} \quad \alpha = \varphi - kx_0$$

* Some authors use $1/\lambda$ instead of $2\pi/\lambda$ in this sense and call $2\pi/\lambda$ the wave parameter.— *Tr.*

(b) An expression describing the displacement distribution of all points at a specified instant ($t = t_0 =$ constant):

$$s = A \cos (kx + \beta), \quad \text{where} \quad \beta = - (\omega t_0 + \varphi)$$

A plot of this function for $\beta = -\pi/2$ is shown in Fig. 56.2.

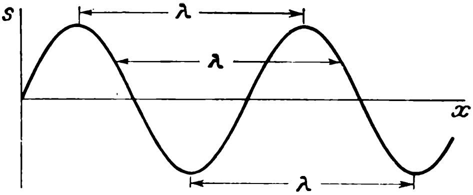


Fig. 56.2

### 56.3. EQUATION OF A SPHERICAL WAVE

1. Suppose a spherical wave set up by a small source (preferably spherical in shape) is propagated in air. We assume that the source radius is $r_0$, the distance from the centre of the source to the microphone is $r > r_0$, the amplitude of the wave near the surface of the source is $A_0$, and the amplitude of the wave at the microphone is $A$.

By moving the microphone from point to point, we shall obtain the same changes in phase as with a plane wave. In the plane-wave experiment, however, the amplitude remained practically unchanged, while the amplitude of the spherical wave decreases as the square of the distance from the sound source to the microphone, although absorption in air is negligible.

2. This result is in agreement with the law of conservation of energy. As a proof, we shall compute the total energy that a wave carries every second through a spherical surface of radius $r$. According to (55.9) and (55.10), we have

$$P = W/t = IS = \frac{1}{2} \rho u \omega^2 A^2 4\pi r^2 \qquad (56.9)$$

On the surface of the source

$$P = 2\pi \rho u \omega^2 A_0^2 r_0^2$$

Equating the two expressions for power, we obtain

$$A = A_0 r_0 / r \qquad (56.10)$$

The equation describing a spherical wave has the form

$$s = A \cos (\omega t - kr + \varphi) = (A_0 r_0 / r) \cos (\omega t - kr + \varphi) \qquad (56.11)$$

3. An expression for the intensity of a spherical wave can be obtained by substituting (56.10) in (55.8):

$$I = \frac{1}{2}\rho u\omega^2 (A_0^2 r_0^2)/r^2 = I_0 r_0^2/r^2 \tag{56.12}$$

Here, $I_0 = \frac{1}{2}\rho u\omega^2 A_0^2$ is the wave intensity on the surface of a spherical source. As is seen, the intensity of a spherical wave varies inversely as the square of the distance to the source.

## 56.4. DOPPLER EFFECT IN ACOUSTICS

1. So far we have taken the source frequency $\omega_0$, the wave frequency $\omega$ and the frequency $\omega'$ received by, say, a microphone to be the same, that is, $\omega = \omega' = \omega_0$. Therefore, we assigned them the same symbol $\omega$. This is valid, however, only if the source and the receiver (or the observer) are stationary relative to the medium in which the wave is propagated. If the source or the observer is moving relative to the medium, then $\omega \neq \omega_0$ or $\omega' \neq \omega_0$. This was discovered by Doppler in 1842.

2. Let the source of a wave motion be moving relative to the medium and the observer be stationary. Let also the velocity of the source relative to the medium be $v$; it should be less than the velocity of wave propagation ($v < u$), since otherwise the observation would be complicated by shock waves. When the source moves away from the observer, the wavelength decreases to $\lambda = (u - v)\, T$; when the source moves towards the observer the wavelength increases to $\lambda = (u + v)\, T$ (see Sec. 30.7, Fig. 30.5). The velocity of the wave is solely determined by the elastic properties of the medium (see Sec. 55.4), and the motion of the source has no effect on it.

The relationship between the observed wavelength, $\lambda$, with the source approaching the observer and the wavelength, $\lambda_0$, emitted by the stationary source may be written as

$$\lambda/\lambda_0 = (u - v)/u \tag{56.13}$$

For the observed wavelength with the source receding from the observer

$$\lambda/\lambda_0 = (u + v)/u \tag{56.14}$$

According to (56.3), $\lambda = 2\pi u/\omega$ and $\lambda_0 = 2\pi u/\omega_0$. Substituting in (56.13) and (56.14) gives the following expressions for the observed radian frequency and the stationary observer:

$$\left.\begin{array}{ll} \omega = \omega_0/(1 - v/u), & \text{with the source approaching} \\ \omega = \omega_0/(1 + v/u), & \text{with the source receding} \end{array}\right\} \tag{56.15}$$

3. If the source is stationary and the observer is moving at a relative velocity $v$, the apparent frequency will likewise change, but for another reason. Now the wavelength is not changed, because the source is stationary, $\lambda = \lambda_0$. However, the velocity of the wave motion relative to the moving observer, $w$, is equal to the algebraic sum of the wave velocity $u$ and the observer's velocity relative to the medium, $v$:

$$\left.\begin{aligned} w &= u + v, \quad \text{with the observer approaching} \\ w &= u - v, \quad \text{with the observer receding} \end{aligned}\right\} \tag{56.16}$$

According to (56.3), the observed radian frequency is

$$\omega' = 2\pi w/\lambda \tag{56.17}$$

From comparison with (56.16) and noting that $\omega_0 = 2\pi u/\lambda$ (because $\lambda = \lambda_0$), we get

$$\left.\begin{aligned} \omega' &= \omega_0 (1 + v/u), \quad \text{with the observer approaching} \\ \omega' &= \omega_0 (1 - v/u), \quad \text{with the observer receding} \end{aligned}\right\} \tag{56.18}$$

4. It is seen that the shift in frequency is caused by both the motion of the source and of the observer, but the mechanisms and the results are somewhat different. The difference is especially noticeable when the velocity of the source or the observer is close to that of the wave motion.

Let, for example, $v = 0.9u$. Then, with the source approaching the stationary observer, we get from (56.15) that $\omega = 10\omega_0$. If, however, the observer is approaching the stationary source, then from (56.18) it follows that $\omega' = 1.9\omega_0 \approx 2\omega_0$ which is one-fifth of the former result.

At first sight, it may appear that the result runs counter to the principle of relativity. One might argue that it makes no difference which of the two, the source or the observer, is moving. What matters, however, is not the motion of the observer relative to the source or the source relative to the observer, but their motion relative to the elastic medim in which the wave motion is set up and to which the reference system is related. Later, in connection with the Doppler effect for electromagnetic waves (Sec. 59.8), we shall see that the same result is obtained with either the source or the observer moving.

### 56.5. REFLECTION AND REFRACTION OF WAVES

1. Experience shows that when a wave travelling in one medium runs into a different medium, part of it in general passes on and undergoes refraction, while part is reflected. The former is known as the *refracted wave* and the latter as the *reflected wave*. Let us find the directions in which the reflected and refracted waves move when the

incident wave strikes the interface at an oblique angle (oblique incidence).

Fig. 56.3 shows two media in which the wave velocity is $u_1$ and $u_2$, respectively. It is customary to measure the directions of the incident, reflected and refracted waves with respect to the normal (perpendicular) to the interface. Thus, the angle of incidence $\alpha_1$, the angle of reflection $\alpha_r$ and the angle of refraction $\alpha_2$ are the angles made by the respective rays (see Sec. 55.1) with the normal.

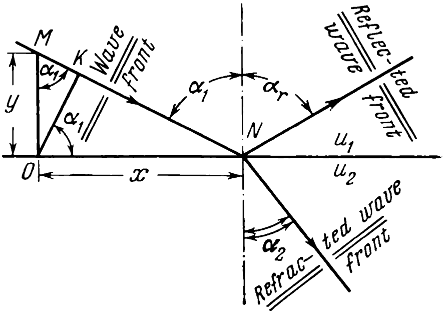


Fig. 56.3

2. The three waves are described by equations of the form

$$\left.\begin{aligned} s_1 &= A_1 \cos(\omega t - k_1 r_1) \\ s_2 &= A_2 \cos(\omega t - k_2 r_2) \\ s_r &= A_r \cos(\omega t - k_r r_r + \varphi) \\ &= \pm A_r \cos(\omega t - k_r r_r) \end{aligned}\right\} \tag{56.19}$$

The phase shift $\varphi$ in the expression for the reflected wave takes care of the fact that reflection may bring about a change in phase of the wave. In the general case, the phase shift may be either $\varphi = 0$ or $\varphi = \pi$. To cater for the two alternatives, two signs are given to the amplitude of the reflected wave (see Sec. 56.6). Referring to Fig. 56.3, for an arbitrary point $M$ on the incident ray the following relation holds:

$$MN = MK + KN$$

But $MN = r_1$, $MK = y \cos \alpha_1$, $KN = x \sin \alpha_1$, and, as a consequence, $r_1 = x \sin \alpha_1 + y \cos \alpha_1$. The same condition applies to the refracted and reflected rays. Thus

$$\left.\begin{aligned} r_1 &= x \sin \alpha_1 + y \cos \alpha_1 \\ r_2 &= x \sin \alpha_2 + y \cos \alpha_2 \\ r_r &= x \sin \alpha_r + y \cos \alpha_r \end{aligned}\right\} \tag{56.20}$$

Then the equations of the three waves will take the form

$$\left.\begin{aligned} s_1 &= A_1 \cos(\omega t - k_1 x \sin \alpha_1 - k_1 y \cos \alpha_1) \\ s_2 &= A_2 \cos(\omega t - k_2 x \sin \alpha_2 - k_2 y \cos \alpha_2) \\ s_r &= A_r \cos(\omega t - k_r x \sin \alpha_r - k_r y \cos \alpha_r) \end{aligned}\right\} \tag{56.21}$$

3. At any point on the interface where the incident wave is split up into a refracted wave and a reflected wave, the displacement of a

point on the refracted wave, $s_2$, should be equal to the algebraic sum of displacements caused by the incident wave, $s_1$, and by the reflected wave, $s_r$:

$$s_1 + s_r = s_2 \tag{56.22}$$

This is the *equation of continuity*.

Substituting (56.21) in (56.22) and noting that at the interface $y = 0$, we get

$$A_1 \cos(\omega t - k_1 x \sin \alpha_1) \pm A_r \cos(\omega t - k_r x \sin \alpha_r) = A_2 \cos(\omega t - k_2 x \sin \alpha_2) \tag{56.23}$$

This relationship should hold at any instant, $t$, and at any point on the interface, that is, at any $x$-coordinate. This is possible only if the arguments in the cosines are the same in all the three terms of the equality

$$\omega t - k_1 x \sin \alpha_1 = \omega t - k_r x \sin \alpha_r = \omega t - k_2 x \sin \alpha_2$$

or, after cancellation,

$$\left.\begin{aligned} k_1 \sin \alpha_1 &= k_r \sin \alpha_r \\ k_1 \sin \alpha_1 &= k_2 \sin \alpha_2 \end{aligned}\right\} \tag{56.24}$$

4. Since the incident and reflected rays are in the same medium, they have the same wave parameter: $k_r = k_1 = \omega/u_1$, (56.6). Then from the first line in equation (56.24) it follows that $\sin \alpha_r = \sin \alpha_1$. Since both angles are acute, the sines can be equal only if the angles are equal

$$\alpha_r = \alpha_1 \tag{56.25}$$

The foregoing may be summed up as follows: the reflected and the incident rays lie in the same plane with the normal to the interface; the angle of reflection always equals the angle of incidence. This is the *law of reflection*.

5. The second line in equation (56.24) may be re-written by expressing the wave parameters in terms of the frequency and velocity of the wave, according to (56.6). It should be noted that in going from one medium into the other, the frequency remains unchanged, because the frequency of forced vibrations is equal to that of the driving force (see Sec. 53.1). Substituting $k_1 = \omega/u_1$ and $k_2 = \omega/u_2$ in (56.24) and cancelling like terms gives

$$\sin \alpha_1 / \sin \alpha_2 = u_1/u_2 \tag{56.26}$$

This is the *law of refraction*. It states that the refracted and incident rays lie in the same plane with the normal to the interface, and that the sines of the angles of incidence and refraction are proportional to the velocities of the wave motion in the respective media.

## 56.6. REFLECTION AND TRANSMISSION COEFFICIENTS

1. Let a wave be normally incident on the interface. Then $\alpha_1 = 0$ and $\alpha_2 = \pi$, and equation (56.23) takes the form

$$A_1 \pm A_r = A_2 \qquad (56.27)$$

The "+" sign in the above expression represents reflection without a change of phase ($\varphi = 0$), while the "—" sign represents a phase reversal ($\varphi = \pi$). To make the problem more specific, we choose the "+" sign.

2. According to the law of conservation of energy, the intensity of the transmitted wave is equal to the difference between the intensities of the incident and reflected waves

$$I_1 - I_r = I_2 \qquad (56.28)$$

Substituting the intensities from (55.8) and cancelling like terms, we obtain

$$\begin{aligned} &A_1 + A_r = A_2 \\ &\rho_1 u_1 (A_1^2 - A_r^2) = \rho_2 u_2 A_2^2 \end{aligned} \qquad (56.29)$$

Solving the two equations simultaneously, we get

$$\left.\begin{aligned} &A_r = A_1 (\rho_1 u_1 - \rho_2 u_2)/(\rho_1 u_1 + \rho_2 u_2) \\ &A_2 = A_1 2\rho_1 u_1/(\rho_1 u_1 + \rho_2 u_2) \end{aligned}\right\} \qquad (56.30)$$

Thus, if $\rho_1 u_1 > \rho_2 u_2$, that is, if a wave is reflected from a medium presenting a lower acoustic impedance, the reflected wave will have a positive amplitude, which corresponds to the meaning of the concept: the phase upon reflection remains unchanged. At $\rho_1 u_1 < \rho_2 u_2$, the reflected wave has a negative amplitude, which implies that upon reflection the wave suffers a phase reversal.

3. The intensity of the reflected wave can be obtained by squaring (56.30):

$$I_r = I_1 R \qquad (56.31)$$

where $R$ is the *reflection coefficient*, defined as

$$R = (\rho_1 u_1 - \rho_2 u_2/\rho_1 u_1 + \rho_2 u_2)^2 \qquad (56.32)$$

If the difference in acoustic impedance between the two media is small, then $R \approx 0$, and practically all of the incident wave will pass from one medium into the other. Conversely, when the difference in acoustic impedance between the two media is considerable, the reflection coefficient tends to unity ($R \approx 1$), and practically all of the incident wave is reflected.

The ratio of the refracted (or transmitted) intensity to the incident intensity gives the transmission coefficient, defined as

$$T = I_2/I_1 = (I_1 - I_r)/I_1 = 1 - R$$

Noting (56.32), we get

$$T = 4\rho_1 u_1 \rho_2 u_2/(\rho_1 u_1 + \rho_2 u_2)^2 \qquad (56.33)$$

# Chapter 57

# INTERFERENCE AND DIFFRACTION

## 57.1. THE PRINCIPLE OF SUPERPOSITION

1. If you touch the water surface with two rods at a time, each will set up a circular wave passing through the other without mutual effect. This is also true of sound waves—if two musical instruments (or any other sound sources) are playing together, the wave from one will be propagated independently of the wave from the other, and each can be made out separately. The same applies to radio waves from two or several radio stations, light waves from several light sources, etc.

Thus, experience shows that waves do not interact, and are propagated independently of one another.

2. Since waves do not interact, each region of space where two or more waves arrive will undergo the vibrations set up by each wave separately. In order to find the resultant displacement at a given point in space, it is necessary to determine the displacement due to each wave, combine them either vectorially, if they occur in different directions, or algebraically, if they occur along the same straight line.

The above formulated rule for finding the resultant displacement is called the *principle of superposition.*

3. The principle of superposition is only applicable to waves of low intensity. Elastic waves of low intensity are sound waves; electromagnetic waves of low intensity include radio waves and light waves from conventional light sources.

The principle of superposition does not apply to shock waves. The point is that a shock wave suddenly changes the parameters of the elastic medium, that is, density, pressure, and temperature. This affects the propagation of other waves and their intensity, which implies the violation of the principle of superposition.

A similar result is observed in optics in the case of high-intensity light waves. That high-intensity light waves would violate the principle of superposition was postulated by Vavilov about 30 years ago.

Lacking strong light sources, he could not verify his hypothesis experimentally. At present, such light beams are readily generated by lasers (see Sec. 79.4). The branch of optics dealing with these effects is called *nonlinear optics*. Although investigations in this field started only recently, very interesting and important results have already been obtained.

## 57.2. STATIONARY WAVES

1. Let us tie one end of a rubber cord to a support and shake it up and down at the other end in regular succession (Fig. 57.1). By trial, find a rate of shaking the cord such that it will be vibrating as shown on the left. If, now, the frequency at which the cord is shaken is exactly doubled, get the pattern shown on the right. In both cases, all semblance of movement along the cord disappears. This is why the waves shown in Fig. 57.1 are called *stationary* or *standing*. In contrast to travelling or progressive waves in which all points are vibrating with the same amplitude but with a difference in phase (see Secs. 56.1 and 56.2), the points on a stationary wave are vibrating all at the same time, but with different amplitudes. The points where the cord remains motionless at all times are called *nodes*; regions between the nodes, where the movement is a maximum, are called *loops* or *antinodes*.

2. A stationary wave results from the superposition of two waves travelling in opposite directions and having identical amplitudes and frequencies. One of these waves is the wave set up by a source and propagated along the $x$-axis; it is described by an equation of the form $s_1 = A \cos(\omega t - kx)$. The other wave comes about because the first wave is reflected from an obstacle. Since it is propagated in the $-x$ direction, the sign of the coordinate in equation (56.8) should be changed. Besides, upon reflection the phase of the wave may be changed. Therefore, the equation describing the reflected wave has the form

$$s_2 = A \cos(\omega t + kx + \varphi)$$

Then, a stationary wave will be described by an equation of the form

$$s = s_1 + s_2 = A \cos(\omega t - kx) + A \cos(\omega t + kx + \varphi)$$

After elementary manipulation, we obtain

$$\begin{aligned} s &= 2A \cos(kx + \varphi/2) \cos(\omega t + \varphi/2) \\ &= B \cos(\omega t + \varphi/2) \end{aligned} \tag{57.1}$$

where

$$B = 2A \cos(kx + \varphi/2) \tag{57.2}$$

is the amplitude of a stationary wave.

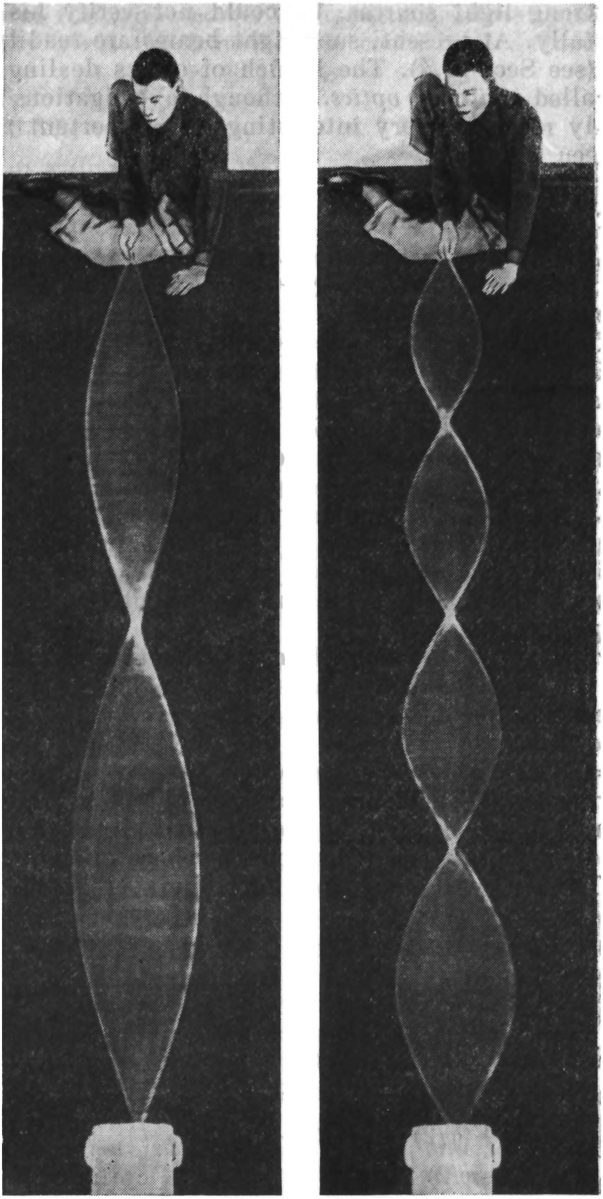


Fig. 57.1

3. As is seen, the amplitude of a stationary wave is a function of position. To make the case more specific, we shall consider a wave reflected from a medium having a higher wave impedance (or, as often but inaccurately said, from a denser medium). In such cases, the wave undergoes a phase reversal upon reflection: $\Delta\varphi = -\pi$ (see Sec. 55.6). This is a reflection with the loss of a half-wave, because at a distance $\Delta x = \lambda/2$, the change of phase is $\Delta\varphi = \pm\pi$. Substituting $\varphi = -\pi$ in (57.1) and (57.2) yields

$$s = B \sin \omega t \qquad (57.3)$$

where

$$B = 2A \sin kx \qquad (57.4)$$

Setting $B = 0$ in (57.4), we find the coordinates of the nodes. From $\sin kx = 0$ it follows that $kx = m\pi$, where $m$ may be any integer ($m = 0, 1, 2, 3, \ldots$). Since $k = 2\pi/\lambda$, (56.7), for the coordinates of nodes we have

$$x_{node} = m\lambda/2 = 2m\lambda/4 \qquad (57.5)$$

The coordinates of loops can be found from $B = \pm 2A$; the negative sign of the amplitude signifies that in passing through the node the stationary wave undergoes a phase reversal. Thus, for loops, $\sin kx = \pm 1$, and, as a consequence, $kx = (2m + 1)\,\pi/2$. Expressing the wave number in terms of wavelength, we get

$$x_{loop} = (2m+1)\,\lambda/4 \qquad (57.6)$$

Thus, the distance between two adjacent nodes or between two adjacent loops is equal to half the wavelength, and the distance between a loop and an adjacent node is a quarter-wavelength.

4. We leave it as an exercise for the reader to show that in reflection without the loss of a half-wave, the nodes and loops will change places in comparison with the case examined above.

## 57.3. NATURAL FREQUENCIES

1. At the beginning of the previous section it is noted that a standing wave will be set up on a cord of a specified length only when it is shaken at certain definite frequencies, called *natural.* Let us determine these frequencies.

Imagine a bar fixed at the ends (Fig. 57.2*a*). In fact, this may be, say, a string, or a column of air in a pipe stopped at both ends, or a beam on two supports. Let the length of the bar be $l$, and the wave velocity in it be $u$. When the bar is set into vibration, a standing wave will form, with nodes appearing at its ends, and one or several loops half-way between them. Since the distance between two nodes is a half-wavelength, the length of the bar will contain a whole number of half-waves:

$$l = m\lambda/2 \quad (m = 1, 2, 3, \ldots) \tag{57.7}$$

Expressing the wavelength in terms of frequency and wave velocity, (56.3), we obtain the following expressions for natural frequencies:

$$\omega = m\pi u/l, \quad \nu = \omega/2\pi = mu/2l \tag{57.8}$$

2. Equations (57.7) and (57.8) are of extremely important significance. They show that a system satisfying certain *end conditions* (for example, that the displacement at the ends of the bar is zero) can only vibrate at definite *discrete* frequencies. As will be shown in Chapter 70, this relationship is utilized in quantum mechanics.

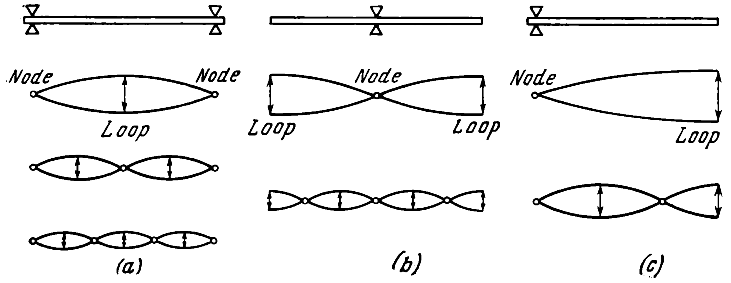


Fig. 57.2

3. We leave it as an exercise for the reader to calculate the natural frequencies of a bar of the same length, but fixed at the middle (Fig. 57.2*b*). Among other things, show that the fundamental frequency is the same as in the previous problem, but the higher harmonics are different, namely in the former case any even harmonics can exist, while in the latter, only odd ones.

Also compute the natural frequencies of the bar fixed at one end (Fig. 57.2*c*). Show that its fundamental frequency is half as great, and that only odd harmonics can exist.

## 57.4. INTERFERENCE

1. The intensity of a travelling wave is the same at all points, because all points are vibrating with the same amplitude. Of course, these points differ in phase, but, according to (55.8), this does not affect the wave intensity. This intensity, $I_0$, is given by

$$I_0 = \frac{1}{2}\rho u\omega^2 A^2 \tag{57.9}$$

In order to derive an expression for the intensity of a standing wave, we substitute the expression for its amplitude, (57.4), in equation (55.8):

$$I=\frac{1}{2}\rho u\omega^2B^2=4I_0\sin^2 kx \tag{57.10}$$

A plot of this relationship is shown in Fig. 57.3.

2. As is seen, the wave intensity at the nodes of the standing wave (points with coordinates $x_{node} = 2m\lambda/4$) is zero over the entire time of observation; at loops ($x_{loop} = (2m + 1)\lambda/4$), the wave intensity is $I_{loop}= 4I_0$. Since a standing wave results from the superposition of the incident and reflected waves having identical intensities, $I_1 = I_2 = I_0$, their total intensity should be $2I_0$.

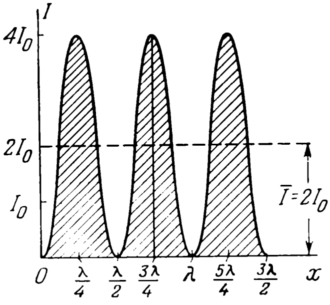


Fig. 57.3

Thus, with the onset of a standing wave, energy in space appears to be transferred from nodes into loops. The energy of vibration at each point on a standing wave is no longer equal to the sum of the energies of both waves. At nodes, the energy is zero (minima), and at loops it is twice the total energy (maxima). It is only the average energy of a standing wave that is equal to the sum of the energies of the superimposed waves, which is in agreement with the law of conservation of energy. Referring to Fig. 57.3, the average wave intensity is given by

$$\overline{I} = 2I_0 = I_1 + I_2 \tag{57.11}$$

3. Consider a region of space where two or several waves interact. If the superposition produces no interference, the intensity of the resultant disturbance at any point is equal to the sum of the component wave intensities. If, in contrast, the wave energy is caused to be redistributed so that intensity maxima appear at some points and intensity minima at others, the superposition is said to have produced interference. Thus, *interference may be defined as the variation of wave energy caused by the superposition of two or more waves meeting certain requirements.*

The time-invariant pattern of intensity maxima and minima caused by interference is called an *interference pattern.*

Thus, the formation of a standing wave is an example of interference; a stable system of nodes and loops is a typical example of an interference pattern.

## 57.5. INTERFERENCE OF WAVES FROM TWO SOURCES

1. Let two balls attached to a single rod be vibrating on the water surface in a large flat pan. Each ball sets up a wave; on meeting, the two waves interfere, and a typical interference pattern appears on the water surface (Fig. 57.4).

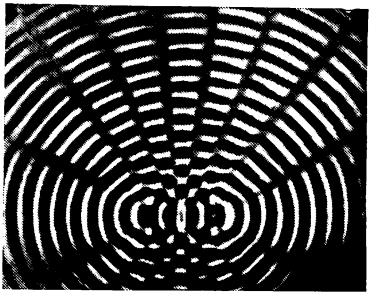


Fig. 57.4

The mechanism by which interference is caused may be elucidated by reference to Fig. 57.5, where $S_1$ and $S_2$ are two sources of spherical waves separated by a distance $S_1S_2 = d$. The point $K$ is half-way between the sources; the distance of the point $K$ to the screen is $KO = l$. Let us compute the intensity of vibration at an arbitrary point, $M$, on the screen; the distance to that point is $MO = y$, $MK = r$, and the observation angle is $\angle MKO = \theta$.

2. If the screen is to display a stable interference pattern it is important that the two waves be of the same frequency. It is only then that the amplitude of the displacement at any point on the screen will be independent of time (see Sec. 49.5). If the waves differ in frequency, beats will be produced instead of a stable interference pattern (see Sec. 50.1). Thus, a necessary condition for an interference pattern to be produced is that the waves must be of the same frequency. *Interference only results from the superposition of waves of*

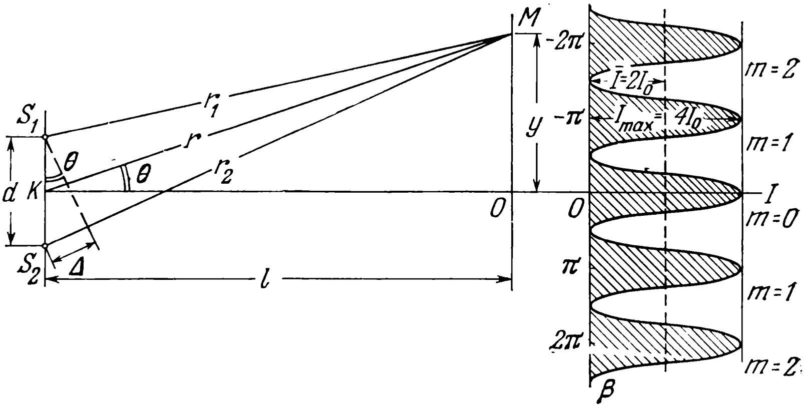


Fig. 57.5

*the same frequency*. If, in addition, the two waves are propagated in the same medium, the waves will then be of the same wavelength and have the same wave parameter.

3. Let us write equations for the two waves

$$\left.\begin{aligned} s_1 &= A_1 \cos(\omega t - kr_1 + \varphi_1) \\ s_2 &= A_2 \cos(\omega t - kr_2 + \varphi_2) \end{aligned}\right\} \tag{57.12}$$

The path length difference (to be defined later), $\Delta = r_2 - r_1$, is only a fraction of each radius-vector. Therefore, $r_2 \approx r_1$, and, according to (56.10), the amplitudes are likewise equal

$$A_1 = A_2 = A$$

The resultant displacement at the point $M$ is

$$s = s_1 + s_2 = 2A \cos\left[\frac{k(r_2 - r_1)}{2} + \frac{\varphi_1 - \varphi_2}{2}\right] \times \cos\left[\omega t - \frac{k(r_1 + r_2)}{2} + \frac{\varphi_1 + \varphi_2}{2}\right] \tag{57.13}$$

This expression may be rewritten as

$$s = B \cos(\omega t + \delta) \tag{57.14}$$

where

$$B = 2A \cos\left[\frac{k(r_2 - r_1)}{2} + \frac{\varphi_1 - \varphi_2}{2}\right] \tag{57.15}$$

is the new amplitude, and

$$\delta = -\frac{k(r_1 + r_2)}{2} + \frac{\varphi_1 + \varphi_2}{2}$$

is the new reference (or initial) phase, or epoch.

Since we are interested in the distribution of intensities in the interference pattern, while the intensity is independent of phase, we shall concentrate on the amplitude of the resultant wave.

4. If the sources emit precisely sinusoidal waves, their epoch angles $\varphi_1$ and $\varphi_2$ are constant quantities, and the phase difference will likewise be a constant quantity. Without any loss of generality, we may set $\varphi_1 - \varphi_2 = 0$ (or $m\pi$). In fact, assigning to this difference, $\varphi_1 - \varphi_2 =$ constant, any other value will only shift the interference pattern on the screen, without affecting the pattern of the intensity distribution. Then, the resultant amplitude will be

$$B = 2A \cos\frac{k(r_2 - r_1)}{2} \tag{57.16}$$

Noting that $k = 2\pi/\lambda$, we have

$$B = 2A \cos\frac{\pi(r_2 - r_1)}{\lambda} \tag{57.17}$$

According to (55.8), the wave intensity is

$$I_1 = I_2 = I_0 = 1/2\rho u\omega^2 A^2$$

and the intensity of the resultant wave is

$$I = \frac{1}{2}\rho u\omega^2 B^2 = 4I_0 \cos^2 \frac{\pi(r_2 - r_1)}{\lambda} \tag{57.18}$$

5. The difference in distance from the point on the screen to the sources

$$\Delta = r_2 - r_1 = d \sin\theta \tag{57.19}$$

is called the *path length difference.*

If the path length difference can hold an even number of half-waves, that is

$$\Delta = r_2 - r_1 = 2m\lambda/2$$

then

$$\cos^2 \frac{\pi(r_2 - r_1)}{\lambda} = \cos^2 m\pi = 1$$

Thus, an interference maximum occurs at point $M$, with an intensity $I_{\max} = 4I_0$. If the path length difference contains an odd number of half-waves, that is

$$\Delta = (2m + 1)\lambda/2$$

then

$$\cos^2 \frac{\pi(r_2 - r_1)}{\lambda} = \cos^2 (2m + 1)\pi/2 = 0$$

that is, an interference minimum occurs at the same point. To sum up, the condition for the occurrence of interference maxima and minima may be written as follows

$$\left.\begin{aligned} &\Delta = 2m\lambda/2 \text{ for a maximum, } I_{\max} = 4I_0 \\ &\Delta = (2m+1)\lambda/2 \text{ for a minimum, } I_{\min} = 0 \end{aligned}\right\} \tag{57.20}$$

As in a stationary wave, the energy is redistributed to form minima and maxima. The average intensity on the screen, as is seen in Fig. 57.5, is equal to the sum of intensities of both waves, that is

$$\bar{I} = 2I_0 = I_1 + I_2$$

which agrees with the law of conservation of energy.

### 57.6. INTERFERENCE OF WAVES FROM SEVERAL SOURCES

1. Imagine a system of $N$ identical wave sources arranged along a common straight line at a distance $d$ from one another. We seek to find the wave intensity at a point removed from the sources for a

distance such that the rays connecting this point to each source may be taken to be practically parallel. With this assumption, the problem reduces to combining $N$ harmonic waves of the same amplitude, whose phases form an arithmetic progression (see Sec. 49.6). What

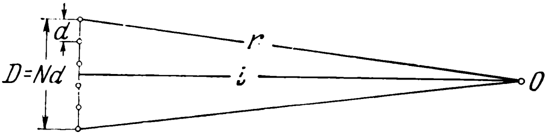


Fig. 57.6

remains to be found is the phase difference for the waves emitted by adjacent sources. According to (57.19), the path length difference is

$$\Delta = d \sin \theta$$

Then, the phase difference is

$$\alpha = k\Delta = kd \sin \theta \tag{57.21}$$

Substituting it in (49.23) gives an expression for the total amplitude

$$A = a \frac{\sin\left(\frac{1}{2} Nkd \sin \theta\right)}{\sin\left(\frac{1}{2} kd \sin \theta\right)} \tag{57.22}$$

where $a$ is the amplitude of the wave from one source and $i_0 = ka^2$ is its intensity.

2. On introducing an auxiliary angle such that

$$\beta = \frac{1}{2} kd \sin \theta = \pi d \sin \theta / \lambda \tag{57.23}$$

expression (57.22) for the amplitude may be given the form

$$A = a \sin N\beta / \sin \beta \tag{57.24}$$

Since the intensity is proportional to the square of the amplitude, the intensity of the resultant wave will be

$$I = i_0 \sin^2 N\beta / \sin^2 \beta \tag{57.25}$$

3. So far we have assumed that the observer is at a large distance from the sources. It is important to define how large this distance may be. Let us refer to Fig. 57.6.

We have assumed that the rays emerging from the sources are practically parallel. Actually, the rays are not parallel, but converge at a point, $O$. However, if the maximum error in the path length difference is only a fraction of a half-wavelength, the error in phase will be a small fraction of $\pi$, and the phases will practically be identical. Thus, our calculation will remain in force, if $r - l \ll \lambda/2$. But $r = \sqrt{l^2 + D^2/4}$, where $D = Nd$ is the length of the array of sources. Substituting it gives

$$\sqrt{l^2 + D^2/4} - l \ll \lambda/2 \tag{57.26}$$

When multiplied through by $l + \sqrt{l^2 + D^2/4} \approx 2l$, expression (57.26) takes the form

$$l^2 + D^2/4 - l^2 \ll 2l\lambda/2$$

Or finally

$$D^2 \ll 4\lambda l \tag{57.27}$$

Thus, the observer is sufficiently far from the sources, if

$$l \gg D^2/4\lambda \tag{57.28}$$

This is the condition under which relationship (57.25) and all of its corollaries are valid.

## 57.7. INTENSITY OF PRINCIPAL MAXIMA

1. The locations of maximum wave intensities can be defined from the condition

$$\beta = m\pi \ (m = 0,\ 1,\ 2,\ \ldots) \tag{57.29}$$

This can be proved as follows. If $\varepsilon$ is a very small number, then $\sin \varepsilon \approx \varepsilon$. At $\beta = m\pi + \varepsilon$, we have

$$\sin\beta = \sin(m\pi + \varepsilon) = \pm \sin\varepsilon \approx \pm\varepsilon,$$

$$\sin N\beta = \sin(Nm\pi + N\varepsilon) = \pm \sin N\varepsilon \approx \pm N\varepsilon,$$

$$\lim_{\varepsilon \to 0} \frac{\sin^2 N\beta}{\sin^2\beta} = \frac{N^2\varepsilon^2}{\varepsilon^2} = N^2$$

Substituting it in (57.25) we find that at principal maxima the intensity is $N^2$ times the intensity of a single wave, or

$$I_{\max} = N^2 i_0 \tag{57.30}$$

If there were no interference, the intensity at any location would be equal to the sum of intensities, that is, $I = Ni_0$. Because of interference, the energy is redistributed (Fig. 57.7) so that at some locations the energy exceeds the sum of the energies from all the sources, and no energy is propagated to other locations. Passing from the auxiliary angle in (57.29) to the solid angle of observation $\theta$ with the aid of (57.23), we get

$$(\pi d \sin\theta)/\lambda = m\pi$$

or

$$\sin\theta_{\max} = m\lambda/d \qquad (57.31)$$

This is the condition for the occurrence of principal maxima for an angle of observation θ.

2. Between two adjacent maxima (Fig. 57.7) there are several minima and secondary maxima whose intensity is only a fraction of that of the principal maxima. A minimum occurs when the numerator in expression (57.25) reduces to zero, while the denominator is non-zero. It is an easy matter to show that it is possible if

$$N\beta = n\pi \qquad (57.32)$$

where $n$ is not a multiple of $N$ (that is, $n \neq mN$), because at $n = mN$ we would arrive at condition (57.29) again and obtain a principal maximum. It is noteworthy that in the interval between the zero-order and first-order principal maxima the number $n$ takes on values from 1 to $N - 1$, so that there are $N - 1$ minima and $N - 2$ secondary maxima.

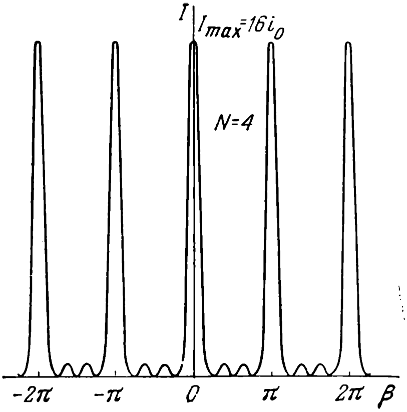


Fig. 57.7

Since the sine of an angle cannot exceed unity, the number of principal maxima can be found from equation (57.31). From $m\lambda/d \leqslant 1$, it follows, that

$$m \leqslant d/\lambda \qquad (57.33)$$

### 57.8. DIFFRACTION

1. Using a long vibrating plate let us set up plane waves on the water surface and see these waves pass through a hole in an obstacle. The experiment will show that if the diameter of the hole, $D$, is less than the wavelength ($D < \lambda$), the waves will bend somewhat and spread out sideways as spherical waves (Fig. 57.8) as if the hole were a point source. With an increase in the hole diameter, a typical interference pattern will be seen beyond it, showing the central principal maximum and the weaker secondary maxima (Fig. 57.9). It is only when the hole diameter is much greater than the wavelength ($D \gg \lambda$) that the bending of the wave round the edge of the hole in the barrier at small distances from it will be hardly discernible.

This bending of waves when they pass near the edge of an obstacle or through small openings is called *diffraction*. In a more general sense, diffraction involves the scattering of waves by well-defined discontinuities in the medium.

2. A rigorous solution of the diffraction problem runs into appreciable mathematical difficulties. Usually, the interference pattern produced by diffraction is analyzed, using the *Huygens-Fresnel principle*. It consists in that the intensity of radiation meeting any point from a spherical wave front is determined by dividing the wave front into half-period elements or zones (*Fresnel zones*) which are then treated as independent and identical sources of waves; the amplitude (and intensity) of the wave at the point of observation is then found as the outcome of the interference of the waves assumed to be produced by the individual zones.

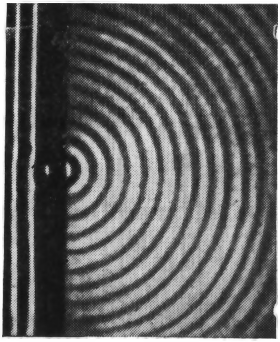


Fig. 57.8

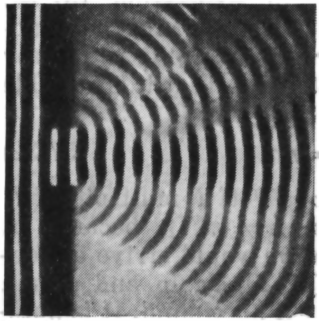


Fig. 57.9

The interference pattern analyzed by this method agrees well with experimental data, provided the aperture diameter is many times the wavelength.

In greater detail, the use of this method will be taken up in connection with the diffraction of light (Ch. 62). For the time being, we shall limit ourselves to a single example, the diffraction of a plane wave on a single slit.

### 57.9. DIFFRACTION THROUGH A RECTANGULAR SLIT

1. Let a plane wave be incident on a long, narrow, rectangular slit cut in an obstacle. The width of the slit is $D$, and its length is $L \gg \gg D$. Divide this slit into narrow zones with their long sides parallel with the slit so that the zone width is $d = D/N$, where $N$ is the number of zones.

Let the observer be at a large distance from the slit, such that $l \gg D^2/4\lambda$, as specified by equation (57.28). Then analysis of the interference pattern will reduce to the interference of waves from $N$ identical sources, examined in Sec. 57.6.

The wave amplitude $A$ at the point of observation can be computed, using equation (57.22), assuming that the amplitude of the wave from an individual zone is $a = A_0/N$, where $A_0$ is the wave amplitude at the aperture.

Substituting it in (57.22) gives

$$A = \frac{A_0}{N} \frac{\sin\left(\frac{1}{2} kD \sin\theta\right)}{\sin\left(\frac{kD\sin\theta}{2N}\right)} \tag{57.34}$$

Using an auxiliary angle such that

$$\alpha = 1/2\ kD \sin\theta = \pi D \sin\theta/\lambda, \tag{57.35}$$

the expression for the wave amplitude at the point of observation may be rewritten as follows

$$A = (A_0/N) \frac{\sin\alpha}{\sin(\alpha/N)} \tag{57.36}$$

2. The accuracy of the solution improves with the number of zones. However, at high values of $N$, the sine of a small angle does not practically differ from the angle expressed in radians, that is, $\sin(\alpha/N) \approx \alpha/N$. Substituting it in (57.36) gives an expression for the amplitude

$$A = A_0 \sin\alpha/\alpha \tag{57.37}$$

and for the wave intensity at the point of observation

$$I = I_0 \sin^2\alpha/\alpha^2 \tag{57.38}$$

3. The bright central disc in the interference pattern is the zero-order principal maximum. For it, $\theta \approx 0$, and the auxiliary angle $\alpha \approx 0$, too. As a consequence, $\sin\alpha \approx \alpha$ and $I = I_0$. With

$$\alpha = m\pi \quad (m = 1, 2, 3, \ldots) \tag{57.39}$$

the numerator in (57.38) reduces to zero, while the denominator is non-zero. As a consequence, the condition stated by (57.39) or the condition equivalent to it

$$\sin\theta = m\lambda/D \quad (m = 1, 2, 3, \ldots) \tag{57.40}$$

defines the locations of minima, that is, the locations where the wave intensity is zero.

It should be noted that expression (57.40) is valid only at $D > \lambda$, because the sine cannot exceed unity. Then, at $D \leqslant \lambda$, our calculation does not apply. Experience (see Fig. 57.8) confirms that no minima occur in such cases, and the aperture radiates waves in all directions.

4. It can be shown that secondary maxima are located about halfway between two minima, that is, at $\alpha \approx (2m + 1)\pi/2$. Their intensity falls off rapidly with increasing order. For example, the intensity of the first-order maximum ($m = 1$) can be found by substituting $\alpha = 3\pi/2$ in (57.38). Then $I_1 = 4I_0/9\pi^2 \approx 0.045 I_0$, that is, its intensity is a mere 4.5% of that at the principal maximum. Accordingly, for the second-order maximum we obtain $I_2 = 4I_0/25\pi^2 \approx 0.016\ I_0$; for the third-order maximum, $I_3 = = 4I_0/49\pi^2 \approx 0.008 I_0$, etc.

In graphic form, the intensity distribution for diffraction through a narrow rectangular slit is shown in Fig. 57.10. The scale has been distorted because two quantities differing by a factor of more than 100 ($I_3 \approx I_0/120$) cannot be conveniently shown on the same drawing.

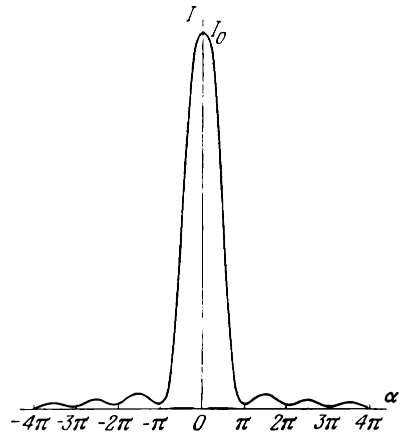


Fig. 57.10

## 57.10. WAVE REFRACTION AND INTERFERENCE

1. In Sec. 56.5, we have examined the behaviour of waves at the interface between two media, using the boundary-condition method widely employed in the theory of wave propagation. On that basis, we have derived the laws of reflection and refraction, equations (56.25) and (56.26). It can be shown that the same results can be obtained using the interference method.

Let a plane wave be incident on the interface between two media. The wave front in the first medium will be designated $XY$ and in the second medium, $PQ$ (Fig. 57.11). The angle of incidence is the angle $\alpha_1$ between the ray and the normal to the interface or the angle between the wave front and the interface; these angles are equal, because their sides are mutually perpendicular.

At the interface, $XQ$, the particles are set into forced vibrations, thereby acting as secondary sources which emit waves in all directions into both the first and the second medium. However, this does not imply that the intensity of the scattered waves will be the same in all directions. In fact, the contrary is true. The scattered waves interfere and cause the energy to be redistributed so that the intensity of the reflected and refracted waves will be a maximum only in certain directions. It can be shown that the principal interference maxima occur precisely in the directions determined by the laws of reflection and refraction.

2. To begin with, we shall find the amplitude (and intensity) of the wave in the second medium in an arbitrary direction determined by the angle $\alpha_2$. For this purpose, we shall use the Huygens-Fresnel principle (see the previous section).

Divide the segment $XQ = D$ into $N$ zones of width $d = D/N$ each, write the equation for the wave emitted in this direction by each zone, and add together these waves. The point of observation is taken to be $l$ distant from the point $X$, the distance being large enough to satisfy condition (57.28).

Let the amplitude of the wave passing into the second medium be $A_0$; then the amplitude of the wave passing through one zone will be $A_0/N$. The wave equation for the first zone will have the form

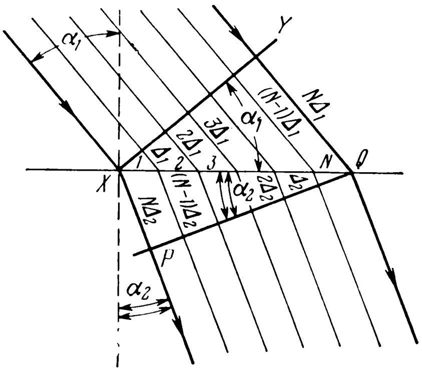


Fig. 57.11

$$s_1 = (A_0/N) \cos (\omega t - k_2 l)$$

The wave emitted by the second zone has to travel an additional distance, $\Delta_1 = d \sin \alpha_1$, in the first medium. At the same time, however, the distance from this zone in the second medium to the point of observation decreases by the amount $\Delta_2 = d \sin \alpha_2$. Then, the equation of this wave takes the form

$$s_2 = (A_0/N) \cos [\omega t - k_2 (l - \Delta_2) - k_1 \Delta_1]$$
$$= (A_0/N) \cos [\omega t - k_2 l + (k_2 \Delta_2 - k_1 \Delta_1)]$$

Accordingly, the equations for the succeeding zones will be

$$s_3 = A_0 \cos [\omega t - k_2 l + 2 (k_2 \Delta_2 - k_1 \Delta_1)]/N$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$s_N = A_0 \cos [\omega t - k_2 l + (N - 1) (k_2 \Delta_2 - k_1 \Delta_1)]/N$$

3. Thus, the problem is one of combining waves of the same frequency, with their phases forming an arithmetic progression. It was solved in Sec. 49.6. According to (49.23), the amplitude of the resul-

tant wave will be described by an equation of the form

$$A=(A_0/N)\frac{\sin[N(k_2\Delta_2-k_1\Delta_1)/2]}{\sin[(k_2\Delta_2-k_1\Delta_1)/2]} \tag{57.41}$$

Substituting the expressions for $\Delta_1$ and $\Delta_2$ gives

$$A=(A_0/N)\frac{\sin\Phi}{\sin(\Phi/N)} \tag{57.42}$$

where

$$\Phi=\frac{1}{2}D(k_2\sin\alpha_2-k_1\sin\alpha_1) \tag{57.43}$$

However, the segment $XQ = D$ may be divided into an arbitrary great number of zones; in other words, as in the previous section we may let $N \to \infty$. Recalling that for a small angle its sine is approximately equal to the angle in radians, $\sin(\Phi/N) \approx \Phi/N$, we obtain expressions (57.37) and (57.38) for the amplitude and intensity of the refracted wave. The intensity distribution is shown in Fig. 57.10.

4. As with diffraction through one slit, the principal maximum of a refracted wave will occur, if the condition is satisfied such that the auxiliary angle $\Phi = 0$; then $A = A_0$ and $I = I_0$. Thus, the principal maximum of a refracted wave will occur in the direction defined by the condition

$$k_2\sin\alpha_2-k_1\sin\alpha_1=0 \tag{57.44}$$

Substituting the wave numbers defined by (56.6), $k_1 = \omega/u_1$ and $k_2 = \omega/u_2$, we obtain the law of refraction as defined by equation (56.26).

5. We leave it as an exercise for the reader to go through the same reasoning for the refractive wave and derive the respective relationship given by equation (56.25). Formally, it stems from (57.44). The point is that since the refractive wave is propagated in the same medium, $k_1 = k_2$, and therefore $\sin\alpha_2 = \sin\alpha_1$ and $\alpha_2 = \alpha_1$.

# Chapter 58

# FUNDAMENTALS OF ACOUSTICS

## 58.1. CHARACTERISTICS OF SOUND

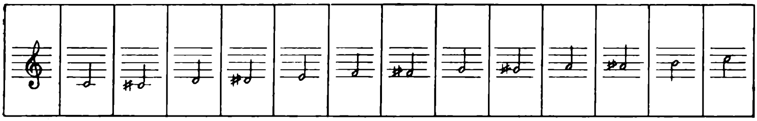
1. We have already used the concept of sound, defining it as an elastic wave of low intensity. In the narrow sense, however, sound implies audible sound, that is, elastic waves that the human ear is capable of picking up. Experience shows that the human ear hears as sound the vibrations in the audible range, that is, with frequen-

cies ranging from 20 Hz to 20 kHz. Elastic waves at frequencies below the audible range are called *infrasonic* (or *subsonic*) and those above the audible range are called *ultrasonic*. Sometimes elastic waves with frequencies of $10^{10}$ Hz and higher, corresponding to Debye thermal waves in liquids or solids (see Sec. 45.3), are referred to as *hypersonic*.

According to their spectrum (see Sec. 50.4), we differentiate between *noise* and *musical tones*. Noise is a jumble of irregularly timed, unrelated, non-periodic vibrations. Noise has a continuous spectrum, that is, a set of frequencies continuously filling a certain interval. In contrast, musical tones have a line spectrum containing multiple frequencies; therefore, they are periodic vibrations.

2. A *pure musical tone* is a sinusoidal sound wave. What is called the *pitch* of a pure musical tone, or its position in the musical scale, is decided mainly by the frequency of the sound wave that strikes the ear. The musical scale currently used is constructed as follows. Each octave, that is, the interval between two sounds having a basic frequency ratio of two, is divided into twelve intervals; on the piano an octave is represented by a group of twelve keys, seven white ones and five raised black ones; the intervals represented by the black keys are marked by the "sharp" sign. Within an octave, the frequency rises in the 2-to-1 proportion, and within an interval, $\sqrt[12]{2} \approx$ $\approx 1.06$ times.

Table 58.1



| Tone | Frequency, Hz | Tone | Frequency, Hz | Tone | Frequency, Hz |
|---|---|---|---|---|---|
| C | 261.63 | F | 349.23 | A sharp | 466.16 |
| C sharp | 277.18 | F sharp | 369.99 | B | 493.88 |
| D | 293.67 | G | 392.00 | C, second octave | 523.25 |
| D sharp | 311.13 | G sharp | 415.31 | | |
| E | 329.63 | A | 440 (precisely) | | |

The tones and frequencies within the compass of the first octave are given in Table 58.1. A *complex musical tone* consists of the lowest-pitch tone, called the *fundamental* (see Sec. 50.4) and a set of *overtones* (or *harmonics*). If two musical tones have the same fundamental pitch, but differ in overtones (that is, in spectrum), they are said to differ in *quality*. Quality is what enables the ear to distinguish the same note played on various instruments or sung by various singers.

3. In addition to pitch and quality, musical tones differ in *loudness*. In the general case, the loudness of a musical tone is directly related to the intensity of the sound wave. Since, however, the human ear responds differently to sounds at different frequencies, this dependence is not so straightforward as it might seem. The human ear is most responsive to tones at frequencies from 700 to 6000 Hz. Within this range, the ear will readily pick up musical tones with an intensity of about $10^{-11}$ to $10^{-12}$ W/m$^2$.

The lowest intensity of a sound wave that the human ear is capable of sensing is called the *threshold of audibility*. The standard threshold of audibility is $I_0 = 10^{-12}$ W/m$^2$ at a frequency $\nu_0 = 1$ kHz.

The minimum effective sound pressure that causes a feeling of pain in the ear forms the upper limit of audibility, called the *threshold of feeling* (or discomfort, tickle or pain). The threshold of feeling is different at different frequencies, ranging from 0.1 W/m$^2$ at 6000 Hz to 10 W/m$^2$ at lower and higher frequencies.

4. As is seen, the human ear is extremely sensitive. The range of intensities from the threshold of audibility to the threshold of feeling is about $10^{12}$ to $10^{13}$. This range of values can conveniently be represented on a logarithmic scale. For this purpose, use is made of what is called the *intensity level*:

$$L = 10 \log_{10} (I/I_0) \qquad (58.1)$$

where $I$ is the intensity of the musical tone investigated and $I_0$ is the standard threshold of audibility.

The intensity of sound, as received at any place, is measured in a unit called a *decibel*. $L = 1$ dB, if $I = 1.26\ I_0$ (in this case, $\log_{10} (I/I_0) = \log_{10} 1.26 = 0.1$).

*Table 58.2*

| Source | $l$, m | $I$, W/m$^2$ | $L$, dB |
|---|---|---|---|
| Whisper | 1 | $10^{-12}$ | 0 |
| Falling droplets of water | 1 | $10^{-10}$ | 20 |
| Ordinary conversation | 1 | $10^{-8}$ | 40 |
| Automobile on asphalt road | 5-10 | $10^{-6}$ | 60 |
| Symphony orchestra | 3-5 | $10^{-4}$ | 80 |
| Rivetting gun | 1 | $10^{-2}$ | 100 |
| Aircraft engine | 10 | 1 | 120 |

For comparison, Table 58.2 shows the rating of some sounds in comparison with the standard threshold of audibility in terms of intensity and intensity level. The distance from the sound source to the ear is in metres.

## 58.2. SOURCES OF SOUND

1. In principle, any body capable of vibrating in the desired frequency interval may serve as a source of sound. In practice, however, use is made of only sources meeting certain requirements. Above all, a source should be a good radiator, that is, it should couple its vibrational energy to the surroundings well. To achieve this, the surrounding medium should be utilized to reinforce vibrations. Besides, the size of a vibrating body should be comparable with the wavelength emitted.

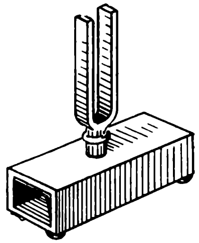


Fig. 58.1

For example, a tuning fork is a poor radiator, even when its prongs are vibrating with a considerable amplitude. However, if a tuning fork is set up on a wooden box open at one end and of a size such that the length of the air column in the box is equal to a quarter of the wavelength of the vibrations set up by the tuning fork, we shall hear a loud tone (Fig. 58.1). This is because, through resonance, the tuning fork sets into vibration both the air column enclosed in the resonating box tuned to its natural frequency and the walls of the resonator. The air column and the resonator wall couple the energy to the surrounding air better than the prongs of the tuning fork.

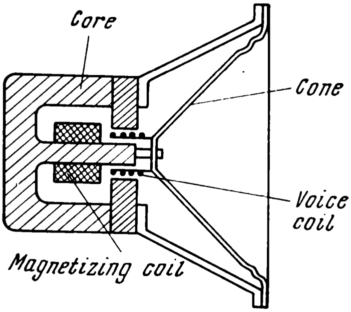


Fig. 58.2

For the same reason, a poor sound comes from a string stretched between, say, two walls. If, however, the same string is stretched over a resonating box or, still better, over the body of a violin or guitar, the sound will at once become remarkably loud. The tone is radiated not by the string, but by the sounding board and the air columns tuned to resonate at the pitch produced by the string. This is why the quality of a musical instrument is decided not so much by its strings as by the quality of its sounding board.

2. A second requirement that a good sound source is often to meet is its ability to reproduce a wide range of frequencies without marked distortion. Examples are a loudspeaker and headphones.

Fig. 58.2 shows a sectional view through a dynamic loudspeaker. It uses a paper diaphragm which has the shape of a truncated cone. The cone is glued to a cylindrical former on which a voice coil is wound with a few turns of fine wire, with its leads taken to two terminals. The voice coil is placed in the air gap of a strong permanent or electric magnet. The voice coil is energized with a signal voltage from an audio-frequency amplifier. Since a current-carrying conductor in a magnetic field is acted upon by Ampere's force, the voice coil and the cone are set into forced vibrations. The vibrations of the cone are transmitted to the surrounding air, and a sound wave is set up in it. In order to reinforce the radiated sound, the loudspeaker is usually mounted on a sounding board, called the *baffle*. This may be the enclosure of a radio or TV set.

If a loudspeaker is not to distort sound, it must respond equally well to all frequencies from several hundred to several thousand hertz. On the other hand the movable system of a loudspeaker should have no natural frequencies of its own in that range or else undesirable resonance might occur. This is why the movable system of a loudspeaker is built to have a low natural frequency (under 100 Hz) and a low $Q$-factor. With this arrangement, the loudspeaker operates within the nearly horizontal part of its resonance curve and all transients in it die out rapidly (see Secs. 53.6 and 53.7), thereby ensuring faithful reproduction.

3. In contrast, if a sound radiator is to operate on a single frequency, it is sought to make the natural frequency of the radiator as close to the frequency to be radiated as practicable. Then resonance will markedly raise the power output and efficiency of the radiator. For example, *ultrasonic transducers*, that is, devices which transform high-frequency electric energy into high-frequency mechanical energy, operate at resonance. At present, only two types of transducers are in common use in the ultrasonic range, namely *piezoelectric elements* and *magnetostrictive elements*.

## 58.3. ULTRASONIC TRANSDUCERS

1. Some crystals, for example, quartz, exibit what is called the *piezoelectric effect*. It consists in that an electric field applied along the $X$-axis of a crystal (Fig. 58.3) causes the crystal either to expand or contract in that direction and to contract or expand along the $Y$-axis, while the size of the crystal along the $Z$-axis remains unchanged.

In sketch form, an ultrasonic transducer immersed in water is shown in Fig. 58.4. The radiating element is a quartz plate, *1*, cut as shown in Fig. 58.3, that is, at right angles to the $X$-axis (see Sec. 32.1), known as an $X$-cut. Steel plates, *2* and *3*, are electrodes through which an alternating voltage is applied to the radiating element from an ultrasonic generator over a cable, *4*. Above the steel plate, *2*, there is an air cushion from which practically all of the ultrasonic energy is reflected (Sec. 56.6), so that all ultrasonic radiation is directed into the water.

For a greater power output and better directivity, the diameter of the radiator should be made as large as practicable. Because of this, instead of a single crystal, use is made of a mosaic of several plates of precisely the same thickness and of precisely the same crystallographic cut.

2. The plates operate at a resonant frequency, so that the amplitude of vibrations is a maximum. The frequency of a plate is determined by its thickness and the velocity of sound through it (see Sec. 57.3). Since the plate thickness can accommodate a whole number of half-waves, its natural frequencies can readily be computed from equation (57.8).

Another most common piezoelectric material used in the manufacture of ultrasonic transducers is polycrystalline barium-titanate ($BaTiO_3$) produced as ceramic plates. The first step in the manufacture of barium-titanate ceramic plates is to grow miniature crystals measuring about one millimetre across. Then they are mixed with a small amount of bonding material (a barium salt), and the mixture is then heated to 1300°-1400 °C to sinter it into specimens of any size and shape. Barium-titanate ceramic plates are then polarized in an electric field with an intensity of about $10^6$ V/m. After removal of the polarizing field, a residual polarization, similar to the residual magnetization of ferromagnetics, is left in barium-titanate, as in all other ferroelectrics.

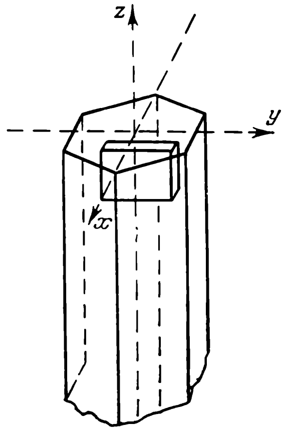


Fig. 58.3

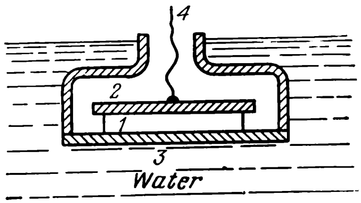


Fig. 58.4

When an alternating electric field is then applied to a pre-polarized polycrystalline specimen in the direction of the polarizing field, longitudinal vibrations, similar to those of an $X$-cut quartz plate, are set up in the same direction.

3. Magnetostrictive transducers (see Sec. 42.6) are manufactured from nickel or nickel-iron-vanadium alloy, like Permendure (49% Fe, 49%Ni, and 2%V), and also from complex oxides known as ferrites, namely manganese-zinc ferrite and nickel-zinc ferrite (like Fer-

roxcube A and Ferroxcube B). A magnetostrictive transducer is assembled from plates in order to reduce eddy-currents and carry a coil energized with a high-frequency current from a generator (Fig. 58.5).

The magnitude of the magnetostriction effect is small. In the most commonly used fields with an intensity of about $5 \times 10^4$ A/m the relative elongation for nickel-iron-vanadium alloy is $\varepsilon = \Delta l/l \approx \approx 5 \times 10^{-5}$. For a magnetostrictive element of the usual height, $l = 65$ mm (operating at a frequency of 25 kHz), the elongation $\Delta l$ is about 3 µm. Therefore, magnetostrictive elements are only operated at their natural frequencies, usually 25 kHz, or, although more seldom, 50 kHz and 100 kHz. At the higher frequences, hysteresis losses increase to a point where a magnetostrictive element becomes ineffective.

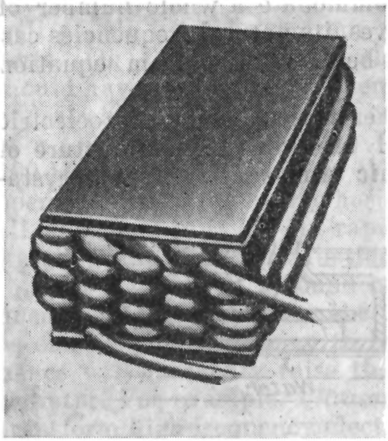


Fig. 58.5

## 58.4. CONVERSION OF SOUND TO ELECTRIC SIGNALS

1. Practically any transducer converting electric energy into sound or ultrasound can perform the reverse function of converting sound or ultrasound back into electric signals.

A device specifically designed to convert sound waves into corresponding electrical variations is called a *microphone*.

One of the types most commonly used today is the *dynamic microphone*. It is similar to the dynamic loudspeaker, except that the heavy paper cone is replaced with a light-weight diaphragm. A sound wave reaching the diaphragm sets it into vibrations, and the diaphragm vibrates together with the voice coil arranged in the air gap of a strong permanent magnet. As a result, a current is induced in the coil vibrating in the magnetic field, which is then fed to an amplifier.

Another type is the *electrostatic microphone* (also called the *capacitor* or *condenser microphone*). The diaphragm and the housing serve as two electrodes that make up a capacitor. The incident sound wave sets the diaphragm in vibration, the diaphragm motion changes the spacing between the two electrodes, and the capacitance of the capacitor is changed in proportion (see Sec. 37.6). As a result, an alternating voltage is developed across the load resistor. The frequency of the output voltage is equal to that of the incident wave, and the amplitude is proportional to the wave amplitude.

In a third type of microphone, called the *carbon microphone*, carbon granules are enclosed in a chamber with one fixed and one movable electrode. The moving electrode is fastened to the diaphragm so that diaphragm motion varies the pressure on the carbon particles. As the pressure varies, it changes the resistance between the granules, thereby modulating the direct current flowing to the carbon chamber. An increase in the pressure brings about a decrease in the resistance and, as a result, an increase in the current; a decrease in the pressure brings about a decrease in the current. As in the previous types, weak variations of current (or voltage) are amplified by a valve amplifier.

2. It should be noted that in all types of microphone, the natural frequency of the moving system is chosen to markedly differ from the frequency of the received vibrations. This is done in order to avoid accentuating any single frequency out of the entire spectrum by resonance. In contrast, receiving ultrasonic transducers operate at a resonant frequency. As already noted, the same ultrasonic transducer may operate as a transmitter and as a receiver in turn.

## 58.5. THE HUMAN EAR

1. The ear, the auditory organ of man (Fig. 58.6), is of rather elaborate structure. Basically, it consists of the outer ear and the inner ear. The outer ear includes the pinna or auricle, often called the ear, and the external auditory meatus or canal leading inward to the tympanum or eardrum which vibrates in response to sound waves. The tympanum, *1*, separates the outer ear from the middle ear, which is a small cavity bridged by three tiny bones, the hammer, anvil, and stirrup, *2*. The hammer is in contact with the tympanum, and the stirrup with the oval vestibular window, *3*, that leads into the inner ear. The middle ear is derived from the pharynx and connected with it by the Eustachian tube. The inner ear, or the labyrinth, is made up of a series of communicating cavities. In the labyrinth, only the cochlea, *4*, containing terminations of the auditory nerve, is

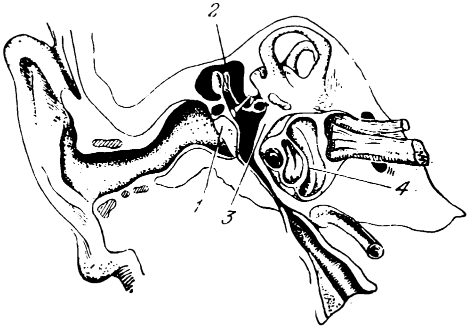


Fig. 58.6

the auditory portion of the ear. The three semicircular canals are responsible for the sense of equilibrium in man.

The cochlea, *4*, consists of a spiral canal filled with liquid (lymph) and making two and a half turns. The central canal encloses Corti's organ, the sensory organ for hearing. It has five rows of what are called hair cells which run the whole length of the cochlea. They make up a total of about 4800 fibres each containing five cells. These cells form the basilar membrane which increases in width as it passes from the base of the cochlea towards the apex.

2. The ear responds to sound as follows. A sound wave, on passing through the external auditory meatus, reaches the tympanum, *1*, and sets it into forced vibrations. In the inner ear, these vibrations pass through bones, *2*, which act as a kind of amplifier, and are allowed to pass through the oval vestibular window, *3*. This oval window sets up vibrations in the lymph and, through it, in the fibres of the cochlea. Of all the fibres those whose natural frequency is identical with that of the incoming sound vibrate most. Because of this, we can tell one musical tone from another and sense the difference in quality. In effect, Corti's organ analyzes the incoming sound wave into its spectral components and transmits appropriate data to the brain for processing.

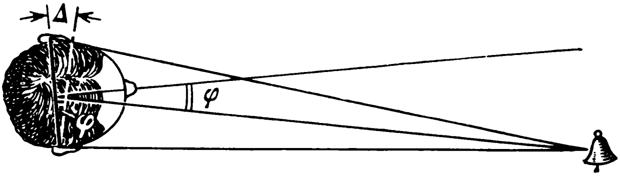


Fig. 58.7

3. With his two ears, man can determine the direction of a sound source (the binaural effect). When a sound source is directly in front of the listener, the sound will arrive at the two ears at the same time; if the sound source is located to one side (Fig. 58.7), the sound will arrive at one ear earlier than at the other, and the listener will recognize this delay as a phase shift.

If the source is offset through an angle $\varphi$, then the path length difference will be $\Delta = d \sin \varphi$, where $d \approx 20$ cm is the separation, or base, between the ears. The time delay is $\tau = \Delta/u = d \sin \varphi/u \approx$ $\approx 5.9 \times 10^{-4} \sin \varphi$. We can unerringly recognize time delay of 0.1 of a cycle (a phase shift of $0.2\pi$). At frequencies near 1000 Hz, this works out to $\tau_{\min} \approx 10^{-4}$ s. Then, from $\sin \varphi_{\min} \approx 10^{-4}/5.9 \times$ $\times 10^{-4} = 0.17$, the minimum angle, $\varphi_{\min}$, will be about 10°.

## 58.6. INFRASONICS AND ULTRASONICS

1. Experience shows that infrasound damps out very slowly. This is why the attenuation of an infrasonic wave is solely due to the redistribution of energy across the rising wave front, if the wave is nearly spherical. If, on the other hand, the source is the sea waves raised by the wind, where the wave front extends for hundreds of metres, the intensity of the infrasonic wave changes with distance very little.

It appears that fish and sea animals can pick up infrasonic waves and can sense the advance of a storm. The strong infrasonic waves accompanying a storm can travel practically with no attenuation for hundreds or even thousands of kilometres, thereby giving a warning of an approaching storm.

2. Among other things, ultrasonic waves have a considerable intensity and can be channelled in a desired direction.

The intensity of an ultrasonic wave is

$$I = \frac{1}{2}\rho u\omega^2A^2$$

Owing to the high frequencies involved, ultrasonic waves may have intensities up to 100 W/cm$^2$ = 10 kW/m$^2$, when produced by a barium-titanate transducer. Usually, use is made of lower power values, in the range 10-20 W/cm$^2$.

Owing to its high intensity, the ultrasonic wave can affect the properties of materials and the path of some industrial processes. For example, strong ultrasonic waves may be utilized to comminute materials into very fine powders, to remove rust or grease films from metallic surfaces, or to remove persistent spots or dirt from fabrics. With an ultrasonic soldering iron, one can readily braze aluminium and some other metals, because ultrasound breaks up the oxide film on the metal surface and enables the solder to make a reliable bond with the base metal. An ultrasonic transducer immersed in a vessel containing two immiscible liquids (for example, water and oil) will turn them into a homogeneous emulsion with a particle size ranging from a split micron to a few microns. This effect is also utilized in the preparation of medicines and drugs in the form of water emulsions of insoluble medicinal substances.

3. Ultrasound can be used to cut, grind, drill and otherwise machine metals and other materials. Ultrasonic cutting consists in that abrasive particles are oscillating together with a vibrating cutting tool and break particles of the material off the workpiece. Using a suitable ultrasonic tool, it is possible to make holes in workpieces widely differing in shape.

4. As a rule, the size of the ultrasonic transducer is many times

the wavelength of the ultrasonic wave in the medium. Because of this, the wave is sent out in a narrow beam whose angular spread is given by equation (57.31) as

$$\sin\theta \approx \lambda/D$$

This property is utilized in echo-sounders and sound navigation and ranging (*sonar*) equipment. In an echo-sounder, the ultrasonic transducer attached to a ship's bottom sends out short ultrasonic pulses (called "pings") with a duration of about 0.1 s. The wave reaches the bottom, is reflected and is picked up either by the same transducer during the intervals between the outgoing pulses or by a separate receiving transducer. The reflected pulses are recorded on a strip chart, and the record is converted to values of depth in metres on a calibrated scale.

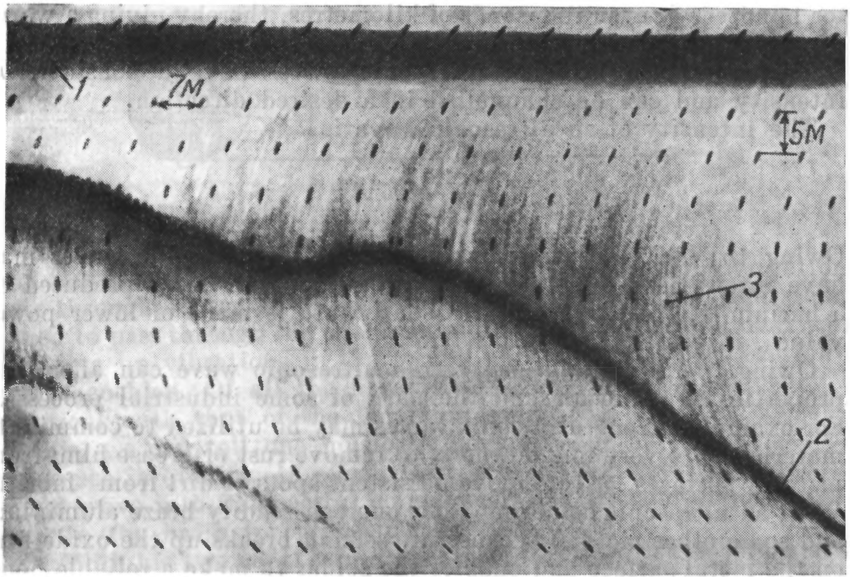


Fig. 58.8

Ultrasound can also be employed to locate shoals of fish, utilizing the fact that the swimming-bladder of fish is filled with air which is a good scatterer of ultrasonic waves. Fig. 58.8 shows a record produced by an ultrasonic fish-finder in which at *3* is the echo-trace from a school of scad, the dark band at *1* is the water surface, and the tilted band at *2* is the sea-floor.

5. Ultrasound is widely employed in the detection of flaws in

castings, forgings, rolled stock, welded seams, and similar products. One such ultrasonic flaw detector developed by Sokolov of the Soviet Union is shown in Fig. 58.9. The generator produces short pulses at a frequency of a few megahertz; these drive a barium-titanate (or quartz) transducer which converts them into ultrasonic pulses coupled into the work-piece being inspected. At the same time a spike is displayed on a CRT screen. On reaching the lower face of the workpiece, the ultrasonic beam is reflected and picked up by the transducer. As a result, a second spike appears on the CRT screen. Should there be any flaw, say a blow-hole, in the way of the ultrasonic beam, it will be reflected back before reaching the other side of the workpiece is reached and an echo will appear shifted from the position in which it would be in the absence of a flaw.

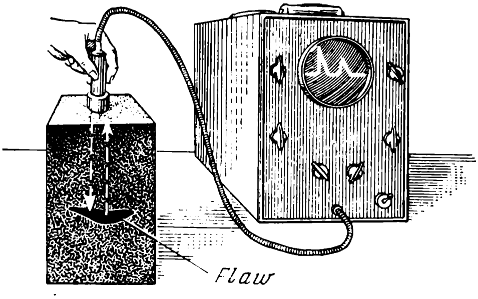


Fig. 58.9

6. Some animals utilize ultrasound, too. For example, bats find their way in flight and hunt for their pray using a kind of *sonar* technique. The bat emits a constant series of short ultrasonic pulses at a frequency of 20 to 60 kHz; the bat's large ears are highly specialized to perceive these sounds and to determine by the elapsed time the direction of objects from which they are reflected and the distance of these objects.

Porpoises and whales and, probably, some other sea animals use ultrasound for the location of objects, too. Using ultrasonic pulses, porpoises can locate schools of fish, avoid collision with obstacles, or even "converse" with one another. The usability of ultrasonic sound to sea animals is explained by the fact that it is absorbed by sea water very little. At a frequency of 50 kHz, the half-thickness is about 2.5 km, while at 100 kHz it is about 100 m. In contrast, light is strongly absorbed even by clear sea water, so that the radius of visibility is limited to a few metres.

# Chapter 59

# ELECTROMAGNETIC WAVES

## 59.1. VELOCITY OF ELECTROMAGNETIC WAVES

1. Working on a theory of electromagnetism between 1864 and 1873, James Clerk Maxwell derived a system of differential equations relating the field vectors to the field sources, namely currents and charges. On the basis of these equations, Maxwell concluded that in a vacuum and dielectrics any disturbances of the electromagnetic field should be propagated as electromagnetic waves.

In 1887-1889, Heinrich Hertz experimentally verified Maxwell's electromagnetic theory. He found that electromagnetic waves could be reflected, refracted, diffracted and focused exactly as had been predicted by Maxwell.

2. According to Maxwell, the velocity of electromagnetic waves in a dielectric, when expressed in terms of permittivity, ε, and permeability, μ, is given by

$$u = 1/\sqrt{\varepsilon\varepsilon_0\mu\mu_0} = c/\sqrt{\varepsilon\mu} \tag{59.1}$$

where $c = 1/\sqrt{\varepsilon_0\mu_0}$ is the velocity of light in a vacuum (see Sec. 40.4). Hence Maxwell believed light to be likewise an electromagnetic wave.

Except ferromagnetics, the permeability of all materials is unity very nearly (see Secs. 42.4 and 42.5). Therefore, setting $\mu = 1$ in equation (59.1) gives an expression for the velocity of electromagnetic waves in dielectrics of the form

$$u = c/\sqrt{\varepsilon} \tag{59.2}$$

## 59.2. PLANE SINUSOIDAL WAVE

1. Away from a source of sinusoidal vibrations at a radian frequency ω, a wave may be assumed to be plane. Let this wave be propagated along the $x$-axis. Then the wave equation can be written as follows:

$$\left.\begin{array}{ll} E_x = 0, & H_x = 0, \\ E_y = 0, & H_y = H_0 \cos(\omega t - kx), \\ E_z = E_0 \cos(\omega t - kx); & H_z = 0 \end{array}\right\} \tag{59.3}$$

Here, $k = \omega/u$ is the wave number, equation (56.6), and $u$ is the wave velocity, equation (59.2). In graphic form the wave is shown in Fig. 59.1.

2. As is seen, in a wave which travels along the $x$-axis, none of the field vectors vary with ($E_x = H_x = 0$). In other words, the ele-

ctromagnetic wave is *transverse*. In this respect, it fundamentally differs from elastic waves which practically always have a longitudinal component.

Well before Maxwell, it had been known that light was a purely transverse wave (see Sec. 64.4). In the wave theory of light advanced by Huygens, Young and Fresnel, which treated light as a process in an elastic ether, this caused a good deal of difficulties because it could not explain why light had no longitudinal component. This difficulty was removed by the electromagnetic theory of light.

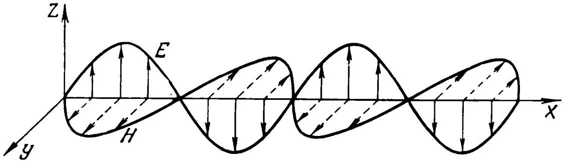


Fig. 59.1

3. From Maxwell's equations it follows that for an electromagnetic wave the magnitudes of the field vectors $E$ and $H$ are related as

$$\mu\mu_0 H^2 = \varepsilon\varepsilon_0 E^2 \tag{59.4}$$

Thus, the electromagnetic wave components have the same volume energy density, that is, $w_e = w_m$ (see Sec. 43.11).

If two quantities are equal, each is equal to the square root of their product. Therefore,

$$w_e = w_m = \sqrt{w_e \cdot w_m} = \sqrt{\frac{\varepsilon\varepsilon_0 E^2}{2} \cdot \frac{\mu\mu_0 H^2}{2}} = \frac{EH}{2} \sqrt{\varepsilon\varepsilon_0 \mu\mu_0} \tag{59.5}$$

Noting expression (59.1) for the wave velocity, we get

$$w_e = w_m = EH/2u \tag{59.6}$$

4. The energy density of a wave is given by

$$w = w_e + w_m = EH/u \tag{59.7}$$

By definition (see Sec. 55.3), the wave intensity is $I = P/S = \overline{w}u$, where $\overline{w}$ is the average energy density. Noting equation (59.7), we get

$$I = \overline{EH} \tag{59.8}$$

Thus, the intensity of an electromagnetic wave is equal to the average product of the field vector magnitudes.

Like an elastic wave, an electromagnetic wave is a carrier of energy, with energy being transferred in the direction of wave propagation. Then, inevitably, an electromagnetic wave should also have a momentum and should bring pressure to bear on bodies. Precisely this conclusion was drawn by Maxwell in his *Treatise on Electricity and Magnetism* published in 1873.

## 59.3. LIGHT PRESSURE

1. Where light pressure comes from can be explained, taking the effect of an electromagnetic wave on a sheet of metal as an example (Fig. 59.2). The electric component of the field causes an electron to move against the electric intensity vector **E** at a velocity $v = (\gamma/en)E$ (see equation 39.25), where $\gamma$ is the electrical conductivity of the metal and $n$ is the concentration (or number density) of conduction electrons. The magnetic component of the field exerts on a moving electron with what is known as the *Lorentz force* given by equation (41.1) as

$$F_m = evB = \mu_0 evH$$

As is seen, the force exerted by the electromagnetic field on each electron is

$$F_m = (\mu_0\gamma/n)\, EH$$

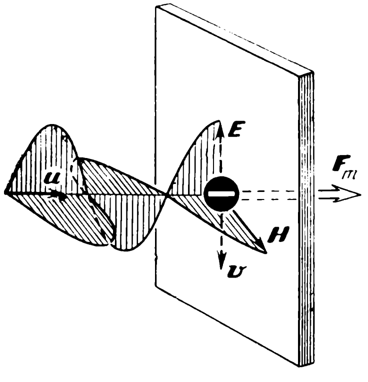


Fig. 59.2

which, rdingacco to (59.7), is proportional to the field energy density. The pressure on a plate is equal to the product of the average force $\overline{F}_m$ by the number of electrons $n_1$ per unit area, that is $p = \overline{F}_m n_1$. To sum up, the pressure due to an electromagnetic wave is

$$p = (\mu_0\gamma n_1/n)\,\overline{EH} = (\mu_0\gamma n_1/n)\,\overline{w}u = K\overline{w} \qquad (59.9)$$

where $K$ is a constant characteristic of a given material.

2. According to Maxwell, the pressure of an electromagnetic wave is

$$p = (1 + R)\,\overline{w} \qquad (59.10)$$

where $R$ is the reflection coefficient. For a reflecting surface, $R_r = 1$ and $p_r = 2\overline{w}$. For a black surface which absorbs all of the incident radiation, $R_{black} = 0$ and $p_{black} = \overline{w}$. Equation (59.10) will be derived in Sec. 68.5.

Thus, although derived by an elementary reasoning, equation (59.9) correctly relates light pressure to the volume energy density of an electromagnetic wave.

3. Some of the leading scientists strongly doubted Maxwell's hypothesis of light pressure, the more so that his reasoning was rather shaky. It was not until 1900 that Maxwell's hypothesis was decisively verified by Lebedev who detected and measured light pressure on solids. In 1907-1910 Lebedev also discovered light pressure on gases.

In his experiments, Lebedev used an extremely sensitive torsion balance in which the moving part was a light-weight frame carrying bright and dark vanes from 0.1 to 0.01 mm thick (Fig. 59.3). Since the light pressure on a dark vane is one-third less than it is on a bright vane, the moving system is subjected to a torque which can be measured from the angle that the filament suspension is twisted through. To measure energy density, Lebedev directed the light beam on a specially designed miniature calorimeter and noticed the rise in temperature. From a series of observations, he concluded that within the limits of experimental error the light pressure was in good agreement with equation (59.10) derived by Maxwell.

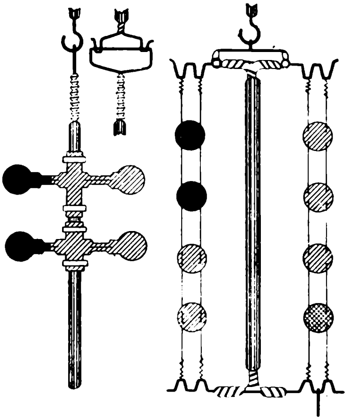


Fig. 59.3

Lebedev's experiments came as a historic proof of Maxwell's idea, notably his discovery that light is an electromagnetic wave. His experiments won Lebedev world renown and came into the history of physics as a classical example of an extremely fine physical experiment.

## 59.4. ELECTROMAGNETIC WAVES DUE TO AN ACCELERATED CHARGE

1. If a charge is at rest or in a uniform and rectilinear motion relative to an inertial reference system, it will set up no electromagnetic waves. The only exception is Cerenkov radiation (Sec. 59.7).

This is because when a charge is at rest relative to a reference system, it gives rise to a time-invariant Coulomb, or electrostatic, field. When a charge is moving at a constant speed, the electric field comes by a magnetic component (see Secs. 40.3 and 40.4). However, the uniform motion of a charge is not accompanied by the emission of electromagnetic waves.

It is only an accelerated charge that can be a source of electromagnetic waves. For simplicity, let us take a charge moving in a vacuum.

2. Let a charge $q$ at a time $t$ be at the pole of a spherical polar

coordinate and be moving with an acceleration $a$ in the negative direction of the $z$-axis (Fig. 59.4). Then, a spherical wave will radiate from the pole as if it were a point source, with the field vectors being functions of time and distance from the source.

According to Maxwell's equations, the electric intensity at any point $M$ is made up of two components. One is due to the usual electrostatic field $E_{electrostat} = q/4\pi\varepsilon_0 r^2$. The other is produced by the accelerated motion of the charge. It is only this component, $E_{wave} = \mu_0 qa \sin\theta/4\pi r$, that represents the wave motion. The electrostatic intensity vector, $E_{electrostat}$, is in the direction of the radius vector ($q > 0$) and at right angles to the vector $E_{wave}$ which is contained in a plane passing through the radius vector and the acceleration vector (see Fig. 59.4).

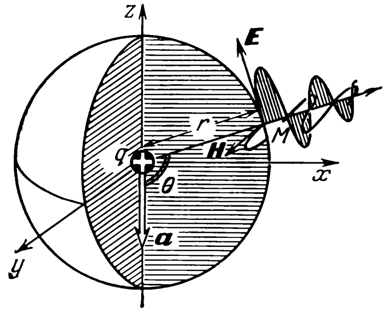


Fig. 59.4

3. Since $E_{wave}$ decreases with distance much more slowly than $E_{electrostat}$, there is a distance $r_0$ where $E_{wave} \gg E_{electrostat}$, or $\mu_0 qa/4\pi r_0 \gg q/4\pi\varepsilon_0 r_0^2$. Hence

$$r_0 \gg 1/\mu_0\varepsilon_0 a = c^2/a \qquad (59.11)$$

The region of space $r \geqslant r_0$ distant from the field source is called a *wave zone*. In a wave zone, the electrostatic field due to a charge may be neglected, and the field vectors of a radiating charge may be written as follows

$$E = \frac{\mu_0 qa \sin\theta}{4\pi r}, \quad H = \sqrt{\frac{\varepsilon_0}{\mu_0}}\, E = \frac{qa\sin\theta}{4\pi cr} \qquad (59.12)$$

### 59.5. ELECTROMAGNETIC WAVES DUE TO AN OSCILLATING CHARGE AND A DIPOLE

1. Let a charge $q$ be oscillating harmonically at the pole of spherical polar coordinates so that $z = A\cos\omega t$, with an acceleration $a = -A\omega^2\cos\omega t$ (see Sec. 49.1). The point $M$ is $r$ distant from the pole, so that $r \geqslant r_0$ (see Fig. 59.4). In order to write an equation for the field at point $M$ at time $t$, the acceleration entering equation (59.12) should be taken at the instant

$$\tau = t - r/c$$

Then

$$\omega\tau = \omega(t - r/c) = \omega t - kr$$

and the expressions for the field vectors will take the form

$$\left.\begin{aligned} E &= \frac{\mu_0 qA\omega^2 \sin\theta}{4\pi r}\cos(\omega t - kr + \pi) \\ H &= \frac{qA\omega^2 \sin\theta}{4\pi c r}\cos(\omega t - kr + \pi) \end{aligned}\right\} \tag{59.13}$$

Equations (59.13) describe the electric and magnetic fields, $E$ and $H$, due to a spherical wave (see Sec. 56.3) whose frequency coincides with that of the oscillating charge.

2. Then condition (59.11) will take the form

$$r_0 \gg c^2/\omega^2 A$$

Noting that $c/\omega = \lambda/2\pi$, we get

$$r_0 \gg \lambda^2/4\pi^2 A \tag{59.14}$$

3. The total power given up by an oscillating charge can be found as follows. To begin with, we find the wave intensity in a given direction from equation (59.8). Then we multiply the wave intensity by an elementary area of the sphere and take a sum over the entire surface. Finally, accurate to the numerical factor, we obtain

$$P \sim \overline{EH}S \sim \frac{\mu_0 qA\omega^2}{4\pi r}\cdot\frac{qA\omega^2}{4\pi c r}\cdot 4\pi r^2 \sim \frac{\mu_0 q^2A^2\omega^4}{4\pi c} \tag{59.15}$$

Precise calculation gives

$$P = \mu_0 q^2 A^2 \omega^4/12\pi c \tag{59.16}$$

As is seen, the radiation power is proportional to the frequency raised to the fourth power. Therefore, at low frequencies, radiation energy is negligibly small, but rapidly increases with the rise in the frequency of oscillation.

4. Equations (59.13) also apply to the case where electromagnetic waves are emitted by a dipole whose electric moment, $p_e = ql$, undergoes harmonic vibrations such that $p_e = p_0 \cos \omega t$. Substituting $p_0$ for $qA$ in (59.13) and (59.16) gives equations describing the fields of a wave and an expression for the total power emitted by an oscillating dipole.

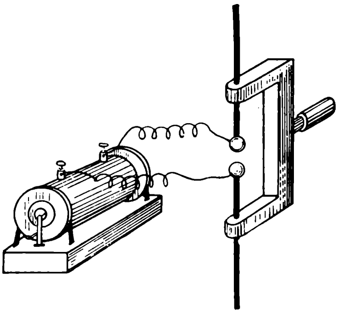


Fig. 59.5

An example of such a dipole is offered by a *half-wave dipole*, sometimes referred to as the *Hertz aerial* (Fig. 59.5). When the rods of the aerial are charged by a Rühmkorff coil (an r.f. transformer) until the air gap between the rods is broken down by a spark which is an r.f. discharge, the dipole emits at a wavelength of $\lambda = 2l$ and at a radian frequency of $\omega = 2\pi c/\lambda = \pi c/l$ (see Sec. 57.3). The radiated waves can be picked up by a similar dipole called a resonator. A similar device was used by Hertz to prove the existence of electromagnetic waves and to investigate their properties.

Hertz used a dipole from 2.5 m to 1 m long, which corresponds to wavelengths from 5 m to 2 m. In 1895, Lebedev built a half-wave dipole about 2.7 mm long and generated electromagnetic waves at a wavelength of about 6 mm. In 1922, Glagoleva-Arkadieva, using a mass radiator, obtained electromagnetic waves with a wavelength of 1 cm to 0.35 mm, thereby closing the gap between radio waves and infra-red rays.

## 59.6. ELECTROMAGNETIC WAVES DUE TO A CHARGE MOVING IN A CIRCULAR PATH

1. If an electric charge is moving in a circle of radius $R$ at constant speed, $v$, it experiences a centripetal (normal) acceleration, $a_n = \omega^2 R$, where $\omega = v/R$ is the angular velocity (see Sec. 4.8). Since any accelerated motion of a charge sets up electromagnetic waves, a charge moving in a circular path should likewise set up waves. This effect can be observed in cyclic particle accelerators. Since for the first time wave emission from charges in a circular path was noted in a synchrotron, it has come to be known as *synchrotron emission*. When accelerated to energies of the order of 100 MeV, an electron will emit waves in the visible region of the spectrum. This is known as a *radiating electron*.

2. The power given up by one electron per unit time can be found from equation (59.16) in which the amplitude is replaced with the radius of the circular path:

$$P = \frac{\mu_0 e^2 R^2 \omega^4}{12\pi c} = \frac{\mu_0 e^2 v^4}{12\pi c R^2} \qquad (59.17)$$

Expressing the radius of the circular path in terms of magnetic induction, equation (41.5), gives

$$P = \mu_0 e^4 v^2 B^2 / 12\pi c m^2 \qquad (59.18)$$

The power lost due to synchrotron emission is made good by an input of energy to the accelerating region (see Secs. 41.4-41.6).

3. Let us apply the above reasoning to an electron travelling in a circular path in an atom. This electron, too, should set up electromagnetic waves, and its energy should decrease. The time during which a charge moving in a circular path loses its energy through emission is approximately equal to the kinetic energy divided by the power of emission:

$$\tau \approx \frac{K}{P} = \frac{mv^2 \cdot 12\pi c R^2}{2 \cdot \mu_0 e^2 v^4} = \frac{6\pi c m R^2}{\mu_0 e^2 v^2} \qquad (59.19)$$

Recalling that when an electron revolves around the nucleus, the force of electrostatic attraction acts as a centripetal force, and setting the charge on the nucleus equal in magnitude to the electron charge, we get

$$mv^2/R = e^2/4\pi\varepsilon_0 R^2$$

Hence, the expression for the velocity of an electron is

$$v^2 = e^2/4\pi\varepsilon_0 mR$$

Substituting it in (59.19) gives the following expression for the sought time:

$$\tau \approx 24\pi^2\varepsilon_0 cm^2R^3/\mu_0 e^4 \tag{59.20}$$

In this equation all terms are known: $\varepsilon_0 = 1/36\pi \times 10^9$ F/m; $\mu_0 = 4\pi \times 10^{-7}$ H/m; $c = 3 \times 10^8$ m/s; $m = 9.1 \times 10^{-31}$ kg; $e = 1.6 \times 10^{-19}$ C; and $R \approx 10^{-10}$ m. Computation gives

$$\tau = \frac{24\pi^2 \times 3 \times 10^8 \times 83 \times 10^{-62} \times 10^{-30}}{36\pi \times 10^9 \times 4\pi \times 10^{-7} \times 6.6 \times 10^{-76}} \approx 2 \times 10^{-10}\ \text{s}$$

4. This is an absurd result. According to it, an electron must lose all of its energy through emission and come to a stop in about $10^{-10}$ s. Then the force of electrostatic attraction would force it to fall into the nucleus, and the atom would be destroyed. However, this runs counter to experience which proves the extremely high stability of atoms.

Thus, classical theory cannot, for fundamental reasons, explain the processes occurring in the atom. This can only be done on the basis of quantum-mechanical concepts.

## 59.7. CERENKOV RADIATION

1. In 1934, Cerenkov discovered that electrons moving with large constant velocity (relativistic electrons) through polarizable media caused a faint bluish glow. Vavilov, who supervised Cerenkov's work, came out with a hypothesis, later verified experimentally, that the glow might be caused by the motion of free electrons through the medium. The correct theoretical explanation of the phenomenon was given in 1937 by Academicians Tamm and Frank. In 1958, Cerenkov, Tamm and Frank were awarded a Nobel prize for their theory of the radiation.

2. According to their theory, Cerenkov radiation is due to the difference between the high velocity of the particle, which may be close to that of light in a vacuum, and the lower velocity of its associated electric and magnetic fields, that is, $v > u$. This phenomenon is similar to the generation of a shock-wave (the Mach wave) set up by an object travelling at a speed exceeding the velocity of

sound (see Secs. 30.7 and 30.8). Indeed, Cerenkov radiation is likewise propagated within a cone (Fig. 59.6), of opening θ given by (30.21):

$$\cos\theta = \sin\alpha = u/v = c/v\sqrt{\varepsilon} \qquad (59.21)$$

On measuring this angle experimentally, one can determine the velocity of hyper-relativistic particles.

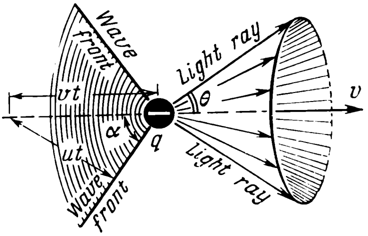


Fig. 59.6

3. Cerenkov radiation is the only case where a charge in a uniform motion emits electromagnetic waves. At first glance it may appear that this phenomenon runs counter to the theory of relativity which states that the velocity of a body can never exceed that of light (see Sec. 12.6). However, this statement is not accurate. In Sec. 12.6 it is clearly stated that the velocity of a body cannot exceed that of light in a vacuum (that is, $v < c$ always), but it is never stated that it cannot exceed that of light in a material, $u = c/\sqrt{\varepsilon} < < c$. Cerenkov radiation occurs when $c/\sqrt{\varepsilon} < v < c$, which in no way contradicts the theory of relativity.

## 59.8. DOPPLER EFFECT IN OPTICS

1. As in acoustics (see Sec. 56.4), the motion of an observer towards or away from a light source or the motion of a source towards or away from a stationary observer causes the apparent frequency, $\omega$, of an electromagnetic wave to be different from the frequency, $\omega_0$, at which this wave is emitted by the source. However, there is a fundamental difference between the Doppler effects in an elastic medium and in an electromagnetic field. The point is that an elastic medium may serve as a reference system, and so the motion of the source relative to the medium differs from that of the observer relative to the same medium, and the two kinds of motion are described by different equations (56.15) and (56.18). In contrast, the electromagnetic field cannot serve as a reference system and one can only speak of the relative motion of the source and the observer.

In Sec. 14.1 it is shown that the longitudinal Doppler effect is due to the relativistic transformation of time as one goes over from one inertial reference system to another. Consider a more general case.

2. Let a reference system, $x_0y_0z_0$, contain a source emitting electromagnetic waves at frequency $\omega_0$. We seek to find the frequency $\omega$ noted by the observer in another reference system, $xyz$, relative to which the source is moving along the $x$-axis at a speed $v$ (Fig. 59.7).

Let the angle between the ray and the direction in which the source is moving relative to the observer reference system be $\theta$, and the angle between the ray and the direction in which the observer is moving in the source reference system be $\theta_0$.

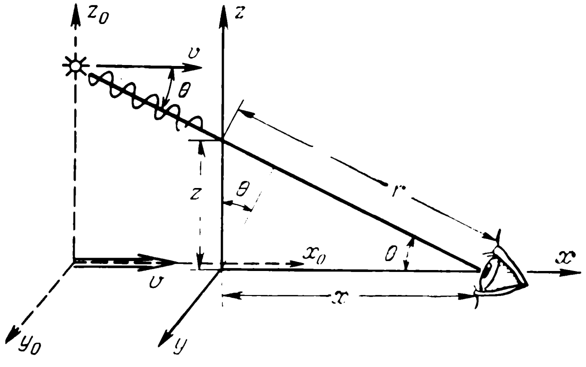


Fig. 59.7

In the wave equation, (56.10), the phase is invariant and so $\omega t - kr = \omega_0 t_0 - k_0 r_0$. It is seen from Fig. 59.7 that $r = x \cos\theta + z \sin\theta$, and from the definition of the wave number, equation (56.6), it is known that $k = \omega/c$. Substituting it in the expression for phase gives

$$\omega\left(t - \frac{x}{c}\cos\theta - \frac{z}{c}\sin\theta\right) = \omega_0\left(t_0 - \frac{x_0}{c}\cos\theta_0 - \frac{z_0}{c}\sin\theta_0\right) \qquad (59.22)$$

By the Lorentz transformation,

$$x = (x_0 + vt_0)/\sqrt{1 - v^2/c^2}, \quad z = z_0, \quad t = (t_0 + vx_0/c^2)/\sqrt{1 - v^2/c^2}.$$

Substituting them in equation (59.22) and noting that $x$, $z$, and $t$ are independent variables and, as a consequence, that equation (59.22) will hold if the coefficients of the variables are equal, we get after simple but tedious manipulation

$$\omega = \omega_0 \sqrt{1 - v^2/c^2}/(1 - v\cos\theta/c) \qquad (59.23)$$

This is an expression for the Doppler effect in optics.

3. An expression for the longitudinal Doppler effect can be derived from the above expression. At $\theta = 0$, this will be equation (14.4), when the source and the observer are approaching each other; at $\theta = \pi$, this will be equation (14.3) when the observer and the source are moving away from each other. At $v \ll c$, they are practically the same as equations (56.15) and (56.18).

The theory of relativity also suggests the existence of a *transverse* Doppler effect, when the source is moving at right angles to the ray, that is, at $\theta = \pi/2$:

$$\omega = \omega_0 \sqrt{1 - v^2/c^2} \tag{59.24}$$

Elastic waves exhibit no transverse Doppler effect.

4. In 1938-1941, H. E. Ives and G. R. Stillwell carried out experiments to observe the transverse Doppler effect, and obtained excellent agreement between experimental findings and theory. In fact, their experiments verified the relativistic law of time transformation described by the Lorentz transformation.

# Chapter 60

# ELEMENTS OF RADIO ENGINEERING

## 60.1. ADVENT OF RADIO COMMUNICATION

1. In 1895, in his report to the Russian Physics and Chemistry Society, A. S. Popov described a storm detector, a device which could sense electromagnetic waves accompanying lightning discharges. In fact, that was the ever first radio receiver. A year later, at a session of the same society, Popov demonstrated radio communication by transmitting the words: "Heinrich Hertz" from one building to another at a distance of 250 m by radio.

At about the same time radio communication was advanced by G. Marconi who did much to make radio a practical proposition. At present, when radio and television contacts are maintained between the Earth and spacecraft near the Moon or the planet Venus or Mars, a radio transmission for a distance of 250 metres or a few kilometres may appear insignificant. Yet those were man's first steps in a new and unexplored realm of science and technology.

2. In order that an aerial (for example, a half-wave dipole) of length $l$ can radiate an appreciable power, radio-frequency oscillations must be excited in it. According to equation (59.16), the radiated power is given by

$$P = (\mu_0 l^2 \omega^2 / 12\pi c)\, I_M^2 \tag{60.1}$$

where $I_M = q\omega$ is the amplitude of current oscillations. Let the aerial length be $l \approx 10$ m, and $I_M \approx 10$ A. We seek to find the frequency required to obtain a radiated power of $P \approx 100$ W. Substituting the numerical values in the above equation gives $\omega \approx 10^7$ s$^{-1}$. This frequency corresponds to a wavelength of $\lambda = 2\pi c/\omega \approx 200$ m.

In the early days of radio engineering, radio-frequency (r.f.) oscillations were generated by means of a radio-frequency spark discharge. However, this method suffered from many drawbacks. For one thing, practically all energy was converted to heat rather than to radio emission so that the efficiency of the oscillator was negligible, and transmission could only cover distances of a few tens of kilometres. For another, the spark discharge could only generate trains of damped pulses rather than sinusoidal waves. These pulses could only be used for radio telegraphy; voice operation was out of the question because damped pulses cannot be modulated.

3. A major breakthrough in radio engineering occurred with the invention of the *triode*, a three-electrode vacuum valve (see Sec. 47.3) which was followed by a variety of multi-electrode valves. With them, one can build circuits for the generation of continuous waves (see Sec. 52.3), their amplification (see Sec. 60.4), modulation and detection (see Sec. 60.5). A detailed discussion of this subject is beyond the scope of this book, and we shall only touch upon a few fundamental ideas.

Since 1950's vacuum valves in many circuits have given way to semiconductor devices, called *transistors* (see Sec. 78.4) which offer a number of important advantages. However, the replacement of a valve with a transistor does not change the essence of the matter, and we shall use circuits built around valves; the reader should remember that in all these circuits the valves can readily be replaced with transistors.

## 60.2. TRANSMISSION AND RECEPTION OF RADIO SIGNALS

1. The heart of a present-day radio transmitter (Fig. 60.1*a*) is a continuous-wave (CW) oscillator built around valves or transistors. The oscillator generates the carrier radio-frequency (or simply the carrier), ω, shown in Fig. 60.2*a*.

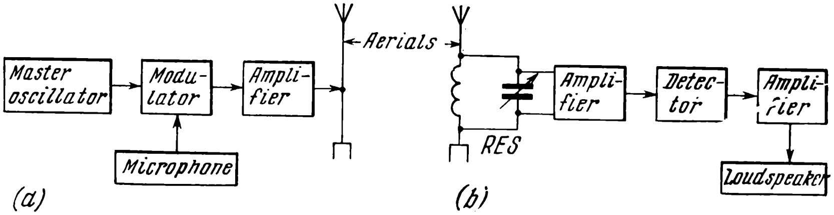


Fig. 60.1

The sound to be transmitted (Fig. 60.2*b*) drives the microphone which converts sound waves into electric oscillations. In the modulator, these electric oscillations are superimposed on the carrier to produce a modulated signal (Fig. 60.2*c*). After amplification, the modulated signal is channelled to the aerial which radiates it as electromagnetic waves.

2. At a receiving station, these waves are picked up by the receiver aerial and set up oscillations in the resonant circuit, RES (Fig. 60.1*b*). The weak r.f. signal then goes to the amplifier from which it proceeds to the detector (see Sec. 60.5). The detector extracts the transmitted intelligence from the modulated signal as an audio-frequency (a.f.) component (Fig. 60.2*e*) which is again amplified and drives the loudspeaker.

3. The resonant circuit of a receiver consists of a coil and a variable capacitor. This combination enables the resonant circuit to be tuned to the frequency of the desired radio station.

If the received signal were a sinusoidal wave, it would be advantageous to use a high-$Q$ resonant circuit because it would have a very high selectivity, that is, the ability to discriminate between the signals of two radio stations operating on closely spaced carrier frequencies (see Sec. 53.4). However, the resonant circuit has to handle a modulated signal consisting of a band of frequencies (see Sec. 50.4) covering a spectrum interval, rather than a single frequency. If the signal is not to be distorted, it is important to reproduce the entire frequency band faithfully, and this calls for a broad resonance curve, which is only possible with a low-$Q$ resonant circuits.

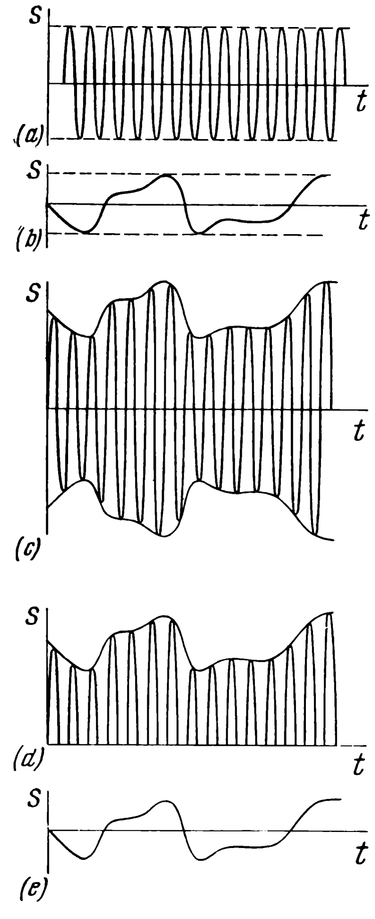


Fig. 60.2

It is seen that the requirement for high selectivity conflicts with that for high fidelity. In practice, the conflict is usually resolved by a reasonable compromise.

It should also be noted that the higher the carrier, the wider the frequency band that can be reproduced faithfully. This is the reason why at present a wide use is made of the H.F., V.H.F., and U.H.F. bands.

## 60.3. TELEVISION

1. The circuits used in television closely resemble those used in radio broadcasting. As regards the transmitter, it differs in that in addition to sound the carrier is modulated also by the vision signal coming from suitable TV camera tubes and also signals intended to synchronize the electron beam in the picture tube of the TV receiver (see Sec. 47.4).

In a TV receiver, the composite r.f. signal is again resolved into the vision signal, the sound signal, and the control signal. After amplification, these signals are routed into the respective circuits to perform their designated functions.

2. The control signals cause the electron beam to scan the screen horizontally and also from line to line. In 1/30 of a second, the electron beam traces out 625 *lines* which make up one *frame*. In the absence of a vision signal, the screen of the picture tube presents a uniformly illuminated pattern of lines, or *raster*.

The amplified signal representing the picture information is fed to the grid of the electron gun so that the intensity of the beam and the brightness of the light spot on the screen varies with the intensity of the original scene, thereby reproducing the transmitted picture.

3. Because a television signal should carry a large amount of information, it occupies a frequency band of the order of 4 or 5 MHz (in a radio broadcast receiver, the signal frequency band is about 10 kHz). This calls for a very high carrier frequency. This is why use is made of frequencies from 50 MHz to 900 MHz (wavelengths of 6 m to 30 cm). Radio broadcasts utilize longer wavelengths, from 1.5 km to tens of metres.

## 60.4. THE VALVE AMPLIFIER

1. Since a radio receiver is a long distance from a transmitter, its aerial can only pick up a negligible fraction of the energy radiated by the transmitter. This is why the weak received signal should be amplified in power. This is done by amplifiers which can be built around valves or transistors. The circuit of a simple amplifier using a vacuum triode is shown in Fig. 60.3.

2. Consider the functions and operating principle of the main elements in this circuit. The weak a.c. signal, $v_g$, which is to be amplified is applied to the valve grid connected by a resistor, $R_g$, to the "—" side of the anode battery. As a result, the grid potential varies by a small amount about the negative potential of the anode supply in step with the signal voltage.

Resistor $R_1$ and capacitor $C_1$ in the cathode circuit provide the so-called cathode bias owing to which the grid potential, $\varphi_g$, is held *below* the cathode potential, $\varphi_k$. This is because inside the triode the flow of current is from anode to cathode and then via resistor $R_1$ to the "—" terminal of the anode battery. From the direction of current flow we learn that $\varphi_k > \varphi_g$. The potential difference between cathode and grid depends on anode supply voltage and the value of resistor $R_1$, owing to which the operating ($Q$-) point of the amplifier can be positioned either within the linear portion of the characteristic curve, (Fig. 60.4*a*) or at its bend (Fig. 60.4*b*).

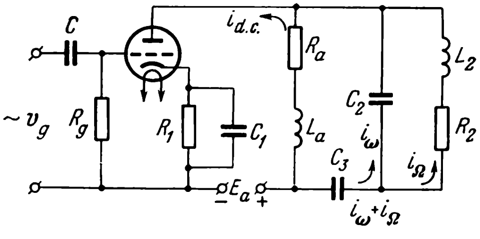


Fig. 60.3

Capacitor $C_1$ of a few tens of microfarads accumulates electric charge, thereby smoothing the ripple in the cathode potential.

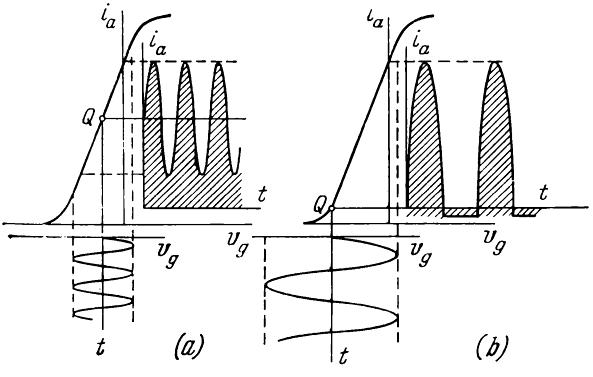


Fig. 60.4

3. If the amplifier operates within the linear portion of the characteristic (Fig. 60.4*a*), then, as follows from the diagram, variations in the grid potential will bring about precisely the same variations in anode current. With a sufficiently great slope of the curve, the amplitude of the alternating voltage $v_a$ across the anode load resistor $R_a$ may be many times the grid voltage.

A similar increase in voltage might be obtained with a transformer. However, there is a good deal of difference between a transformer and an amplifier. In a transformer, an increase in voltage is accompanied by a proportionate decrease in current; in other words, a transformer cannot build up a signal in power, because it contains no source of additional energy. In an amplifier, the power of oscillations in the anode circuit is many times that of the oscillations in the grid circuit—the necessary energy input is provided by the anode supply battery, $E_a$ (see Fig. 60.3). Thus, a valve amplifier boosts the signal in power as well as in voltage.

4. At present, instead of a triode, valve amplifiers use tetrodes (valves with two grids) and pentodes (valves with three grids). These valves have a higher amplification factor than a triode, but an amplifier using a tetrode or a pentode operates on precisely the same principle as an amplifier using a triode.

## 60.5. DETECTION (DEMODULATION)

1. As already noted, the aerial of a radio receiver picks up modulated r.f. signals (see Fig. 60.2*c*). Before the sound spoken into the microphone at the transmitting station can be heard again, it is necessary to extract the transmitted intelligence as an audio-frequency (a.f.) component from the modulated signal. This is called *demodulation*. Demodulation is accomplished by a *detector*, a device showing unidirectional conduction. It can, for example, be a vacuum diode, that is, a two-electrode valve (see Sec. 47.2), a crystal diode (see Sec. 78.3), or generally any non-linear element, that is, an element which does not obey Ohm's law.

To obtain an insight into the part played by a non-linear element in demodulation, we shall refer to the circuit of Fig. 60.3, assuming that the *Q*-point is positioned on the bend of the characteristic curve (Fig. 60.4*b*). The grid is fed a sinusoidal signal, and a pulsating current appears in the anode circuit.

In a first approximation, it may be assumed that within the operating portion the characteristic has the form of a parabola, and the anode current is a quadratic rather than a linear function of grid voltage:

$$i_a = i_0 + \alpha v_g + \beta v_g^2 \tag{60.2}$$

where $\alpha$ and $\beta$ are constants to be defined later.

2. Let the grid be fed a modulated signal,

$$v_g = A\ (1 + k \cos \Omega t) \cos \omega t,$$

where $k < 1$ is the depth of modulation, $\Omega$ is the modulating frequency, and $\omega \gg \Omega$ is the carrier frequency (see Sec. 50.2, Fig. 50.2).

The anode current is given by

$$i_a = i_0 + \alpha A\ (1 + k \cos \Omega t) \cos \omega t$$
$$+ \beta A^2\ (1 + 2k \cos \Omega t + k^2 \cos^2 \Omega t) \cos^2 \omega t \tag{60.3}$$

Noting that $2 \cos^2 \alpha = 1 + \cos 2\alpha$, it can be shown that the anode current, equation (60.3), is a sum of three terms, namely a direct current $i_{dc}$, an r.f. component $i_\omega$ of frequencies $\omega$ and $2\omega$, and an a.f. component $i_\Omega$ of frequencies $\Omega$ and $2\Omega$, that is,

$$i_\Omega = (\beta A^2/2)(1 + 2k \cos \Omega t + k^2 \cos 2\Omega t/2) \qquad (60.4)$$

As is seen, the a.f. component of anode current is a sufficiently precise replica of the modulated signal. Although detection has produced a component at twice the modulating frequency $k^2 \cos 2\Omega t/2$, it can be neglected at $k \ll 1$.

3. Let us see how the a.f. signal of interest, $i_\Omega$, can be separated from the remaining two components. We turn again to the circuit of Fig. 60.3. Here, the impedance of the voice coil in the speaker is $Z = \sqrt{R_2^2 + L_2^2\Omega^2}$. Connected in parallel with the coil is a capacitor, $C_2$.

The direct component of the detected signal cannot pass through the blocking capacitor, $C_3$; instead, it flows through the choke $L_a$, resistor $R_a$, and the valve. In contrast, the a.c. components $i_\omega$ and $i_\Omega$ are practically free to pass through the d.c. blocking capacitor $C_3$ (its capacitive reactance $X_{C3} = 1/C_3\omega$ is small), but cannot pass through the choke $L_a$, the inductive reactance of which $X_L = L_a\omega$ is high (see Secs. 54.5 and 54.6).

The r.f. signal, $i_\omega$, at frequencies $\omega$ and $2\omega$ then passes through capacitor $C_2$, because at r.f. the capacitor reactance $X_{C2} = 1/C_2\omega$ is a small fraction of the inductive reactance $X_{L2} = L_2\omega$. It is obvious that for the a.f. signal at frequency $\Omega < \omega$ the reverse is true, because of which it is free to pass through the voice coil.

4. In practice, instead of square-law detection, use is made of other forms, nor is the modulating signal sinusoidal. Yet, the example discussed above gives a true idea of how an r.f. signal is demodulated by a detector.

# Chapter 61

# INTERFERENCE OF LIGHT

## 61.1. ELECTROMAGNETIC SPECTRUM

1. The following units are used to measure the wavelengths of light in the infrared, ultraviolet, and X-ray regions of the spectrum:

1 micrometre ($\mu$m) $= 10^{-6}$ m;

1 nanometre (nm) $= 10^{-9}$ m;

1 Ångstrom (Å) $= 10^{-10}$ m.

Before the introduction of the International System of Units (SI), the micrometre was called the micron (μ), and the nanometre was called the millimicron (mμ). The Ångstrom is an off-system unit.

2. The term *visible light* applies to the part of the electromagnetic spectrum lying in the wavelength range from $\lambda_r = 7800$ Å $= 780$ nm (red light) to $\lambda_v = 4000$ Å $= 400$ nm (violet light).

In its physical nature, however, visible light does not differ from other electromagnetic waves, namely radio waves, infrared, ultraviolet, X- and gamma-rays. Therefore, the term "light" is used in a broader sense, meaning electromagnetic waves in general. Incidentally, the same goes for the term "sound wave" which may be applied to any elastic waves of low intensity, and not only to audible sound. Among other things, speaking of the velocity of light in a vacuum, $c$, we had in mind not only visible light, but any electromagnetic waves.

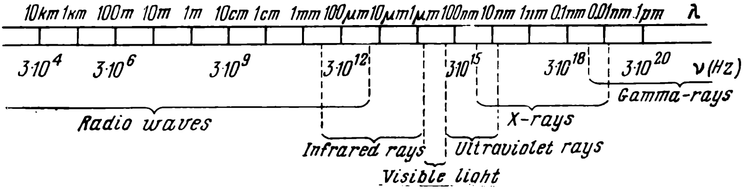


Fig. 61.1

3. Fig. 61.1 shows the electromagnetic spectrum. It extends over a huge range of frequencies, from a few oscillations per second to $10^{22}$ Hz (respectively, wavelengths from several hundred thousand kilometers down to $10^{-4}$ Å).

4. *Radio waves* extend from $10^6$ m to 1 mm. Basically, they are classed into long waves with wavelengths in excess of 1 km, medium waves from 1 km down to 100 m, short waves from 100 m down to 10 m, and ultra-short waves from 10 m down to 1 mm.

The ultra-short wave (USW) band is immediately adjacent to the *infrared region*. The boundary between them is arbitrary and is only decided by the method of generating the respective waves. Ultrashort radio waves are generated by suitable oscillators (radio methods), while infrared rays are emitted by hot bodies.

Beyond the visible region of the spectrum extends the *ultraviolet region* with wavelengths from 4000 Å down to 10 Å. Ultraviolet rays are produced by glow discharges (see Sec. 48.6), usually in mercury vapour.

The far ultraviolet region is adjacent to the *X-ray region* (see Sec. 73.3), with wavelengths from 10 Å down to 0.1 Å. The X-ray region is followed by the *gamma-ray region* (see Sec. 81.10) with wavelengths shorter than 1 Å. The X- and gamma-ray regions partly overlap, and these waves can be descriminated by the manner in which they are produced rather than by their properties. X-rays are generated in suitable tubes (see Sec. 73.3), and gamma-rays are emitted by the radioactive nuclei of some elements.

## 61.2. WAVE TRAIN. LIGHT VECTOR

1. The mechanism by which visible light, infrared, ultraviolet and X-rays are emitted will be discussed in detail in Chapters 67 through 74. For the time being, it may be stated that an excited atom with excess energy, in jumping to a lower energy state, emits an electromagnetic wave. This transition occupies a time interval of $\tau \approx 10^{-8}$ s, and the radiation lasts as long. Thus, an atom emits an *interrupted sine wave bundle* or a *wave train* (Fig. 61.2). The length of a wave train in a vacuum is $l = x_2 - x_1 = c\tau \approx 3$ m, while the wavelength of light is about $10^{-6}$ m. Thus, a single wave train accommodates several million wavelengths.

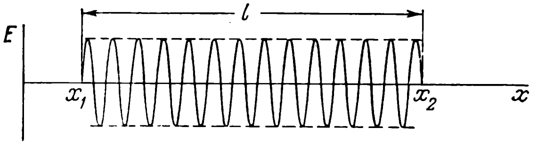


Fig. 61.2

2. Fig. 61.2 only shows vibrations of the electric intensity vector of an electromagnetic wave; the magnetic component is not shown. In our further discussion, we shall show only one component of an electromagnetic wave, assuming that the other component is present in a plane perpendicular to that shown (see Fig. 59.1). The reason why we have chosen as the *light vector* the **E** vector and not the **H** vector is that the electric component of an electromagnetic wave has a greater effect on free and conduction electrons than the magnetic component.

This can be proved as follows. According to equation (59.4), in a vacuum the wave vectors are related as $H = E\sqrt{\varepsilon_0/\mu_0}$, whence for the magnetic induction vector we have $B = \mu_0 H = E\sqrt{\varepsilon_0\mu_0} = E/c$. The electric force is $F_e = eE$, and the magnetic force is $F_m = evB = (v/c)\, eE = F_e\, v/c$. However, the velocity of electrons is only a fraction of that of light, and so the magnetic force is only a fraction of the electric force.

3. Experiments have proved this calculation. Among other things, it has been established that the photoelectric effect (see Sec. 68.1), photochemical reactions (see Sec. 68.4), the effect of light on the eye, photoluminescence (see Sec. 79.2) and some other phenomena are decided by the vector **E**. Yet, it should be remembered that in an electromagnetic wave both field vectors, **E** and **H**, are inseparable, and in no circumstances is it possible to obtain a wave containing only one field component

### 61.3. UNCERTAINTY RELATIONS FOR POSITION AND WAVE NUMBER

1. A wave train is a non-sinusoidal wave; in fact, it is not unlike quasi-harmonic pulses (see Sec. 53.7) and may be approximated by beats (see Sec. 50.1). For this purpose, we shall replace an interrupted sine wave packet of length $l$ by a sum of two sinusoidal waves of closely spaced radio frequencies ($\omega_1 = \omega - \Delta\omega$, and $\omega_2 = \omega + \Delta\omega$) and closely spaced wave numbers ($k_1 = k + \Delta k$ and $k_2 = k - \Delta k$). Then

$$E_1 = E_0 \cos(\omega_1 t - k_1 x + \varphi), \quad E_2 = E_0 \cos(\omega_2 t - k_2 x + \varphi) \tag{61.1}$$

The resultant wave (see Sec. 50.1) will be a nearly sinusoidal wave of the form

$$E = B \cos(\omega t - kx + \varphi) \tag{61.2}$$

where the time-varying amplitude is given by

$$B = 2E_0 \cos(\Delta\omega t - \Delta kx) \tag{61.3}$$

2. At a particular instant, the amplitude will be only a function of position

$$B = 2E_0 \cos(\Delta kx) = B_0 \cos(\Delta kx) \tag{61.4}$$

This wave is made up of spatial beats whose wave shape at the selected instant is shown in Fig. 61.3*a*.

The coordinates of the nodes can be obtained by setting $\cos(\Delta kx) = 0$ in equation (61.4). Hence, $\Delta kx_m = (2m + 1)\,\pi/2$, and finally,

$$x_m = (2m + 1)\,\pi/2\Delta k. \tag{61.5}$$

The length of a beat, that is, the separation between adjacent nodes is

$$L = x_{m+1} - x_m = \pi/\Delta k \tag{61.6}$$

This expression is similar to the period of a beat (see Sec. 50.1).

3. Spatial beats are not wave trains. Yet, the difference between them is not so marked as it may seem. The point is that any instru-

ment has a certain threshold sensitivity, and should the beat amplitude be below this threshold, the instrument will not be able to detect a wave train. Instead, it will register interrupted beats (Fig. 61.3*b*), and these will not practically differ from an interrupted sine wave or a wave train.

Nor can a wave train be distinguished from beats because it has a constant amplitude, while that of beats is variable. Discussing the reception of quasi-harmonic pulses in Sec. 53.7, we have seen that the transients set up each time the resonator is turned on or off may markedly affect the pulse wave shape.

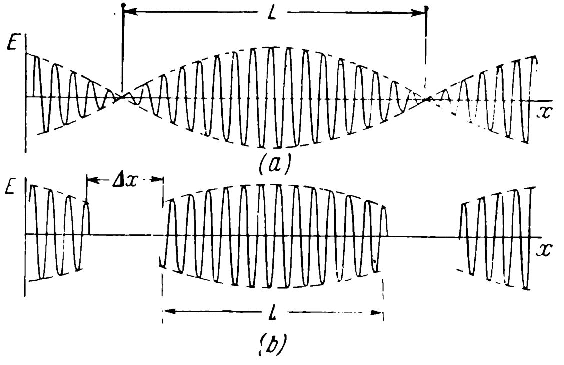


Fig. 61.3

4. Let us evaluate the uncertainty in position, $\Delta x$, for a wave train. Suppose that a beat satisfactorily represents the properties of a wave train if the amplitude at the end of this beat is at least half the maximum amplitude, that is, $B/B_0 \approx 0.5$. The intensity will then fall off to not less than one-quarter ($I/I_0 = B^2/B_0^2 = 0.25$). Using equation (61.4), we have

$$B/B_0 = \cos(\Delta k \cdot \Delta x) \approx 1/2$$

Hence, $\Delta k \cdot \Delta x \approx \pi/3 \approx 1$, and finally

$$\Delta k \cdot \Delta x \approx 1 \qquad (61.7)$$

This is the uncertainty relation between position and wave number.

5. The meaning of this relation is as follows. Since a wave train is not an infinite sinusoid, but a part of a sinusoid, instead of one definite wave number it has a range of wave numbers $\Delta k$ wide. Also, we cannot know exactly the length of a wave train; what we can determine is the length interval, $\Delta x$. This implies that the position of a wave train in space can only be determined with an uncertainty $\Delta x$.

According to (61.7), the uncertainties in position and in wave number are inversely proportional to each other. That is, if one tries to improve accuracy in determining the wave number (or the wavelength), the result will be an increased uncertainty in the position of the wave train on the coordinate axis. Conversely, if one tries to locate the wave train to a better accuracy, the result will be an increased uncertainty in the value of the wave number (and the wavelength, because $\lambda = 2\pi/k$). The uncertainty principle described by equation (61.7) plays a very important part in quantum mechanics (see Sec. 70.2*ff*).

## 61.4. MONOCHROMATIC RADIATION

1. Light from ordinary sources is a non-sinusoidal wave. It is shown in Chapter 50 that any non-sinusoidal wave may be analyzed into a spectrum (see Sec. 50.4), that is, it can be expressed as an infinite series of sinusoidal components or harmonics. In the succeeding sections, we shall examine the physical methods by which one can analyze a light wave into its components (Secs. 61.8, 61.9, 62.2, and 66.12).

Light at a single frequency (or at one wavelength) is called *monochromatic* (from the Greek "*monos*" for "one" and "*chroma*" for "colour"). In any spectroscope, its spectrum has the form of a narrow line.

2. Since, however, an emission act occupies a finite time interval $\tau \approx 10^{-8}$ s, a given spectral line has a wavelength range $\Delta\lambda$ (or a frequency range $\Delta\nu$) instead of a single wavelength (or frequency).

The wavelength (or frequency) range due to the finite emission time is called the *intrinsic line width*. The intrinsic line width can be obtained from the uncertainty relation for time and frequency given by equation (53.22), namely, $\Delta\omega\,\Delta t \approx 1$. Since the time range, $\Delta t$, is about equal to the time of emission, $\tau$, then from equation (53.22) it follows that

$$\Delta\omega \approx 1/\tau \approx 10^8 \text{ s}^{-1} \tag{61.8}$$

3. A more convenient measure of the spread in the spectral line is the *relative line width*. It is defined as the ratio of the line width to the wavelength (or, respectively, frequency) at which the spectral line has a maximum intensity. That $\Delta\omega/\omega = \Delta\nu/\nu = \Delta\lambda/\lambda$ can be shown as follows.

If the wavelength corresponding to frequency $\nu$ is $\lambda = c/\nu$, then the wavelength corresponding to frequency $\nu_1 = \nu - \Delta\nu$ is $\lambda_1 = \lambda + \Delta\lambda = c/(\nu - \Delta\nu)$, and the wavelength corresponding to frequency $\nu_2 = \nu + \Delta\nu$ is $\lambda_2 = \lambda - \Delta\lambda = c/(\nu + \Delta\nu)$. The equality $\lambda_1\nu_1 = \lambda_2\nu_2 = c$ may be rewritten as follows

$$(\lambda + \Delta\lambda)(\nu - \Delta\nu) = (\lambda - \Delta\lambda)(\nu + \Delta\nu)$$

Opening the brackets and collecting the like terms gives the sought equality for the relative line width in terms of wavelength or frequency. Substituting $\Delta\omega \approx 1/\tau \approx 10^8$ s$^{-1}$ and $\omega \approx 10^{15}$ s$^{-1}$ gives

$$\Delta\omega/\omega = \Delta\nu/\nu = \Delta\lambda/\lambda \approx 10^{-7} \qquad (61.9)$$

4. The spread in spectral lines may be caused by other factors, such as the thermal motion of atoms (Doppler broadening) and collisions of atoms (collisional broadening). In the latter case, atoms in gases under a pressure of 20 to 30 atmospheres and a temperature of 500 to 600 K collide at a high rate. Because of the collisions, the time $\tau$ during which an atom radiates is reduced, and its spectral line is broadened in proportion.

5. As is shown in Sec. 59.8, the motion of a light source is accompanied by a change in the frequency of the emitted light wave. The frequency of the wave emitted by an atom moving towards the observer increases so that $\nu_1 = \nu + \Delta\nu$, while the frequency of the wave emitted by an atom moving away from the observer decreases so that $\nu_2 = \nu - \Delta\nu$.

The average velocity of the thermal motion of atoms can be derived from the equation $\bar{\varepsilon} = m_0\overline{v^2}/2 = 3/2kT$ (see Sec. 26.5):

$$\bar{v} = \sqrt{3kT/m_0} = \sqrt{3RT/M}$$

where $M$ is the molecular mass of the gas, and $R = 8.31 \times 10^3$ J/(kilomole K) is the universal gas constant (see Sec. 26.9). The thermal velocity is a maximum for hydrogen atoms having a molecular mass of $M = 1$ kg/kilomole. At $T = 6000$ K, which corresponds to the temperature on the Sun's surface, $\bar{v} = \sqrt{3 \times 8.3 \times 10^3 \times 6 \times 10^3/1} = 1.2 \times 10^4$ m/s. This velocity is only a small fraction of that of light ($v/c \approx 10^{-4}$), so that an approximate equation, $\nu' = \nu\,(1 \pm v/c)$ may be used for calculations. Hence

$$\Delta\nu/\nu = (\nu' - \nu)/\nu = \pm v/c \qquad (61.10)$$

For hydrogen atoms on the Sun's surface the relative Doppler broadening of the spectral line is

$$\Delta\nu/\nu = v/c = 1.2 \times 10^4/3 \times 10^8 \approx 10^{-4}$$

which is 1000 times the intrinsic line width.

As is seen, a purely monochromatic light cannot be obtained for fundamental reasons. So, *light is a non-monochromatic wave.*

6. It was maintained for a long time that the intrinsic line width could not be reduced. However, in the monochromatic light emitted by lasers, the line width is less than the intrinsic line width (see Secs. 79.3 and 79.4).

## 61.5. INTERFERENCE OF LIGHT

1. In Sec. 57.5, we examined the interference of sinusoidal waves. It was shown that sinusoidal waves of the same frequency emitted by two sources will interfere destructively at some points and constructively at other points to produce an interference pattern. No interference pattern can, however, be obtained if we use two independent light sources. At the point where the two beams meet we shall only observe that the two intensities are added together.

2. This is not to say, however, that light waves cannot interfere at all. Back in 1675, Newton observed light interference (the so-called Newton rings, Fig. 61.4) in a suitable experimental set-up, but he could not explain why or how the concentric, alternately bright and dark rings he observed had been produced.

In 1801, Thomas Young observed light interference, using the experimental arrangement shown in Fig. 61.5. In this arrangement, a bright light source $LS$ illuminates a narrow slit $S$. The light wave bends around the edges of the slit (this is diffraction) and illuminates two narrow slits, $S_1$ and $S_2$. Owing to diffraction, the two waves emerging from the two slits partially overlap. The overlapping light waves interfere and produce a series of interference maxima and minima which can be observed as distinct light and dark bands when the pattern is received on a screen, $M$. Young was correct in explaining these bands as due to the interference of light waves, and determined the wave length to be $\lambda \approx 5 \times 10^{-7}$ m.

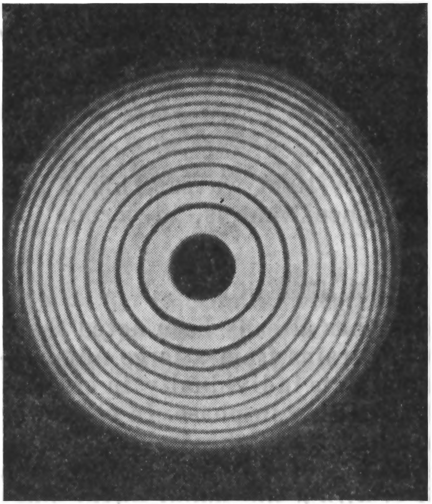


Fig. 61.4

Apart from Young's experimental set-up, there are other arrangements in which light waves can be caused to interfere. Some of them will be discussed later.

3. Now let us remove the slit $S$ from Young's set-up. Then light from the source will strike the slits $S_1$ and $S_2$ directly. Although removal of the slit $S$ has not changed the frequency of light, and the two slits, $S_1$ and $S_2$, still pass light waves at the same frequency, no interference will be produced.

Thus, it is seen that while the equality of frequencies is sufficient for *sinusoidal waves* to interfere, this condition is insufficient for ordinary light waves. This is because ordinary light waves lack in coherence, that is, in correlation between the phases.

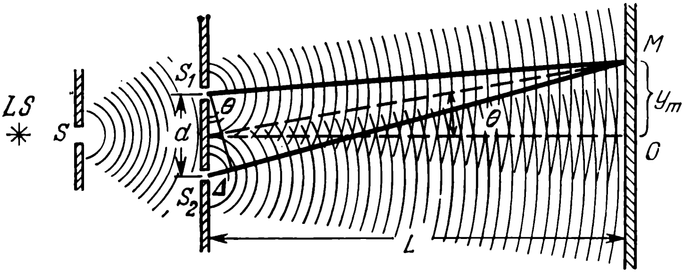


Fig. 61.5

## 61.6. COHERENCE

1. It has already been noted that a light wave consists of individual wave trains from a large number of atoms or molecules. Since each atom acts individually in creating a wave train, the resulting wave trains differ in phase. In other words, light is a quasi-harmonic wave motion whose phase varies at random. Therefore, in the expression, $s = A \cos(\omega t - kx + \varphi)$, the reference phase $\varphi$ is not constant (as it is with a sinusoidal wave), but fluctuates irregularly with time.

2. In Sec. 57.5, we derived an expression, equation (57.15), for the amplitude of the resultant wave produced by superimposing two waves having the same frequency. Noting that the wave intensity is proportional to the square of the amplitude and designating the intensity of the waves being superimposed as $I_1 = I_2 = I_0 = kA^2$, and the intensity of the resultant wave as $I = kB^2$, where $k$ is a proportionality factor, we get

$$I = 4I_0 \cos^2\left[\frac{k(r_2 - r_1)}{2} + \frac{\varphi_1 - \varphi_2}{2}\right] = 4I_0 \cos^2\left(\frac{k\Delta}{2} + \frac{\delta}{2}\right) \qquad (61.11)$$

where $\Delta = r_2 - r_1$ is the path length difference and $\delta = \varphi_1 - \varphi_2$ is the phase difference.

In sinusoidal waves, the reference phases $\varphi_1$ and $\varphi_2$ are constant; therefore $\delta = \varphi_1 - \varphi_2$ is also constant, and the intensity of the resultant wave is solely determined by the path length difference, $\Delta = r_2 - r_1$ (see Sec. 57.5).

3. When light waves are superimposed, the situation is complicated by the fact that $\varphi_1$ and $\varphi_2$ fluctuate irregularly with time. In principle, two cases are possible:

(a) The two phases vary in the same manner, and the phase difference is $\delta = \varphi_1 - \varphi_2 = 0$. This case has been examined already in Sec. 57.5.

Waves having the same frequency and a constant (most often, zero) phase difference, $\delta =$ constant ($\delta = 0$) are called *coherent*. Only coherent waves can interfere.

(b) The two phases fluctuate irregularly with time and independently of each other, and so, the phase difference, $\delta = \varphi_1 - \varphi_2$, also varies at random. These are non-coherent waves. It can be shown that non-coherent waves do not interfere, that is, they produce no interference patterns.

The duration of a light wave train is $\tau \approx 10^{-8}$ s. As a consequence, the phase fluctuates irregularly at a rate of more than 100,000,000, per second. Since light detectors (the eye, a photoelectric cell, photographic film, etc.) respond with a definite time lag, they can only register the light intensity averaged over a time interval $t \gg \tau$.

$$\overline{I} = 4I_0 \overline{\cos^2 (k\Delta/2 + \delta/2)}$$

It is known from trigonometry that $2 \cos^2 \alpha = 1 + \cos 2\alpha$. Since the argument in the cosine, ($\alpha = k\Delta + \delta$), fluctuates irregularly, the average value of the cosine over a large time interval will be zero: $\overline{\cos (k\Delta + \delta)} = 0$. Hence, the mean square of the cosine is

$$\overline{\cos^2 (k\Delta/2 + \delta/2)} = \frac{1}{2} [1 - \overline{\cos (k\Delta + \delta)}] = 1/2$$

Hence, the average intensity of the resultant wave is

$$\overline{I} = 4I_0 \overline{\cos^2 (k\Delta/2 + \delta/2)} = 2I_0 = I_1 + I_2$$

4. Thus, the superposition of non-coherent waves does not produce an interference pattern; the average intensity of a wave at any point is equal simply to a sum of the intensities of the component waves. Thus, again, an interference pattern will appear only when the light waves being superimposed are coherent.

This explains why the slit $S$ is important in Young's experiment (Fig. 61.5). In his experimental set-up, the two slits, $S_1$ and $S_2$, are on the same wave front and are illuminated by the same wave train emerging from the slit $S$. Therefore, the light waves emerging from the two slits are in phase, that is, they are coherent waves which, when received on a screen, give an interference pattern. With the slit $S$ removed, the slits $S_1$ and $S_2$ are illuminated by wave trains from different areas on the light source. The waves emerging from the two slits are non-coherent, and no interference pattern is produced.

## 61.7. SEPARATION BETWEEN INTERFERENCE MAXIMA

1. In the previous section it is shown that in calculating the interference pattern produced by interfering coherent waves, one may use the results applicable to sinusoidal waves, namely equation (57.18) for intensity, $I = I_0 \cos^2 (\pi\Delta/\lambda)$, and equation (57.20) describing the conditions for the occurrence of maxima, $\Delta = 2m\lambda/2$, and minima, $\Delta = (2m + 1)\lambda/2$.

It should, however, be stressed that while the path length difference for two sinusoidal waves may be arbitrarily great, with light waves a constant phase difference may only be maintained if the path length difference does not exceed the length of a single wave train.

2. In Sec. 57.5, the path length difference was defined as the difference in path length from the sources to the point where interference is observed, that is, $\Delta = r_2 - r_1$. This, however, is only true if a wave travels in a vacuum at velocity $c$. If a wave is propagated through a medium other than vacuum, the definition of the path length difference should be re-stated to be applicable to cases where the waves travel along the paths $r_1$ and $r_2$ with different velocities.

The velocity of light through a medium other than vacuum is $u = c/\sqrt{\varepsilon}$ (see Sec. 59.1). Now, we designate $\sqrt{\varepsilon} = n$ and call it the *refractive index* (see Sec. 63.1). Then, $u = c/n$. The wave numbers for each of the media are $k_1 = \omega/u_1 = \omega n_1/c$ and $k_2 = \omega n_2/c$. Then, the expression for the amplitude of the resultant wave, equation (57.16), will take the form

$$B = 2A \cos \frac{k_2 r_2 - k_1 r_1}{2} = 2A \cos (\omega/2c)(n_2 r_2 - n_1 r_1)$$

From a comparison with (57.17), we obtain an expression for the *optical* path length difference (the optical path being defined as the product of the refractive index $n$ and the geometrical distance $r$):

$$\Delta = n_2 r_2 - n_1 r_1 \tag{61.12}$$

3. To find the distance between interference maxima (or minima), let us refer to Fig. 61.5. Here, $L$ is the distance from the sources to the screen, $d$ is the separation between the sources, $y_m$ is the distance from the centre of the interference pattern to the maximum of order $m$, and $\theta$ is the angle of observation.

It is obvious that at point $O$ is the zero-order maximum because the waves from both sources arrive at this point in phase. At other points on the screen, a maximum will occur if the condition expressed by (57.20) is satisfied, that is, $\Delta = 2m\lambda/2$, where $m$ is the order of the maximum. At $L \gg d$ and at small angles of observation, we have $\Delta = d \sin\theta$ and $y_m = L \tan\theta \approx L \sin\theta$, and, as a consequence, $\Delta/y_m \approx d/L$. Substituting the expression for the path length difference, we get

$$y_m = m\lambda L/d \tag{61.13}$$

Equation (61.13) expresses the distance from the zero-order (principal) maximum to the $m$th-order maximum. The distance between two adjacent maxima is

$$b = y_{m+1} - y_m = \lambda L/d \tag{61.14}$$

All quantities entering equation (61.14), except the wavelength, can be measured directly. Then, using Young's experimental set-up, one can measure the wavelength of light. In one of the experiments the following figures were obtained: $L = 3$ m, $d = 1$ mm, and $b = 2.1$ mm. The source emitted red light. The wavelength was found to be

$$\lambda = bd/L = 1/3 \times 2.1 \times 10^{-3} \times 10^{-3} = 7 \times 10^{-7}\ \text{m} = 700\ \text{nm}$$

4. It follows from equation (61.13) that the position of an interference maximum depends on wavelength. This implies that interference is accompanied by the dispersion of a non-monochromatic (non-sinusoidal) light wave into sinusoidal components (see Ch. 50). Thus if the light source in Young's arrangement emits white light, then only the zero-order (principal) maximum will be white; the remaining maxima will be tinted. Since the orange-red region of the spectrum has longer wavelengths ($\lambda_r \approx 780$ to 600 nm), and the blue-violet region has shorter wavelengths ($\lambda_v \approx 480$ to 420 nm), then, according to equation (61.13), these waves must be diffracted through different angles.

Finally, it should be noted that the angular distribution of maxima and minima in an interference pattern is independent of light intensity or of the illumination of the screen. If, in the same experimental set-up, we obtain an interference pattern on a photographic plate, using a bright light source and a short exposure time, then another interference pattern at a reduced light intensity and with the exposure time increased in proportion, the two photographic records will be identical. This property of interference will be taken up again in Sec. 68.7 in connection with the quantum-mechanical properties of light.

### 61.8. THE MICHELSON INTERFEROMETER

1. The diagram of the Michelson interferometer is shown in Fig. 61.6. Here, $S$ is a highly-monochromatic light source; $P_1$ and $P_2$ are two highly polished glass plates of the same thickness, the plate $P_1$ has a thin deposit of silver applied so that it operates as a beam splitter by allowing half of the original beam to pass straight through

and by reflecting the other half (the so-called half-silvered mirror); $M_1$ and $M_2$ are two metal mirrors whose position can be adjusted by micrometer screws $W_1$ and $W_2$; and $T$ is a telescope. The original beam of light from the light source $S$ is allowed to strike the plate $P_1$ which splits it into two beams. One beam is thrown upon the mirror $M_1$, is reflected back, passes again through the plates $P_2$ and $P_1$, and reaches the telescope. The other beam, on passing through the beam splitter, reaches the mirror $M_2$, is reflected back, goes again through the plate $P_1$, and, on being reflected from the beam splitter, enters the telescope. Since the two beams split up from the original wave train by the half-silvered mirror are coherent, the observer looking into the telescope will see an interference pattern appearing as a series of well-defined, alternating dark and bright bands ("fringes").

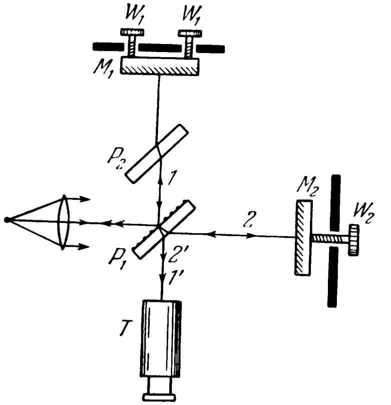


Fig. 61.6

2. The path length difference is twice the difference in distance from the centre of mirror $P_1$ to mirrors $M_1$ and $M_2$, that is, $\Delta = 2(l_1 - l_2)$. As is seen, the Michelson interferometer uses a large path length difference, which calls for a highly monochromatic light source. If the mirror $M_1$ is moved a quarter-wavelength, the path length difference will change by a half-wavelength, and a maximum in the interference pattern will take the place of a minimum, and vice versa. This displacement of interference fringes can clearly be seen by the observer. Practically, in a well-adjusted interferometer the observer can note a shift in an interference maximum by 0.1 the distance between adjacent bands, which corresponds to a displacement of the mirror $M_1$ for $l = \lambda/20 \approx 500\ \text{nm}/20 = 25\ \text{nm}$.

Thus, the Michelson interferometer can be used for the precision measurements of length to an accuracy of 20 to 30 nm. It is noteworthy that the standard of length, the metre, is at present measured by interference methods accurate to the ninth decimal place. According to the definition adopted in the International System of Units (SI), one metre is equal to 1 650 763.73 wavelengths of orange radiation from the crypton isotope of atomic mass 86 in a vacuum, which corresponds to a transition between the $2p_{10}$ and $5d_5$ levels.

3. In 1881, Michelson, and in 1887 Michelson and Morley carried out an experiment to determine the difference in the velocity of light parallel to and across the Earth's orbit.

Essentially, the experiment consisted in that one arm of an interferometer was aligned parallel to, and the other across the Earth's orbit, and the movable mirror $M_1$ of the interferometer was adjusted to set the interference pattern to zero. Then the interferometer was turned through 90° so as to interchange the orientation of the two arms. Like all physicists of his time, Michelson believed that the classical law of velocity addition (see Sec. 2.5) was applicable to light and that the velocity of light was different parallel to and across the Earth's orbit. If this were so, the interference pattern could only be set to zero by making the interferometer arms different in length, and turning the whole apparatus through 90° would be accompanied by a shift in the interference fringes.

However, no difference whatever could be detected between the observations in the two positions. A further refinement in the experiment (by Kennedy and Thorndike in 1932) showed that if the classical law of velocity addition were applicable to light, the shift of interference fringes would be observed already at a velocity of $v = 2$ km/s (the Earth's orbital velocity is 30 km/s). But no displacement of interference fringes was observed. The negative result of Michelson's and all other subsequent experiments led ultimately to the recognition of the idea that the velocity of light was the same in all inertial reference systems. As already noted (see Ch. 12), this idea together with the principle of relativity was incorporated in the foundation of the relativity theory by Einstein.

## 61.9. APPLICATION OF OPTICAL INTERFERENCE

1. Apart from the precise measurement of small distances mentioned in the previous section, interference phenomena are widely utilized in other fields of science and technology.

Above all, an interferometer can be used to test ground and polished surfaces for the quality of finish. One such instrument, known as the *Linnik interference microscope*, is a combination of a small-size interferometer and a microscope. The interferometer does not practically differ from the Michelson instrument, except that one of the mirrors is replaced by the surface being inspected. Should there be a scratch or an indentation on the surface, the observer will immediately see a pattern of interference bands or fringes in the microscope (Fig. 61.7). An idea about the size of the scratch can be obtained from the manner in which the interference bands are distorted.

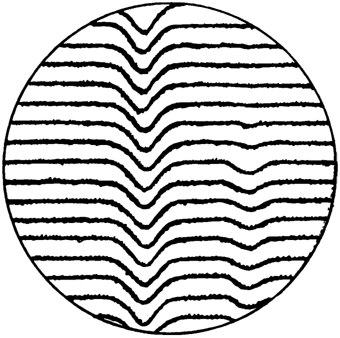


Fig. 61.7

2. A further application for interferometers is to measure the coefficients of linear expansion for solids, and also changes in the dimensions of ferro-magnetics in a magnetic field or ferro-electrics in an electric field (the magneto-striction and electro-striction effects). In all of these effects, the resultant change in dimension is extremely small and can be detected only by interference methods.

3. Very important applications for interference are the inspection of lenses and mirrors for surface finish; the measurement of refractive indexes for many materials, notably gases; the measurement of extremely small concentrations of substances in gases and liquids. In astronomy, an interferometer designed to be placed on a telescope measures the angular diameter of stars.

Unfortunately, because of the limited scope of the book, we cannot take up in detail the above listed and other uses for the interferometer.

# Chapter 62

# LIGHT DIFFRACTION

## 62.1. DIFFRACTION THROUGH A SINGLE APERTURE

1. Since light is a wave, it must be diffracted, that is, bend around the edges of opaque obstacles (see Secs. 57.8 and 57.9). However, the diffraction of light is difficult to observe, because a light wave has a negligibly short wavelength in comparison with the objects around which it bends.

The distance $l$ at which an observer should be, when the object has a size $D$ and the wavelength is $\lambda$, can be found by equation (57.28): $l \approx D^2/4\lambda$. Thus, a well-defined diffraction pattern through a hole with a diameter of $D \approx$ $\approx 2$ mm illuminated by green light with a wavelength of $\lambda = 500$ nm $=$ $= 5 \times 10^{-7}$ m will be seen at a distance of $l \approx 4 \times 10^{-6}/4 \times 5 \times$ $\times 10^{-7} = 2$ m. Of course, diffraction can also be caused by much larger obstacles but it can only be observed at very great distances from the obstacle.

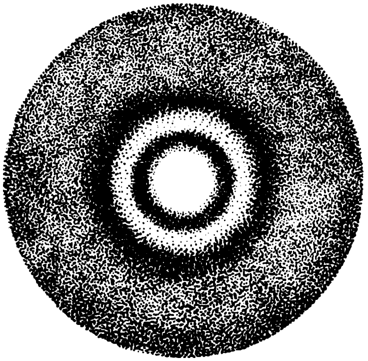


Fig. 62.1

2. Diffraction can be observed by an unaided eye, using a sheet of dense black paper in which either a fine slit is cut with a razor blade or a hole is pricked with a pin. The paper sheet should be placed

at 1.5 to 2 m from a bright lamp (100-200 W), and the observer's eye should be 0.5 to 1 m distant from the paper. The diffraction pattern formed by a round pinhole has the shape shown in Fig. 62.1.

The position of diffraction minima (dark rings) can be determined by equation (57.40): $\sin\theta = m\lambda/D$, where $\theta$ is the angle of observation and $D$ is the diameter of the pinhole. Although equation (57.40) has been derived for a rectangular slit, with a negligible error it may be applied to a pinhole.

## 62.2. DIFFRACTION GRATING

1. A *diffraction grating* is a series of $N$ very fine, closely spaced parallel slits made by ruling fine scratches with a diamond point on glass. Fig. 62.2 shows a side elevation of a part of a diffraction grating at high magnification. Sometimes, use is made of *reflecting gratings* likewise ruled with a fine diamond point on a polished metal. Reflecting gratings with a total length of $l \approx 150$ mm and with a total number of slits $N \approx 10^5$ (ruled with a density of $n = 600$ scratches/mm) are common. However, the ruling of a diffraction grating is a complicated and expensive process. As an alternative, it has been found possible to flow certain plastic solutions, such as gelatine, over an original grating. Then, after the evaporation of the solvent, a film can be removed which has all of the lines of the grating impressed upon it. The cost of a *replica grating*, as it is called, is minor compared to that of an original grating.

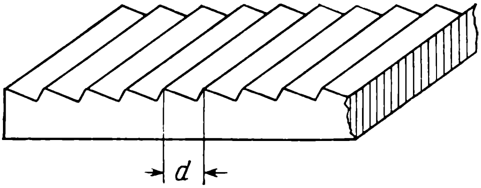


Fig. 62.2

A diffraction or reflecting grating is specified in terms of the total number of slits, $N$, the density of ruling, $n$ (the number of slits to the millimeter), and the width of a single grating ruling $d = 1/n$ (or the grating constant).

2. Since a diffraction grating is illuminated by a single wave front, its $N$ slits may be looked upon as the number $N$ of coherent sources. Therefore, we may apply to them the theory of interference from many identical sources discussed in Secs. 57.6 and 57.7.

According to (57.25), the light intensity is

$$I = i_0 \sin^2 N\beta / \sin^2 \beta \qquad (62.1)$$

where $\beta = \pi d \sin \theta / \lambda$ and $i_0$ is the light intensity through one slit. Maxima occur if the condition (57.31) is satisfied:

$$\sin \theta_{\mathrm{M}} = m\lambda / d \quad (m = 0, 1, 2, 3, \ldots) \qquad (62.2)$$

The light intensity at principal maxima is

$$I_{\mathrm{M}} = N^2 i_0 \qquad (62.3)$$

3. It follows from equation (62.2) that it is advantageous to make diffraction gratings with a small ruling width. Then one can readily obtain large angles for the diffracted light emerging from the grating and, as a consequence, a wide diffraction pattern.

For example, at $d = 1/600$ mm and $\lambda = 600$ nm $= 6 \times 10^{-4}$ mm, the first interference maximum occurs if $\sin \theta_1 = \lambda/d = 6 \times 10^{-4} \times 600 = 0.36$, that is, if the angle of the incident light with respect to the grating is $\theta_1 \approx 21°$; the secondary maximum at $\sin \theta_2 = 2\lambda/d = 0.72$, that is, at $\theta_2 = 46°$. Maxima of higher orders do not occur because at $m \geqslant 3$, $m\lambda/d > 1$.

4. It also follows from (62.2) that a diffraction grating separates white light into spectral components because it sends light of different wavelengths in different directions at any maximum, except the zero-order maximum. If a diffraction grating is illuminated with white light, only the zero-order maximum will be white in colour; the remaining maxima will show all colours of the rainbow. For example, in the problem given above, red light ($\lambda_r \approx 760$ nm) will be diffracted at the first maximum through an angle of $\theta_r = 27°$, and violet light ($\lambda_v = 400$ nm) through an angle of $\theta_v = 14°$.

A diffraction grating may be employed to measure the wavelength of light. For this purpose, we should only know the width of an individual grating ruling and measure the angle through which a ray of the light in question is diffracted. Then equation (62.2) will give the wavelength.

### 62.3. ANGULAR WIDTH OF PRINCIPAL MAXIMA

1. With a great number, $N$, of slits in a diffraction grating, the light intensity at principal maxima increases considerably, because, according to equation (62.3), it is proportional to $N^2$. However, any increase in energy at principal maxima should inevitably be accompanied by a decrease in energy in the remaining regions of the spectrum, because the total energy must remain unchanged. This is possible only if with an increase in $N$, principal maxima become narrower.

The solid angle subtended by a principal maximum between adjacent minima is called the *angular width*, $\gamma$, of a principal maximum. For simplicity, we choose the zero-order principal maximum ($m = 0$). Setting $n = 1$ in equation (57.32) we obtain an expression for the auxiliary angle defining the location of a minimum:

$$\beta_1 = \pi/N$$

This auxiliary angle corresponds to a solid angle

$$\theta = \gamma/2$$

Substituting it in equation (57.23) gives

$$\pi/N = \pi d \sin(\gamma/2)/\lambda$$

whence

$$\sin \gamma/2 = \lambda/Nd$$

2. As is seen, an increase in $N$ brings about a decrease in the angular width of the major maximum. At $Nd \gg \lambda$, which is always true for light, $\sin(\gamma/2) \approx \gamma/2$. Then

$$\gamma = 2\lambda/Nd \tag{62.4}$$

In the diffraction grating mentioned above for green light ($\lambda = 500$ nm), we have

$$\gamma = 2 \times 5 \times 10^{-4} \times 600 \div 10^5 = 6 \times 10^{-8} \text{ radians}$$
$$= 6 \times 10^{-6} \times 180 \times 3600 \div \pi \approx 1.2''$$

## 62.4. RESOLVING POWER OF A DIFFRACTION GRATING

1. As interference maxima become narrower, the diffraction grating gains in *resolving power*, that is, in the ability to resolve or separate two spectral lines of wavelengths $\lambda_1$ and $\lambda_2$.

In spectroscopy, an instrument is said to resolve, that is, separate. two spectral lines, if their images can be seen separately; if the two images merge together, the instrument is said to be unable to resolve them.

2. Fig. 62.3 shows the intensity distribution obtained by the superposition of two closely spaced spectral lines of wavelengths $\lambda_1$ and $\lambda_2$ when a beam of light is diffracted by a grating having a small total number of rulings. The width of the major maximum is much greater than the angular separation of the interference maxima in both lines. This means that the two lines cannot be seen separately; instead, they merge into a single wide band. If the same beam of light is allowed to pass through a grating having a greater total number of rulings, it will produce an interference pattern like that shown in Fig. 62.4. In this pattern, the maxima are narrow and

bright, and there is a marked gap between them, so that each spectral line can be seen or recorded separately.

Two lines are resolved, if their angular separation is at least half the line width, that is, $\theta_2 - \theta_1 \geqslant \gamma/2$. Or, stated differently, the principal maximum of one pattern falls on the first minimum of the other pattern (the *Rayleigh criterion of resolving power*).

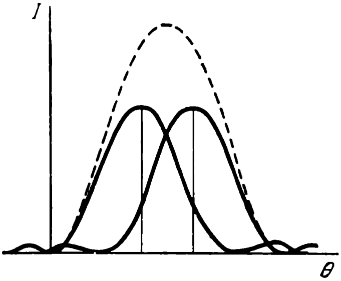


Fig. 62.3

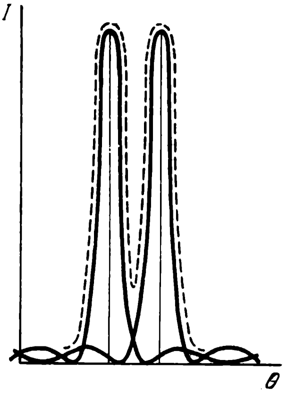


Fig. 62.4

It follows from equation (62.2) that

$$\sin\theta_2 = m\lambda_2/d$$

$$\sin\theta_1 = m\lambda_1/d$$

At small angles,

$$\theta_2 = m\lambda_2/d, \quad \theta_1 = m\lambda_1/d$$

Since, under the conditions of the problem, $\theta_2 - \theta_1 \approx \gamma/2$, then

$$\gamma/2 \approx m\Delta\lambda/d \tag{62.5}$$

However, according to (62.4), $\gamma = 2\lambda/Nd$, where $\lambda = \lambda_1 \approx \lambda_2$. Substituting it in (62.5) gives

$$\lambda/\Delta\lambda = mN \tag{62.6}$$

3. The quantity $\lambda/\Delta\lambda$, symbolized $A$, is called the *resolving power* of a spectral instrument. The higher the resolving power of an instrument, the closer together the spectral lines may be that it can separate.

For example, the spectrum of sodium vapours shows a bright yellow line with a wavelength of $\lambda = 589$ nm. After the advent of high-resolution spectral instruments (with $A \geqslant 1000$), it has been found that this is a double line (a doublet) with $\lambda_1 = 5890$ Å and $\lambda_2 = 5896$ Å.

The diffraction grating mentioned in Sec. 62.2 has a very high resolving power. It has $N = 10^5$, and it is capable of displaying spectra of the first and second orders ($m \leqslant 2$). Therefore, its maximum resolving power is $A = mN = 2 \times 10^5$. In other words, within the green region of the spectrum ($\lambda = 5000$ Å), the grating can resolve two lines with a wavelength difference of $\Delta\lambda = \lambda/A \approx 0.025$ Å.

### 62.5. DIFFRACTION OF X-RAYS

1. In Chapters 32 and 33, we examined the structure of crystals in detail. It was shown (see Sec. 32.3) that the structure of a crystalline lattice can be revealed by X-ray analysis. Now we shall discuss in detail the mechanism by which X-rays are diffracted by crystals.

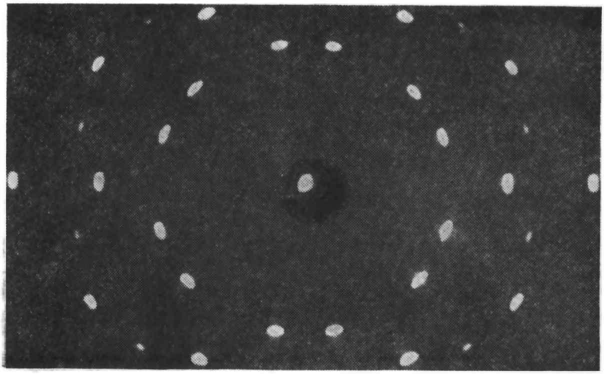


Fig. 62.5

In 1895, Roentgen discovered a new kind of rays he called X-rays. Their origin remained unclear for a long time. In 1897, Stokes, came out with a hypothesis that X-rays were short electromagnetic waves produced when electrons were suddenly retarded in a material (*bremsstrahlung*), which was in agreement with Maxwell's theory (see Sec. 59.4). However, all attempts to observe regular reflection, refraction or diffraction of X-rays failed.

2. In 1912 von Laue and his co-workers discovered the diffraction of X-rays by passing a narrow beam of X-rays through a single crystal which acts as a three-dimensional grating. The interference maxima due to the diffraction of X-rays on the lattice sides could be seen on a fluorescent screen or photographed (Fig. 62.5).

From the lattice constant (about 1 Å) and the angles of diffraction (about 10° to 20°), the experimenters found, using equation (62.2), the wavelength of X-rays to be about 1 Å. This explained why all the early experiments to observe the reflection, refraction or diffraction of X-rays had failed—the grating used had been too coarse.

## 62.6. DIFFRACTION BY A CRYSTAL LATTICE

1. Diffraction by the three-dimensional grating formed by the crystal lattice somewhat differs from diffraction in the plane ruled grating examined in Sec. 62.2. In a plane ruled grating, interference maxima occur on any wavelength, the relationship between the wavelength $\lambda$ and the diffraction angle (the *Bragg angle*) $\theta$ being: $\sin\theta = m\lambda/d$, as given by equation (62.2). In a crystal lattice, maxima can only be observed for some of the wavelengths properly related to the lattice constant.

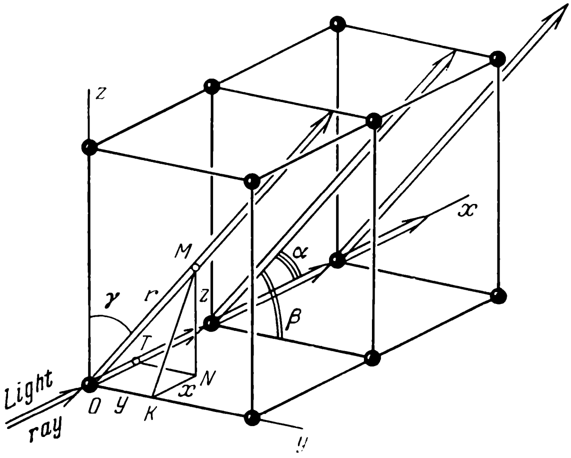


Fig. 62.6

For simplicity, consider diffraction by a simple cubic crystal (Fig. 62.6). Entering the crystal, the wave is scattered by atoms or other particles occupying the sites of the crystal lattice. The secondary waves produced by scattering are coherent and can therefore interfere.

2. Imagine that a wave is incident upon a crystal at right angles to the $yz$ face. Consider the secondary waves scattered in the direction shown in Fig. 62.6. Here, the rays make an angle $\alpha$ with the $x$-axis, an angle $\beta$ with the $y$-axis, and an angle $\gamma$ with the $z$-axis. Let us find

the condition under which the waves scattered in a given direction produce an interference maximum.

The sites arranged along the $y$-axis make up a linear grating with a grating space $d$. Maxima will occur on satisfying the condition (62.2) which we shall somewhat modify by replacing the angle $\theta$ between the ray and the normal to the grating by the angle $\beta$ between the ray and the $y$-axis. However, $\beta = (\pi/2) - \theta$, and so, $\sin\theta = \cos\beta$, and the condition (62.2) takes the form

$$d\cos\beta = m_1\lambda \quad (m_1 = 0,\ 1,\ 2, \quad .) \tag{62.7}$$

Similarly, for the sites arranged along the $z$-axis:

$$d\cos\gamma = m_2\lambda \quad (m_2 = 0,\ 1,\ 2. \quad .) \tag{62.8}$$

The sites arranged along the $x$-axis make up a system of point sources like the one discussed in Sec. 57.6. It is seen from Fig. 62.6 that the path length difference is $\Delta = d - d\cos\alpha = d\,(1 - \cos\alpha)$. Then, the condition for an interference maximum takes the form

$$\Delta = d\,(1 - \cos\alpha) = m_3\lambda \quad (m_3 = 0,\ 1,\ 2,\ \ldots) \tag{62.9}$$

3. The above three equations are not always simultaneously solvable, because one more condition is imposed upon them. As is seen from Fig. 62.6, the projections of the radius vector $\mathbf{r}$ on the coordinate axes are

$$OT = x = r\cos\alpha; \qquad OK = y = r\cos\beta; \qquad MN = z = r\cos\gamma \tag{62.10}$$

Using the Pythagorean theorem, it is an easy matter to show that $x^2 + y^2 + z^2 = r^2$. Noting (62.10), we get

$$\cos^2\alpha + \cos^2\beta + \cos^2\gamma = 1 \tag{62.11}$$

4. Substituting $\cos\beta = m_1\lambda/d$, $\cos\gamma = m_2\lambda/d$ and $\cos\alpha = 1 - (m_3\lambda/d)$ in equation (62.11), we get after simple manipulation

$$\lambda = 2m_3 d/\,(m_1^2 + m_2^2 + m_3^2) \tag{62.12}$$

Thus, interference maxima will only be produced by the wavelengths satisfying the condition (62.12); all other waves will simply be scattered in all directions. It is only at the central (zero-order) maximum satisfying the condition $m_1 = m_2 = m_3 = 0$ that we can encounter all waves, irrespective of their wavelengths.

5. This is, in effect, the Laue method of X-ray analysis of the crystal structure. In the Laue method, a beam of X-rays of all wavelengths is passed through a thin slice of the crystal, and the wavelength satisfying condition (62.12) produces an interference pattern as a series of spots, which is received on a photographic plate. With the lattice constant known, it is an easy matter to determine the

wavelength. The same method may be employed to obtain a monochromatic beam of X-rays, that is, to derive a beam falling within a narrow band of wavelengths from a broad spectrum.

### 62.7. X-RAY ANALYSIS OF CRYSTAL STRUCTURE

1. At present, the Laue method is only of historical interest. For practical purpose, use is made of either the Debye-Scherrer-Hull method or the rotating crystal method.

In the Debye-Scherrer-Hull method proposed in 1926, a beam of X-rays is directed on a powder sample of the material (hence it is also called the powder method) and the diffracted beams are received on a photographic plate. This method provides a convenient tool for the investigation of polycrystalline materials (see Sec. 32.2). This is extremely valuable, because to grow a single crystal of a substantial size is a complicated procedure, and very often the desired single crystal cannot be obtained at all.

The rotating-crystal method is only applicable to single crystals, which is a major disadvantage of the method. Yet, it reveals a more detailed picture of the crystal lattice than the Debye-Scherrer-Hull method.

2. Let a monochromatic beam of X-rays of wavelength $\lambda$ be indicent on a crystal so that the beam makes an angle $\theta$ with the crystal plane (Fig. 62.7). If the wavelength satisfies the condition (62.12), interference maxima will be produced in some directions. As is seen from Fig. 62.7, the path length difference between the interfering reflected beams is given

$$\Delta = 2d \sin \theta$$

The condition for an interference maximum is

$$\Delta = 2d \sin \theta = m\lambda \tag{62.13}$$

This condition was established in 1913 by Bragg and is known as the *Bragg equation* or the *Bragg law*. At about the same time and independently, this law was also established by Wulf of Moscow University.

3. The arrangement used in the rotating-crystal method is shown in Fig. 62.8. A single crystal of the material under investigation is mounted on a stage which is caused to oscillate at a slow rate by a clockwork. A fine monochromatic beam of X-rays is allowed to pass through stops $D_1$ and $D_2$ to fall on the crystal and be diffracted. Interference maxima are received on a photographic plate, $P$.

As the crystal is rotated, the condition for an interference maximum is upset, and the intensity of the scattered beam is decreased. However, the wave with the reduced intensity will be received at another point on the photographic plate, which is shown by the dotted line in Fig. 62.8. Thus, using the rotating-crystal method, one can determine not only the spacings of the atomic planes (the lattice constant) of a crystal, but also, on the basis of the intensity distribution, the size of the particles occupying the lattice sites.

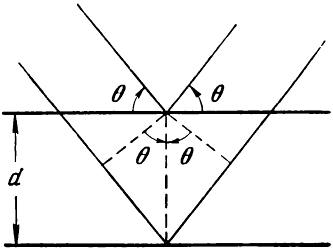


Fig. 62.7

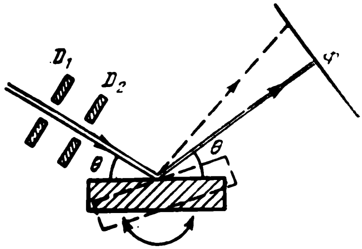


Fig. 62.8

4. In order to obtain an X-ray powder photograph, a narrow beam of monochromatic X-rays is directed onto a bar compressed from the fine-grained powder to which the material under investigation has

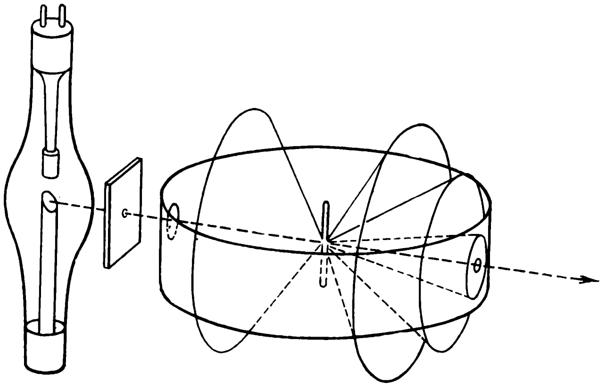


Fig. 62.9

been crushed (Fig. 62.9). In a fine-grained sample, there are many crystallites oriented at random, but with the correct orientation for each of the *d*-spacings. Consequently, they will satisfy the Bragg condition, and will produce interference maxima. The first-order maximum will be 2θ distant from the zero-order maximum, the second-order maximum 4θ distant, etc. Since there are no preferred directions in a fine-grained powder specimen, the condition $2d \sin \theta = m\lambda$ will be satisfied by all rays along the generators of cones of openings $4\theta$, $8\theta$, etc, eak h $d$-spacing producing a separate concentric cone. These cones appear on the film as a series on concentric circles with a zero-order maximum at the centre (Fig. 62.9).

An X-ray powder photograph reveals the structure of not only polycrystalline substances, but also high-molecular-weight compounds, namely fibres and giant molecules (of the protein type), and changes in the structure of a material accompanying phase transitions (see Ch. 36).

## 62.8. SCATTERING OF LIGHT

1. In equation (62.12), the $d$-spacing of a crystal lattice should be greater than the wavelength, that is, $d > \lambda$. If the $d$-spacing is less than the wavelength, only a zero-order interference maximum can be produced. This is so because equality (62.12) at $d < \lambda$ can only be satisfied, if $m_1 = m_2 = m_3 = 0$. Referring to expressions (62.7-62.9), it is seen that from $m_1 = m_2 = m_3 = 0$ it follows that $\beta = \gamma = \pi/2$, $1 - \cos\alpha = 0$, that is, $\alpha = 0$. This implies that a wave incident upon a crystal at right angles to its $yz$-face will be propagated along the $x$-axis without being scattered off-axis.

In such a case, the crystal behaves like a homogeneous medium free from scattering centres. Thus, the inequality $d < \lambda$ is a condition for the optical homogeneity of a crystal with a spacing $d$ and a wavelength $\lambda$.

It should be stressed once more that one and the same crystal may be optically homogeneous to some wavelengths and inhomogeneous to others. The point is that the $d$-spacings in a crystal are a few Ångstrom units. Therefore, for visible light with wavelengths of about 4000 to 7000 Å, the crystal will be an optically homogeneous medium. For X-rays with wavelengths of about 1 Å to 0.1 Å, the same crystal will act as a three-dimensional grating with well-defined periodic discontinuities.

2. Thus, for visible light an ideal crystal is perfectly homogeneous optically, and light must not be scattered in it. Any real crystal, however, has a number of defects (see Sec. 32.4) which disturb its homogeneity. Waves scattered by these defects are non-coherent and travel in all directions. A theory of light scattered by optical defects was advanced in 1907 by Academician Mandelshtam.

In materials optical discontinuities can also be produced by fluctuations in density (see Secs. 28.10 and 28.11), this being true not only of crystals, but also of liquids and gases. This is known as *molecular scattering*. The theory of molecular scattering was developed by Einstein and Smolukhovsky in 1908-1910.

Furthermore, light may be scattered by microscopic inclusions measuring up to $0.1\lambda$ across, provided they are spaced so that $d > \lambda$.

These may be smoke particles, small droplets of fat in water (milk), droplets of moisture in air (fog), etc. A medium containing these discontinuities is called *turbid.*

3. Experience shows that short-wave radiation (violet and blue light) is scattered more, and long-wave radiation (orange and red light) is scattered less. This difference was explained in 1899 by Rayleigh. According to him, the electromagnetic wave incident upon scattering centres induces in them forced vibrations at the frequency of the wave. Then, a scattering centre may be treated as a miniature oscillating dipole radiating secondary waves whose intensity, according to (59.16), is proportional to the fourth power of frequency, that is, inversely proportional to the fourth power of wavelength (the *Rayleigh law of scattering*):

$$i_{scat} \sim \nu^4 \sim 1/\lambda^4 \tag{62.14}$$

This explains why the sky is blue and the sunrise is red. When the Sun is high above the horizon, what we can see, looking away from it, is not direct sunlight, but the light scattered by fluctuations in the density of air, in which, according to Rayleigh's law, the short-wave colours of the spectrum are predominant (blue and violet). In contrast, we see direct sunlight at the sunrise or sunset. Since its short-wave component is scattered, the transmitted light contains a greater proportion of long waves, that is, orange and red light. Incidentally, because the Moon has no atmosphere, its sky appears pitch-dark and shows no glow at either the sunrise or sunset.

4. Rayleigh's law explains why crystals do not practically scatter sonic or ultrasonic waves, but strongly scatter hypersonic waves associated with the propagation of heat (see Sec. 45.3). Sound has wavelengths in excess of 10 cm, and ultrasound in excess of 1 mm, which is much greater than the *d*-spacing of the crystal lattice and also the spacing between crystal defects. Therefore, for these wavelengths a crystal is a homogeneous medium. It is only hypersound with wavelengths of the order of 10 to 100 Å that is scattered strongly.

# Chapter 63
# DISPERSION AND ABSORPTION

## 63.1. REFRACTIVE INDEX FOR LIGHT

1. In deriving the laws of reflection and refraction of waves (see Sec. 56.5), we based ourselves on the general wave equation, (56.8), which also holds for electromagnetic waves. Conversely, the laws of reflection and refraction of waves expressed by equations (56.25)

and (56.26) are applicable to light. Experience, too, bears out the validity of this reasoning.

Let a light wave be incident from a vacuum upon a dielectric as shown in Fig. 63.1. The velocity of a wave in a vacuum is $c$, while in the dielectric it is $u < c$; then, according to (56.26), we have

$$\sin\alpha_0/\sin\alpha = c/u \qquad (63.1)$$

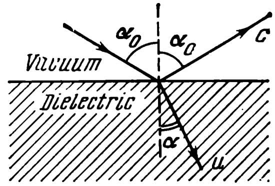


Fig. 63.1

The ratio of the sine of the angle of incidence, $\alpha_0$, to the sine of the angle of refraction, $\alpha$, as light passes from a vacuum into a material is called the *absolute refractive index* of that material (see Sec. 61.7). Sometimes, it is simply called the refractive index. Yet it is important to remember that the absolute refractive index is meant. So,

$$\sin\alpha_0/\sin\alpha = n \qquad (63.2)$$

Equation (63.2) is called the *Snell law* (or *Snell's law*), although Snell formulated it in terms of cosecants and not in terms of sines. In the form expressed by equation (63.2), the law appears to be formulated in 1630 by Descartes independently of Snell whose work had not yet been published at that time.

2. Comparing expressions (63.1) and (63.2) gives

$$n = c/u \qquad (63.3)$$

Thus, the refractive index of a material is equal to the ratio of the velocity of light in a vacuum to the velocity of light in the material.

Noting that, according to (59.2), the velocity of light in a material is $u = c/\sqrt{\varepsilon}$, we have

$$n = \sqrt{\varepsilon} \qquad (63.4)$$

Now we have expressed the refractive index of a material in terms of its permittivity (Maxwell's equation).

3. The fact that the velocity of light in an optically homogeneous material is lower than it is in a vacuum is explained as follows. On entering a material, an electromagnetic wave sets electrons into forced vibrations, and the electrons radiate secondary waves. Since in an optically homogeneous medium the spacing between particles is shorter than the length of the light wave, the primary and secondary waves are coherent. They interfere, and the wave velocity changes according to the frequency of the primary wave.

The interference pattern in this case is extremely involved to calculate, and we shall therefore examine a simpler, although a somewhat formal, theory of dispersion in Sec. 63.5.

## 63.2. COEFFICIENTS OF REFLECTION AND TRANSMISSION

1. Let us see how the energy is distributed between the reflected and refracted waves. For simplicity, we shall limit ourselves to the normal incidence of a wave at the interface of two media with refractive indexes $n_1$ and $n_2$. Let the two media be dielectrics.

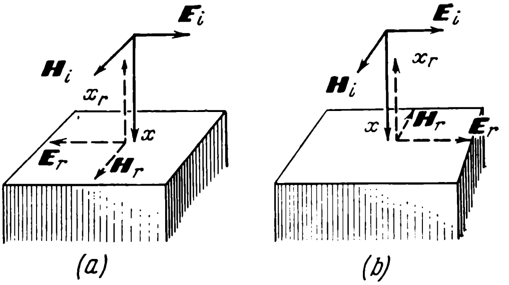


Fig. 63.2

Since the field vectors **E** and **H** are parallel to the interface (Fig. 63.2), they are continuous at the interface. The continuity equation can be written as follows

$$E_i \pm E_r = E_{refr}, \quad H_i \mp H_r = H_{refr} \qquad (63.5)$$

In other words, the algebraic sum of the intensity vectors of the incident wave, $E_i$ and of the reflected wave, $E_r$, is equal to the field intensity of the refracted (transmitted) wave, $E_{refr}$. The same holds for the magnetic component.

2. In the reflected wave, one of the field vectors changes sign. That is, this component of the reflected wave undergoes a phase reversal. The phase reversal occurs because the field vectors **E** and **H** bear a right-hand-screw relationship to the direction of wave propagation (in Fig. 59.1, this is the $x$-axis); now that, on reflection, the wave travels back on itself, that is, changes sign, one of the field vectors should change sign, too. This may be either **E** (Fig. 63.2*a*) or **H** (Fig. 63.2*b*). Since from (59.4) it follows that $H = \sqrt{\varepsilon\varepsilon_0/\mu\mu_0}E$, we may rewrite equation (63.5) as follows

$$E_i + E_r = E_{refr}, \quad (E_i - E_r)\sqrt{\varepsilon_1\varepsilon_0/\mu_1\mu_0} = E_{refr}\sqrt{\varepsilon_2\varepsilon_0/\mu_2\mu_0} \qquad (63.6)$$

3. For dielectrics, $\mu_1 = \mu_2 = 1$, $n_1 = \sqrt{\varepsilon_1}$ and $n_2 = \sqrt{\varepsilon_2}$. Solving equations (63.6) simultaneously for the above conditions, we obtain

$$E_r = E_i(n_2 - n_1)/(n_2 + n_1), \quad H_r = -H_i(n_2 - n_1)/(n_2 + n_1) \quad (63.7)$$

According to (59.8), the reflected wave intensity, $I_r = \overline{E_r H_r}$, may be expressed in terms of the incident wave intensity, $I_i = \overline{E_i H_i}$, as follows

$$I_r = I_i\,[(n_2 - n_1)/(n_2 + n_1)]^2 \tag{63.8}$$

The transmitted (refracted) wave intensity at normal incidence is

$$I_r = I_i - I_r = I_i\,[4n_2 n_1/(n_2 + n_1)^2] \tag{63.9}$$

4. By analogy with Sec. 56.6, we define the coefficients of refraction and transmission as:

$$R = I_r/I_i = \left(\frac{n_2 - n_1}{n_2 + n_1}\right)^2; \quad T = I_{refr}/I_i = 1 - R = 4n_2 n_1/(n_2 + n_1)^2 \tag{63.10}$$

These coefficients are fully analogous to those derived in Sec. 56.6 for elastic waves.

### 63.3. DISPERSION

When he derived equation (63.4), Maxwell could not verify it experimentally because the permittivities of most materials were either unknown altogether or had been measured with gross errors. Meanwhile, the verification of this relationship would prove the validity of Maxwell's theory as forcefully as experiments to measure light pressure (see Sec. 59.3). The first investigator to measure permittivities on a systematic basis was Boltzmann in 1872-1874. Interesting results were obtained by Stoletov's followers, namely Schiller in 1874 and Zilov in 1875.

2. For comparison, the permittivities and refractive indexes of some materials are listed in Table 63.1. The frequencies at which the permittivity was measured are given for water and ice. All refractive indexes are given for the yellow region of the spectrum (sodium line, $\lambda_d = 5896$ Å or helium line, $\lambda_d = 5876$ Å; at $\nu = 5.1 \times 10^{14}$ Hz). Where the frequency is not given, the results apply to practically all frequencies.

3. As is seen, for air (which exemplifies gases), the relationship $n^2 = \varepsilon$ is satisfied very well. The same is true of liquids and solids with covalent bonds between atoms (benzene, carbon tetrachloride, paraffin, diamond, etc.).

For materials with an ion lattice (quartz, salt, glass) the relationship $n^2 = \varepsilon$ is satisfied poorly. The values of $n^2$ and $\varepsilon$ are entirely different for materials with polar molecules (water and ice). However, this in no way disproves Maxwell's theory. The point is that the values of $n^2$ and $\varepsilon$ have been measured at different frequencies and they may well not check in, a fact stressed by Maxwell himself.

*Table 63.1*

| Material | $n$ | $n^2$ | $\varepsilon$ | $\nu$ |
|---|---|---|---|---|
| Air | 1.000292 | 1.000584 | 1.000576 | — |
| Benzene | 1.50 | 2.25 | 2.2836 | — |
| Carbon tetrachloride | 1.46 | 2.13 | 2.24 | — |
| Paraffin (molten) | 1.48 | 2.19 | 2.2 | — |
| Diamond | 2.4195 | 5.855 | 5.68 | — |
| Fused quartz | 1.4587 | 2.129 | 3.8 | — |
| Rock salt | 1.5412 | 2.374 | 6.0 | — |
| Glass (borosilicate crown) | 1.5097 | 2.280 | 5.3 | — |
| Glass (extra dense flint) | 1.7004 | 2.890 | 6.9 | — |
| Ice (—5 °C) | 1.31 | 1.716 | 75 | 0 |
| | | | 25 | 10 kHz |
| | | | 5 | 50 kHz |
| | | | 3 | 1 MHz |
| Water (20 °C) | 1.333 | 1.773 | 80 | up to 1 GHz |
| | | | 64 | 10 GHz |
| | | | 44 | 19 GHz |
| | | | 35 | 24 GHz |
| Water (vapour) | 1.000252 | 1.0005 | 1.0060 | 30 GHz |

4. Experience shows that both the permittivity and refractive index of a material depends on the frequency and, therefore, wavelength of light. The velocity of light in a material is likewise dependent on frequency, according to (63.3). The dependence of the velocity of a wave on its frequency is called *dispersion*. By the same token, the dependence of wave velocity on wavelength may be called *dispersion*, too

## 63.4. DISPERSION AND LIGHT SPECTRUM

1. The result of dispersion is that non-monochromatic (say, white) light is separated into a sequence of colours which are the sinusoidal components of the original light beam. The cause of this spreading process, also called *dispersion*, is that the various colours are refracted by slightly different amounts, which means that the index of refraction is slightly different for each colour. It follows, on the wave theory, that what we perceive as a colour is really a specific wavelength, or, which is the same, a specific frequency of light.

Let, for example, a beam of white light be incident upon the water surface from the air at an angle of $\alpha = 80°$. We seek to determine the angle of refraction for red light having a wavelength of $\lambda = 6708$ Å and a refractive index of $n_r = 1.33$, and for violet light with a wavelength of $\lambda = 4047$ Å and a refractive index of $n_v = 1.34$. By the law of refraction, equation (63.2), the angles of refraction are

$$\sin \alpha_r = \sin 80° \div 1.33 = 0.7405, \quad \alpha_r = 47°46',$$

$$\sin \alpha_v = \sin 80° \div 1.34 = 0.7350, \quad \alpha_v = 47°19'$$

As is seen, violet rays are refracted more than red light, because of which the original white light appears to be spread out into the component colours, or harmonics.

2. It appears that the ancients were already aware of the brilliant hues produced when sunlight passes through transparent gems and crystals. However, nobody before Newton had taken up the problem in earnest. In 1666, Newton, while still a university student, became concerned with the undesirable fringes of colour that surrounded the images in refracting telescopes, and he performed a simple experiment that revealed the true character of colour. Holding a triangular glass prism in the path of a narrow beam of sunlight, he found that the rays were fanned out into a band of colours after passing through. The sequence was the same as the one seen in the rainbow, with red at one end merging gradually into orange, then yellow, green, blue and violet. His further experiment was to see if the colour coming through the prism could be further broken up by a second prism. This did not happen. Thus Newton showed that "a colour of one kind" (that is, monochromatic waves) cannot be resolved into further components any longer.

Newton held that the various colours were refracted by different amounts because red light consisted of larger particles (corpuscles) while violet light consisted of smaller ones. Attracted by the material, the less massive violet particles were reflected more than the more massive red particles. Although in the light of present-day theory, dispersion is explained in an entirely different way, Newton's achievements in the field of optics, as a scientist who discovered and investigated a wide range of novel optical phenomena and sought to develop a consistent theory, are great.

### 63.5. ELECTRON THEORY OF DISPERSION

1. In a vacuum, all electromagnetic waves are propagated at the same velocity, $c$, irrespective of their frequency; dispersion only appears in a material. Thus, dispersion, like other properties of any material, should be explained on the basis of the structure of that material.

The dispersion theory based on the classical electron theory was developed by Lorentz at the end of the 19th and the beginning of the 20th century. The underlying principles of his theory are as follows.

2. The difference between $\varepsilon$ and $n^2$ becomes appreciable only in materials with polar molecules when $\varepsilon$ is measured on frequencies markedly different from those of light waves. If the refractive index of a material is measured in the optical region (which encompasses visible light and the adjacent infrared and ultraviolet regions of the spectrum), it will be about the same for materials with both polar and non-polar molecules (see Table 63.1). Hence, the orientational polarizability, characteristic of materials with polar molecules (see Sec. 38.6), occurs only in a static field and also in alternating fields with a relatively low frequency. In high-frequency fields, the molecular dipole has no time to turn, the degree of orientation of such dipoles decreases, and so does the permittivity. For ice, it decreases by a factor of 25 as the frequency changes from zero to 1 MHz.

Thus, in the optical region of the spectrum it is no longer important whether a molecule has a dipole moment or whether the particles of a crystal or liquid are held together by ionic or covalent bonds. Now the material is polarized only due to the deformation of the electron cloud (see Sec. 38.5).

3. In order to explain the dependence of the refractive index on frequency (that is, dispersion), let us examine the mechanism by which an atom or a molecule is polarized in the electromagnetic field of a light wave. Let the light vector (see Sec. 61.2) undergo vibrations such that

$$E = E_0 \cos \omega t \tag{63.11}$$

The electric force $F = eE = eE_0 \cos \omega t$ sets the electrons of the atoms into forced vibrations (see Sec. 53.1). As is shown in Sec. 61.2, the action of the magnetic component may be neglected. The forced vibrations of the electron cloud are described by the following equation (see Sec. 53.1):

$$l = A \cos \omega t = \frac{F_m}{m(\omega_0^2 - \omega^2)} \cos \omega t = \frac{eE_0}{m(\omega_0^2 - \omega^2)} \cos \omega t \tag{63.12}$$

Here, $\omega_0$ is the natural radian frequency of the electron cloud, $\omega$ is the radian frequency of the wave. The change in the reference phase from $\varphi = 0$ at $\omega < \omega_0$ to $\varphi = \pi$ at $\omega > \omega_0$ is taken care of by the sign of the amplitude.

4. The field-induced dipole moment of a molecule undergoes vibrations in a similar manner

$$p_e = el = \frac{e^2 E_0}{m(\omega_0^2 - \omega^2)} \cos \omega t = \frac{e^2}{m(\omega_0^2 - \omega^2)} E \tag{63.13}$$

Comparing this expression with equation (38.18), the polarizability of a molecule can be described as follows:

$$\alpha = p_e/\varepsilon_0 E = \frac{e^2}{m\varepsilon_0 (\omega_0^2 - \omega^2)} \tag{63.14}$$

According to (38.20), the permittivity $\varepsilon = n^2$ is

$$n^2 = \varepsilon = 1 + n_0\alpha = 1 + \frac{e^2 n_0}{m\varepsilon_0 (\omega_0^2 - \omega^2)} \tag{63.15}$$

This expression describes the *law of dispersion*. The concentration of molecules is designated by $n_0$ in order that it can be distinguished from the refractive index $n$.

5. With gases, the electric susceptibility $\varkappa = n_0\alpha$ is appreciably less than unity, and so the expression (63.15) may be simplified. Thus, $n^2 - 1 = \varkappa$ or $(n - 1)(n + 1) = \varkappa$. But $n + 1 \approx 2$, and so $n - 1 = \varkappa/2$, or

$$n = 1 + \varkappa/2 = 1 + \frac{e^2 n_0}{2m\varepsilon_0 (\omega_0^2 - \omega^2)} \tag{63.15'}$$

## 63.6. NORMAL AND ANOMALOUS DISPERSION

1. Expression (63.15) offers an explanation for dispersion. As is seen, the refractive index is a function of the radian frequency at which the electromagnetic field vibrates in a light wave.

In deriving the law of dispersion, we took into account only the effect of the applied electromagnetic field on the electron cloud of a molecule and ignored intermolecular interactions. Therefore, the above formulated law of dispersion is rigorously applicable only to gases. Qualitatively, however, equation (63.15) may also be applied to liquids and solids to obtain an insight into the mechanism (but not for calculations).

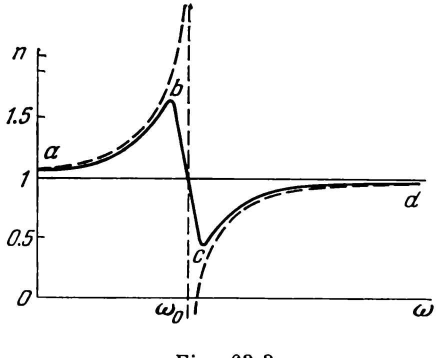


Fig. 63.3

2. In graphic form, the law of dispersion described by equation (63.15′) is shown in Fig. 63.3 by the dotted line. As is seen, when the wave frequency is lower than the natural frequency of the electron cloud ($\omega < \omega_0$), the refractive index exceeds unity. At $\omega \to \omega_0$, the refractive index tends to infinity. If, on the other hand, the wave frequency exceeds the natural frequency of the electron cloud ($\omega > \omega_0$), the refractive index is less than unity and at $\omega \to \omega_0$ it would become a negative quantity, which has no sense.

This forces us to conclude that expression (63.15) has sense only if the wave frequency markedly differs from the natural frequency of the electron cloud. When $\omega \approx \omega_0$, we obtain an absurd result. This is fully corroborated by the analysis carried out in Sec. 53.2, according to which expression (63.12) and its corollary (63.15) cannot be used at resonance. One should also take into account attenuation which, expressed in terms of the $Q$-factor, defines the amplitude of vibrations at resonance.

3. A more rigorous theory of dispersion, which will not be presented here, takes into account this attenuation and gives a correct relationship between the refractive index and frequency over the entire frequency range. A plot of this function appears in Fig. 63.3 as the solid line. This is the *dispersion curve.*

As is seen, in the regions *ab* and *cd*, where the wave frequency markedly differs from the resonant frequency, the refractive index rises with increasing frequency. These are the regions of *normal* dispersion. In the region *bc*, which is near the resonant frequency, the refractive index decreases with increasing frequency. This is the region of *anomalous* dispersion.

It is relevant to note that from the viewpoint of the electron theory there is nothing anomalous about it. The region *bc* is as normal as the regions *ab* and *cd.* Simply, until the mid-19th century, such regions had not been observed within the dispersion curve, and physicists since Newton's time had customarily believed that the refractive index should increase with increasing frequency (or with decreasing wavelength). Therefore when in 1862 Le Rous discovered that the refractive index of iodine vapours decreased with increasing frequency, he called this anomalous dispersion.

## 63.7. LIGHT ABSORPTION

1. In Sec. 55.4, we investigated the attenuation of waves and derived a law, the Bouguer law, which describes the attenuation of a plane wave, $I = I_0 e^{-\mu x}$ (see equations (55.16) and (55.17)). Since the Bouguer law is independent of the mechanism of wave propagation it is applicable to electromagnetic as well as elastic waves.

2. Experience shows that the absorption coefficient is a function of frequency. In other words, in addition to the dispersion of the refractive index, there is the dispersion of the absorption coefficient. Fig. 63.4 relates the absorption coefficient to wavelength, as $\lambda = 2\pi c/\omega$. For comparison, the dotted line respresents the dispersion curve for the refractive index versus wavelengths. The normal and anomalous dispersion regions are marked by the same letters as in Fig. 63.3.

As is seen, the absorption of light waves is a maximum in the anomalous dispersion region, *bc*. This phenomenon stems from the general property of forced vibrations, namely that the system absorbs the highest power at resonance (see Sec. 54.8). In greater details, the mechanism of light absorption by atoms and molecules will be discussed in chapters on quantum mechanics (Chapters 71 and 72).

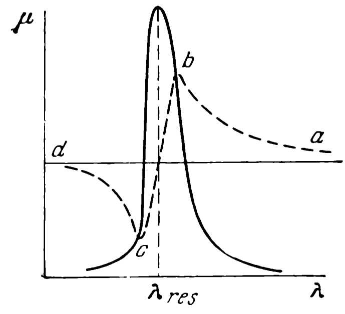


Fig. 63.4

3. Atoms and molecules oscillate at a set of natural frequencies, rather than at a single frequency (see Sec. 74.4). In the vicinity of each of these frequencies, there is a sudden increase in the absorption coefficient. By measuring the absorption coefficient, one can readily determine the natural frequencies of atoms, molecules, crystals, etc.

In solids and solutions the strong interaction between atoms or molecules broadens the anomalous dispersion regions into absorption bands between which lies a range of frequencies which are absorbed but little. This property is utilized in *light filters*. A light filter is a plate of glass containing some salts, or plastic films doped with certain organic dyes, or solutions of dyes in water, alcohol and other solvents. According to its chemical composition, a light filter transmits a certain range of frequencies and absorbs the remainder.

## 63.8. PHASE AND GROUP VELOCITIES

1. It has already been noted that light is a multitude of wave trains rather than a single sinusoidal wave; and it has been shown that in a first approximation a wave train may be visualized as a chopped beat (see Sec. 61.3). However, associated with each chopped beat, that is, with a wave train, is a range of frequencies (and a range of wave numbers). It is legitimate then to ask how one can define the velocity of a wave train in a medium.

We have learned that in a vacuum light waves are propagated at the velocity $c$, irrespective of the frequency, and this is also true of a wave train. In a medium, however, the two sinusoidal components of a wave train will differ in velocity because of dispersion, and so the definition of the velocity of a wave train is not immediately obvious.

2. To begin with, we shall define the velocity with which a point of constant phase is propagated in a progressive sinusoidal wave at a single frequency (an unmodulated wave) as the *phase velocity*, *u*. From equation (56.6) it follows that

$$u = \omega/k \qquad (63.16)$$

The velocity associated with a wave train (or a wave packet or a group of waves) and equal to the velocity at which energy or intelligence is propagated by means of a wave train will be called the *group velocity*, $U$.

To determine the group velocity, we shall turn to equations (61.2) and (61.3) which describe the manner in which a chopped beat or, which is the same, a wave train is propagated. From equation (61.3) it follows that the *beat amplitude* is propagated as a wave which is called a *beat amplitude wave*

$$B = 2E_0 \cos(\Delta\omega t - \Delta kx) \tag{63.17}$$

The velocity at which a beat amplitude wave is propagated is the group velocity. By analogy with (63.16), the expression for the group velocity is

$$U = \Delta\omega/\Delta k \tag{63.18}$$

Since energy is proportional to the square of the amplitude, the velocity at which energy is propagated is equal to that at which a beat amplitude wave is propagated, that is, to the group velocity.

3. Since the phase velocity, $u = c/n$, (63.3), it follows that in the high-frequency range where $n < 1$ (see Fig. 63.3), this velocity appears to exceed the velocity of light in a vacuum. At first sight, it may appear that signals could exist for which $u > c$, which runs counter to the basic postulate of the theory of relativity stating that nothing can move faster than light in a vacuum (see Sec. 12.6).

However, there is no contradiction whatsoever, for *any signal*, that is *an intelligence-carrying wave*, is a *modulated wave* (Secs. 50.2 and 60.2), and the envelope of an amplitude-modulated wave is propagated at the group, not phase velocity. Then expression (63.18) takes the form

$$U = \frac{c}{n + \omega(\Delta n/\Delta\omega)} \tag{63.18$'$}$$

In the normal dispersion region, $\Delta n/\Delta\omega > 0$, which is seen from Fig. 63.3. Here $n + \omega(\Delta n/\Delta\omega) > 1$, and so $U < c$. Thus, in the normal dispersion region, the group velocity is less than the velocity of light in a vacuum, which is in full agreement with the theory of relativity.

In the anomalous dispersion region, the concept of group velocity loses all sense. Yet, a rigorous theory shows that in this case, too, the signal velocity is less than the velocity of light in a vacuum.

### 63.9. MEASUREMENT OF THE VELOCITY OF LIGHT

1. In order to determine the velocity of light, it would suffice to know the time $\tau$ in which light covers a distance $l$ and to divide the distance by the time. Since, however, light is propagated at a fantastically

high velocity, one would have to measure either astronomic distances (which would involve high values of $\tau$) or extremely small time intervals (so that the distance $l$ can be scaled down to the Earth's proportions).

2. For the first time the fact that light is propagated through space at a finite, though a very large, velocity and not instantaneously, as previously supposed, was discovered in 1676 by Olaf Roemer. He made this discovery by observing that the eclipses of the satellites of Jupiter would show at a later time with the Earth receding from Jupiter or at an earlier time with the Earth approaching Jupiter. He accounted for this irregularity by assuming that the differences were due to the fact that the light took a longer or shorter time to come from Jupiter to the Earth as the distance between these two planets changed due to their individual rotations about the Sun.

Let us designate the sum of delays in the eclipses over the half year when the Earth is receding away from Jupiter as $t_1$. If the number $N$ of observations is made in the meantime and $T$ is the orbital period of the satellite, then $t_1 = NT + (D/c)$, where $D$ is the diameter of the Earth's orbit. The sum of all advances in the times of the eclipses with the Earth approaching Jupiter will be $t_2 = NT - (D/c)$. Subtracting gives

$$\tau = (t_1 - t_2)/2 = D/c \qquad (63.19)$$

Roemer found that $\tau = 1320$ s. From the diameter of the Earth's orbit, $D = 298.9 \times 10^6$ km, we obtain for the velocity of light:

$$c = D/\tau = 298.9 \times 10^6 \div 1320 = 227 \times 10^3 \text{ km/s}$$

In 1909, Samison found the time delay to be equal to $\tau = 997.6$ s, whence

$$c = D/\tau = 298.9 \times 10^6 \div 997.6 \approx 300 \times 10^3 \text{ km/s}.$$

3. Experimentally, the velocity of light may be determined either by the chopped-beam method or by the rotating mirror method. With the chopped-beam method (Fig. 63.5), light from a source $S$ is interuppted into pulses by a chopper $Ch$, travels a distance $L$ and, on being reflected from a mirror $M$, goes back to chopper. The reflected beam will be able to pass through the chopper only if the time during which the light pulse completes a round trip journey, $\tau = 2L/c$, is equal to the period, $T$, of the chopper, that is, if $T = \tau = 2L/c$. Then, on passing through the chopper, the light beam will be reflected from a semitransparent mirror $A$ into a telescope. Thus, the velocity of light is

$$c = 2L/\tau = 2L/T = 2L\nu \qquad (63.20)$$

where $\nu = 1/T$ is the chopper frequency.

In 1849, Fizeau measured the velocity of light by passing a beam through a rapidly rotated toothed wheel to a distant mirror and returning the light through the same mirror. This result, $c = 315 \times \times 10^3$ km/s, was an accomplishment for his time. At present, electronically controlled optic shutters are employed as choppers. This has markedly improved the accuracy of measurements. In 1950, Bergstrand, using a base of $L = 7$ km, has computed the velocity of light (in a vacuum) as

$$c = 299\,793.1 \pm 0.25 \text{ km/s}$$

4. A rotating mirror for experimental measurement of the velocity of light was for the time used by Foucault in 1862. With a mirror

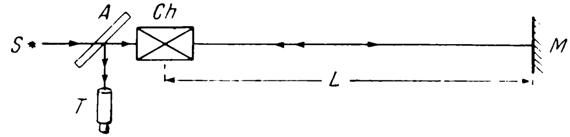


Fig. 63.5

rotating at a speed of 800 rps, the base $L$ could be reduced to 4 m. By passing light to a water-filled pipe, Foucault showed that in water light travelled more slowly than it did in air, which was in full agreement with the wave theory of light (see (63.3)).

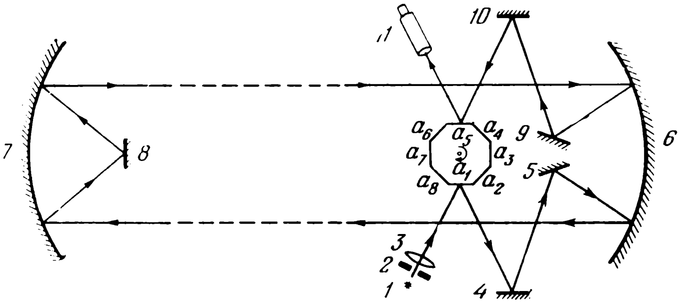


Fig. 63.6

Michelson further improved the rotating mirror method. His experimental arrangement is shown in Fig. 63.6. A beam of light from a source, *1*, passes through a stop, *2*, and a focusing system, *3*, to strike the face, $a_1$, of a rotating octahedral mirror. Reflected off the face $a_1$, the light beam is directed onto another mirror, *4*, and then, as is seen from the drawing, travels through a system of mirrors, *4* → *5* → *6* → *7* → *8* → *7* → *6* → *9* → *10*, until it strikes another face, $a_5$, of the octahedral mirror. Reflected off that face, the light beam enters a telescope, *11*.

The concave mirrors, *6* and *7*, were set up on two mountains separated by a distance of $L = 35\,373.21$ m. The octahedral mirror was spun at $\nu = 528$ rps. Over the time $\tau$ during which the light beam made a round trip between the mountains, the mirror rotated through 1/8 of a revolution, so that instead of the face $a_5$, it struck the face $a_6$, and the light beam could be seen in the telescope, *11*. The time occupied by one-eighth of a revolution was $\tau = 1/8\nu = 1 \div 8 \times 528$ s, and the distance travelled by the light beam was $l = 2L = 2 \times \times 35\,373.21$ m. This worked out to a velocity of light of $c = l/\tau = = 16L\nu$. Allowing for the experimental error, the figure that Michelson published for the velocity of light in 1926 was:

$$c = 299\,796 \pm 4 \text{ km/s}$$

5. In all experimental measurements of the velocity of light, the light signal is broken up into pulses. In the Roemer method, this is done by eclipses, in the Michelson method by a rotating mirror, etc. Thus, what we measure is the velocity of a wave packet, that is, a group velocity. The phase velocity of light through the same medium can be computed, having measured the refractive index $n$ and the dispersion $\Delta n/\Delta\omega$ by an interferometer (see Secs. 61.8 and 61.9). From the phase velocity and refractive index, it is an easy matter to find the velocity of light in a vacuum.

6. A new method for measuring the velocity of light appeared with the advent of lasers (see Sec. 79.4). Its idea is very simple: the wavelength and the associated frequency are measured separately, and the velocity of light is then found by equation (56.3): $c = \lambda\nu$.

The wavelength can conveniently be measured in the optical region by means of an interferometer (see Sec. 61.8), while the frequency can accurately be measured in the radio-frequency region, using the frequency of a cesium laser (with a wavelength of 3.27 cm) as the frequency standard. The optical frequency of interest is compared with this standard by means of non-linear converters similar to the detector discussed in Sec. 60.5.

As measured by this method in 1972, the published figure for the velocity of light in a vacuum was

$$c = 299\,792\,456.2 \pm 1.1 \text{ m/s}$$

## Chapter 64

## POLARIZATION OF LIGHT

### 64.1. POLARIZED AND UNPOLARIZED LIGHT

1. In contrast to a longitudinal elastic wave in which the particles vibrate along the direction of propagation, that is, in the direction of the ray, in a transverse wave the particles vibrate at right angles

to the direction of propagation, that is, at right angles to the ray (Sec. 55.1). If we imagine a ray directed along, say, the $x$-axis, then in a longitudinal wave all directions in the $yz$-plane will be equal, while in a transverse wave there will be a preferential direction differing from other directions in physical properties, namely in that the light vector, that is, the electric intensity vector, vibrates along this direction (see Sec. 61.2).

The plane in which the electric vector **E** vibrates is called the *plane of vibration.* For historical reasons, the plane in which the magnetic vector **H** vibrates is called the *plane of polarization.* The plane of vibration of a wave from a vibrating charge or dipole (see Fig. 59.4) contains both the ray and the acceleration vector. By specifying the ray and the plane of vibration, we thereby automatically specify the plane of polarization.

2. An electromagnetic wave in which the electric intensity **E** is everywhere parallel to some plane and the magnetic intensity **H** is everywhere perpendicular to that plane is referred to as *linearly polarized* or *plane-polarized.*

3. Ordinary light is composed of radiation from large numbers of atoms or molecules. Since each atom or molecule acts individually in creating the light waves (see Sec. 61.2), the resulting radiation is of a random, statistical nature.

In Sec. 61.6, it is shown that the statistical mechanism of light emission causes the wave phase to change likewise at random, because of which light is usually non-coherent. Because of this randomness, each wave train has its own plane of vibration or, which is the same, its own plane of polarization, and these planes in a light wave as a collection of individual wave trains fluctuate irregularly with time.

Thus, ordinary light consists of many separate components, each polarized at some arbitrary angle, and the overall effect of such combinations shows none of the properties of a simple polarized wave. More specifically, the field vectors are vibrating in all directions at right angles to the ray rather than in a fixed direction. Light in which the electric (and, also, the magnetic) vector is vibrating in directions changing at random so that in a plane perpendicular to the ray they are equi-probable is called *unpolarized.*

## 64.2. ANALYZER. MALUS COSINE-SQUARED LAW

1. Imagine a rubber cord along which a transverse wave is propagated. Even if the cord were invisible, we still could locate the plane of vibration, using the arrangement shown in Fig. 64.1. When the plane of vibration coincides with the plane of the slot formed by parallel boards, the wave will be free to pass through the arrangement; if, on the other hand, the boards are turned through 90° the

vibrations will be suppressed, and the wave will not be able to travel through the slot.

Similarly, using a dipole aerial tuned to resonance with a wave, it is possible to locate the plane of vibration in a plane polarized

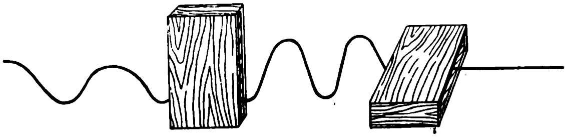


Fig. 64.1

electromagnetic wave. If the dipole is arranged so that it is in the plane of vibration (Fig. 64.2*a*), the lamp will illuminate. This is because the vibrations of the electric vector directed along the

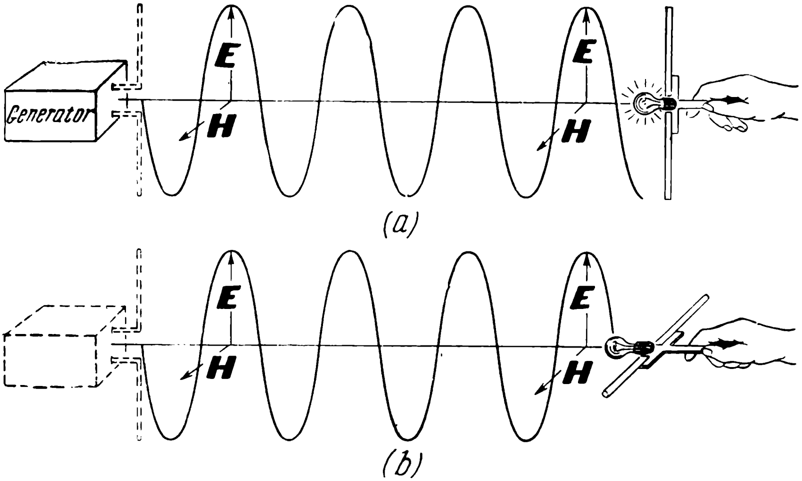


Fig. 64.2

conductor induce in the conductor forced oscillations, that is, an r.f. current. In flowing through the lamp filament, the current raises it to white heat, and the lamp glows. If the dipole is turned through 90°, the electric intensity vector will be at right angles to the conductor (Fig. 64.2*b*), no current will then be induced in the conductor, and the lamp will go out.

A similar experiment for polarized light will be discussed in Sec. 64.6.

2. Any device which passes only that component of light which is polarized in a particular plane and thus locates the plane of vibra-

tion of the polarized wave is called an *analyzer*. In the above experiment with an elastic wave on a rubber cord, the analyzer was the slot formed by two boards. In the experiment with the electromagnetic wave, the analyzer was a dipole aerial. The operating principle and arrangement of an analyzer for the optical region of the spectrum will be discussed in Sec. 64.6.

3. If the analyzer aerial in the experiment shown in Fig. 64.2 be rotated slowly from position *a* to position *b*, the glow of the lamp will gradually decrease from a maximum in position *a* to zero in position *b*. With further rotation of the dipole, the glow will again increase untill it reaches a maximum with the dipole turned through 180°. This can be explained as follows.

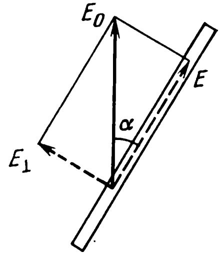


Fig. 64.3

Let the electric vector be vibrating in a vertical direction with an amplitude $E_0$ and let the direction of the analyzer dipole make an angle $\alpha$ with the plane of vibration (Fig. 64.3). Resolve the vector $E_0$ into two components, one along the aerial, $E$, and the other at right angles to it, $E_\perp$. Referring to the figure

$$\left.\begin{aligned} E &= E_0 \cos\alpha \\ E_\perp &= E_0 \sin\alpha \end{aligned}\right\} \tag{64.1}$$

Only one of these components, namely the vector $E$ directed along the aerial, will induce oscillations in the aerial. In contrast, the vector $E_\perp$ will set up no oscillations in the aerial.

4. As will be recalled, the intensity is proportional to the square of the amplitude (see Sec. 55.3). Designating the wave intensity as $I_0 = kE_0^2$ and the intensity of oscillations in the aerial as $I = kE^2$, we get

$$I/I_0 = E^2/E_0^2 \tag{64.2}$$

From a comparison of this proportion with the first line in equations (64.1) we obtain

$$I = I_0 \cos^2\alpha \tag{64.3}$$

Thus, the intensity of the wave passing through an analyzer is proportional to the squared cosine of the angle between the plane of vibration of the wave and the axis of the analyzer. This is the *Malus cosine-squared law*.

### 64.3. BIREFRINGENCE (DOUBLE REFRACTION)

1. If an unpolarized ray of light strikes the surface of a certain crystal, it will be divided into two transmitted (refracted) rays differing in properties. Such crystals are called *double-refracting* or *birefringent*. They include calcite, $CaCO_3$ (also known as Iceland spar) and quartz.

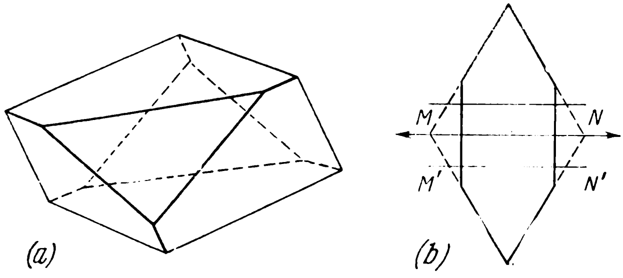


Fig. 64.4

Calcite is a hexagonal crystal (see Sec. 33.2); the axis of symmetry of the hexagonal prism that makes up its unit cell (see Figs. 33.4 and 32.3) is called its *optic axis*.

2. Let us cut a plate from a calcite crystal so that its faces are at right angles to its optic axis as shown in Fig. 64.4. A side elevation of this plate and the optic axis are shown in Fig. 64.4*b*. It should be noted that the optic axis is not a line; it is a direction in the crystal. Any straight line, such as $M'N'$, parallel to the $MN$ direction, will likewise be an optic axis. Let us call a plane which passes through the ray of light and the optic axis of the crystal its *principal plane*.

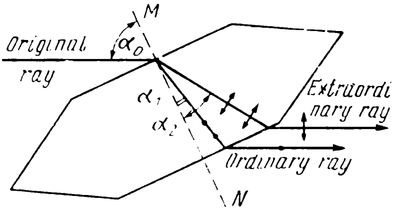


Fig. 64.5

3. If an unpolarized ray of light be directed along the optic axis, it will suffer no *double refraction*, or *birefringence*, because the structure of the crystal is symmetrical about that direction. If, on the other hand, an unpolarized ray of light be directed at an angle to the optic axis, it will be split up into two rays (Fig. 64.5). One is called the *ordinary ray*, and the other, the *extraordinary ray*. They differ in several respects as follows.

For one thing, the law of refraction, $\sin \alpha_0 = n \sin \alpha$, (63.2), is satisfied differently by the two rays. With the ordinary ray, the refractive index is independent of the angle of incidence, that is, $n_o = 1.658$ for any angles of incidence. As a consequence, the velocity of the ordinary ray, $u_o = c/n_o$, will be the same in all directions.

With the extraordinary ray, the refractive index $n_e$ depends on the direction in which light is propagated. More specifically, along the optic axis, the refractive index is the same for the extraordinary and the ordinary ray, $n_e = n_o = 1.658$, while at right angles to the optic axis, $n_e = 1.486$.

Hence, the velocity of a light wave along the optic axis is the same for the ordinary and the extraordinary ray, while in all other directions, the velocity of the extraordinary ray propagated in calcite is greater than that of the ordinary ray. The difference in velocity between the two rays is a maximum in the direction perpendicular to the optic axis. In this direction, the velocity of the extraordinary ray is 11.5% higher than that of the ordinary ray.

4. It should be noted that all of the data given above apply to yellow light with a wavelength of $\lambda = 5893$ Å. The values of refractive index for other wavelengths are listed in Table 64.1.

*Table 64.1*

| Light | $\lambda$, Å | Calcite | | Quartz | |
|---|---|---|---|---|---|
| | | $n_o$ | $n_e$ | $n_o$ | $n_e$ |
| Red | 6708 | 1.6537 | 1.4843 | 1.5415 | 1.5505 |
| Yellow | 5893 | 1.6584 | 1.4864 | 1.5443 | 1.5534 |
| Green | 5086 | 1.6653 | 1.4895 | 1.5482 | 1.5575 |
| Blue | 4800 | 1.6686 | 1.4911 | 1.5501 | 1.5594 |
| Violet | 4047 | 1.6813 | 1.4969 | 1.5572 | 1.5667 |

## 64.4. CAUSE OF BIREFRINGENCE

1. Using an analyzer, let us determine the direction in which the vector (that is, the electric vector **E**) is vibrating in the extraordinary and ordinary waves. We find that in the ordinary wave the vector **E** is vibrating at right angles to the principal plane, as shown by dots on the ray in Fig. 64.5. In the extraordinary wave, the electric vector is vibrating in the principal plane of the crystal, as shown by arrows in Fig. 64.5. The difference in velocity between the two waves, and, as a consequence, in their refractive indexes, may be explained on this basis as follows.

In anisotropic bodies (except cubic crystals) the force of interaction between the electron cloud and the lattice is different in different crystallographic directions. Because of this, the natural frequency of the electron cloud is likewise dependent on the direction in which the electrons are caused to vibrate by the incident light wave. According to (63.15), this results in different refractive indexes and different velocities for different directions.

2. Let the electron cloud in a crystal be caused to vibrate by a wave which is oscillating at right angles to the principal plane; in Fig. 64.5 this is the ordinary wave. Assume that in this case the natural frequency of the electron cloud is independent of the direction of wave propagation. Then, according to (63.15) the refractive index will not depend on the direction of the wave, either. Then the velocity of light propagation will be the same in all directions. Precisely such properties are displayed by the ordinary ray.

3. Now, let the electron cloud be caused to vibrate by a wave which is oscillating in the principal plane; in Fig. 64.5 this is the extraordinary wave. Experience shows that in this case the refractive index is dependent on the direction of the ray. This may be explained by assuming that the natural frequency of the electron cloud depends on the direction in which the vector **E** exciting these oscillations is vibrating.

According to (63.15), the inequality in natural frequency leads to an inequality in refractive index, that is, if $\omega_0^e \neq \omega_0^o$, then $n_e \neq n_o$. The same applies to the velocities of these waves. Thus, the electron theory provides an explanation, at least qualitative, of what causes birefringence in crystals.

4. The phenomenon of double refraction, or birefringence, in calcite was discovered by Erasmus Bartholinus in 1669. In 1690, Huygens came out with a formal theory of Bartholinus' discovery, based on the hypothesis that the two rays differed in velocity; however, he could not explain why this was so. The idea that the difference in velocity between the two rays is due to some internal characteristics of the rays themselves belongs to Newton who advanced it in his famous *Opticks* (1704). In 1808, Malus revived Newton's concept by ascribing the behaviour of the double refracted rays to their "polar" properties similar to the poles of a magnet. Quite appropriately, he suggested the term "polarization of light", still used today.

Young and Fresnel, who had convincingly explained interference and diffraction by the wave behaviour of light, realized that the polarization of light could be explained satisfactorily if they imagined light waves to be transverse. Fresnel backed his hypothesis by experiment and reasoning. Since transverse vibrations are possible only in solids, physicists were forced to accept the concept of an elastic solid ether. However, they failed to come up with a consistent theory of an elastic solid ether sustaining only transverse waves and free from longitudinal waves. In short, the elastic solid theory of the ether failed to explain convincingly the polarization of light. The contradictions were removed only with advent of Maxwell's electromagnetic theory of light. In fact, his theory made superfluous the very concept of the ether. This became clear, however, after relativity theory had been developed.

## 64.5. DICHROISM

1. Some crystals absorb the ordinary and the extraordinary rays differently. When a beam of unpolarized light strikes a plate of a tourmaline crystal just a few millimetres thick at right angles to the optic axis, the ordinary ray is completely absorbed, and only the extraordinary ray emerges from the crystal (Fig. 64.6).

This difference in the absorption of the ordinary and extraordinary rays is called the *anisotropy of absorption*, or *dichroism*.

2. To explain this phenomenon, it is necessary to recall that electromagnetic waves are strongly absorbed in the anomalous dispersion region, that is, when the wave frequency is close to the natural frequency of the electron cloud (see Sec. 63.7). However, it is shown in the previous section that for anisotropic materials $\omega_0^o \neq \omega_0^e$. Hence, if the wave frequency is close to the natural frequency of the electron cloud in the direction perpendicular to the principal plane (that is, if $\omega \approx \omega_0^o$), then the ordinary wave will be strongly absorbed. Conversely, the extraordinary wave, for which $\omega \neq \omega_0^e$, will be absorbed less.

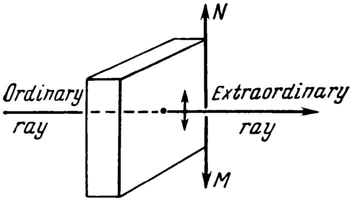


Fig. 64.6

Thus, on the basis of the electron theory we have been able to explain qualitatively both the anisotropy of refraction (double refraction) and its corollary, the anisotropy of absorption, or dichroism. To a different degree, all double refracting crystals should obviously display dichroism, but they do this differently.

## 64.6. POLAROID AS POLARIZER AND ANALYZER

1. Polarizing devices, or *polarizers*, turn ordinary light into plane-polarized. For their effect, they all depend on the separation of the ordinary from the extraordinary ray. An example of a nature-made polarizer is a plate of tourmaline (see Sec. 64.5). At present, wide use is made of man-made polarizing materials, such as "Polaroid". This material consists of a very thin film of an organic substance made up of long-chain molecules. By a special process, the molecules are lined up to impart the material the dichroic property such that it absorbs practically all of the extraordinary ray. A polarizer made from this material is a piece of film mounted in a frame on a glass disk. These film polarizers are inexpensive and can be made in large

sizes, because of which they have come into wide use in practice. Yet, they suffer from disadvantages, namely they produce a tinted light and can only operate within a relatively narrow region of the spectrum.

2. Using a film polarizer, it is possible to analyze plane polarized light, that is, to determine the direction in which its electric vector is vibrating (the plane of vibration).

Let a plane-polarized wave strike a film polarizer so that the direction in which the electric vector is oscillating coincides with the optic axis of the film polarizer (Fig. 64.7*a*); then the plane of vibration will coincide with the principal plane of the crystal. This implies that the wave propagated in the crystal (see Sec. 64.4) is an extraordinary one, which is absorbed very little. The light will pass through the film polarizer and will be seen by the observer. Now turn the film polarizer through 90° (Fig. 64.7*b*). The same wave will now be ordinary, because the plane of vibration is at right angles to the principal plane of the crystal. Because the film polarizer strongly absorbs the ordinary wave, no light will pass through the polarizer.

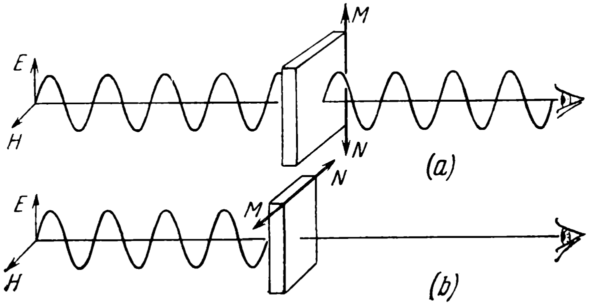


Fig. 64.7

3. We leave it as an exercise for the reader to take up a case in which the plane of vibration of a plane-polarized wave makes an angle $\alpha$ with the principal plane of the film polarizer. By resolving the wave into an ordinary and an extraordinary component, it is possible to derive the Malus cosine-squared law for light waves (see Sec. 64.2). Thus, a film polarizer may be used as both a polarizer and an analyzer.

## 64.7. ROTATION OF THE PLANE OF POLARIZATION

1. If we place two film polarizers in succession in the path of a light beam so that their optic axes are at right angles to each other (so-called *crossed polarizers*), no light will pass through. This is because

the first polarizer converts natural light into plane-polarized, and the second absorbs it (Fig. 64.8). Now if we place a cell holding a solution of sugar in the path of the light beam, we shall see that the field of view has become bright. If we turn the polarizer right and through a certain angle $\alpha$, the field of view will again become dark.

This experiment leads us to the conclusion that when a beam of plane-polarized light passes through a sugar solution, the light remains plane-polarized, but the plane of vibration and, as a consequence, the plane of polarization (see Sec. 64.1) is rotated through a certain angle.

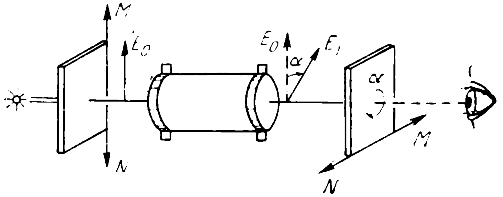


Fig. 64.8

2. Substances that cause the plane of polarization to rotate are called *optically active*. Optically active crystals include quartz. If a beam of light is propagated along its optic axis, the plane of polarization will be rotated through different angles at different wavelengths as follows.

| Light | λ, Å | Specific rotation, [α], mm⁻¹ |
|---|---|---|
| Red | 6563 | 17.32° |
| Yellcw | 5893 | 21.72° |
| Blue | 4861 | 32.76° |
| Violet | 4340 | 41.92° |

A complete rotation of the plane of polarization is proportional to the thickness of the plate, $d$, or

$$\alpha = [\alpha]\, d \tag{64.4}$$

3. Optical activity is manifested not only by crystals, but also by some liquids (turpentine, nicotine), and also by solutions of some substances in water, for example, saccharosa ($C_{12}H_{22}O_{21}$), glucose ($C_6H_{12}O_6$), tartaric, malic and mandelic acids; solutions of camphor, brucine and strychnine in alcohol, etc. The angle

through which the plane of polarization is rotated is given by

$$\alpha = [\alpha]\, cd \qquad (64.5)$$

where $c$ is the concentration of the optically active material, as the number of grams per 100 ml of the solution.

Since the angle of rotation of the plane of polarization is proportional to the concentration of the optically active material in solution, it is possible to determine this concentration from the angle of rotation by means of specially designed instruments called *polarimeters* or *saccharimeters*.

4. Optical activity is displayed by substances whose molecules lack a centre or a plane of symmetry; among them are molecules of most organic compounds. These substances remain optically active as crystals, melts or solutions.

Sometimes optical activity is due to the screw-like arrangement of the crystal lattice rather than to the properties of the molecules. In such cases, the melts or solutions of these substances will not be optically active. For example, quartz will not rotate the plane of plane-polarized light until its atoms take up an orderly arrangement in a crystal; fused quartz (where its atoms are in an amorphous state) is optically inactive.

## 64.8. OPTICAL ACTIVITY IN NATURE

1. Optically active crystals always occur as *optical antipodes*, that is, as two kinds of the same compound composed of the same atoms and atomic linkages which differ in their structural formulas only in that one is the mirror image of the other. It would seem that optical antipodes should likewise occur among optically active organic compounds. Yet experience shows that a sugar solution always rotates the plane of polarization to the right, or clockwise, if viewed against the ray. This property is shown not only by sugar, but by all other products of life processes, namely proteins, amino acids, nucleic acids, etc.

2. No purely synthetic chemical product (for example, sugar) will be optically active. In a synthetic product, there is a mixture of as many right-handed as left-handed molecules. Generally, in inorganic nature all substances with assymmetrical molecules exist as such mixtures.

If this mixture is fed to a living organism, it will only assimilate one of the structures answering the optical activity of that organism. For example, certain bacteria feeding on sugar when placed in a solution of synthetic sugar will assimilate only the right-handed form. Some time later, the amount of left-handed sugar in the solution will exceed that of right-handed sugar, which may be learned from the rotation of the plane of polarization. Finally, the bacteria will have destroyed all of the right-handed sugar and go starving, although the solution still holds a large amount of left-handed sugar. This is because the bacteria cannot assimilate it.

3. The asymmetry of optical activity is characteristic of only life processes or their products. The fact that petroleum shows optical activity is a convincing argument in favour of its organic origin.

The entropy of an optically active medium is less than that of a mixture. This is because the entropy of a system is a maximum when the thermodynamic probability is a maximum, and this goes with the equal number of left-handed and right-handed molecules (see Ch. 28). Thus, the optical activity of life processes or their products reflects the general orderliness of living matter which is always in a state of non-equilibrium, whose entropy is far from a maximum.

The cause of asymmetry in the optical activity of life processes or their products is not clear. It is not unlikely that this asymmetry came about by chance and has since been perpetuated by the mechanism of heredity.

# Chapter 65

# GEOMETRICAL (RAY) OPTICS

## 65.1. BASIC LAWS OF GEOMETRICAL OPTICS. BEAM AND RAY

1. All problems of optics can be handled on the wave theory of light as has been shown in Chapters 61 through 64. However, this approach involves the use of a large body of mathematics. Well before the wave theory of light was established, physicists had used geometrical methods in the construction of optical images formed by mirrors and lenses and in designing optical instruments. These methods comprise what is known as *geometrical* or *ray optics*.

2. Geometrical optics treats light as if it were actually composed of rays diverging in various directions from the source and abruptly bent by refraction or turned back by reflection into paths determined by well-known laws. The idea that light travels in straight lines is uppermost, while its wave character and other physical aspects are disregarded. Since the laws of reflection and refraction are corollaries of the wave nature of light (see Secs. 56.5 and 63.1), they may be applied without any limitations. As regards the statement that light travels in straight lines it may be applied only with certain qualifications.

The point is that in optical instruments light always travels through openings known differently as *pinholes*, *diaphragms* or *apertures*, which cut a certain area out of the wave front. As will be recalled, this is accompanied by diffraction (see Secs. 57.8, 57.9, and 62.1). Therefore, diffraction sets a limit to the applicability of the assumption that light travels in straight lines, and thereby a limit to the applicability of geometrical optics.

3. The basic concepts of geometrical optics are the *beam* and the *ray*. The meaning of these concepts will be clear from the following experiments.

Cover a window with a sheet of perforated cardboard and let a little smoke into the air in the room. We shall see that some sunlight passes through the holes in the cardboard as narrow cylindrical shafts. A conical shaft of light will be produced, if a small lamp is enclosed in an opaque box with a hole in it. The cylindrical or conical shafts within which light is propagated are called *light beams*. The lines indicating the direction in which light travels (including those on the surface and at the axis of light beams) are called *light rays*.

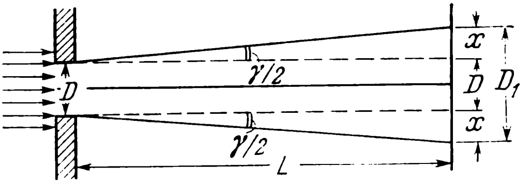


Fig. 65.1

4. Because of diffraction, a light wave bends around the edges of an obstacle, and light is no longer propagated in straight lines. Instead of a beam with a well-defined surface, as would be expected by the laws of geometrical optics, we obtain a diverging beam lacking any well-defined boundary. A marked glow is observed within a cone (Fig. 65.1) with a vertex half-angle $\gamma/2$ defined as $\sin(\gamma/2) \approx \lambda/D$ (see Sec. 62.3), where $D$ is the diameter of the aperture (pinhole) in the diaphragm (an opaque screen).

Thus, with an aperture (and any optical instrument uses at least one) light does not travel in straight lines any longer. In some cases, however, the beam broadening $x = (D_1 - D)/2$ (Fig. 65.1) is small in comparison with the aperture diameter, and diffraction may then be neglected in a first approximation.

Referring to Fig. 65.1, it is seen that $x = L\tan(\gamma/2)$. At small angles, we may set $\tan(\gamma/2) \approx \sin(\gamma/2) \approx \lambda/D$. Thus, $x \approx L\lambda/D$. The condition $x \ll D$ takes the form $L\lambda/D \ll D$, whence

$$D \gg \sqrt{L\lambda} \tag{65.1}$$

This is a criterion for the applicability of geometrical optics.

5. Sometimes it is said that geometrical optics applies if the aperture is many times the wavelength of light. We see that this is an insufficient criterion, because it does not take into account the distance $L$ from the screen (point of observation) to the diaphragm. With a considerable value of $L$, condition (65.1) is not satisfied, and the pattern observed in the experiment differs drastically from that calculated with geometrical optics, although the aperture may be sufficiently large. For example, let the aperture diameter be $D = 1$ mm. This is 2000 times the wavelength of green light ($\lambda = 5000$ Å $= 5 \times 10^{-7}$ m). As is seen, $D \gg \lambda$. However, according to (65.1), the laws of geometrical optics will remain valid to $L \ll 10^{-6}/5 \times 10^{7}$ m, that is, to $L \ll 2$ m. Already at a distance of about 1 m from the diaphragm, we shall see a diffraction pattern, and geometrical optics will prove inapplicable.

Nor can we look upon a ray of light as a narrow beam produced by "stopping down", that is, by using a smaller aperture. Infinitesimally narrow light beams are non-existent; a beam of light is always of a finite width. The ray is, in effect, the axis of a beam, and not the beam itself. A ray of light is a purely geometrical concept representing the direction in which energy is propagated.

## 65.2. REFRACTION OF LIGHT. TOTAL INTERNAL REFLECTION

1. In Sec. 63.1, we examined the refraction of light as it passed from vacuum into a material. Now we shall generalize this law to a case where light passes from one medium into another. For this purpose, we shall use expression (56.26), substituting in it the velocity of light in a medium from equation (63.3). Then we get

$$\sin\alpha_1/\sin\alpha_2 = u_1/u_2 = cn_2/n_1c = n_2/n_1 = n_{21} \qquad (65.2)$$

The quantity

$$n_{21} = n_2/n_1 \qquad (65.3)$$

is called the *relative refractive index* of the second medium, with the first medium taken as a reference.

2. Equation (65.2) may be rewritten as follows

$$n_1 \sin\alpha_1 = n_2 \sin\alpha_2 \qquad (65.4)$$

In this form, the law of refraction can conveniently be used in solving problems. Expression (63.2) is a special case of the more general form, (65.4), of the law, if we set $n_0 = 1$ for a vacuum.

Of the two media in which light is propagated at different velocities, that in which the light velocity is lower and the refractive index is greater will be called optically denser. For example, glass ($n = 1.5$ to $1.7$) is optically denser than water ($n = 1.33$).

3. At the interface of two media, the incident light beam is separated into two beams, *reflected* and *refracted* (see Secs. 56.6 and 63.2). The intensities of the two beams add up to the intensity of the incident beam. As regards the intensity of each beam, it depends in a very elaborate way on the angle of incidence and the relative refractive index, and we shall not derive the respective equations. We shall only take up a very interesting case where light passes from an optically denser medium into a less dense medium. At a small angle of incidence (Fig. 65.2), the intensities of the reflected and refracted beams can approximately be found by equation (63.10).

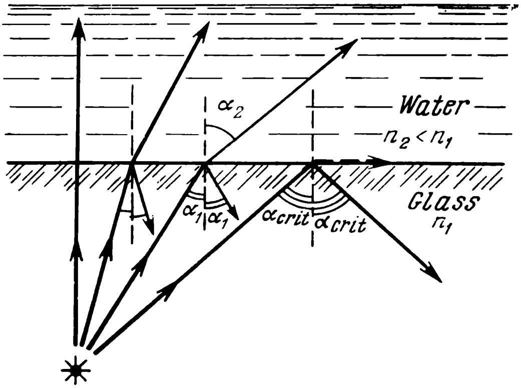


Fig. 65.2

Since the ratio of the refractive indexes is usually less than 2, the intensity of the refracted beam is considerably less than 12%, while often, at $n_1 \approx n_2$, the intensity of the reflected beam is zero very nearly. If the angle of incidence be increased, the angle of refraction will increase at a faster rate. At the same time the intensity of the reflected beam will increase, and that of the refracted beam will decrease.

4. When the angle of incidence, $\alpha_1$, approaches a certain value, $\alpha_{crit}$, the angle of refraction tends to the right angle, $\alpha_2 \to \pi/2$, and the intensity of the refractive beam tends to zero at a very fast rate. At all angles of incidence exceeding $\alpha_{crit}$, which is called the *critical angle*, there will be no refracted beam, and light will be totally reflected from the interface, as if it were an ideal mirror. This is *total internal reflection* (TIR) or simply *total reflection.*

5. In order to determine the critical angle, we set in equation (65.4) $\alpha_1 = \alpha_{crit}$ and $\alpha_2 = \pi/2$, and obtain

$$\sin \alpha_{crit} = n_2/n_1 \qquad (65.5)$$

If light passes from a medium with a refractive index $n_1 = n$ into air where $n_2 = 1.000292 \approx 1$, expression (65.5) takes the form

$$\sin \alpha_{crit} = 1/n \tag{65.5'}$$

It should be noted that at an angle of incidence exceeding the critical angle, the law of refraction has no sense. This is because from $\alpha_1 > \alpha_{crit}$ it follows that $\sin \alpha_1 > \sin \alpha_{crit} = n_2/n_1$. But then from (65.2) or (65.4) it should follow that $\sin \alpha_2 = n_1 \sin \alpha_1/n_2 > 1$, which cannot be. Physically, this implies that there is no refraction and that all light is reflected.

6. A rigorous solution of this problem on the wave theory yields a somewhat different result. As it turns out, a light wave can pass into a less dense medium even when the angle of incidence exceeds the critical angle. In such a case, however, the refracted wave damps out very rapidly, and at a distance of few wavelengths from the interface its intensity is practically zero. This result can be verified experimentally. With a light beam incident at an angle exceeding the critical angle on a glass prism ($n_1 = 1.7$) immersed in a solution of fluorescein ($n_2 = 1.34$), a faint glow will be noted in the thin layer of fluorescein under the prism although, according to the laws of geometrical optics, no light can be propagated there.

Similar penetration of particles into a region forbidden by classical physics will be discussed in Sec. 70.6.

### 65.3. THE PRISM

1. Imagine a parallel beam of light striking a face of a prism. Let the prism be made from a material denser than the surroundings, for example a glass prism in air. Then the light beam, on being refracted twice in the prism, will be deflected from the original direction through some angle ε towards the base of the prism. The ray path for this case is shown in Fig. 65.3*a*.

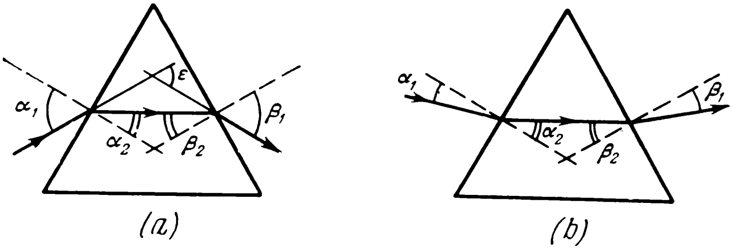


Fig. 65.3

If the prism is made from a material less dense optically than the surroundings, the beam will be deflected towards the vertex of the prism, as shown in Fig. 65.3*b*. We leave it for the reader to check the ray path by calculation.

2. So far we have assumed that the beam strikes a second face of the prism at an angle $\beta_2$ which is less than the critical angle. If, however, $\beta_2 \geqslant \alpha_{crit}$, then the light will be totally reflected from the second face. Total internal reflection (TIR) prisms are widely used in optical instruments instead of mirrors. As an example, trace the ray path in the glass prism shown in Fig. 65.4*a*. The beam strikes the first face at right angles (normal incidence), and so it is not refracted. The angle of incidence at the second face is $\alpha = 45°$, exceeding the critical angle; according to (65.8), at the interface between glass and air the critical angle is $\alpha_{crit} = \arcsin (1/1.5) = 42°$. Therefore, the light beam suffers total internal reflection from the second face and proceeds as shown in the drawing.

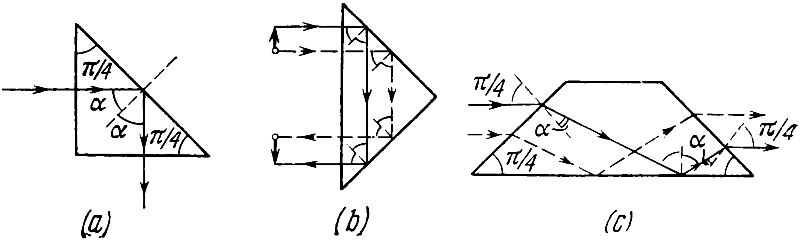


Fig. 65.4

Total internal reflection is utilized in erecting prisms for which the ray path is traced in Fig. 65.4*b* and *c*. We leave it for the reader to prove this construction by calculation.

3. Let us find the angle $\varepsilon$ through which the light beam will be deflected by a prism with a vertex angle $\varphi$. For simplicity, we set that the light beam strikes the first face of the prism at normal incidence (Fig. 65.5*a*). As is seen, the angle of incidence on the second face is equal to the prism angle, that is $\alpha_1 = \varphi$. The angle of refraction is $\alpha_2 = \varphi + \varepsilon$. From the law of refraction, (65.4), it follows that

$$\sin (\varphi + \varepsilon) = n \sin \varphi \tag{65.6}$$

4. It can be shown that calculations based on interference will yield the same results. For this prupose, we shall trace the ray path through a Michelson echelon. This is a highly specialized form of diffraction grating devised by Michelson. It consists of a row of glass plates of exactly equal thickness, packed together to form a miniature stairway of equal risers (Fig. 65.5*b*). It is obvious that increasing the number of risers without bound and decreasing their size in proportion would turn the Michelson echelon into a prism.

Light enters normally to the highest plate at one end, and emerges at various deviations through the lower risers. Because of interference, however, the intensity of the successive emergent beams is different. The direction of the principal interference maximum can be found by the method used to calculate interference from $N$ coherent sources used in Secs. 57.6 through 57.10.

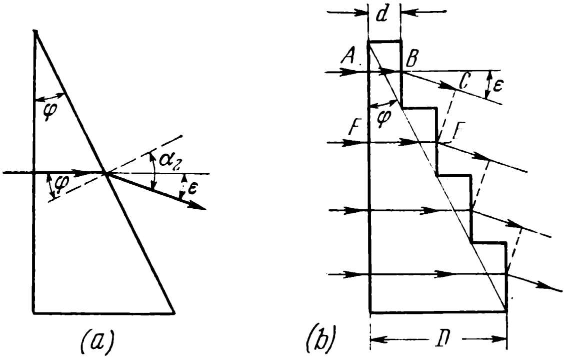


Fig. 65.5

Let the width of the entire echelon be $D$, the width of one riser $d = D/N$, and the height $h = d \cot \varphi$. By analogy with expression (57.41) we may write the following expression for the amplitude of a light wave in the direction defined by the angle of deviation, ε

$$A_\varepsilon = (A_0/N) \frac{\sin (Nk\Delta/2)}{\sin (k\Delta/2)}$$

Here Δ is the optical path difference between the rays $ABC$ and $FE$. As in Secs. 57.9 and 57.10, the direction of the principal maximum is determined by the condition such that $\Delta = 0$.

According to (61.12), the optical path difference may be written as

$$\Delta = ABn + BC - FEn = BC - nd$$

where $n$ is the refractive index of glass. It is an easy matter to prove that $BC = [d \sin (\varphi + \varepsilon)]/\sin \varphi$. Then the optical path difference will be

$$\begin{aligned}\Delta &= [d \sin (\varphi + \varepsilon)/\sin \varphi] - nd \\ &= (d/\sin \varphi) [\sin (\varphi + \varepsilon) - n \sin \varphi] = 0\end{aligned}$$

We have obtained expression (65.6) again.

5. Thus, as in Sec. 57.10, the laws of geometrical optics are a limiting case for the interference laws. Therefore, when tracing

the path of rays in lenses and generally in any optical instruments, we may use the simpler rules of geometrical optics instead of the rather elaborate interference computations.

It should be borne in mind, however, that geometrical optics is a simplified method of calculation and that actually an image in any optical instrument is interference-dependent and that secondary maxima cannot always be neglected. This matter will be taken up again in Sec. 66.8.

6. The angle $\varepsilon$ through which a beam of light is deflected by a prism (Fig. 65.5) depends on the refractive index. On the other hand, it is known that the refractive index depends on frequency (this dependence is known as dispersion, see Ch. 63). Then, if a non-monochromatic (say, white) light strikes a prism, the rays representing the various colours (that is, the various frequencies), will be deflected through different angles. This was discovered by Newton in 1666 when he observed that a beam of white light passing through an opening and then through a glass prism was separated into a spectrum, with red light deflected least and blue and ultraviolet, most of all.

## 65.4. THE LENS

1. As a rule, a lens is a piece of glass with its two sides ground to spherical form. Sometimes, use is made of lenses with cylindrical, parabolic or other surfaces, but they lie outside the scope of this book.

The straight line passing through the centres of the two lens surfaces is called its *principal axis.*

2. Let a pencil of light parallel with the principal optical axis be incident on a lens. If the width of the pencil is a small fraction of the radii of curvature of the lens, it will be called *paraxial.* After refraction in the lens, the pencil is collected at a point on the principal axis, called the *principal focus* of the lens. Let us determine the position of the principal focus.

To do this, refer to Fig. 65.6. The ray $SK$ is first refracted at point $K$, then at point $L$, and passes through the focus $F$. Extend the rays $SK$ and $LF$ until they meet at point $M$, and let a plane, $MC$, perpendicular to the principal optical axis, pass through that point. The plane $MC$ is called the *principal plane* of a lens. Each lens has two principal planes. If, however, a lens is sufficiently thin, the two principal planes will practically merge together. In our further discussion we shall limit ourselves to thin lenses and construct only one principal plane for them.

3. The point $C$ at which the principal plane of a thin lens intersects its principal axis is called the *optical centre* of the lens. The

distance from the optical centre to the principal focus, $CF = f$, is called the *focal length* of the lens.

The reciprocal of the focal length of a lens is called its *focal power*:

$$\Phi = 1/f \qquad (65.7)$$

and is a measure of the converging or diverging effect of the lens. It is commonly expressed in *diopters*, or reciprocal metres; thus,

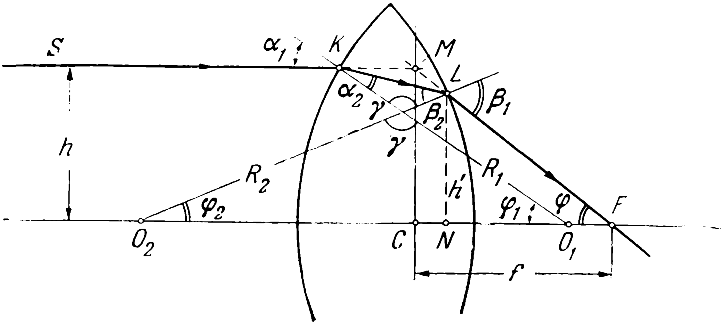


Fig. 65.6

if the local length of a lens is 1 m, its focal power is 1 diopter. To find the focal power of a lens in diopters, the focal length in equation (65.7) should be expressed in metres.

4. Let us go back to Fig. 65.6. If the light beam is sufficiently narrow (such that $h \ll R_1 \approx R_2$), the angles of incidence and refraction are very small, and their sines and tangents will not practically differ from the angles in radians. Then the law of refraction, (65.2), may be written as

$$\alpha_1 = n_{21}\alpha_2, \quad \beta_1 = n_{21}\beta_2 \qquad (65.8)$$

Referring to the figure, it is seen that $\alpha_1 = \varphi_1$ as corresponding angles formed by parallel lines:

$$\alpha_2 + \beta_2 + \gamma = \varphi_1 + \varphi_2 + \gamma = 180^\circ$$

whence it follows that

$$\alpha_2 + \beta_2 = \varphi_1 + \varphi_2$$

and finally

$$\beta_1 = \varphi_2 + \varphi$$

as an external angle of a triangle. Substituting in (65.8) and adding together the two equalities, we get

$$\alpha_1 + \beta_1 = n_{21}(\alpha_2 + \beta_2) \quad \text{or} \quad \varphi_1 + \varphi_2 + \varphi = n_{21}(\varphi_1 + \varphi_2) \qquad (65.9)$$

whence

$$\varphi = (n_{21} - 1)(\omega_1 + \varphi_2) \tag{65.10}$$

Since at small values, angles may be replaced by their sines or tangents, we get

$$\tan \varphi = \varphi = h/f, \quad \sin \varphi_1 = \varphi_1 = h/R_1, \quad \sin \varphi_2 = \varphi_2 = h'/R_2 = h/R_2$$

Substituting in (65.10) and cancelling $h$'s, we obtain a final expression for the focal power of a thin lens

$$\Phi = 1/f = (n_{21} - 1)(1/R_1 + 1/R_2) \tag{65.11}$$

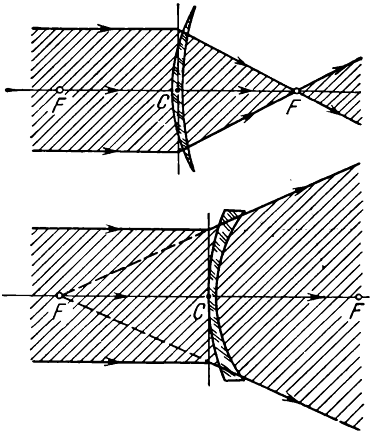


Fig. 65.7

5. In expression (65.11) the radii should be taken with an appropriate sign, namely positive for convex surfaces and negative for concave surfaces; a plane should be treated as a surface with an infinite radius of curvature. Noting this rule, the reader will easily prove that convex lenses have a positive focal power and concave lenses a negative focal power. This is the reason why a concave lens is said to be negative and to have a *virtual focus*. The effect of a convex and a concave lens with a refractive index greater than that of the surroundings is illustrated in Fig. 65.7. Quite appropriately, the former is also called *converging*, and the latter, *diverging*.

We leave it for the reader to prove that, if less optically dense than the surroundings, a convex lens will have a negative focal power, and a concave lens a positive focal power.

## 65.5. FORMATION OF AN IMAGE BY A THIN LENS

1. Let point $A$ be $d > 2f$ distant from a thin convex lens (Fig. 65.8). The lens cuts out part of the beam, shown shaded in the figure, and collects it at point $A'$ which is an image of point $A$. To locate point $A'$, we select two rays of known paths. One is the ray $AM$ parallel with the principal optical axis; after refraction in the lens, it passes through the posterior focus $F'$ of the lens. The other is the ray $AN$ passing through the anterior focus $F$ of the lens; after refraction, it will be parallel with the principal optical axis of the lens. Point $A'$ is at the intersection of the two rays.

We leave it for the reader to show that the ray $AC$ passing through the optical centre of the lens likewise passes through point $A'$. To this end, it will suffice to prove that the straight lines $AC$ and $A'C$ make the same angle with the optical axis.

2. Thus we have three characteristic rays whose paths are known. These are the ray parallel with the principal optical axis of the lens, which passes through the focus of the lens after refraction; the ray passing through the focus of the lens and directed parallel with its principal optical axis after refraction; and the ray which passes through the centre of the thin lens (the auxiliary or secondary optical axis) and proceeds on without being refracted. Using any two of the three characteristic rays, we can readily construct an image of any point and, by the same token, of any object as a collection of points.

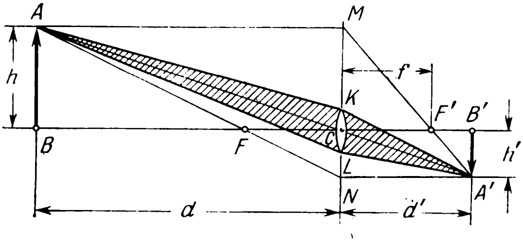


Fig. 65.8

As is seen, to construct the image of a point formed by a lens, we need not know the position of its refracting surfaces and a detailed path of rays in the lens. It will suffice to know the position of its principal optical axis and focuses. Of course, energy is transferred from point $A$ to point $A'$ by the shaded portion of the light beam, but without the above construction it would remain unclear how the rays $AK$ and $AL$ are directed after refraction by the lens.

3. Let us construct the path of a ray other than characteristic, that is, one incident on a lens at an arbitrary angle (Fig. 65.9). Draw the secondary optical axis $KC$ parallel with the ray $MN$ and the focal plane $ab$ passing through the principal focus parallel with the principal optical axis. Much as a paraxial beam of light parallel with the principal optical axis closes down to a point at the principal focus $F'$, so the beam of light parallel with the secondary optical axis $KC$ closes down to a point at the secondary focus $F_s$ lying in the focal plane. As a consequence, the ray $MN$ will pass through point $F_s$ after being refracted.

We leave it for the reader to construct an image of a point on the principal optical axis of a lens, using the above method. Also, as an exercise, set up a ray incident on a concave (diverging) lens at an arbitrary angle.

4. Let us construct an image of a point as it is formed by a diverging lens (Fig. 65.10). After refraction, a ray, *AM*, parallel with the principal optical axis will proceed in the direction of *MK* because its extension passes through the focus. A second ray, *AC*, (the secondary optical axis) gets clear through the lens without refraction. As is seen, the merging rays (shown shaded in the figure) are spread apart. Obviously, a diverging beam cannot form an image, and a diverging lens cannot produce an image. If, however, an observer's eye is placed as shown in Fig. 65.10, the refracted rays will seem to come from a point, *A'*, on the extension of the rays. This is a *virtual focus*, and the resultant image is likewise *virtual* in contrast to the real image formed by a converging lens (see Fig. 65.8).

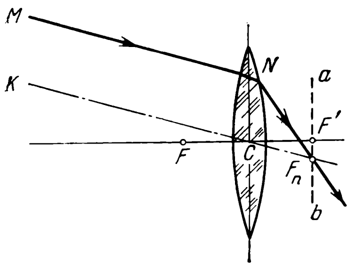


Fig. 65.9

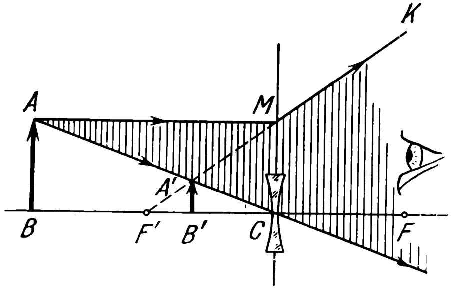


Fig. 65.10

A fundamental difference between the two images is as follows. With a real image, the energy of light waves is concentrated at the focus, and this concentration can be detected by objective means, such as a photoelectric cell, a thermo-couple, photographic paper and the like. With a virtual image, the rays of light only seem to come from a point where nothing can be detected by objective means.

We leave it for the reader to form an image of a point lying between a thin converging lens and its focus.

## 65.6. THIN LENS FORMULA

1. Experience shows that as the object distance *d* changes, the image distance *d'* changes too. To establish the relation that exists between the two distances, let us go back to Fig. 65.8 again. From the similarity of triangles *ABC* and *A'B'C'* we have

$$h'/h = d'/d,$$

and from the similarity of triangles $MCF'$ and $A'B'F'$

$$h'/h = (d' - f)/f$$

Since the left-hand sides of the equations are the same, we may write

$$d'/d = (d' - f)/f \qquad (65.12)$$

Hence,

$$d'f = d'd - fd \text{ or } fd + fd' = dd'$$

Dividing this equality by the product $fdd'$ and cancelling, we get

$$1/d + 1/d' = 1/f \qquad (65.13)$$

This is the *thin-lens* (or *simple-lens*) *formula*.

2. It is important to follow the sign convention adopted for the terms of Eq. (65.13). The object distance $d$ is taken to be positive always, while the focal length $f$ is taken positive for a converging lens and negative for a diverging lens. If the image distance $d'$ turns out positive, the image is real and on the far side of the lens; if the image distance turns out negative, the image is virtual and on the near side of the lens. We leave it as an exercise for the reader to check this statement, using suitable examples.

## 65.7. ABERRATIONS OF LENSES

1. In the preceding discussion we have tacitly assumed that the lenses acted in an ideal fashion, that is, the image produced by a lens is a true representation of the object, point for point. But there are many ways in which actual lenses fall short of this ideal. These shortcomings are known by the general name *aberrations*. Knowledge of these defects is important because it suggests ways and means for overcoming them.

2. *Spherical Aberration*. In deriving the formula for the focal power of a lens (65.13), we assumed that the paraxial beam of light incident on the lens has a width $h$ which is a small fraction of the radius of curvature of the lens, $h \ll R$. Under this condition, the central zone of the lens causes the incident beam to converge to exactly the same point, the focus. With a broad beam (Fig. 65.11), the outer annular zone of the spherical surface of the lens does not cause the outer rays to converge to the focus, thus producing a blurring of the resultant image, known as *spherical aberration*.

One of the methods for correcting spherical aberration is to narrow the light beam by placing an *aperture stop* in the beam path ("stopping down"). Unfortunately, this inevitably reduces the energy of the light beam, which is not always desirable. Another method is by combining two lenses, one converging and the other diverging (see Fig. 65.11). In this combination, the spherical aberration associated with one element tends to cancel that associated with the other. The two lenses may be matched so that their aggregate focal power will be non-zero while spherical aberration will be reduced to a considerable extent.

3. *Chromatic Aberration.* In Chapter 63 it has been shown that glass or any other substance produces dispersion (that is, refracts light of different colours by different amounts). Because of this, in a converging lens the focal length is greater for red light than for violet, while in a diverging lens the reverse is true. As a result, the image of a white point will be either red with a blue-violet fringe, or violet with a yellow-red fringe (depending on the location of the viewing screen). This defect is known as *chromatic aberration.*

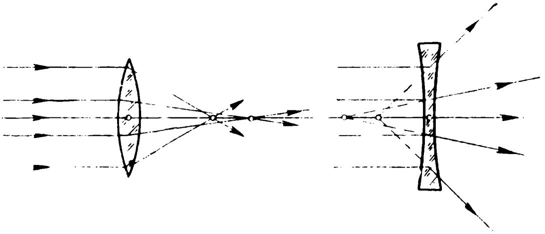


Fig. 65.11

Since chromatic aberration in a converging and a diverging lens is opposite in its effect, it is possible to correct it by a compound lens, that is a combination of a concave and a convex lens cemented together (Fig. 65.12), made from different glasses with properly matched refractive indexes and ground to suitable radii of curvature. Incidentally, compound lenses correct both chromatic and spherical aberrations. Freedom from chromatic aberration is known as *achromatism*, and lenses made for the purpose are called *achromatic lenses* or, simply, *achromats.* They are used as the objectives of telescopes, field glasses and other optical instruments.

4. *Astigmatism.* Still another aberration of lenses is known as *astigmatism.* It consists in that the image of a point lying widely off the optical axis of the lens is formed by a ray at a large angle of incidence. No point image is formed, whatever the position of the viewing screen. Instead, the image is a pair of short lines normal to each other and at slightly different distances from the lens. This defect may be shown even by lenses corrected for spherical and chromatic aberrations.

Astigmatism can be corrected by a combination of lenses matched so that all will cancel out the effect of each. Such a combination is called an *anastigmatic system* or an *anastigmat*. For example, good-quality cameras use anastigmats.

5. *Distortion*. It may so happen that the image formed by a lens is geometrically dissimilar to the object. This kind of aberration is called *distortion*. For example, the square grid of Fig. 65.13*a* may be imaged as shown in Fig. 65.13*b*—this is *pin-cushion distortion*, or as shown in Fig. 65.13*c*—this is *barrel distortion*. It may be corrected by suitably matching lenses in which distortion produces opposite effects. Lenses corrected for distortion are used in aerial photographic cameras.

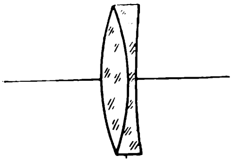
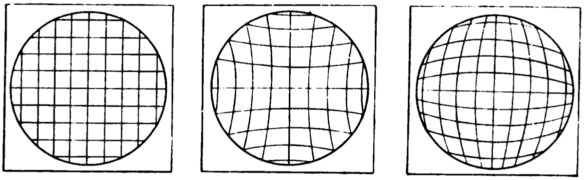


Fig. 65.12



(a) (b) (c)

Fig. 65.13

## 65.8. THE SPHERICAL MIRROR

1. Let a bundle of rays parallel to the diameter of a spherical mirror passing through its vertex, $C$, strike it on the concave side (Fig. 65.14). Experience shows that the rays will converge to a point, $F$, a focus. Let us find the focal length, $CF = f$.

The normal at the point of incidence of the ray $KM$ is the radius $OM = R$. The angles of incidence and reflection are equal:

$$\angle KMO = \angle OMF = \alpha$$

Besides,

$$\angle FOM = \angle KMO = \alpha$$

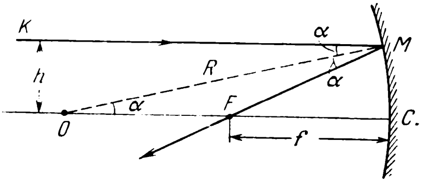


Fig. 65.14

as the alternate angles formed by parallel lines. Hence, the triangle $OFM$ is an isosceles triangle, and the straight-line segment $OF = OM/2 \cos \alpha = R/2 \cos \alpha$. Therefore, the focal length is

$$f = CF = OC - OF = R - R/2 \cos \alpha = (R/2)(2 - 1/\cos \alpha)$$

Since $\sin \alpha = h/R$, we finally get

$$f = (R/2)(2 - 1/\cos \alpha) = (R/2)(2 - 1/\sqrt{1 - h^2/R^2}) \qquad (65.14)$$

2. As is seen, a spherical mirror suffers from spherical aberration too: the focal length is different for the rays at different distances from the optical axis. Other aberrations, except chromatic aberration, take place as well.

Yet, with a paraxial beam ($h \ll R$), all rays are caused to converge to the same point, and the focal length of the mi ror for them is

$$f = R/2 \tag{65.15}$$

and its focal power is

$$\Phi = 1/f = 2/R \tag{65.16}$$

We leave it for the reader to check that at $h \leqslant 0.1R$ expressions (65.15) and (65.16) are accurate to within 0.5%.

3. From a comparison of (65.16) and (65.11) it is seen that a concave mirror may be treated as a lens which has a refractive index $n_{21} = -1$, one surface of an infinite radius of curvature, and the concave surface, according to the sign convention (see Sec. 65.4), of a radius $R_1 = -R$. Then,

$$\Phi = (n_{21}-1)(1/R_1+1/R_2) = (-1-1)(-1/R+0) = 2/R$$

Hence, we may conclude that all rules of image formation set forth in Sec. 65.5 apply to spherical mirrors, provided reflected rays are taken instead of refracted ones.

For a spherical mirror, the following characteristic rays may be singled out (compare with Sec. 65.5):

—the ray parallel with the principal optical axis (diameter $OC$ in Fig. 65.14), which passes through the focus after reflection;

—the ray passing through the focus, which travels parallel with the principal optical axis after reflection;

—the ray passing through the centre of curvature (the secondary optical axis), which is reflected and returns upon itself.

4. We leave it as an exercise for the reader to construct images of objects formed by a concave and a convex mirror. It may be noted that a concave mirror behaves like a convex lens made from an optically denser material, and a convex mirror like a concave lens.

Reasoning along the same lines as in Sec. 65.6, the reader will readily derive a mirror formula of the form

$$1/d + 1/d' = 1/f = 2/R \tag{65.17}$$

By the same sign convention, the image will be virtual when $d'$ is negative, and real when $d'$ is positive.

## 65.9. FOCUSING OF ELECTRON BEAMS

1. One of the methods of focusing electron beams has been examined in Sec. 41.7. It has been shown that the electrons emitted by a small cathode and propagated in a uniform magnetic field are brought to a focus some distance from the cathode (Fig. 41.5, Vol. I). Thus, a uniform magnetic field acts like a magnetic lens.

It should be noted that this magnetic lens suffers from the same shortcomings as a light lens. This is because it can cause convergence only to the electrons in a paraxial beam, that is, the electrons propagated at an angle, $\alpha$, to the magnetic lines of force such that $\cos\alpha \approx 1$. Obviously, a broad electron beam will be poorly focused. Thus, as regards beam focusing, a magnetic lens is fully identical with a thin lens (see Sec. 65.4) and a spherical mirror (see Sec. 65.8). Similarly, this defect is called the spherical aberration of a magnetic lens.

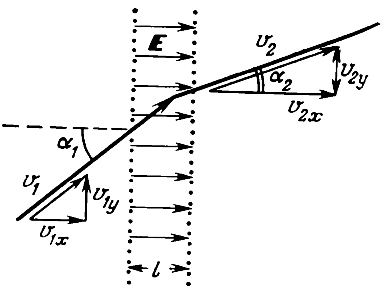


Fig. 65.15

Poor focusing may also be due to the fact that electrons are emitted from the cathode at different velocities. By analogy with light beams, a beam containing electrons differing in velocity is called *non-monochromatic*, and lack of focusing due to the spread in velocity is called the chromatic aberration of a magnetic lens.

2. As an alternative, a beam of electrons can be focused by what is called an *electrostatic lens*. For an insight into its action, we shall discuss the behaviour of an electron passing through an inter-grid space in which a uniform electrostatic field of intensity $E$ is set up (Fig. 65.15).

The electron travels from left to right at velocity $v_1$, strikes the grid at an angle $\alpha_1$, and leaves the field at an angle $\alpha_2$ and at velocity $v_2$. The $y$-component of the velocity remains unchanged, because no forces act on the electron in that direction: $v_{2y} = v_{1y}$. Along the $x$-axis, the electron is acted upon by a force, $F = eE$. By the law of the conservation of energy, for non-relativistic electrons we have

$$mv_1^2/2 + eEl = mv_2^2/2$$

or

$$v_2/v_1 = \sqrt{1 + 2eEl/mv_1^2}$$

However, the work done by the field forces is

$$eEl = eu$$

where $u$ is the difference in potential between the electrodes. Denoting the initial kinetic energy of an electron as

$$K = mv_1^2/2$$

we finally get

$$v_2/v_1 = \sqrt{1 + eu/K} \qquad (65.18)$$

3. Referring to Fig. 65.15, for the sines of the angles of incidence and refraction we have

$$\sin\alpha_1 = v_{1y}/v_1, \quad \sin\alpha_2 = v_{2y}/v_2$$

Since, however, $v_{2y} = v_{1y}$, then, by virtue of (65.18), the law of refraction for an electron beam takes the form

$$\sin\alpha_1/\sin\alpha_2 = v_2/v_1 = \sqrt{1 + eu/K} \qquad (65.19)$$

This expression is fully analogous to that for the law of refraction of light given by (65.2). As with light, dispersion takes place, that is, the refractive index

$$n_{21} = \sqrt{1 + eu/K} \qquad (65.20)$$

varies with the kinetic energy of the electrons in the beam. In other words, slow electrons are refracted more than fast electrons.

This simple scheme is not used for electron beam focusing, the more so that it is difficult to realize experimentally. It has been taken up only to show the refraction of an electron beam in an electric field in simple terms.

4. In practice, use is made of magnetic and electrostatic lenses maintaining a strongly non-uniform field. Their construction and operating principle lie outside the scope of this book, however.

Electrostatic and magnetic lenses are used to focus electron beams in oscilloscopes, television receivers, electron microscopes and similar apparatus.

# Chapter 66

# OPTICAL INSTRUMENTS

## 66.1. PHOTOMETRY

1. Of the total electromagnetic spectrum, the human eye only sees a very narrow region called *visible light* (Sec. 61.1). The eye's sensitivity is different for different wavelengths. It is a maximum at $\lambda = 555$ nm and falls off rapidly to zero on moving away from this maximum corresponding to green light. The plot of Fig. 66.1 shows

the relative spectral sensitivity of the human eye, $K_\lambda$ which is defined as the ratio of the sensitivity at a given wavelength to that at $\lambda = 555$ nm.

Because of this property of the human eye, we are forced to evaluate the quantitative characteristics of light beams in terms of visual sensation rather than in terms of the energy they carry. Consider the quantities and units involved.

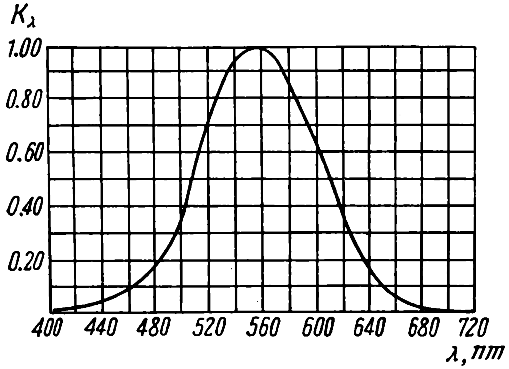


Fig. 66.1

2. In our examination, we shall need one more geometrical concept—that of a *solid angle* which is a measure of the area cut out by a cone on a sphere. As will be recalled, a plane angle $\alpha$ is defined as the ratio of the intercepted arc in a circle, $l$, to the radius, $r$, of that circle, or $\alpha = l/r$ (Fig. 66.2*a*). Similarly, the solid angle, $\Omega$ (Fig. 66.2*b*), can be defined as the ratio of the area of a spherical segment, $\sigma$, to the squared radius of the sphere:

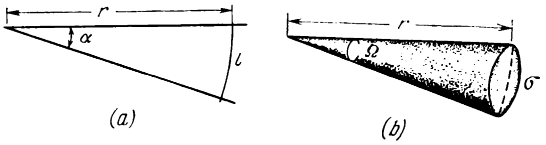


Fig. 66.2

$$\Omega = \sigma/r^2 \tag{66.1}$$

The unit used for measuring a solid angle is the *steradian* (abbreviated to sr). A steradian is a solid angle whose vertex is at the centre of a sphere and which cuts out on that sphere an area equal to the square of its radius:

$$\Omega = 1 \text{ steradian}$$

if $\sigma = r^2$.

It is an easy matter to prove that the total solid angle about a point is $4\pi$ steradians. This result can be obtained by dividing the area of a sphere by the square of its radius.

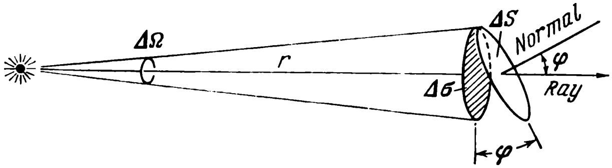


Fig. 66.3

Let a small solid angle $\Delta\Omega$ subtend a small area $\Delta S$ the normal to which makes an angle $\varphi$ with the incident ray (Fig. 66.3). Then the elementary spherical area will be

$$\Delta\sigma = \Delta S \cos \varphi$$

and the solid angle will be

$$\Delta\Omega = \Delta\sigma/r^2 = \Delta S \cos \varphi/r^2 \qquad (66.2)$$

3. *Luminous Flux.* Imagine a uniform point source of light, that is, one whose dimensions are negligible in comparison with the distance to the observer, located at the vertex of a solid angle (see Fig. 66.3) and emitting electromagnetic radiation (*radiant flux*) in all directions. The capacity of this radiation to evoke the brightness attribute of visual sensation in the normal eye is called *luminous flux*, $\Phi$. It is expressed in *lumens* (lm). For monochromatic light corresponding to the visibility maximum ($\lambda = 555$ nm), the luminous flux is 683 lumens if the radiant power is 1 W. For other wavelengths, the luminous flux can be found from the plot of Fig. 66.1.

An incandescent lamp emitting within a wide spectral region has a *luminosity* (the ratio of the luminous flux to the corresponding radiant power) of 14 lumens per watt, and a fluorescent lamp has a luminosity of about 43 lumens per watt.

4. The *luminous intensity*, $I$, of a point source is defined as the ratio of the luminous flux, $\Delta\Phi$, in a given direction to the solid angle $\Delta\Omega$ containing this direction:

$$I = \Delta\Phi/\Delta\Omega \qquad (66.3)$$

If a point source is emitting equally in all directions, then

$$I = \Phi_{total}/4\pi \qquad (66.4)$$

where $\Phi_{total}$ is the total luminous flux emitted by the source.

The unit of luminous intensity is the *candela* (cd), determined with a suitable standard source. In the International System of Units, the candela is a fundamental unit. According to (66.3), we have

1 lumen = 1 candela × 1 steradian

5. The *illuminance*, $E$, of an area is defined as the ratio of the luminous flux, $\Delta\Phi$, to that area, $\Delta S$:

$$E = \Delta\Phi/\Delta S \qquad (66.5)$$

The unit of illuminance is the *lux* (lx). It is defined as the illuminance of an area of 1 square metre over which a luminous flux of 1 lumen is uniformly distributed:

$$1 \text{ lux} = 1 \text{ lumen}/1 \text{ m}^2$$

As an alternative, use may be made of the *phot* defined as a luminous flux of 1 lumen distributed over an area of 1 cm$^2$:

$$1 \text{ phot} = 1 \text{ lumen}/1 \text{ cm}^2 = 10^4 \text{ luxes}$$

If an area is illuminated by a point source, the illuminance at each point on the area may be different. To find the illuminance in such a case, substitute the solid angle, (66.2), in (66.3):

$$I = \Delta\Phi r^2/\Delta S \cos\varphi = Er^2/\cos\varphi$$

whence the illuminance due to a point source is given as

$$E = I \cos\varphi/r^2 \qquad (66.6)$$

This expression is the *inverse square law*. If an area is illuminated by a nearly parallel beam of light, its illuminance is

$$E = E_0 \cos\varphi \qquad (66.7)$$

where $E_0$ is the illuminance that would exist with the normal incidence of light upon the area, and $\varphi$ is the angle of incidence of light (the angle that the ray of light makes with the normal to the surface).

6. *Luminance and Brightness.* If a source is other than a point one, it is described in terms of luminance and brightness (Fig. 66.4).

*Luminance*, $R$, is defined as the ratio of the total luminous flux, $\Delta\Phi$, emitted by a surface element $\Delta S$ in all directions (that is, inside a solid angle of $2\pi$ steradians) to that surface element:

$$R = \Delta\Phi/\Delta S \qquad (66.8)$$

As with illuminance, the units are the lux and the phot.

*Brightness*, $B$, in a given direction is defined as the ratio of the luminous intensity, $\Delta I$, within an elementary solid angle enclosing a surface element $\Delta S$, to the area $\Delta\sigma$ of its projection on a plane normal to the direction of the ray:

$$B = \Delta I/\Delta\sigma = \Delta I/\Delta S \cos\varphi \qquad (66.9)$$

For most sources, the brightness is different in different directions. It is the same in all directions only for a black-body radiator (see Sec. 67.2) and perfect diffuse radiators such as frosted glass. These sources are said to obey the Lambert law

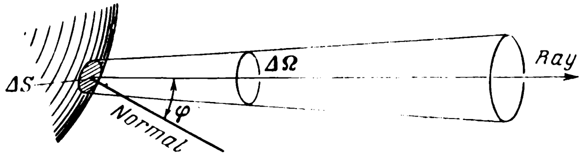


Fig. 66.4

$$B = R/\pi \qquad (66.10)$$

The unit of brightness is the *candela per square metre*. It is defined as the luminance of a surface emitting or reflecting light at the rate of one candela per square metre in a direction normal to that surface. Previously, it was known as the *nit*: 1 nit = 1 cd/m$^2$. Use is also made of the *stilb*:

$$1 \text{ stilb} = 1 \text{ cd}/1 \text{ cm}^2 = 10^4 \text{ cd/m}^2$$

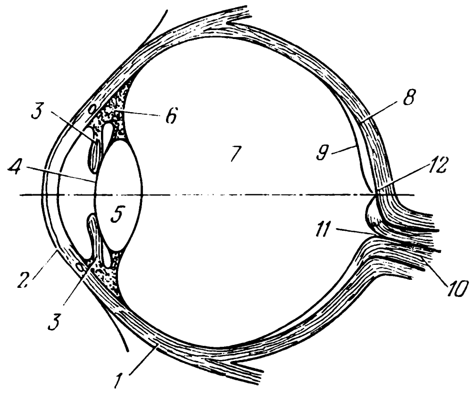


Fig. 66.5

## 66.2. THE HUMAN EYE

The organ that enables man to form visual images is called the eye. In diagrammatic form, it is shown in Fig. 66.5.

The eyeball is enclosed in the *sclera*, *1*, which protects the insides of the eye and gives it the necessary stiffness. At the front, the sclera extends into a thin and transparent *cornea*, *2*, acting as an entrance lens for the eye. Back of the cornea is the *iris*, *3*, a pigmented muscular ring which can expand and contract to control the size of its central opening called the *pupil*, *4*, and as a result, the amount of light that can enter the eye. Back of the iris is the "*crystalline*" *lens*, *5*, flexible like rubber and suspended by a *ciliary muscle*, *6*, which can be stretched taut or slackened to vary the curvature of the "crystalline" lens and, as a consequence, its focal power (see Sec. 65.11). The chamber in front of the lens and back of the cornea is filled with a watery fluid called the *aqueous humour*, while the chamber behind the lens contains a jelly called the *vitreous body* or *humour*, *7*. The cornea, the aqueous humour, the "crystalline" lens and the vitreous humour make up an optical system equivalent to a lens with a focal power of about 58.5 diopters ($f = 17.2$ mm). The optic centre of the system is about 5 mm distant from the cornea; in Fig. 66.5 the optic axis of the image is shown by the dashed line.

The chamber holding the vitreous body is lined with a sensory layer, the *retina*, *9*. It is the shape of a hemisphere and contains light receptors called rods and cones. The human eye has a total of 125 million rods and 6.5 million cones. These light-sensitive cells are located on the rear surface of the retina supported by a *vascular layer*, *8*. Sidewise to the optic axis, the nerve cells of the retina converge to the *optic nerve*, *10*, entering the eye. At the point where the optic nerve enters the eye, there are no rods or cones, and this portion of the retina is called the *blind spot*, *11*. In contrast, a depression in the retina at the axis of the eyeball, called the *fovea centralis*, *12*, is the point of most acute vision. Here, cones responsible for colour vision are concentrated. The remaining areas of the retina are occupied mainly by rods.

2. In rods, light entering the eye causes the visual purple, a compound of a form of vitamin A (retinen) and the protein of the retina (opsin), called rhodopsin, to change from the cis- into the trans-isomer. Owing to this reaction, each rod generates a nerve impulse which is conveyed by the optic nerve to the brain. The impulses are generated owing to the energy stored by the light receptors, and light stimulation only triggers the reaction. This explains why rods are highly sensitive to light—each rod is capable of responding to a unit quantum of light (see Sec. 68.3).

Rods are responsible for the so-called scotopic (seeing in the dark) vision which perceives only the size and shape and not the colour of objects.

3. Colour is perceived by daylight vision which is the function of cones. This is why colours are perceived only if the image of an object is thrown onto the central area of the retina.

A consistent theory of colour vision is still lacking. Probably there are at least three different kinds of cones, some perceiving green, others red, and still others blue. Intermediate colours are perceived when two or three types of cones are stimulated simultaneously. Depending on the amount of excitation each type of cones receives, different trains of pulses reach the brain, and it interprets them as different colours.

### 66.3. ACCOMMODATION. BINOCULAR VISION

1. The eye must see objects at different distances equally well. Whatever the object distance, $d$, a sharp image must be formed on the retina. According to the lens equation, (65.13), this is possible only if the focal length of the optical system, $f$, changes in proportion. As noted in the previous section, the focal power, $\Phi$, and focal length, $f = 1/\Phi$, of the eye are adjusted by varying the curvature of the crystalline lens. This is known as *accommodation*.

The eye accommodates itself involuntarily. As the eye is shifted from one object to another, the sharpness of the image is upset, and an appropriate signal is sent to the brain. The response from the brain to the ciliary muscle causes it to contract or extend until the image is again brought into a sharp focus. The nearest point on which the eye is focused when the ciliary muscle is fully relaxed is called the *far point of the eye*. In contrast, the nearest point on which the eye is focused with the ciliary muscle fully contracted is called the *near point of the eye*. For the normal eye, the far point is at infinity, and the near point is about 15 to 20 cm distant.

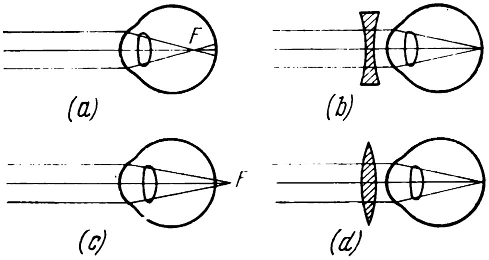


Fig. 66.6

2. In a near-sighted eye, the far point is at a finite distance; in fact, the distance may be very short in a very near-sighted eye. The near point moves in, too (Fig. 66.6*a*). This is why a near-sighted person will hold the object he wants to see close to his eyes. The fault may be in an eyeball that is too long front to back or in too great a convexity of the lens because of a spasmodic contraction of the ciliary muscle. The remedy is to wear spectacles with diverging (concave) lenses as shown in Fig. 66.6*b*.

In a far-sighted eye, the near point is moved away (Fig. 66.6*c*). The cause may be too short (shallow) an eyeball or the failure of the eye to accommodate itself properly.

As a rule, far-sightedness is a companion of old age, when the crystalline lens loses much of its elasticity, although the defect may be in-born. It may be corrected by means of glasses having convex lenses (Fig. 66.6*d*).

3. We have two eyes, and each sees the same thing a little differently, so that two different images are formed on the retinas. Through long experience, the brain blends the two images and interprets the resulting sensation so that the scene stands out in solid relief, or stereoscopically. The perception of depth in space occurs because in setting the eyes on the same thing we cause the muscles to turn the eyeballs so that their optic axes converge on the thing, with one eye seeing the thing more from the left and the other from the right.

The angle between the optic axes of the eyes is called the *angle of convergence*, $\alpha$. The separation between the pupils of the eyes (interpupillary distance) is $b = 5$ cm and the object distance is $d > 25$ cm. Thus, the angle of convergence $\alpha \approx b/d$ can vary from zero (for the far point) to 10° (for the near point).

Accommodation coupled with convergence of the eyes enables us to get an accurate estimate of distance and to perceive depth in space better and more accurately than with one eye. This faculty may be further enhanced by using binoculars or stereoscopic range finders which increase the interpupillary distance artificially.

## 66.4. ANGLE OF VIEW. RESOLVING POWER OF THE EYE

1. The size of the image formed on the retina is solely determined by the angle of view, $\varphi = h/f$, with its vertex at the optic centre of the eye and with the sides pointing at the extreme points of the object (Fig. 66.7). The angle of view may be increased by moving the object closer to the eyes. This, however, overstrains the ciliary muscle, and the eye experiences fatigue. The eye is especially difficult to accommodate when an object is placed close to the near point.

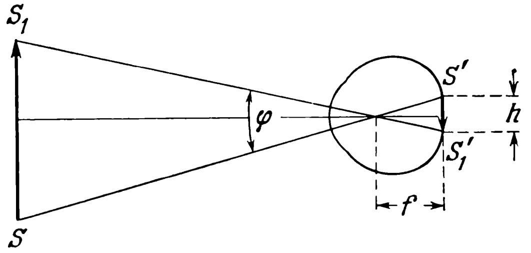


Fig. 66.7

The distance of the maximum visual acuity is the object distance at which the angle of view is a maximum while the effort of accommodation is small enough not to cause eye strain. For a normal eye, the distance of maximum visual acuity is about 25 cm. The distance is shorter for near-sighted persons and longer for far-sighted individuals.

2. Two points on an object will be perceived separately if their images are formed on different light receptors of the retina. Otherwise, light from them will stimulate only one sensory cell, and the two points will be seen as one. In such a case, the eye is said *not to resolve* (that is, separate) the two points. Conversely, the ability of the eye to separate the images of two points which are close together is called the *resolving power* of the eye. It is estimated in terms of the minimum angle of view, $\varphi_0$, at which two points adequately illuminated are seen separately.

3. According to experiments, the minimum angle of view is about one minute of arc ($\varphi_0 \approx 1'$) for an illuminance of 5 luxes. This checks well with the fact that the separation between two adjacent rods or cones is about 5 μm ($h_0 \approx 5 \times 10^{-3}$ mm). Referring to Fig. 66.7, the minimum angle of view is $\varphi_0 = h_0/f$, where $f = 17.2$ mm is the focal length of the eye. Hence,

$$\varphi_0 \approx h_0/f \approx 5 \times 10^{-3} \times 180 \times 60 \div 17.2\pi = 1'$$

The resolving power of the eye decreases with decreasing illumination, that is, the eye loses some of its *visual acuity*. Visual acuity is a quantity which is the reciprocal of the minimum angle resolved at a given illuminance and expressed in minutes:

$$B = 1/\varphi_0$$

It ranges from 0.3 at an illuminance of less than 0.1 lux to 1.3 at an illuminance of over 100 luxes.

## 66.5. THE MAGNIFIER

1. The simplest optical instrument that increases the angle of view is the *magnifier*, a low-power lens placed between the eye and the object being viewed as shown in Fig. 66.8.

With an unaided eye, we can view a small object $AB = h$ placed at the distance of maximum visual acuity, $D = 25$ cm, at an angle of view, $\varphi_0$, the tangent of which is $\tan \varphi_0 = h/D$. If the same object be placed near the focus of a magnifier, the eye will see it at an angle of view $\varphi$ which is determined by the condition $\tan \varphi = h/f$ where $f$ is the focal length of the lens. As a result, the image $ab$ of the object formed on the retina of the eye aided by the magnifier will be greater than the image formed on the retina of the unaided eye. It will appear as if we see a large object, $A_1B_1$, and not the small object $AB$.

2. The ratio of the tangent of the angle $\varphi$ at which an object is seen in a magnifier to the tangent of the angle $\varphi_0$ at which the same

object is seen by an unaided eye at the distance of maximum visual acuity is called the ***angular magnification***, $\gamma$, of the magnifier. Since, however, $\tan \varphi = h/f$ and $\tan \varphi_0 = h/D$, it follows that

$$\gamma = \tan \varphi / \tan \varphi_0 = D/f \tag{66.11}$$

In practice, use is made of magnifiers with focal lengths from 10 cm to 1 cm. They give a magnification of 2.5 to 25 diameters.

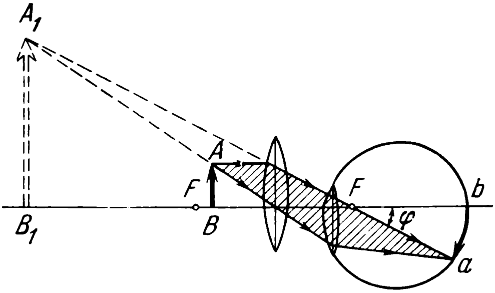


Fig. 66.8

Since short-focus magnifiers suffer from spherical aberration, astigmatism and distortion (see Sec. 65.7), magnifiers are usually limited to a magnification of five to ten diameters.

## 66.6. THE MICROSCOPE

1. Large angular magnifications (of the order of several hundred) are provided by *microscopes*. A microscope is a combination of two short-focus lens systems, an *objective* and an *eyepiece*, or *ocular* (Fig. 66.9).

The object $h$ being viewed is placed near the focus, $F_1$, of the objective; a real image, $H$, is formed beyond the objective, near the focus, $F_2$, of the eyepiece. It is an easy matter to see that the distance from the first image to the focus of the objective is approximately equal to the distance between the foci of the objective and eyepiece. Using this relation, we can readily determine the linear dimension of the first image. Referring to Fig. 66.9,

$$H/h \approx \Delta/f_{ob} \tag{66.12}$$

2. The angular magnification of a microscope can be found in the same manner as that of a magnifier. Then,

$$\gamma = \tan \varphi / \tan \varphi_0 = H/f_{oc} \div h/D = h\Delta/f_{oc}f_{ob} \div h/D$$
$$= D\Delta/f_{oc}f_{ob} \tag{66.13}$$

For a good-quality microscope, $f_{ob} \approx 2.5$ mm, $f_{oc} \approx 15$ mm, and $\Delta \approx 160$ mm. Noting that $D \approx 250$ mm, we get

$$\gamma \approx 160 \times 250 \div (2.5 \times 15) \approx 1000$$

As will be shown in Sec. 66.8, nothing is gained by increasing the magnification of microscopes above 1000; as a rule, the limit is set at 500 to 600 diameters.

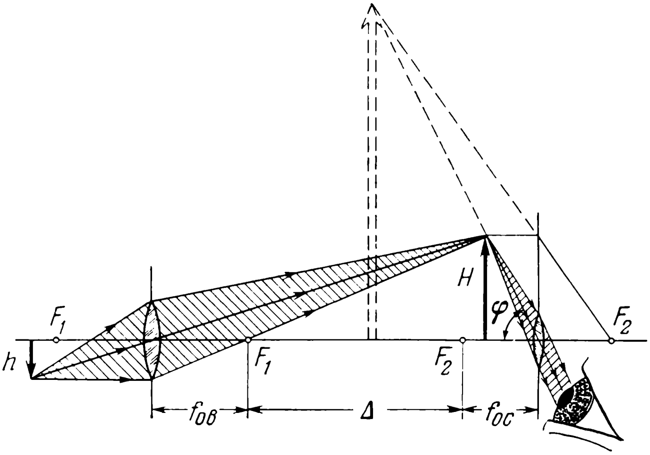


Fig. 66.9

## 66.7. THE TELESCOPE

1. The function of a telescope is to increase the angle of view when distant objects are watched. A great number of telescope designs exist; we shall limit ourselves to the *Keplerian telescope*, more commonly known as the *astronomical telescope* (Fig. 66.10).

Let the object $AB$ being viewed be far away from the objective with a focal length $f_{ob}$, so that the point $B$ is on the optical axis of the system, and the point $A$ above the axis. The unaided eye sees the object at an angle of view $\varphi_0$. The image of the object, $A_1B_1 = h$, is formed practically in the focal plane of the objective. Now we shall place the eyepiece with a focal length $f_{oc}$ so that its front focus coincides with the rear focus of the objective. Then the eyepiece will act as a magnifier, and a parallel beam of light will enter the eye at an angle of view $\varphi > \varphi_0$.

Hence, the angular magnification of the telescope is

$$\gamma = \tan \varphi / \tan \varphi_0 = h/f_{oc} \div h/f_{ob} = f_{ob}/f_{oc} \qquad (66.14)$$

In order to obtain high magnifications, telescopes use objectives with a long focal length and eyepieces with a short focal length.

2. All of the luminous flux emerging from the eyepiece should preferably reach the retina. It is therefore necessary to match the diameters of the objective and the eyepiece so that the light beam leaving the telescope will fill all or a greater part of the eye's pupil.

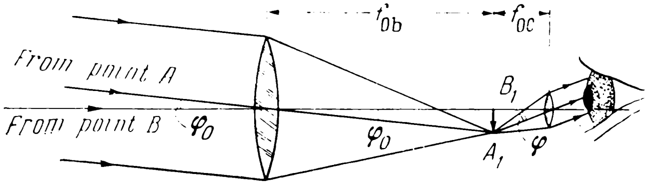


Fig. 66.10

If the light beam is wider than the pupil, some of the emergent beam will be scattered by the iris, and only part of the objective will actually form an image. Approximately, the diameter of the eyepiece, $D$, may be set equal to that of the pupil. For night observations, this works out to $D_{oc} \approx$ 6 to 8 mm, and for daylight observations, 2 or 3 mm. Hence, we can find the diameter of the objective. Referring to Fig. 66.10,

$$\gamma = f_{ob}/f_{oc} = D_{ob}/D_{oc} \tag{66.15}$$

For example, for a magnification of 20X, a telescope for night observations should have an objective with a diameter $D = 20 \times 8 \text{ mm} = 160 \text{ mm}$. Since the focal length of the eyepiece is $f_{oc} \approx 20$ mm, it follows from (66.15) that $f_{ob} = \gamma f_{oc} = 20 \times 20 \text{ mm} = 400 \text{ mm}$. Thus, the overall length of the telescope will be 420 mm.

3. At present, there are two broad classes of telescopes, refracting and reflecting. A *refracting telescope* uses a lens for gathering light and forming a real image of a distant object, and a *reflecting* telescope uses a concave mirror for the same purpose.

Refracting telescopes use objectives with a diameter of up to one metre. Lenses of a larger diameter are difficult to fabricate. It is much easier to cast a glass blank (a boule) for a mirror. A boule need not be optically homogeneous as is the case with that for a lens, and one needs to grind only one side. In large-diameter mirrors, the reflecting surface is ground to a parabolic rather than spherical shape in order to reduce spherical aberration, and is given a coat of aluminium in a vacuum.

The first reflecting telescope was built by Newton in 1671-72. Its optical arrangement is shown in Fig. 66.11. A parallel beam of light from a distant object falls on a mirror, $M$. On being reflected from the mirror and from another mirror $C$ which turns the light beam through 90°, the beam is focused at the point $F$ where it forms a real image. The eyepiece $O$ operates as it does in a refracting telescope.

The world's largest reflecting telescope with a mirror 5 m in diameter and 16.5 m long is installed at the Mount Palomar observatory in the United States. In the USSR, work is under way to make a 6-metre mirror for a still bigger telescope.

4. The name telescope implies that it is a device for "seeing at a distance". Let us see how it does its job in various cases. When one views a distant object on the earth or a planet, a telescope increases the angle of view and this brings about an increase in the resolving power. An unaided eye can only see a few dark spots on the Moon and perceives the planet Mars as a reddish point. Using a 5-metre reflecting telescope, one can discern details as small as 1 m across on the Moon and down to 100 m on Mars. Yet the brightness of the objects remains the same as when they are viewed with an unaided eye.

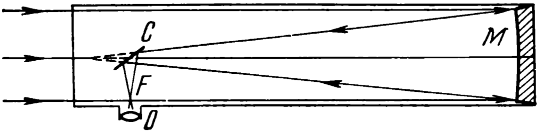


Fig. 66.11

The situation is different when stars are viewed in a telescope. The stars are extremely distant, and their angle of view is so small that even after magnification by a telescope it remains smaller than the angle resolved by the human eye, $\varphi_0 \approx 1'$ (see Sec. 66.4). As a result, even with the most powerful telescope, the image of a star is formed on a single sensory cell of the retina, and we perceive it as a luminous point. However, the illuminance of the point is increased over that perceived by an unaided eye in proportion to the ratio of the objective area to the pupil area. During daytime observations with a 5-metre objective this increase is about $(5000\ \text{mm}/3\ \text{mm})^2 \approx 3 \times 10^6$. As a result, the illuminance of the retina increases several million times. This is why a telescope can bring out very faint and extremely distant stars not visible to an unaided eye.

### 66.8. RESOLVING POWER OF AN OPTICAL INSTRUMENT

1. So far it has been tacitly assumed that the image of a luminous point in an optical instrument (say, a lens) is likewise a point. Strictly speaking, it is not so, even though all aberrations have

been corrected. This is because a lens cuts a limited area out of the wave front, and the image of the point is a complex diffraction pattern (see Secs. 57.8 and 57.9). The principal maximum occurs at the centre surrounded by alternate bright and dark rings. The first minimum is observed at an angle determined by (57.40). Setting $m = 1$, we get

$$\sin\theta = \lambda/D \qquad (66.16)$$

where $\lambda$ is the wavelength of the incident light and $D$ is the diameter of the lens.

Let a lens be illuminated by light from several distant point sources. If the angular separation between any two sources is small, the diffraction patterns due to them will partly overlap, and it may so happen that we cannot tell one from the other (Fig. 66.12 shows photographs of three point sources. In *a*, two of them are seen to merge together.). In such a case, the lens is said not to *resolve* (*separate*) the images of two points. Any amount of subsequent magnification will not help; if the images of two points are not resolved by at least one lens in a system, they will remain unresolved. The resolving power can, however, be increased by using a lens of a larger diameter (Fig. 66.12*b* and *c*).

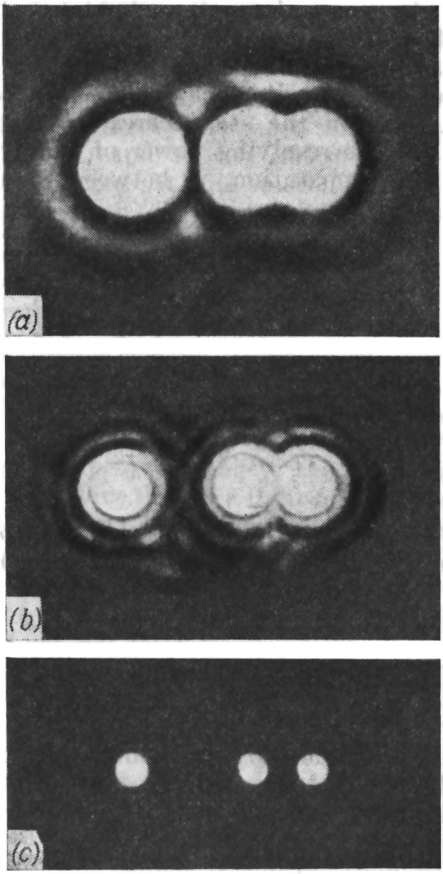


Fig. 66.12

2. In quantitative terms, the resolving power of the objective of an optical instrument can be evaluated, using Rayleigh's criterion (see Sec. 62.4) which states that the images of two points will be resolved if the principal maximum of one pattern falls on the first minimum of the other. Then the angle $\varphi$ between the directions of the two point sources should not be less than the angle $\theta$ set by (66.16). Hence,

$$\sin\varphi \geqslant \sin\theta = \lambda/D$$

Since the lens diameter is always a fraction of the light wavelength it follows that the angles $\varphi$ and $\theta$ are very small, and the sines may be replaced with the angles in radians

$$\varphi \geqslant \theta \approx \lambda/D \qquad (66.17)$$

As is seen, large-diameter telescopes have a high resolving power, $A = 1/\varphi \approx D/\lambda$. The minimum angular resolution is $\varphi_{\min} \sim 10^{-7}$ radians $\approx 0.02$ second of arc. This angle is subtended by a match seen from a distance of 600 km. Yet this is not enough to resolve details on the star nearest to the Earth.

3. The resolving power of a microscope is customarily expressed as the separation, $\varepsilon$, between the nearest points that can still be seen separately. A rigorous theory of the matter is outside the scope of the book; in approximate terms we may reason as follows.

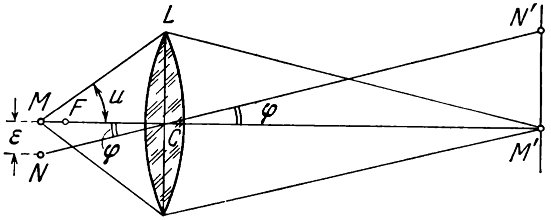


Fig. 66.13

Let two points, $M$ and $N$, set a distance $\varepsilon$ apart, be placed near the focal plane of an objective (Fig. 66.13). Their images, $M'$ and $N'$, will be seen separately if the condition (66.16) is satisfied. Setting the object distance $MC = d$, we have

$$\varepsilon = d \tan \varphi > d \sin \varphi \geqslant \lambda d/D \qquad (66.18)$$

The angle $LMC = u$ subtended by the radius of the lens at point $M$ is called the *aperture angle* or *angular aperture*. Referring to the figure,

$$R/d = \tan u$$

where $R$ is the lens radius. Substituting in (66.18) gives the minimum resolvable distance as

$$\varepsilon \geqslant \lambda/2 \tan u \qquad (66.19)$$

4. By a rigorous theory, the sine, and not the tangent of the aperture angle, must be used:

$$\varepsilon \geqslant \lambda/2 \sin u \qquad (66.20)$$

In a good-quality microscope, the object is placed near the focal plane, that is, $d \approx f = 2$ mm; the objective radius is $R \approx 2$ or

3 mm. Then, $\sin u \approx 0.9$, and the microscope can resolve details about a half-wavelength in size.

If the space between the objective and the lens be filled with *immersion oil*, the numerator of (66.20) will be the wavelength of light in the oil, $\lambda = \lambda_0/n$, where $\lambda_0$ is the light wavelength in a vacuum and $n$ is the refractive index of the oil. Then

$$\varepsilon \geqslant \lambda_0/2n \sin u \qquad (66.21)$$

At $n = 1.5$ (cedar oil) and $\sin u = 0.9$, the minimum separable distance will be $\varepsilon \geqslant 0.37\ \lambda_0$. The human eye is most sensitive to light at $\lambda = 555$ nm (see Sec. 66.1), and so $\varepsilon \approx 200$ nm. Since a living cell measures over 1000 nm across, it will be seen in a light microscope, while viruses measuring from 275 nm to 10 nm cannot; they are investigated in electron microscopes.

## 66.9. THE ELECTRON MICROSCOPE

1. The resolving power of a microscope can be increased by using radiation of progressively shorter wavelengths. For example, ultraviolet microscopes using quartz optics operate in the ultraviolet region, that is, at wavelengths of about 250 nm which is half as long as the wavelength of visible light. Accordingly, the resolving power of an ultraviolet microscope is twice that of a light microscope.

A further marked increase in resolving power could be obtained with X-rays. Unfortunately, the refractive index (see Eq. (63.15) at $\omega \rightarrow \infty$) for these rays is practically unity, and mirrors and lenses that could be used in the X-ray region cannot be built.

2. As will be shown in Chapter 69, the particles of matter display wave-like properties. The theoretical value of the wavelength for an electron has been worked out by de Broglie to be (see Eq. 69.2)

$$\lambda = h/mv$$

where $h$ is Planck's constant, $m$ and $v$ are the mass and velocity of the electron. This is a very short wavelength. For example, the wavelength of electrons accelerated in an electric field with a potential difference of 1000 V is $\lambda = 0.4$ Å, which is shorter than the wavelength of X-ray. Unlike X-rays, electron beams can be readily controlled; they may be focused with electrostatic or magnetic lenses (see Sec. 65.9). This is the basis of the electron microscope.

3. The arrangement of a magnetically focused electron microscope is shown in Fig. 66.14, in comparison with that of a light microscope. The eyepiece of the electron microscope forms a real image which can be photographed or observed on a fluorescent viewing screen.

While the sine of the aperture angle of a light microscope is about 0.9, that of an electron microscope is 0.01 or 0.02. Therefore, the limit of resolution for an electron microscope is of the order of 10 to 20 Å. Although atoms and small molecules still lie outside the limits of vision, larger protein molecules such as viruses can be observed (see Fig. 33.14).

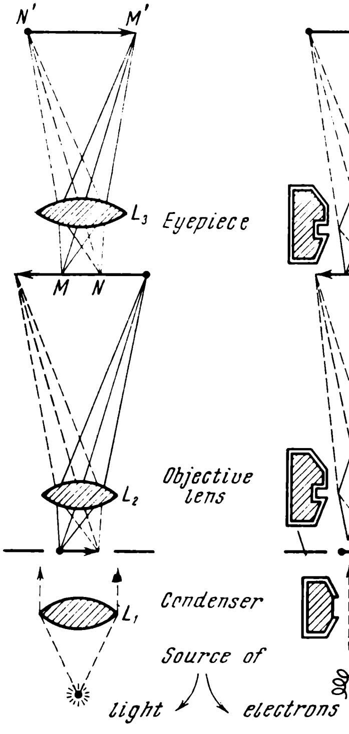


Fig. 66.14

## 66.10. THE PHOTOGRAPHIC CAMERA. PROJECTORS

1. A lens can form a real image of an object that can be seen on a photographic plate or a viewing screen (see Fig. 65.8). This is utilized in photographic cameras, slide and cine projectors, and similar applications.

A photographic camera is optically very similar to the human eye. Its lens forms a reduced real image of the scene on photographic film or plate. A diaphragm is provided for varying the size of the lens opening and so regulating the brightness of the image. The image is brought into a sharp focus by extending or retracting the lens (Fig. 66.15). The action of the light on the film produces molecular changes in the sensitive material. Subsequent chemical treatment (development) brings out a visible image which is a negative of the original scene (it is dark where the original was light, and light where it was dark). Finally, a positive print is obtained by passing uniform light through the negative onto sensitized paper which is chemically treated to give the finished picture.

2. A slide projector forms an enlarged image of drawings, sketches or photographs on a viewing screen. With transparent slides, or transparencies, the process is called *diaprojection* (from the Greek "dia" for "through"). With opaque slides, the process is called *epiprojection* (from the Greek "epi" for "surface").

3. The eye has what is known as the *persistence of vision*. The sensation in the retina persists for about 0.1 s after the stimulus is removed. If the stimuli follow at intervals less than 0.1 s, the visual sensation will be a continuous one. This is the basis of the motion picture, or cinematograph.

Pictures of a moving object are made in rapid succession (usually, 24 frames per second) on a continuous length of film. Each picture records the position of the subject at the instant of exposure. When these same pictures are projected in the same order and at the same rate, they will merge into one another and the movements of the subject will be reproduced.

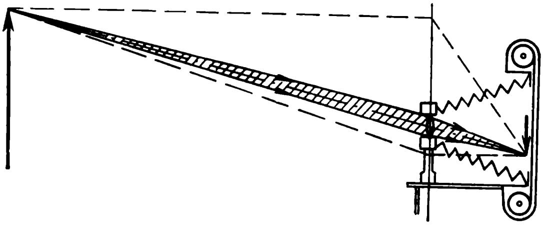


Fig. 66.15

Motion pictures are shown by means of motion-picture, or cine, projectors. They differ from slide projectors in that they have a film-transport mechanism which advances the film at 24 frames per second. As the film is advanced one picture at a time, the projection lens is closed by a shutter (at $S$ in Fig. 66.16).

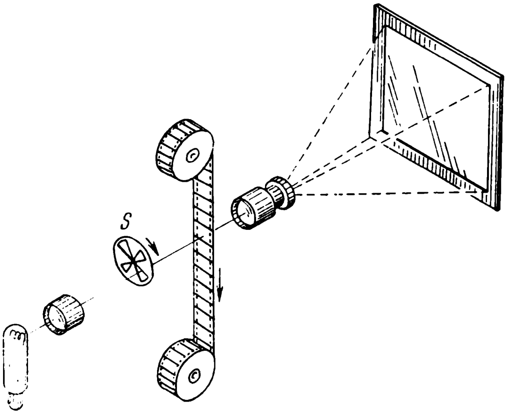


Fig. 66.16

Sometimes one may need a slow-motion record of a very fast process or a fast-motion record of a process that takes an extremely long time to complete. This is done by special photographic techniques. Slow-motion records are obtained by what is known as *high-speed photography*; with this technique, pictures are taken up to several thousand frames per second and projected at the normal rate (24 frames per second). As a result, an action normally occurring in one second can be expanded into a clearly followed "slow-motion" sequence of about three minutes. The observer may well follow the break-up of an armour plate by an artillery shell.

A fast-motion record of a very slow process is produced by so-called *time-lapse techniques*. With time-lapse photography, processes that normally proceed very slowly are photographed at fixed intervals so that the action of several hours, days or weeks can be condensed into a smoothly flowing sequence of several minutes. As an example, this technique can be used to follow the opening of a flower from a closed bud to full blossom.

### 66.11. THE FIELD ION MICROSCOPE

1. As already noted, the resolving power of an electron microscope is 10 to 20 Å. Finer details, the size of an atom, (1.2-1.6 Å) can be investigated with the *field-ion microscope* invented by E. Müller in 1950-51.

In sketch form, the instrument is shown in Fig. 66.17. A small amount of helium, hydrogen or neon is admitted into a flask where a pressure of $10^{-5}$ mm Hg and a temperature of 20 to 40 K is maintained. The atoms are subjected to the action of a strongly non-uniform electric field set up between a sharp field-emission tip prepared from the metal under investigation and rounded to a radius of 1000 Å, and a viewing screen 10 cm distant from the tip. The electric field intensity near the tip is about $2.2 \times 10^{10}$ to $4.5 \times 10^{10}$ V/m.

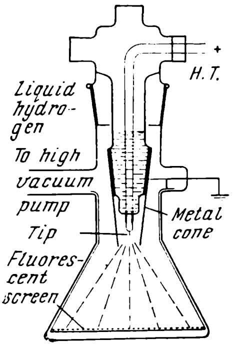


Fig. 66.17

In the field the atoms are polarized, that is, they come by an induced dipole moment owing to which they are pulled into the strong-field region (see Sec. 38.5). On reaching the tip, the atoms are field-ionized, the electrons tunnelling into the tip and the ions being driven to the screen. Gaining kinetic energy, the ions strike the phosphor screen, yielding a highly magnified electrical image of the tip.

2. The surface of the field-emission tip is not perfectly smooth. In the direction of the crystallographic axes the curvature of the surface is different from that in other directions. This difference stems from the anisotropy of crystals (see Sec. 32.1). Besides, the

atoms are ionized differently in different areas of the tip. As a result, different numbers of ions strike different areas of the screen, producing pattern detail, and the pattern is characteristic of the crystal structure.

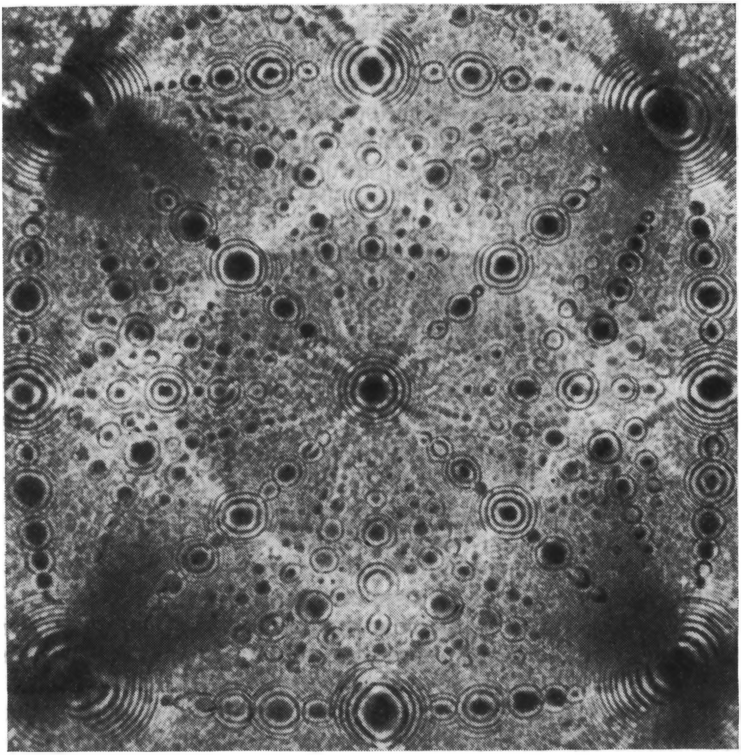


Fig. 66.18

Fig. 66.18 is a photograph of the crystal structure for a platinum tip with a radius of 2000 Å, obtained with a field-ion microscope. As is clearly seen, platinum has a face-centred cubic structure.

From a comparison of field-ion photomicrographs with X-ray diffraction analysis data (see Sec. 62.7), a further insight can be gained into the crystal structure. The field-ion microscope is at present the only tool available for detecting lattice (especially, point) defects such as vacancies, interstitials, and substitutions (see Sec. 32.4) directly undetectable by other means.

## 66.12. SPECTROSCOPY

1. A *spectroscope* is an instrument for producing and viewing the spectrum of the light emitted by a source. If the instrument produces a permanent record of the spectrum on, say, a photographic plate it is called a *spectrograph.*

The light emitted by the substance under investigation is separated into a spectrum (see Sec. 50.4) either by a diffraction grating (see Secs. 62.2 through 62.4) or by a prism (see Secs. 63.4 and 65.3). Spectroscopic instruments intended for studies in the visible region use glass optics, and those for investigations in the ultraviolet and infrared use optics from quartz, fluorite or rock salt.

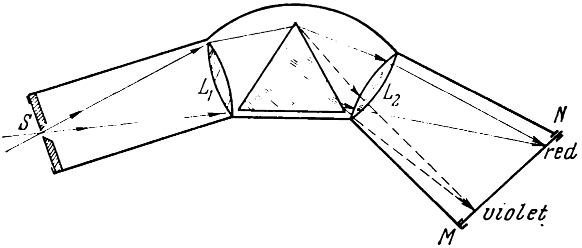


Fig. 66.19

2. The essential features of a spectrograph are shown in Fig. 66.19. These are a collimator, $S$-$L_1$, on the left of the sketch, a slit, $S$, arranged to be in the focal plane of an objective, $L_1$, and a telescope between $L_2$ and $MN$ for viewing the spectrum at an eyepiece. The collimator renders the light from the slit parallel before it enters the prism. Owing to dispersion, the parallel beams emerging from the prism are deviated through various angles according to the wavelength. The emergent beams form a multiplicity of slit images in the focal plane of another lens, $L_2$, each image corresponding to a particular wavelength.

If the light incident on the slit is a mixture of several monochromatic waves, a *line spectrum* appears on the plate $MN$—a series of narrow lines separated by dark intervals. If, on the other hand, the slit is illuminated by white light, a *continuous spectrum* is produced with a gradual transition from one colour to another.

3. In spectroscopes, the lens $L_2$ usually has a short focal length and forms a real image of the spectrum in its focal plane $MN$, viewed in an eyepiece. Also placed in this plane is an index coupled to a micrometric screw and a graduated knob for calibrating the spectroscope against standard light sources.

## 66.13. HOLOGRAPHY

1. In 1948 D. Gabor came out with a fundamentally new photographic technique he called wavefront reconstruction. This invention won Gabor the Nobel prize for physics in 1971.

His wave-front reconstruction technique boils down to the following.

The images obtained with conventional optical instruments (a photographic camera, a magic lantern, a motion-picture projector and the like) register only the intensity of a wave, that is, the square of its amplitude (see Sec. 55.3), while the wave phase is lost. With Gabor's technique, both the frequency and phase attributes of a wave are recorded as a blurred diffraction pattern; this is the first step. In the second step, the blurred photograph is inserted in an optical device which is capable of restoring it into a reconstructed image of the primary object. Since the blurred diffraction pattern contains almost all the available information concerning the sample, it is called a *hologram* (from the Greek "holos" for "total" and "gramma" for "record"), and the technique has finally come to be known as *holography*.

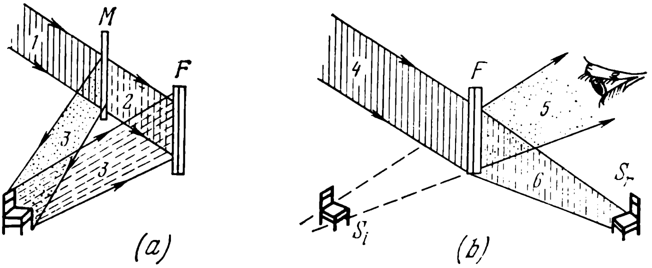


Fig. 66.20

2. To obtain a hologram (Fig. 66.20*a*), a beam of light, *1*, is thrown on a semitransparent mirror, *M*, which splits the beam into two beams, a reference beam, *2*, and an object beam, *3*. The reference beam is allowed to reach a photographic plate *F* directly; the object beam illuminates the object, or scatterer, *S*. Some of the scattered light reaches the photographic plate and interferes with the reference beam, producing a diffraction pattern which is caught by the photographic emulsion. This is a hologram.

It should be noted that a hologram is not a distinct image of the primary object; it consists merely of a confusing mixture of diffraction fringes and lines only roughly resembling the details of the specimen and looking more like a system of Newton rings (see Fig. 61.4 or 61.7).

3. It should be stressed that there is a marked path length difference between the beams *2* and *3*; it may be from tens of centimetres to a few metres. This poses certain difficulties in making holograms.

The semitransparent mirror splits the incident beam into two beams which should produce an interference pattern on re-uniting. But an interference pattern can only be produced if the two beams meeting at a given point in space belong to the same original wave train. This relation should be maintained over the time interval comparable with the propagation time of the wave train, while the length of the wave train, $L$, should be tens or even hundreds of times the path length difference, $\Delta$ (Fig. 66.21*a*). Then the wave trains in the two beams will arive at the point of observation in phase, the waves will remain coherent, and an interference pattern will be produced.

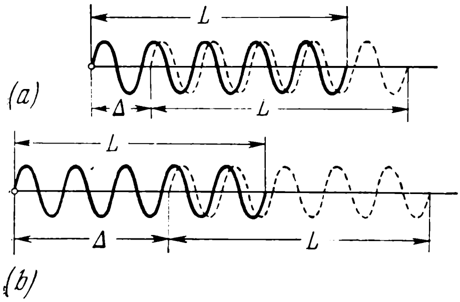


Fig. 66.21

If, on the other hand, the path length difference is very close to the length of a wave train (Fig. 66.21*b*), the wave trains in the reference beam will arrive independently of those in the object beam, and no interference pattern will be produced.

Setting $\Delta \approx 1$ m, $L \approx 30\ \Delta \approx 30$ m, we get for the time of wave-train emission

$$\tau \approx L/c \approx 30 \div (3 \times 10^8) = 10^{-7}\ \text{s}$$

According to (61.8), the uncertainty in frequency will then be

$$\Delta\omega \approx 1/\tau \approx 10^7\ \text{s}^{-1}$$

and the fractional uncertainty in frequency, (61.9), at a light frequency of $\omega \approx 10^{15}\ \text{s}^{-1}$, will be

$$\Delta\omega/\omega \approx 10^7/10^{15} \approx 10^{-8}$$

From a comparison with Secs. 61.4 through 61.6 it is seen that the coherence and monochromaticity of conventional light sources is insufficient for making holograms. This is why holography was dormant, in Gabor's words, for over a decade. It was not until 1962-63 that a new impetus was given to holography with the advent of lasers (see Sec. 79.4) capable of emitting highly monochromatic light with a wave train length several thousand times that of wave trains from conventional light sources (say, mercury-vapour lamps). The coherent light from lasers produces high-quality holograms.

4. The set-up used for image reconstruction from a two-beam hologram is shown in Fig. 66.20*b*. The reconstruction beam of coherent light, *4*, is incident on the hologram at the same angle as the reference beam, *2*, fell on the recording plate. Scattered by the interference rings recorded on the hologram, the light is converted into two beams, diverging, *5*, and converging, *6*.

The converging beam, *6*, forms a real three-dimensional image, $S_r$. As is seen from the figure, a major drawback is that it is a mirror image of the object, which is not always convenient.

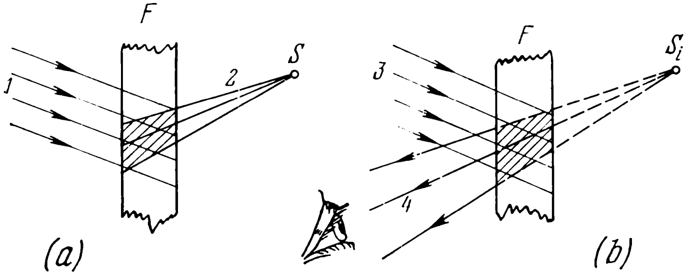


Fig. 66.22

As a rule, the diverging beam, *5*, is utilized for observations. The observer's eye in the beam's path looks through the hologram and sees a virtual image, $S_i$, which is an exact representation of the original object.

5. In 1962, Yu. N. Denisyuk proposed a method for making colour holograms, based on Lippmann's colour process using thick high-resolution photographic emulsion. With Denisyuk's technique (Fig. 66.22*a*), a reference wave, *1*, and an object wave, *2*, are incident on a thick photographic emulsion, *F*, from two sides and give rise to interference in stationary waves (see Sec. 57.2).

To reconstruct the hologram it is illuminated with a reconstruction wave, *3*, at the same angle as in making the hologram. The wave is scattered by the silver layers corresponding to the antinodes, and the observer watching the scattered wave, *4*, sees a virtual image, $S_i$ (Fig. 66.22*b*). A distinction of this type of hologram is that the antinodes associated with different wavelengths produce silver deposits in different layers. Therefore, when the hologram is illuminated with white light, the silver layers will reinforce by interference the reflected light of the same wavelength as that to which the plate was exposed, and the observer will see a three-dimensional image of the object in colour.

This discovery has won Yu. N. Denisyuk a Lenin prize.

6. What are then the advantages of holography which is developing so rapidly now? Let us discuss some of them.

(a) In an ordinary photograph, every area of emulsion represents a part of the original object. Therefore the information contained in one area is in no way related to that in any other area. The destruction of a part of the photograph leads to the loss of the respective information. In a hologram, each part contains information about the entire object, permitting its reconstruction from any small portion of the hologram, although the reconstructed image may be not so well defined and clear. The situation is similar to using a fragment of a lens for making an image.

Thus, as a medium for data storage, a hologram is more reliable than an ordinary photograph.

(b) As compared with an ordinary photograph, a hologram can store an amazingly greater amount of information. While a 6 × 9 mm photograph can hold one printed page, a single hologram plate of the same size can store 100 to 300 such pages, depending on the emulsion quality. Now that the problem of storage media for the rapidly growing printed information is becoming more and more urgent, holography can offer a way out.

(c) Holography may be adapted to fill the need for 3-D colour motion pictures and television.

(d) If the wavelength used for restoring the hologram into a reconstructed image of the primary test object, $\lambda'$, is greater than that used for obtaining the primary diffraction pattern, $\lambda$, the reconstructed image will be magnified, the degree of magnification being proportional to the ratio between the two wavelengths, $\lambda'/\lambda$. A remarkably high magnification can be obtained along with an increase in resolving power. However, there is a limit, too. The wavelength of the restoring beam, $\lambda'$, should be a fraction of the spacing between the interference fringes. Otherwise, the emulsion will be an optically homogeneous medium for the restoring wave, and no holographic effect will be obtained.

(e) There is a good deal of attraction in acoustical holography. Coherent acoustic waves are easy to produce, while sound (or ultrasound) is readily propagated in liquids or solids. With them, making a three-dimensional acoustical hologram of an opaque object will present no problem. By restoring the hologram in visible light, we shall be able to see the internal structure of, say, a metal bar, a concrete girder, or a living organism. This is a strikingly new opportunity for both technology and medicine.

A major snag of acoustical holography is the difficulty of catching an acoustical hologram. Yet, research is under way, and some approaches are being tried out.

7. We have only taken up some of the applications for holography. Many of its potential uses have not yet been discovered theoretically; still fewer have found embodiment in practice. Yet, it will be no wild guess to say that future holds much in store for holography.

# PART SEVEN BASIC QUANTUM PHYSICS OF ATOMS, MOLECULES AND SOLIDS

## Chapter 67
## THERMAL RADIATION

### 67.1. THERMAL RADIATION DEFINED

1. All bodies when heated emit radiation. Solids in a hot state first emit red light. Most substances when raised to a very high temperature vaporize (or change in chemical composition); yet they keep emitting visible light. Incandescent lamps emit yellow light when heated up to 3000 °C. Some substances will give up white light when hot.

The radiation emitted by hot bodies is called *thermal radiation*. Any hot body is a source of thermal radiation. It would be wrong, however, to think that thermal radiation is emitted at elevated temperatures only. This happens at room temperature, too. The only difference is that the intensity of thermal radiation is lower and its spectrum is different. When moderately hot, bodies emit red light with a wavelength of about 860 nm and, mainly, infrared rays which occupy a wide region from $10^6$ to $10^3$ nm in the electromagnetic spectrum.

2. In practice, the invisible infrared rays are revealed by their thermal effect. Illuminating a body, infrared rays raise its temperature. Let us make the following experiment. Place an electric heating element at the focus of a parabolic mirror and let the current raise the heater to a point where it emits infrared rays. If we place a piece of cotton wool (preferably, black) at the focus of a second parabolic mirror, the rays emitted by the heating element will cause it to burst into flame.

3. In ordinary surroundings bodies not only emit but receive thermal radiation. Otherwise the piece of cotton wool in the above experiment would not catch fire; it is only because its temperature was raised to a very high point by the thermal radiation from the heater that the cotton wool caught fire. Experience shows that

a body emitting a particular wavelength at a given temperature will eagerly absorb the same wavelength at the same temperature. It follows that a good radiator is a good absorber.

4. The quantity describing the capacity of a body to emit light at a particular frequency $\nu$ (rather, in the frequency interval from $\nu$ to $\nu + \Delta\nu$) at a given temperature $T$ is called the *spectral emissive power* of that body, symbolized as $E_{\nu T}$. It gives a measure of light at frequency $\nu$ radiated from that body per unit time per unit area of radiating surface. For radiation of all wavelengths we have the *total emissive power* of a body, $E_T$.

5. The quantity describing the capacity of a body to absorb incident light at a particular frequency (or wavelength) is called the *spectral absorptive power* or *absorptivity* of that body, symbolized as $A_{\nu T}$. The absorptivity is the ratio of radiation at a frequency $\nu$ absorbed in a body of material per unit time per unit area to the radiation incident upon it.

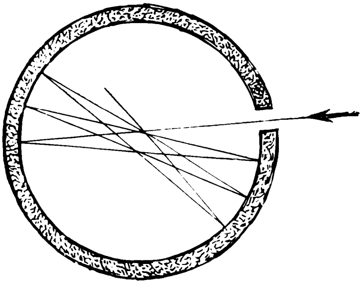


Fig. 67.1

An ideal body which would, if it existed, absorb all and reflect none of the radiation falling upon it at a temperature that would not destroy it is called a *black body*. Its absorptivity would be $A_{\nu T}^{black} = 1$. Black bodies are non-existent; they are physical abstractions. The materials that come closest to this abstraction are black velvet, black paper and soot. The nearest approach to the ideal black body, experimentally, is not a sooty surface, however. The laboratory type is an *isothermal enclosure*, which is a spherical cavity blackened inside and completely closed except a narrow slit on the side (Fig. 67.1). White light or other radiation entering by the slit is almost completely trapped inside by multiple reflections from the walls, which may be of any material, so that the opening usually appears intensely black.

6. In 1859 Kirchhoff formulated a law of radiation, later named after him, which states that *the ratio between the absorptivity and emissive power is the same for each kind of rays for all bodies at the same temperature and is equal to the emissive power of a black body at that same temperature.*

Denoting the emissive power of a black body as $\varepsilon_{\nu T}$, the Kirchhoff radiation law may be written as

$$E_{\nu T}/A_{\nu T} = \varepsilon_{\nu T} \qquad (67.1)$$

The absorptivity, $A_{\nu T}$, of a body cannot be greater than unity. Therefore the emissive power, $E_{\nu T}$, of any body cannot be greater than that of a black body, $\varepsilon_{\nu T}$, at the same temperature, $T$.

A black body would be the strongest source of thermal radiation. At a given temperature, a black body would emit more energy per unit time per unit area of radiating surface than any other body. Experimentally, this may be proved as follows. Pour hot water into a box-shaped vessel with two sides painted black and with the remaining two sides white, and place identical thermal detectors at equal distances from the blackened and white walls—they will show that more energy comes from the blackened than from the white surfaces, although their temperature is the same.

### 67.2. LAWS OF BLACK BODY RADIATION

1. In 1884 Boltzmann deduced theoretically that *the total emissive power of, or the total radiation from, a black body is proportional to the fourth power of the absolute temperature of the black body*

$$\varepsilon_T = \sigma T^4 \qquad (67.2)$$

Five years before him, the same law was discovered empirically by Stefan; this is why it has come to be known as the *Stefan-Boltzmann law*. The constant of proportionality, $\sigma$, is called the *Stefan-Boltzmann constant*. For a black body it has experimentally been determined as

$$\sigma = 5.672 \times 10^{-8} \text{ W/m}^2\,\text{K}^4$$

From the Stefan-Boltzmann law it follows that the radiation of a black body is solely determined by its temperature. With the temperature of a black body doubled, its total radiation increases 16-fold.

2. A body consists of a huge number of atoms each of which acts as an oscillator (see Sec. 59.5), each oscillating at a frequency of its own. This is why the radiation from a hot body contains all possible frequencies or wavelengths or, which is the same, is emitted in a continuous spectrum.

The efforts to establish the energy distribution of black-body radiation between the various wavelengths, the spectral distribution of black-body radiation, have led to the formulation of many of the fundamental concepts of present-day physics, notably quantum physics.

The spectral energy distribution of black-body radiation has been thoroughly investigated experimentally. In graphic form, it is represented by the plots of Fig. 67.2. The area bounded by the curve and the $x$-axis represents the total radiation from a black body per unit area per unit time. The area increases rapidly with rising temperature because it is proportional to $T^4$.

3. It will be noted that all curves have a peak and that this peak is shifted, or *displaced*, towards the shorter wavelengths as the temperature rises. This is why with a continuous rise in temperature, an incandescent body changes from red through orange to white.

The experimental curves of Fig. 67.2 represent a simple dependence on absolute temperature

$$\lambda_{\max} = b/T \tag{67.3}$$

This is a mathematical expression for the *Wien displacement law* which states that *the wavelength of the spectral distribution for which the radiation has the greatest intensity is inversely proportional to the absolute temperature of the black body.*

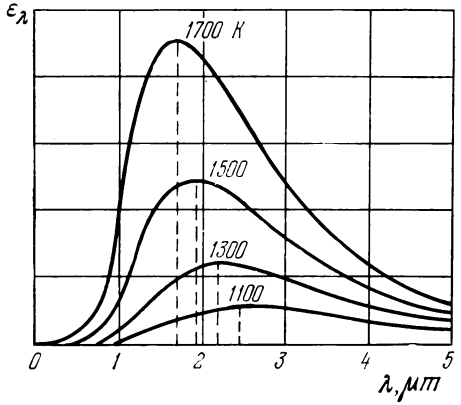


Fig. 67.2

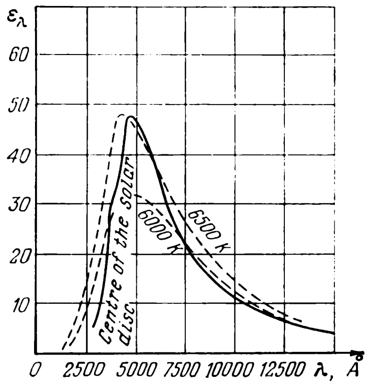


Fig. 67.3

The value of the displacement constant, $b$, in (67.3) is

$$b = 2.898 \times 10^{-3} \text{ m K}$$

The spectral energy distribution of the Sun's radiation is very close to that of a black body. This is seen from the dotted curves of Fig. 67.3 plotted for the radiation from a black body at 6000 K and 6500 K. The peak in the energy distribution of the radiation from the Sun occurs at about 4700 Å. If we assume the Sun to be a black body and use the Wien displacement law, the temperature of the Sun's outer layers must be close to 6200 K.

4. The classical theory of black-body radiation failed in deducing theoretical equations that would describe the experimentally observed spectral energy distribution shown in Fig. 67.2. All attempts to deduce the relation $\varepsilon = \varepsilon_T(\lambda)$ theoretically proved futile. In fact, they finally led to difficulties of a fundamental nature whose

importance outgrew that of the whole problem of thermal radiation. These difficulties are outside the scope of this book. It may only be stated that the consistent application of classical physics to the spectral energy distribution of black-body radiation invariably ends up in absurd results running counter to the law of the conservation of energy.

## 67.3. PLANCK'S INVESTIGATION OF BLACK-BODY RADIATION. PLANCK RADIATION FORMULA

1. A way out of the difficulties that had arisen in the study of black-body radiation was found in 1900 by Max Planck, an outstanding physicist. Classical physics treated the emission of light as a continuous process, that is, one in which a radiator is emitting continuously and its energy takes a continuity of values. By the same token, electromagnetic waves incident on a body were thought to be absorbed continuously. Planck felt that precisely these classical concepts were responsible for the inadequacy of the theory of thermal radiation.

2. Planck introduced the concept that a black body can only emit or absorb radiation in *quanta*, or increments of energy proportional to the frequency of the corresponding radiation. As Planck deduceld theoretically, a quantum of energy has the magnitude given by

$$\varepsilon_0 = h\nu \tag{67.4}$$

where $\nu$ is the frequency of the radiation and $h$ is *Planck's constant* (see Sec. 14.2). This is an absolute constant found to be $h = 6.62 \times \times 10^{-34}$ J s. According to Planck, a radiator always emits an energy $\mathscr{E}$ equal (for any frequency) to $\mathscr{E} = \varepsilon_0 n$, where $n$ is any positive integer. The quantum formula, (67.4), and Planck's universal constant have proved crucial in overcoming the dilemma of black-body radiation and maintaining agreement between theory and the experimental thermodynamic relations established for black-body radiation.

3. Planck found it possible to develop a consistent theory of black-body radiation and to deduce the relation $\varepsilon = \varepsilon_\nu (T)$ theoretically by assuming *discontinuous* energy emission. At the same time, Planck sought to tie in his hypothesis with the electromagnetic theory of light which was that the continuous flow of energy through space in the form of electromagnetic waves required that both absorption and emission of such energy should obey the laws of classical wave optics (see Ch. 63). Combining his concept of discontinuous energy emission with statistical mechanics, he found what is now known as the *Planck radiation law*:

$$\varepsilon_\nu (T) = \frac{2\pi h\nu^3/c^2}{e^{h\nu/kT} - 1} \tag{67.5}$$

where $c$ is the velocity of light in a vacuum, $h$ is Planck's constant, $k$ is Boltzmann's constant (see Sec. 26.9), $T$ is the absolute temperature, and $\nu$ is the frequency of the radiation.

The Planck radiation law admirably checks with the experimental data on the spectral energy distribution of black-body radiation.

4. Planck's idea of discontinuous emission and absorption of light gave a tremendous impetus to the further progress of physics. Before Planck, the existing concept was that the energy of any body could be changed continuously and by any arbitrary amount. In general, classical physics stood entrenched by the idea that all physical processes must be continuous. Planck's quantum idea gave these views their first shattering blow.

5. On the basis of the Planck radiation law, the laws of black-body radiation (see Sec. 67.2) could be proved theoretically and the constant $h$ could be related to the Stefan-Boltzmann constant $\sigma$, the displacement constant $b$ and the Boltzmann constant $k$:

$$h = \pi k \sqrt[3]{2\pi^2 k/15c^2\sigma}\ ; \qquad h = 4.965bk/c$$

where $c$ is the velocity of light in a vacuum. Using these relations, one can readily find Planck's constant. The first was used by Planck to find the value of $h$. The value of $h$ given by the second checks with that given by the first, and both agree with the values of $h$ found by other methods (see Sec. 68.3).

# Chapter 68
# BASIC QUANTUM OPTICS

## 68.1. THE PHOTOELECTRIC EFFECT

1. In 1887 Heinrich Hertz observed that a spark would pass an air gap more easily if ultraviolet light fell on the negative pole of the gap. However, Hertz failed to give a plausible explanation to his observation. It was through the experiments of Wilhelm Hallwachs in 1888 and, especially Professor Stoletov of Moscow University in 1888-89, that the effect was found to be due to knocking negative charges out of the metallic cathode by light.

The experimental set-up used by Stoletov is shown in Fig. 68.1. There was a capacitor consisting of a positive plate made from a copper grating or gauze, $C$, and a negative zinc plate, $D$, connected to a voltage source, $B$. When light from a source, $S$, struck the *negative* plate $D$, a current began to flow around the circuit. When the battery connections were interchanged so that the plate $D$ was charged *positively* and the grid $C$ negatively, the galvanometer registered no flow of current in the circuit.

2. Stoletov's experiments showed that light incident on a metal electrode caused it to lose negatively charged particles. Later measurements of the specific charge, $e/m$, of these particles (see Sec. 41.7) proved them to be electrons.

The knocking of electrons out of solids and liquids by light has come to be known as the *outer photoelectric effect* or, more accurately, *photoemissive effect*. The electric current appearing in the circuit of Fig. 68.1 when the plate $D$ is illuminated is called the *photo(electric) current*.

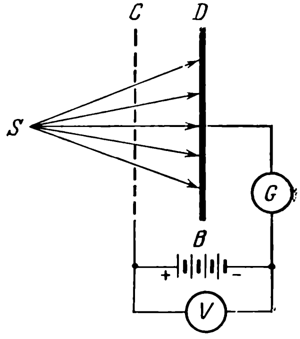


Fig. 68.1

There is also the *inner photoelectric*, or *photoconductive*, *effect*, which will be described in Sec. 78.5.

3. It might be attempted to describe the photoelectric effect on the electromagnetic theory of light by assuming that an electromagnetic wave incident on the metal causes the oscillations of its electrons to build up to a point where they can break their bond to the metal atoms. The theory of forced oscillations would then require the velocity of a released electron to increase with the amplitude of the incident light wave. By the same reasoning, the velocity and kinetic energy of the escaping electrons would be required to depend on the amplitude of vibration of the electric field intensity vector in the electromagnetic wave, that is, on the wave intensity (see Sec. 59.2). Experiments have not verified this.

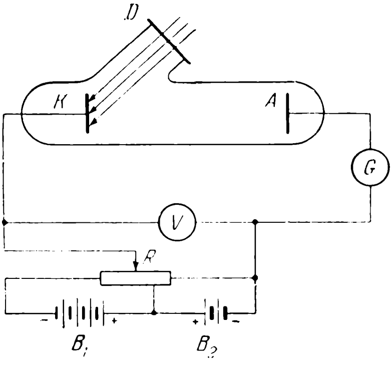


Fig. 68.2

## 68.2. LAWS OF THE PHOTOEMISSIVE EFFECT

1. The magnitude of the photoemissive current is determined by the number of electrons (called *photoelectrons*) released from the metal in unit time. The photoelectric current has been found to vary according to the chemical purity of the metal and the state of the surface. Even the minutest amount of contamination on the surface can radically change the conditions for the release of electrons and, as a consequence, the magnitude of photoelectric current.

2. The photoemissive effect can conveniently be investigated with a vacuum tube shown in Fig. 68.2. The cathode, $K$, is given

a coat of the metal whose photoemissive effect is to be investigated. The window covered by a quartz glass plate, $D$, admits ultraviolet light to the cathode. On striking the cathode, the ultraviolet light knocks electrons out of its surface. The released electrons are accelerated by the field existing between the cathode and the anode, $A$. The voltage, $u$, between anode and cathode can be adjusted with a potentiometer, $R$, and is measured on a voltmeter, $V$. The two batteries, $B_1$ and $B_2$, connected in opposition, enable the experimenter to control not only the absolute value of the voltage $u$ with the potentiometer, but also its polarity. With the accelerating voltage adjusted to a sufficient value, all electrons escaping from the cathode, will be collected by the anode, and the galvanometer, $G$, will register the maximum current obtainable at a given illuminance and a given temperature of the cathode. As already noted, its magnitude is determined by the number of electrons escaping from the cathode surface per unit time. This is the *saturation current*; it is the main quantitative characteristic of the photoemissive effect.

3. The electrons released from the cathode have some kinetic energy and can therefore do some work against the retarding electric field in the case of the negative voltage between cathode and anode. In this case, too, electrons may reach the anode, and there will be a photoelectric current flowing around the circuit. If the maximum initial velocity of an electron of mass $m$ is $v_{\max}$, then its kinetic energy is $mv_{\max}^2/2$. Owing to this energy, the electron can overcome the retarding electric field. If the maximum retarding voltage at which the photoemissive effect still takes place is $(-u_0)$, then

$$mv_{\max}^2/2 = eu_0 \tag{68.1}$$

At $u \geqslant |u_0|$, there will be no photoelectric current flowing.

As the applied voltage is raised, the photocurrent $I$ will rise gradually, as an increasing number of electrons is collected by the anode. The maximum value of current will be the saturation photocurrent, $I_s$. It corresponds to the value of applied voltage, $u$, at which the anode collects all the electrons escaping from the cathode

$$I_s = en \tag{68.2}$$

where $n$ is the number of electrons released by the cathode in unit time.

4. Experimentally, the following three laws have been developed for the photoemissive effect:

(1) The maximum initial velocity of photoelectrons is determined by the frequency of the incident light and is independent of its intensity.

(2) Each substance has a definite minimum frequency, $\nu_0$, of the incident light, called the *threshold frequency* at which the photoemissive effect is still possible.

(3) The number of electrons liberated from the cathode per unit time (the saturation photoelectric current) is directly proportional to the intensity of the incident beam.

It has also been found that photoelectric emission has *practically no time lag*; it is observed the same instant as a beam of light strikes the photoemissive cathode, provided that the frequency of the incident light is $\nu \geqslant \nu_0$ and that the material is capable of photoelectric emission.

5. It is relevant to note that the first and second laws of the photoemissive effect run counter to the explanation given in Para. 3 of Sec. 68.1 by the electromagnetic theory of light.

The contradictions in explaining photoelectric emission by the wave theory of light were removed by Einstein in 1905 when he used Planck's *quantum concept*. Einstein's reasoning falls outside the scope of this book. It may only be noted that his findings stemmed from a study into black-body radiation.

## 68.3. QUANTUM THEORY OF LIGHT. QUANTUM-MECHANICAL EXPLANATION OF PHOTOELECTRIC EMISSION

1. Although Planck introduced the concept of discontinuous emission and absorption of light, he clearly did not conceive of the quantization of radiation. In 1905, Einstein extended Planck's theory to say that not only are absorption and emission discontinuous, but even the energy of a light beam travels through space in quanta, called *photons*. That was a major departure from the classical wave theory.

According to Einstein, light is a stream of photons travelling always at velocity $c$ (the velocity of light in a vacuum). In a beam of monochromatic light of frequency $\nu$, all photons have the same energy given by $h\nu$. The absorption of light consists in that photons impart to the atoms and molecules they encounter all their energy. This is why the absorption of light is *discontinuous*.

2. The quantum theory of light offers an explanation of thermoelectric emission from metals, which differs from that based on the electromagnetic theory. As will be recalled, to escape from a metal, an electron should overcome the potential barrier at the metal-vacuum interface. This calls for an input of energy equal to the work function of the metal, $A_0$ (Sec. 44.9). Let the electron absorb a photon. Then its gain in energy will be $h\nu$. If $h\nu \geqslant A_0$, the electron will escape from the metal. By the law of conservation of energy, the maximum kinetic energy of a photoelectron is

$$mv_{\max}^2/2 = h\nu - A_0 \qquad (68.3)$$

or,

$$h\nu = A_0 + mv_{\max}^2/2 \qquad (68.3')$$

This is the *Einstein photoelectric equation.* By combining (68.1) and (68.3), we may also write

$$eu_0 = h\nu - A_0 \tag{68.4}$$

3. The Einstein photoelectric equations, (68.3) and (68.3′), adequately explain the laws of photoelectric emission. Thus, according to (68.3), the maximum kinetic energy of a photoelectron and, as a consequence, its maximum initial velocity is determined by the frequency $\nu$ of the incident light and the binding energy of the electron (or, which is the same, the work function $A_0$ of the material), but is independent of the intensity of the incident light. This is the first law of thermoelectric emission. It also follows that photoelectric emission can take place only if $h\nu \geqslant A_0$. The photon energy should be at least sufficient to break an electron's bond to the atom, without imparting it any kinetic energy ($v_{\max} = 0$). Denoting the threshold frequency (the "red boundary") of photoelectric emission* as $\nu_0$, we have

$$h\nu_0 = A_0 \tag{68.5}$$

or

$$\nu_0 = A_0/h \tag{68.5′}$$

The threshold frequency of photoelectric emission is solely decided by the electronic work function, that is, the chemical nature and the state of the surface of the metal specimen. This is an explanation of the second law of photoelectric emission.

Lastly, the total number $n$ of photoelectrons escaping from the surface of a metal specimen per unit time should be proportional to the number $n'$ of photons incident on the specimen per unit time, that is, $n \backsim n'$. Designating the illuminance of the cathode surface as $E$, which is proportional to the light intensity, the number of photons incident on the surface every second will be $n' \backsim E/h\nu$. This is the third law of photoelectric emission.

The foregoing applies to the absorption of single photons by single electrons. At high light intensity, such as provided by the laser (see Sec. 79.3), a single electron can absorb two identical photons. In this situation, the laws of photoelectric emission may be violated, notably that of the threshold frequency (68.5). The point is that two identical photons of energy $2h\nu$ will act as a single photon of twice the original frequency, for $2h\nu = h\,(2\nu)$; the frequency of the incident light doubles, as it were. In many-photon absorption, the apparent frequency of the incident light becomes $n\nu$ (where $n$ is an integer), i.e. the condition for the threshold frequency is violated.

* This boundary is called *red* because at $\lambda > \lambda_0 = c/\nu_0$, that is, at wavelengths "more red" than $\lambda_0$, no photoeffect occurs.

4. Equations (68.4) and (68.5) may be given a form convenient for the experimental verification of the Einstein photoelectric equation:

$$eu_0 = h(\nu - \nu_0) \tag{68.6}$$

Equation (68.6) can be verified by deriving Planck's constant from it

$$h = eu_0/(\nu - \nu_0)$$

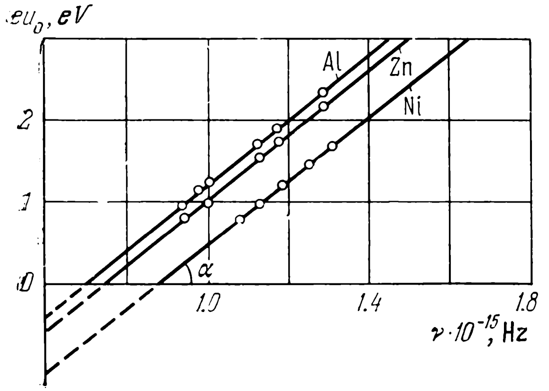


Fig. 68.3

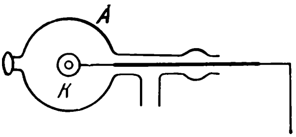


Fig. 68.4

In an experiment, the first step is to determine the voltage ($-u_0$) at which the photoelectric current just ceases to flow. The next step is to plot $eu_0$ as a function of $\nu$. In Fig. 68.3 this relation is represented for three metals, aluminium, zinc and nickel. The dots give the values of $eu_0$ measured at different frequencies. As is seen, the straight lines are all parallel, and their slope is independent of the kind of metal. Planck's constant can be found from the angle between the straight lines and the $x$-axis:

$$\tan \alpha = hK$$

where $K$ is the scale factor for the $eu_0$ and $\nu$ axes.

In the more rigorous experiments made by P. I. Lukirsky and S. S. Prilezhayev in 1928, the vacuum tube of Fig. 68.4 was a spherical capacitor. The glass envelope silvered on the inside was the outer plate of the capacitor and acted as the collecting electrode, or anode, $A$. The cathode, $K$, was a sphere made from the material under investigation. In this set-up, the anode collects all electrons with an initial velocity $v_0$ such that $mv_0^2/2 \geqslant e\,|u|$, where $u < 0$ is the retarding voltage. With this set-up, both the maximum velocity of photoelectrons, $v_0$, and Planck's constant can be determined to a very high degree of accuracy.

The average value of $h$, obtained by the most precise experiments with photoelectric emission, has been found to be $6.543 \times 10^{-34}$ J s, which agrees well with the results obtained by other methods. This is a further proof of the Einstein photoelectric equation and of Einstein's concept of the quantized interaction of light with electrons in photoelectric emission.

5. The fact that photoelectric emission is free from time lag is likewise due to the quantized nature of light. If one were to use the wave theory of light alone, one has to allow a finite time for the incident electromagnetic wave, even of a very high intensity, to build up the oscillations of an electron in a metal so that it can break loose from its atom (see Sec. 53.6).

The quantum properties of light, that is, the fact that radiation energy is carried in discrete quanta, raise the value of light energy. As an example, compare the energy of a photon of visible light at frequency $\nu \approx 10^{15}$ Hz with the average kinetic energy of the random thermal motion of a gas molecule. According to (26.8) given in volume I, the average kinetic energy of a molecule per degree of freedom is $kT/2$, where $k$ is Boltzmann's constant and $T$ is the absolute temperature. From $h\nu = kT/2$ it follows that a gas molecule would have the same energy as a quantum of visible light only at a temperature as high as $10^5$ K.

The quantum properties of light stand out with particular clarity in the interaction of short-wave light with matter.

## 68.4. PHOTOCHEMICAL EFFECTS

1. The light absorbed by substances may give rise to chemical changes in them. Such changes are called *photochemical reactions*. Among them is the dissociation of molecules. When bromine vapours are exposed to light, the bromine molecule, $Br_2$, dissociates into two atoms, Br. Light causes the molecule of silver bromide, AgBr, to dissociate into atoms of silver and bromine. In the green parts of plants, light brings about the photochemical break-up of $CO_2$. According to Timiryazev, chlorophyll, which gives the green parts of plants their specific colour, eagerly absorbs the long-wave red light from the Sun, and this causes the dissociation of carbon dioxide absorbed from the air. This reaction is then followed by a long chain of changes which result in the synthesis (photosynthesis) of carbohydrates essential to sustaining life, both vegetable and animal. Thus, photosynthesis is an important link in the chain of transformations that carbon undergoes in nature.

2. According to Bunsen and Roscoe, *the amount of chemical change, $m$, is proportional to the energy $\mathscr{E}$ of the absorbed light*

$$m = C\mathscr{E} \tag{68.7}$$

where $C$ is a proportionality factor dependent on the kind of photochemical reaction and the frequency of light, $\nu$.

For each photochemical reaction there is a threshold frequency, that is, the lowest frequency, $\nu_0$, at which light is still chemically active, that is, capable of inducing the reaction. Einstein explained the existence of the threshold frequency (as for thermoelectric emission, see Sec. 68.2) by the fact that light is absorbed in quanta. For a molecule to undergo a photochemical change, it must absorb an amount of energy, $\mathscr{E}_a$, called the *activation energy*. A quantum of light will induce a chemical change if its energy is $h\nu \geqslant \mathscr{E}_a$. Hence, the threshold frequency $\nu_0$ of a chemically active light can be found from

$$h\nu_0 = \mathscr{E}_a$$

as

$$\nu_0 = \mathscr{E}_a/h \tag{68.8}$$

3. In some cases, the photochemical effects of light obey the *Stark-Einstein law of the photochemical equivalence* which states that *each molecule taking part in a chemical reaction induced by exposure to light absorbs one quantum of radiation causing the reaction.* Hence the number of molecules taking part in a light-induced reaction on absorbing unit energy of light ($\mathscr{E} = 1$) is inversely proportional to the energy of one photon, $h\nu$ (provided $\nu \geqslant \nu_0$), or

$$N \backsim 1/h\nu = \lambda/hc \tag{68.9}$$

where $\lambda = c/\nu$ is the wavelength of the radiation.

The Stark-Einstein law of the photochemical equivalence is often violated: one absorbed photon may induce chemical change in several molecules. For example, when $H_2$ and $Cl_2$ are exposed to light, a chain reaction is induced, producing HCl and ending up in an explosion. That is, light only triggers the reaction which then proceeds spontaneously.

The theory of chain reactions (to be discussed on Sec. 82.9) was developed by Academician N. N. Semyonov in 1928-34. In 1956, he was awarded the Nobel prize for chemistry.

## 68.5. MASS AND MOMENTUM OF THE PHOTON. LIGHT PRESSURE IN THE QUANTUM THEORY OF LIGHT

1. In presenting Einstein's quantum concept, we have only mentioned that a photon has an energy given by $\varepsilon_{ph} = h\nu$. To be consistent, however, we must of necessity assume that, since photons behave like a stream of particles, each photon should have mass and momentum.

In the theory of relativity (see Sec. 16.1), energy and mass are generally related as

$$\mathscr{E} = mc^2 \text{ or } m = \mathscr{E}/c^2$$

For a photon, $\mathscr{E} = \varepsilon_{ph} = h\nu$, and so its mass is

$$m_{ph} = h\nu/c^2 \tag{68.10}$$

A photon differs fundamentally from macroscopic and elementary particles (discussed in Chapter 83) in that it has no rest mass, $m_0$. In other words, photons at rest are non-existent. For, if a beam of light is "stopped", light ceases to exist, that is, photons are absorbed by the atoms and molecules of matter, with the energy of photons changed to other forms. For example, when a metal specimen absorbs light, some of the energy of photons is imparted to photoelectrons, and these escape from the metal as in photoelectric emission.

It should be stressed that in any medium photons travel at the same velocity equal to that of light in a vacuum, $c$, as long as they move between atoms and molecules. In colliding with a particle, the photon is either scattered by the particle and remains in vacuum, or absorbed by the particle.

2. At first sight, this statement runs counter to the result obtained in Chapter 63, namely that the phase velocity of light in a substance is $1/n$ of its value in a vacuum: $u = c/n$, where $n$ is the absolute refractive index of the substance ($n > 1$). Actually, there is no contradiction because light is propagated in matter so that photons can be absorbed and re-emitted by the particles of the medium (of course, other forms of interaction of light with matter are possible). The "re-emitted" photons correspond to the secondary waves discussed in Sec. 63.1. The slow-down of the phase velocity of light in matter by a factor $n$ in comparison with its velocity in vacuum is due to the fact that the atoms and molecules re-emit the absorbed photons with a phase delay.

3. The fact that the photon has no rest mass, $m_0$, shows that Einstein's quantum ideas are no return to Newton's corpuscular theory of light. According to Newton, the corpuscles of light are ordinary mechanical particles. If they existed, then, by present-day theory, they would have to have a rest mass ($m_0 \neq 0$). In his criticism of Newton's views as regards the nature of light, Lomonosov justly noted that if the corpuscular theory were true, the corpuscles would collide and there would be "a confusion of light rays".

4. The mass of a photon should be looked upon as a field mass. That is, a photon has a mass related to the electromagnetic field of the light wave. Since, according to relativity theory, $\mathscr{E} = mc^2$, then, if we assume that $\mathscr{E}$ is the energy of the electromagnetic field associated with a light wave, $m$ is the mass of the electromagnetic

field of the wave. This brings us to a very important result, namely that, like material substances the electromagnetic field should have energy and mass, too. For, substance is a form of the existence of matter dealt with in physics and other natural sciences. Another form of the existence of matter dealt with in physics is represented by fields, and so the electromagnetic field is among them. The fact that the electromagnetic field has energy and mass, the most important attributes of matter, is a forceful proof that the electromagnetic field is a material entity.

5. In addition to energy and mass, a photon has a momentum, $p_{ph}$. The energy and momentum of a photon are connected by the general equation of relativity theory (see Sec. 16.3):

$$\varepsilon_{ph}=c\sqrt{p_{ph}^2+m_0^2c^2}$$

Since for a photon, $m_0 = 0$, then

$$p_{ph}=\varepsilon_{ph}/c=h\nu/c=m_{ph}c \tag{68.11}$$

where $m_{ph}$ is the mass of a photon.

Equation (68.11) looks like that from classical mechanics which states that if a body of mass $m$ is moving at a velocity $v$, its momentum is $p = mv$. From (68.11) it is seen that the momentum $p_{ph}$ and velocity $c$ of a photon are connected in a similar manner.

Using the relativistic relation between mass and energy, $m = \varepsilon/c^2$, the relativistic momentum may be written as

$$p_{ph}=(\varepsilon/c^2)\,v$$

Since, however, for a photon $v = c$, we get

$$p_{ph}=\varepsilon_{ph}/c=h\nu/c$$

In some cases, the momentum of a photon is stated differently. Recall that in optics the wave parameter $k$ is defined (Sec. 56.2) as the reciprocal of the wavelength, or $k = 2\pi/\lambda$. Then (68.11) may be given the form

$$p_{ph}=h\nu/c=h/\lambda=hk/2\pi=\hbar k \tag{68.12}$$

where $\hbar = h/2\pi = 1.05 \times 10^{-34}$ J s is Planck's constant. Its value has been found to be

$$\hbar=(1.05450\pm 0.00005)\times 10^{-34}\ \text{J s}$$

The momentum is a vector quantity. To determine the direction of the vector momentum of a photon, we shall use a *wave vector* **k** numerically equal to $k = 2\pi/\lambda$, and pointing in the direction of light propagation. Then (68.12) may be re-written in vector form as follows

$$\mathbf{p}_{ph}=\hbar\mathbf{k} \tag{68.12$'$}$$

6. Now that we have established that a photon, like any other moving particle, has energy, mass and momentum, we may re-write equations for these physical quantities which may be called the *corpuscular characteristics* of a photon:

$$\varepsilon_{ph} = h\nu; \quad m_{ph} = h\nu/c^2; \quad p_{ph} = h\nu/c \tag{68.12''}$$

Their name implies that they are shared by all bodies and particles. It should be noted that the three corpuscular characteristics of a photon are connected by the most important characteristic of light, its frequency $\nu$. This connection is not incidental; behind it is a deep-lying cause which will be discussed in Sec. 68.7.

7. Experimentally the fact that a photon has both mass and momentum is proved by light pressure (see Sec. 59.3). Previously, we have explained light pressure by the wave theory of light. As forcefully, it can be explained by the quantum theory: the pressure of light on the surface of a body is due to the fact that on striking the surface each photon imparts to it its momentum, much as the pressure of a gas on the walls of a vessel is due to the transmission of momentum from gas molecules to the walls of the vessel (see Sec. 17.5).

Let a beam of light strike the surface of a body in one direction, say, normally to it. Let also a number $n$ of photons strike unit area of the surface per unit time. Some photons will be absorbed by the surface, and each will hand to it its momentum, $p_{ph} = h\nu/c$. Some will be reflected, and each reflected photon will move away from the reflecting surface in the reverse direction with a momentum $-\mathbf{p}_{ph}$. Therefore, the total momentum imparted to the wall will be $p_{ph} - (-p_{ph}) = 2p_{ph} = 2h\nu/c$.

The pressure of light on a surface is numerically equal to the momentum imparted by all $n$ photons striking unit area per second. Designating the reflection of light from an arbitrary surface as $R$, then the number of reflected photons will be $Rn$, and that of absorbed photons $(1 - R)\,n$. Thus, the light pressure is

$$p = Rn\frac{2h\nu}{c} + (1-R)\,n\frac{h\nu}{c} = (nh\nu/c)\,(1+R)$$

The product $nh\nu = \mathscr{E}$ is the energy of all photons incident on unit area per unit time. By definition, this is light intensity (see Sec. 55.3). The ratio $\mathscr{E}/c = w$ is the volume energy density of incident light (see Sec. 59.2). Thus, finally, the pressure of light is given by

$$p = w\,(1 + R) \tag{68.13}$$

For the first time, this equation was derived by Maxwell in his electromagnetic theory of light and experimentally verified by Lebedev (see Sec. 59.3).

## 68.6. THE COMPTON EFFECT

1. In 1923, Compton observed the scattering of X-rays (see Sec. 61.1) of a definite wavelength $\lambda$ by the electrons of graphite, paraffin and other light-weight substances.

In Compton's experimental set-up shown in Fig. 68.5, a monochromatic beam of X-rays excited in an X-ray tube, *A*, is limited by an aperture, *B*, and the emergent narrow beam is focused to strike a specimen of light-weight scatterer, *C*. The rays scattered through an angle $\theta$ are caught by an X-ray detector, *D*, which is an X-ray spectrograph measuring the wavelength of the scattered X-rays.

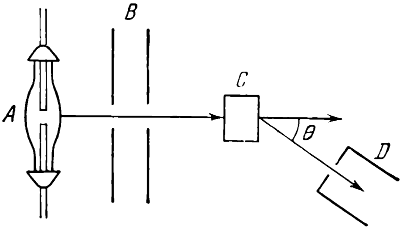


Fig. 68.5

Compton's experiments showed that the wavelength, $\lambda'$, of the scattered X-rays was longer than that, $\lambda$, of the incident rays. The shift in wavelength, the *Compton shift*, $\Delta\lambda = \lambda' - \lambda$, was found to depend solely on the scattering angle, $\theta$, and to be independent of either the properties of the scatterer or the wavelength of the incident radiation:

$$\Delta\lambda = \lambda' - \lambda = 2\lambda_C \sin^2 \theta/2 \tag{68.14}$$

The quantity $\lambda_C$ is the same for all substances and equal to $\lambda_C = 2.43 \times 10^{-12}$ m. It is called the *Compton wavelength*, and the elastic scattering of photons by electrons has come to be known as the *Compton effect.*

2. It follows from Eq. (68.14) and from the constancy of the Compton wavelength .that the Compton shift will be a maximum at $\theta = \pi$, that is, when the scattered photon retraces its path. This is because at $\theta = \pi$,

$$\Delta\lambda = \Delta\lambda_{\max} = 2\lambda_C$$

At the same time, the scattering electron, called the *Compton recoil electron*, receives a maximum kinetic energy.

3. The Compton effect could not satisfactorily be explained by the wave theory of light according to which (see Sec. 62.8) an electromagnetic wave incident on a specimen should produce in it a secondary wave of the *same* wavelength. By the quantum theory, the Compton effect, like photoelectric emission and photochemical processes, is caused by the interaction of incident photons with the electrons of atoms or molecules. If the scattering of photons by electrons or other particles does not result in the generation of new

photons of the same wavelength (as happens in lasers, see Sec. 79.4), then the energy $h\nu$ of an incident photon should partly be expended to support some other process apart from the production of a scattered photon of energy $h\nu'$, where $\nu' = c/\lambda'$ is the frequency of the scattered radiation. By the law of conservation of energy,

$$h\nu = h\nu' + \mathscr{E}_1 \tag{68.15}$$

where $\mathscr{E}_1 > 0$ is the reduction in the energy of the photon through non-radiant processes.

In the Compton effect, X-ray quanta are scattered by the electrons of the lighter atoms of low atomic number, $Z$ (see Sec. 71.1). The energy $\mathscr{E}_1$ is taken by an electron which may for all practical purposes be treated as a free one or at least loosely bound to its atom. From Eq. (68.15) it follows that $h\nu > h\nu'$, that is, $\nu > \nu'$ and, as a consequence, $\lambda < \lambda'$. Thus, the scattering of X-rays by electrons should bring about an increase in the wavelength of the scattered radiation, and this was found by Compton.

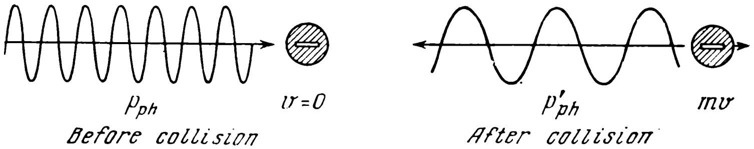


Fig. 68.6

4. The Compton wavelength, $\lambda_C$, may be found by applying the proper relativistic formulae for momentum and energy. For simplicity it may be assumed that prior to collision the electron is stationary in the selected reference system, while after collision the electron and the photon fly in opposite directions (a central impact, Fig. 68.6). Then the equations for momentum and energy can be written as

$$p_{ph} = -p'_{ph} + p, \quad \varepsilon_{ph} + \mathscr{E}_0 = \varepsilon'_{ph} + \mathscr{E} \tag{68.16}$$

where $p$ is the momentum of the electron, $\mathscr{E}_0$ and $\mathscr{E}$ are its energies before and after the impact. Squaring the two equations and noting that the energy and momentum of a photon are connected by the relation $\varepsilon_{ph} = p_{ph}c$ and those of an electron by the relation $\mathscr{E}^2 = \mathscr{E}_0^2 + p^2c^2$, we get after some manipulations

$$4\varepsilon_{ph}\varepsilon'_{ph} = 2\mathscr{E}_0 (\varepsilon_{ph} - \varepsilon'_{ph}) \tag{68.17}$$

Recalling that $\varepsilon_{ph} = hc/\lambda$, and $\mathscr{E}_0 = m_0c^2$, where $m_0$ is the rest mass of an electron, Eq. (68.17) can be given the form

$$(\lambda'/hc - \lambda/hc)\, m_0c^2 = 2$$

or, finally

$$\Delta\lambda = \lambda' - \lambda = 2h/m_0c \tag{68.18}$$

From a comparison with (68.14), we obtain the following expression for the Compton wavelength

$$\lambda_C = h/m_0c \tag{68.19}$$

5. It can be shown that a free electron cannot absorb the energy of a photon. By contradiction, assume that a *stationary free electron* does absorb an incident photon. Then two laws of conservation must be satisfied simultaneously

$$\begin{aligned} (m - m_0)\, c^2 &= h\nu, \\ mv &= h\nu/c \end{aligned} \tag{68.20}$$

The first line of (68.20) expresses the law of conservation of energy, and the second, the law of conservation of momentum. Here, $m$ is the relativistic mass of an electron, $m_0$ is its rest mass, and $v$ is its velocity after the absorption of a photon. It is an easy matter to show that both lines of (68.20) cannot be satisfied simultaneously at arbitrary finite values of $\nu$. We leave it as an exercise for the reader to prove the point.

The physical significance of the fact that the two lines of Eq. (68.20) cannot be satisfied simultaneously is that free electrons cannot absorb light. Light can only be absorbed by bound electrons, such as those in the atoms of gases and solids or other systems of particles. Then the equations for the conservation of energy and momentum should somewhat differ from (68.20):

$$(m - m_0)\, c^2 + W = h\nu, \quad mv + p = h\nu/c \tag{68.21}$$

They differ by the term $W$ which is the electron binding energy, and the term $p$ which is the momentum transferred to the system in the case of photoelectric emission. Eqs. (68.21) have univalued solutions at arbitrary finite values of $\nu$. To sum up, bound electrons can absorb photons.

## 68.7. THE WAVE-PARTICLE DUALITY OF LIGHT

1. In this chapter we have taken up a number of light phenomena which support the particle (or, rather, quantum) theory of light. In the preceding chapters (Ch. 61 through 64), on the other hand, where we discussed the effects of diffraction, interference and polarization, the particle concept did not fit while the wave theory did. One may then legitimately ask what light finally is? In the words of Sir William Bragg, should "we teach the wave theory on Mondays, Wednesdays and Fridays, and the corpuscular theory on Tuesdays,

Thursdays and Saturdays"? In short: is light electromagnetic waves emitted by a light source or a stream of photons travelling through space at the velocity of light in vacuum?

At first glance it may appear that the electromagnetic wave theory and the particle (quantum) theory are mutually exclusive, for some of the wave and particle properties are diametrically opposite. For example, moving photons are precisely localized in space, while an advancing wave cannot be pinned down to a particular point in space. Because one is forced to treat light as waves in some cases and as particles in others, one might think that the present-day conception of light is incomplete or that this duality is an artificial device and that either one or the other theory may well explain all that is known about light.

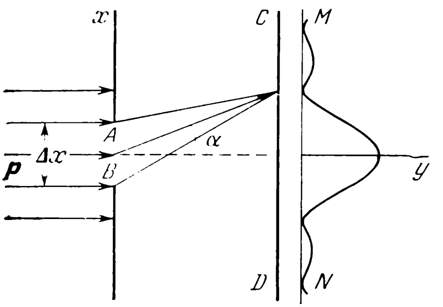


Fig. 68.7

2. When we review the advances of optics and allied fields, we realize that the continuous nature of electromagnetic waves cannot be set against the discontinuous nature of photons. For light is neither waves nor corpuscles; it is both. This is what is meant by the duality of light. It is reflected in equations (68.12″) describing the principal characteristics of photons. As follows from these equations, the corpuscular characteristics of photons, energy $\varepsilon_{ph}$, momentum $p_{ph}$, and mass $m_{ph}$, are tied in with the wave concept through its frequency, $\nu$.

There is a kind of order in how light manifests its dual nature. With long-wave radiation (for example, infrared rays), the particle properties are hard to observe while the wave behaviour stands out clearly. This is why interference, diffraction and polarization of light are readily explained on the wave theory. As we move, however, across the electromagnetic spectrum from the longer to the shorter wavelengths, the wave properties become less noticeable in comparison with the particle aspects. This is borne out by the existence of threshold frequency for photoelectric emission and for photochemical processes. As has been shown in Sec. 62.5, the diffraction of short-wave X-rays can be detected only with the crystal lattice used as a diffraction grating; otherwise the wave nature of X-rays would hardly be noticed.

3. Consider the tie-in between the wave and particle properties of light, taking the passage of light through a slit in an opaque screen as an example (Fig. 68.7). Let a parallel bundle of monochro-

matic light rays pass through the slit $AB$ along the $y$-axis. In terms of the dual theory of light, both a stream of light particles, or photons, and an electromagnetic light wave are passing through the slit at the same time. Since the light beam is monochromatic, a diffraction pattern is produced on the screen $CD$ located behind the slit (see Sec. 62.1). Each point on the screen has an illuminance, $E$, proportional to the light intensity at that point. On the right of Fig. 68.7, is a plot of light intensity distribution over the screen. As will be recalled, the light intensity is proportional to the square of the amplitude, $A$, of the light wave. Therefore, the illuminance, $E$, at each point of the screen is proportional to the square of the amplitude of the light wave at that point, or $E \backsim A^2$.

In terms of the quantum theory of light, the diffraction pattern forming on the screen as light passes through the slit implies that photons are redistributed in space, so that different numbers of photons reach different points on the screen. The illuminance, $E$, of the screen at a given point is decided by the total energy of photons incident at that point per unit time. In turn, this energy is proportional to the number $n_0$ of photons delivering this energy. Thus, $E \backsim n_0$.

Let the radiant flux falling on the slit (Fig. 68.7) be so faint that it might be taken as consisting of a small number of photons. In the limit, it might be taken as consisting of photons passing through the slit one by one. Each photon should then display itself at the point on the screen it strikes. However, experiments have shown that, no matter how the radiant flux or the light intensity is reduced, the diffraction pattern remains unchanged. The relation between the light and dark areas on the screen characteristics of diffraction by a particular obstacle is the same even with a faint luminous flux.

In a practical experiment, a luminous flux made of single photons following one after another cannot be produced. The best we can do is to imagine that photons are allowed to hit repeatedly a particular point on the screen a great number of times and that the trials are observed over a sufficiently long time interval. Then the results would be the same as with a luminous flux consisting of a large number of photons, arriving at the obstacle all at the same time. The diffraction pattern will represent the actual distribution of light and dark areas characteristics of diffraction by a given obstacle.

4. It is instructive to compare the two expressions for illuminance derived above. From them it follows that $A^2 \backsim n_0$, or in words, *the square of the amplitude of a light wave at a point in space is proportional to the number of photons arriving at that point.* Or stated differently, *the square of the amplitude of a light wave is a measure of the probability that photons will hit that point.*

As is seen, the wave and quantum properties of light supplement rather than exclude one another; they represent the actual picture of how light is propagated and interacts with matter. The quantum properties stem from the fact that the energy, momentum and mass of radiation are concentrated in quanta of light, or photons. The wave properties of light, notably the amplitude of a light wave, determine the probability that a photon can be found at a particular point in space.

From the foregoing it follows that wave properties are displayed not only by a collection of a large number of photons travelling all at the same time, but also by *each individual photon*. The wave properties of a single photon are manifested in that one cannot pin down a point on the screen where a photon can turn up after passing through the slit (Fig. 68.7). One can only speak of the probability of a photon hitting a particular point.

This interpretation of the relation between the wave and particle properties of light was advanced by Einstein. It has made a huge contribution to the advance of present-day physics.

# Chapter 69
# THE WAVE PROPERTIES OF ELEMENTARY PARTICLES

## 69.1. THE WAVE-PARTICLE DUALITY OF ELEMENTARY PARTICLES

1. In 1924, de Broglie announced that it is "essential to consider corpuscles and waves simultaneously, if it were desired to reach a single theory, permitting of the simultaneous interpretation of the properties of light and those of matter...".

His reasoning may be stated in the following simple terms (although actually, it was not at all simple). We have seen that as the frequency, ν, of light rises, its wave properties become progressively less observable. With gamma-rays (see Sec. 61.1), the shortest electromagnetic waves, one will hardly suspect that light has wave properties. On the other hand, elementary particles are decidedly corpuscular in their properties; yet, is it not possible that there are still shorter waves associated with material corpuscles? The hypothesis that de Broglie put forward postulated that the electron, in addition to the usual corpuscular properties (charge and mass) should also have wave properties and that, given certain conditions, it should behave like a *wave*, or in his words, "...it will no longer be possible to consider the material corpuscles, or electrons, as discrete isolated entities...; they are accompanied by a wave which is bound up with their own motion".

2. He used the relation for the momentum of a photon, Eq. (68.12), from which it follows that

$$\lambda = h/p_{ph} \tag{69.1}$$

His idea was that equation (69.1) must be of a universal nature, applicable to any wave processes. Then the wave associated with any particle having momentum, $p$, should have a wavelength given by

$$\lambda = h/p = h/mv \tag{69.2}$$

where $p = mv$ is the momentum of a particle of mass $m$ and velocity $v$, and $\lambda$ is the *de Broglie wavelength.*

It should be noted from the outset that the *de Broglie waves*, also called the *electron waves*, are *not* electromagnetic waves. Their specific nature will be explained later.

3. After de Broglie's announcement, Einstein commented that if de Broglie's hypothesis was valid, electrons must be diffracted. In 1927, Davisson and Germer (of the Bell Telephone Company, New York) conducted a research to see whether electrons might be reflected from a crystal.

Their technique was along the lines used by Compton in the study of the elastic scattering of X-rays (see Sec. 68.6). The experimental set-up is shown in Fig. 69.1. The electron gun at $A$ generated a beam of electrons whose energy and velocity was controlled by the accelerating voltage inside the gun. The narrow beam of electrons was then directed at a controlled speed onto a target, $B$, which was a single crystal of nickel. The nickel target could be rotated about an axis normal to the plane of the drawing, as could an electron detector, $C$, which collected the electrons reflected from the target through different angles. If the electrons were classical material particles, they would be reflected from the target according to the laws of geometrical optics. Actually, in addition to a distinct "regular" reflection, the experimenters found a series of diffraction maxima and minima. That was a proof that electrons experience diffraction.

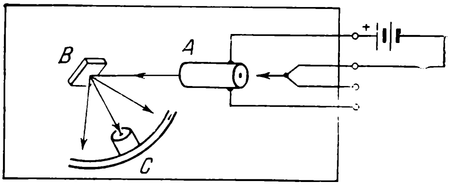


Fig. 69.1

The plots of Fig. 69.2 show the distribution in angle of the electrons reflected from the target $B$ for two angles of incidence of the electron beam, $N$. The radius $r$ drawn from the centre of the target is proportional to the number of electrons reflected at a given angle. It is clearly seen that there are maxima and minima in the number of electrons scattered at different angles.

Using the technique usually employed for the diffraction of X-rays on single crystals (see Sec. 62.7), Davisson and Germer could experimentally find the wavelength associated with a moving electron.

Since, in an accelerating field (with a potential difference $\Delta\varphi$), an electron of charge $e$ gains a kinetic energy, $W$, we may write

$$W = mv^2/2 = e \cdot \Delta\varphi$$

whence the velocity of the electron is

$$v = \sqrt{2e\Delta\varphi/m}$$

Substituting the above expression in that for the de Broglie wavelength, (69.2), gives

$$\lambda = h/mv = h/\sqrt{2em\Delta\varphi} \qquad (69.3)$$

Substituting the numerical values of $h$, $e$ and $m$, in (69.3) gives the final expression for the wavelength of the wave associated with an electron moving in an electric field with a potential difference $\Delta\varphi$:

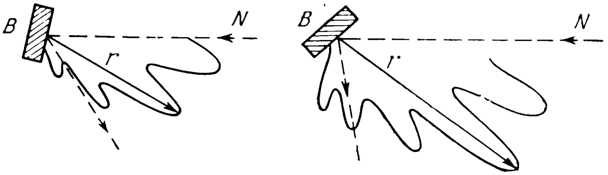


Fig. 69.2

$$\lambda = (12.25/\sqrt{\Delta\varphi})\ \text{Å} \qquad (69.4)$$

If the potential difference in the above expression is in volts, the wavelength will be in Ångstrom units.

The value of $\lambda$ found by Davisson and Germer was in good agreement with that given by Eq. (69.4).

4. Soon after Davisson and Germer, de Broglie's hypothesis was experimentally verified again by P. S. Tartakovsky at Leningrad University and, independently, by G. P. Thomson at Aberdeen. They found the diffraction of electrons by passing electron beams through thin films (about $10^{-7}$ m) of metals having polycrystalline structure. The diffraction patterns were strikingly similar to those obtained with polycrystalline powders for X-rays (see Sec. 62.7). Photographs of the diffraction patterns produced by X-rays and a beam of electrons passed through thin specimens of the same

material are shown on the right and left of Fig. 69.3, respectively. Using this technique, Thomson found the de Broglie wavelength from Eq. (69.4) and then, from the known relations for diffraction by three-dimensional structures, the lattice dimensions of the metal specimens through which electrons were passed. The results tallied with those obtained by X-ray diffraction analysis.

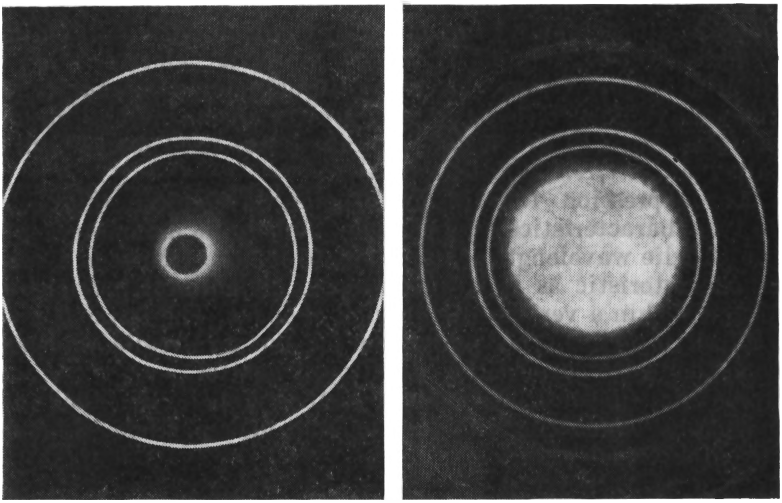


Fig. 69.3

5. In 1949, L. M. Biberman, N. G. Sushkin and V. A. Fabrikant of the Soviet Union conducted a research to find diffraction for single electrons travelling in a succession. The intensity of the electron beam in their experiment was so low that only one electron could reach a thin metal film at a time. In principle, this was the embodiment of the mental experiment suggested in Sec. 68.7 for photons. As we have seen, with the trials repeated many times, the optical diffraction pattern of the slit must be the same as with a luminous flux containing a multitude of photons. This finding also applies to the repeated bombardment of a film specimen with single electrons travelling in a succession. Although the individual electrons pass through the specimen singly and their behaviour is independent of one another, the experiment repeated many times produces the same diffraction pattern as electron beams with an intensity tens of times as great. This is an indication that single electrons passing through a specimen produce a diffraction pattern as well. The probability that the electrons passing through a specimen will hit the same point on the screen is determined by their

wave properties, that is, by the de Broglie wave associated with them.

6. The diffraction of electrons, as a manifestation of their wave properties, will take place only if the de Broglie wavelength of moving electrons is of the same order of magnitude as the interatomic distances in the crystals used for diffraction. This condition is the basis of *electron diffractometry*, an instrumental technique for analyzing crystal structures from the diffraction of electrons by crystals.

Electron diffractometry has much in common with X-ray diffraction analysis (see Sec. 62.7). Since, however, electrons are less penetrating than X-rays, electron diffractometry is mainly used in the study of surface structures or the effects of environments on surfaces, such as oxidation. The apparatus used for the purpose is called an *electron diffractometer*.

7. The expression for the de Broglie wavelength, (69.2), connects two characteristics of a particle, namely its momentum $p$ and the de Broglie wavelength, $\lambda$, ascribed to it. The former is a corpuscular characteristic as it is an attribute of material particles concentrated within a very limited region of space. The latter is a wave characteristic. Thus, the expression for the de Broglie wavelength reflects the wave-corpuscular duality of matter.

Present-day physics has made a further step in the dual theory of matter by connecting the total energy of a particle, $\mathscr{E}$, and its de Broglie wavelength:

$$\mathscr{E} = h\nu = \hbar\omega \tag{69.5}$$

where $\hbar = h/2\pi$ and $\omega = 2\pi\nu$ is the radian frequency. This expres sion has been borrowed from optics, Eq. (67.4), where it connects the energy of a photon to its frequency. In other words, the relation between frequency and energy given by (69.5) is a universal one, applicable to particles having a rest mass as well as to photons.

That the relation (69.5) is valid for any particles is confirmed by the good agreement of the results obtained by present-day atomic and nuclear physics with experiment. Before we can take up some of them, it is important to examine in greater detail the wave properties of elementary particles and the far-reaching implications of their dual wave-corpuscular nature.

### 69.2. WAVE PROPERTIES OF NEUTRONS, ATOMS AND MOLECULES

1. According to de Broglie, any particles having a mass $m$ and a velocity $v$, and not only electrons, have a definite wavelength ascribed to them, that is, the de Broglie wavelength. This can be seen from Eq. (69.2) which in no way singles out electrons as special particles.

Among other things, experiments have found the diffraction of neutrons, particles which are part of the atomic nucleus (see Sec. 80.3) and which can be produced by nuclear reactors (described in Sec. 82.10).

The experimental set-up used to observe the diffraction of neutrons is shown in Fig. 69.4. The neutrons produced in a nuclear reactor are slowed down by passing through a thermal column, a heavy bed of graphite. On repeatedly colliding with the carbon nuclei, the neutrons are slowed down to become so-called thermal neutrons (see Sec. 82.5). Emerging from the thermal column, the neutrons have velocities and energies corresponding to the Maxwell velocity distribution at the temperature $T$ of the graphite (see Sec. 25.2), and so one can readily determine their velocities in a beam of thermal neutrons. On passing through a narrow slit, the neutrons are diffracted by a crystal target and collected by a detector. The crystal target and the detector are aligned so as to keep the angles of incidence and reflection, $\theta$, for the neutrons constant. Since the velocity $v$ and mass $m$ of the neutrons are known, we can readily determine the de Broglie wavelength associated with moving neutrons, using Eq. (69.2). On the other hand, for a given angle $\theta$, the reflecting neutrons should satisfy the Bragg law (see Sec. 62.7). More specifically, neutrons will be reflected only if $\lambda$ and $\theta$ are connected by the relation

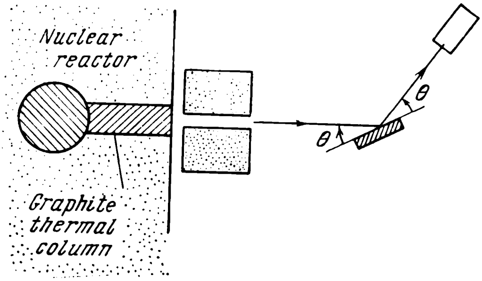


Fig. 69.4

$$2d \sin \theta = n\lambda \tag{69.6}$$

where $d$ is the distance at which the planes of atoms in the crystal are spaced apart, and $n$ is an integer. With $d$ known in advance, Bragg's law gives the wavelength, $\lambda$. The good agreement of $\lambda$ computed and found by experiment using the reflection of neutrons is a convincing proof that neutrons have wave properties.

2. The diffraction of neutrons is utilized as a tool for analyzing the structure of hydrogen-containing solids, especially crystals,

because neutrons strongly interact with atomic nuclei, notably hydrogen nuclei. The scattering of neutrons by hydrogen-containing atoms in crystals immediately reveals the presence and location of hydrogen atoms in the crystal lattice.

X-ray and electron diffractometry is unsuitable for the interpretation of the structure of such substances. The point is that X-rays interact with atomic electrons, but there is only one in a hydrogen atom. Electrons passing through a specimen experience electromagnetic interaction with the electrons of the atoms and with the protons of their nuclei, but this interaction is weak in hydrogen-containing crystals. As a result, the scattering of X-rays and electrons by hydrogen is not enough to serve as a basis for X-ray or electron diffraction analysis of hydrogen-containing crystals. In contrast, *neutron diffractometry* offers a powerful tool.

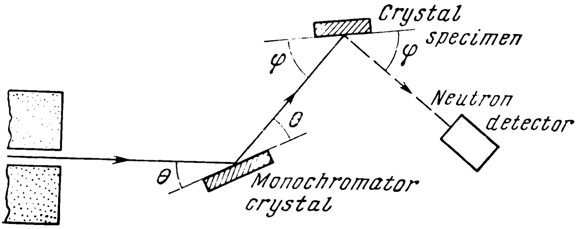


Fig. 69.5

3. The arrangement of a *neutron diffractometer* is shown in Fig. 69.5. Basically, it is similar to the apparatus shown in Fig. 69.4. A narrow beam of thermal neutrons is focused to fall on a crystal arranged to be at a fixed angle relative to the beam, which is the angle of incidence $\theta$. The crystal can only reflect the neutrons for which the de Broglie wavelength satisfies the Bragg condition. The neutron beam reflected by the crystal is monochromatic in that all of the neutrons in the beam have the same de Broglie wavelength, $\lambda$. Quite aptly, the crystal is called a monochromator. Then the monochromatic beam of neutrons reflected by the crystal falls on the crystal specimen under investigation at an angle $\varphi$. This crystal specimen and the neutron detector are rotated in synchronism so that the angles of incidence and reflection, $\theta$, of neutrons are maintained equal. The detector collects neutrons only if their angle $\varphi$ and wavelength satisfy the Bragg condition, Eq. (69.6). Then, with $\lambda$ and $\varphi$ known in advance, one can readily determine the interplanar distance $d$ of the crystal, as with X-ray diffraction.

4. Soon after the discovery of wave properties of electrons in 1929, O. Stern and his co-workers found wave properties in a beam of neutral atoms and molecules. If the molecules or atoms in a beam

travel at a velocity corresponding to the temperature $T$, one can readily determine the velocity, $v$, for the molecules. The mass of a molecule or an atom, $m$, can be found from the molar mass, $M$, namely

$$m = M/N_A$$

where $N_A$ is Avogadro's number. Then the wavelength associated with the moving atoms or molecules can be found by Eq. (69.2). For example, at 300 K the wavelengths for hydrogen ($M = 2$ kg/kmole) and helium ($M = 4$ kg/kmole) are 1.3 Å and 0.9 Å, respectively. These de Broglie wavelengths are of the same order of magnitude as the interplanar spacings of the crystal lattices of solids. Therefore, if the neutral molecules and atoms of hydrogen and helium have wave properties, the reflection of the atomic beams from crystals should produce diffraction patterns. That is, in addition to the "regular" reflection of atoms (or molecules) at an angle equal to that of incidence, a series of maxima and minima should be observed at other angles.

Stern's experiments fully bore out the above reasoning. Fig. 69.6 shows a plot of the number of atoms (or molecules) reflected by a crystal at different angles. As is seen, in addition to the principal maximum corresponding to the "regular" reflection, there are minor diffraction maxima. These minor maxima are caused by the wave properties of neutral particles and can be fully explained on the basis of the de Broglie wavelength and the relations for a two-dimensional diffraction grating which is the surface of a crystal.

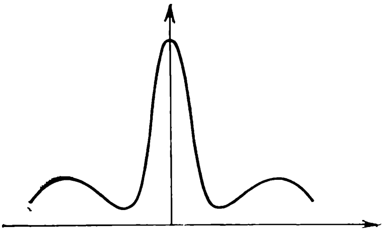


Fig. 69.6

## 69.3. PHYSICAL SIGNIFICANCE OF DE BROGLIE WAVES

1. From the last two sections it is seen that de Broglie's wave theory of matter has been experimentally verified for both charged particles (electrons) and neutral particles such as neutrons, atoms and molecules. We are then prompted to ask, and seek an answer for, two questions:

(a) can we find wave properties in the megascopic (large-scale) bodies we are dealing with in everyday practice?

(b) what is the physical significance of the waves associated with moving particles of matter?

2. To answer the first question, we should recall the relation of the mass, $m$ and the velocity, $v$, of a particle to its de Broglie wave-

length. Planck's constant, $h$, has a very small value: $h = 6.62 \times \times 10^{-34}$ J s $= 6.62 \times 10^{-27}$ erg s. For a body of mass $m = 1$ g, moving at a velocity $v = 1$ cm/s, the associated de Broglie wavelength will be $\lambda = 6.62 \times 10^{-27}$ cm. This wavelength lies far outside the capabilities of diffractometry because nature-made diffraction gratings with an interplanar distance of $d = 10^{-27}$ are non-existent. The chance of detecting wave properties in megascopic bodies becomes progressively meagre as their mass increases.

The situation is entirely different with moving particles of a mass comparable with that of an electron or a proton. As has been shown in the previous section, the de Broglie wavelength for these particles is a few Ångstrom units, and the wave properties of such particles can readily be brought out experimentally.

3. To answer the second question, we should recall the relation between the corpuscular and wave properties of light discussed in Sec. 68.7. It has been shown that the square of the amplitude of a light wave at a point in space is proportional to the number of photons falling on that point. By analogy, we may apply the concept of amplitude to the waves associated with moving particles of matter (matter waves). Then an insight into the nature of these waves can be gained by revealing the physical significance of their amplitude or intensity.

It has been noted that the intensity of a wave is proportional to the square of its amplitude. Experiments with the reflection of electrons and other particles (see Secs. 69.1 and 69.2) show that the number of reflected particles is a maximum at certain angles. From the view-point of the wave theory, the existence of maxima at some angles implies that the waves associated with the reflected particles have a maximum intensity at those angles. In other words, the intensity of the de Broglie wave increases with the number of particles at a given point in space. Conversely, *the intensity of the de Broglie waves at a given point of space determines the number of particles arriving there*. Thus, the waves associated with moving particles are statistical, or probabilistic in nature. *The square of the amplitude of the de Broglie wave at a given point in space is a measure of the probability of finding a particle at that point.*

The probabilistic interpretation of the de Broglie waves was advanced by Max Born.

It should be stressed once more that the waves associated with moving particles (matter waves) have nothing to do with the propagation of an electromagnetic field or electromagnetic waves. Among the electromagnetic, acoustical and any other waves known in classical physics, there is no counterpart of the "probability waves" associated with the moving particles of matter.

4. Consider some properties of the de Broglie waves. Let us determine the phase and group velocities of the de Broglie waves associat-

ed with particles having a mass $m$ and a velocity $v$. The phase (or wave) velocity is given by (63.16):

$$u = \omega/k$$

Multiplying the numerator and the denominator by $\hbar$ and using Eq. (69.5) and also (68.12′) valid for de Broglie waves, we get

$$u = \omega/k = \hbar\omega/\hbar k = \varepsilon/p = mc^2/mv = c^2/v > c \tag{69.7}$$

It is seen that the phase velocity of the de Broglie waves exceeds that of light in a vacuum because $v < c$. As has been found in Sec. 63.8, this does not contradict the relativity theory.

The group velocity of the Broglie waves can be determined by Eq. (63.18):

$$U = \Delta\omega/\Delta k$$

Multiplying the numerator and the denominator by $\hbar$ and using the same equations as for the phase velocity, we get

$$u = \Delta\omega/\Delta k = \hbar\Delta\omega/\hbar\Delta k = \Delta\varepsilon/\Delta p = v \tag{69.8}$$

In deriving the above expression, we have also used Eq. (16.10)

$$\Delta\varepsilon = v\Delta p$$

As is seen, the group velocity of the de Broglie wave is equal to that of the particle. This result underscores the tie-in of the de Broglie waves with moving particles.

5. The discovery that the moving particles of matter have wave properties has been an important contribution to the advance of present-day physics. Together with the quantum nature of the laws describing interatomic processes, irrefutably verified by experiment, the discovery of wave properties in the particles of matter has laid the foundation for *quantum mechanics*, a division of present-day theoretical physics dealing with the motion of particles such as atoms, molecules, atomic nuclei and the elementary particles electrons, protons, neutrons, mesons and others, with sizes in the range $10^{-10}$-$10^{-15}$ m.

# Chapter 70

# AN OUTLINE OF QUANTUM MECHANICS

## 70.1. THE CONCEPT OF THE WAVE FUNCTION

1. Classical mechanics proved inadequate to describe the wave properties displayed by the particles of matter and the probabilistic nature of the de Broglie waves. In Newton's classical mechanics,

the motion of a body or a particle under the action of a force is described by Newton's second law. In Chapter 8, we have seen that if the initial position (that is, initial coordinates) and the initial velocity are specified, the position and velocity of a body (or particle) at any next instant can be found by Newton's second law. Thus giving the position and velocity of a body at any instant chosen as reference ($t = 0$) will fully specify the state of the body. It is assumed that the accuracy with which the initial position and velocity of bodies (or particles) can be specified is solely dependent on the quality of the instruments measuring position and velocity.

2. The experiments that led to the discovery of the wave properties in particles of matter and the physical significance of the de Broglie waves call for a different approach to describing the state of a particle. From what we have learned it follows that in quantum mechanics it is only legitimate to speak in terms of the probability that a particle can be found at a particular point in, or rather infinitesimal region of, space at a particular instant.

In contrast to classical mechanics, the most that quantum mechanics can do in describing the state of a particle is to specify the infinitesimal element of configuration space, $\Delta V$, where the particle has the probability of being found at a time $t$. According to Sec. 69.3, this probability is given by the square of the amplitude of the de Broglie waves. Accordingly, quantum mechanics introduces a function, $\psi(x, y, z, t)$, called the *wave function*, or the psi-function, in which three variables determine the position of a particle ($x$, $y$, $z$), and the fourth is time, $t$. It is defined as follows: the probability, $\Delta w$, that a particle will be in an element, $\Delta V$, of configuration space, is proportional to the square of the wave function, $|\psi|^2$, and the element, $\Delta V$, of configuration space:

$$\Delta w = |\psi|^2 \Delta V \tag{70.1}$$

where $|\psi|^2$ is the square of the absolute value of the wave function. From Eq. (70.1) it is seen that it is the square of the wave function, and not the wave function itself, that has any physical meaning:

$$|\psi|^2 = \Delta w/\Delta V$$

Since $|\psi|^2$* determines the probability of finding a particle at a particular point in space at a particular time, it defines the intensity of de Broglie waves. Thus, the wave function that determines the position of a particle in space is a statistical quantity.

3. In 1926, Erwin Schrödinger proposed that the de Broglie wavelength be substituted in the classical wave equation, and from this beginning he derived a wave equation, now called the *Schrödinger wave equation*, which is the basic equation of quantum mecha-

* The quantity $|\psi|^2$ is called the *probability density*.

nics. To quantum mechanics, the Schrödinger equation is what Newton's second law is to classical mechanics. Much as Newton's law is involved in solving the problem related to the motion of megascopic bodies, so in quantum mechanics the Schrödinger equation comes in where one is dealing with the motion of microparticles.

The Schrödinger equation is especially important in describing the motion of electrons in atoms, molecules and crystals of solids. With this equation, one can readily prove that the energy of electrons in atoms, molecules and crystals can only take definite discrete values, or, in terms of quantum mechanics, the electrons can only occupy definite energy states.

It is outside the scope of this book to examine the Schrödinger equation and its corollaries in detail. Yet, in a later chapter we will note that in certain cases some physical quantities, such as energy and angular momentum, associated with moving particles can only take discrete values.

## 70.2. HEISENBERG PRINCIPLE OF INDETERMINANCY

1. The wave properties of particles and the fact that the state of a particle can be described only by giving the probability of finding it at a particular point at a particular time set a limit to using the very concepts of position (coordinates) and velocity (or momentum) in quantum mechanics. Even in classical mechanics, the concept of coordinates may sometimes be unsuitable for describing the position of an object in space. For example, there is no sense in saying that an electromagnetic wave is at a particular point in space or that the wave front in water has the coordinates $x$, $y$, $z$. The wave-particle duality of entities dealt with in quantum mechanics makes it impossible, even in the classical sense, to describe a particle by giving its position in space (coordinates) and velocity (or momentum).

2. In Sec. 61.3 it is shown that a finite wave packet (or train of waves) with a length $\Delta x$ along the $x$-axis cannot be monochromatic. That is, instead of a single frequency, the waves making up the packet should occupy a range of frequencies, $\Delta\omega$. Instead of a frequency range $\Delta\omega$, we may use a range of wave parameters, $\Delta k$. As will be recalled, the wave parameter (or wave number) is $k = 2\pi/\lambda$. According to (61.7), $\Delta x$ and $\Delta k$ are connected as

$$\Delta x \, \Delta k \geqslant 1 \tag{70.2}$$

This relation* is valid for any wave processes. Let us apply it to the de Broglie wave associated with a particle moving along the

* In a more rigorous derivation, the "$\approx$" sign is replaced with the "$\geqslant$" sign.

$x$-axis and having a momentum $p_x = p$*. From (69.2) it follows that

$$p = h/\lambda$$

or, introducing $\hbar = h/2\pi$,

$$p = 2\pi\hbar/\lambda = k\hbar \tag{70.3}$$

A similar relation should exist between $\Delta p$ and $\Delta k$, incremental changes in $p$ and $k$:

$$\Delta p = \Delta k\hbar$$

and so,

$$\Delta k = \Delta p/\hbar = \Delta p_x/\hbar$$

Substituting in (70.2) gives

$$\Delta x\, \Delta p_x \gg \hbar \tag{70.4}$$

If a particle is moving along the $y$- or $z$-axis so that the rectangular components of its momentum are $p_y$ and $p_z$, we get similar relations:

$$\Delta y\, \Delta p_y \gg \hbar, \tag{70.5}$$

$$\Delta z\, \Delta p_z \gg \hbar \tag{70.6}$$

Expressions (70.4) through (70.6) are called the *Heisenberg indeterminancy* (or *uncertainty*) *relations* after Werner Heisenberg who deduced them in 1927. In these equations, $\Delta x$, $\Delta y$ and $\Delta z$ stand for the ranges of coordinates along the three coordinate axes where the particle to which a particular de Broglie wavelength is ascribed can be found. The rectangular components of momentum are contained within the ranges $\Delta p_x$, $\Delta p_y$ and $\Delta p_z$.

The uncertainty relations show that the $x$, $y$, $z$ coordinates of a particle and the respective rectangular components of its momentum, $p_x$, $p_y$, $p_z$, cannot simultaneously be equal to $x$ and $p_x$. $y$ and $p_y$, $z$ and $p_z$, or, which is the same, the quantities $\Delta x$ and $\Delta p_x$, $\Delta y$ and $\Delta p_y$, and $\Delta z$ and $\Delta p_z$ connected by the uncertainty relations cannot be zero simultaneously. In other words, the coordinates and momentum components of a particle that may be determined simultaneously will have a range of values, or will be subject to the uncertainties resulting from (70.4) through (70.6).

3. The Heisenberg indeterminancy principle can be illustrated, taking the passage of an electron beam through a narrow slit as an example. Let this slit be arranged in an opaque screen so that it is parallel with its $x$-axis and has a width $AB$, and that the electron beam incident in the direction of the $y$-axis has a velocity $v$

* If a particle is moving in an arbitrary direction, its momentum has three rectangular components, $p_x$, $p_y$, $p_z$. In moving along the $x$-axis, $p_y = p_z = 0$, and so $p_x = p$.

(see Fig. 68.7). On the left side of the screen, each electron has a definite momentum, $p = p_y = mv$, and so $\Delta p_y = 0$. The $x$- and $z$-components of the momentum are zero, $p_x = p_z = 0$. The $y$-coordinate of each electron may be any, that is from $-\infty$ to 0. From Eq. (70.5) it follows that $\Delta y \geqslant \hbar/\Delta p_y$ so that $\Delta y = \infty$ at $\Delta p_y = 0$. This gives an entirely indeterminate position of an electron on the $y$-axis. At the instant when an electron is passing through the slit, it is contained between its edges, and the range of the $x$-coordinate is limited by the width $AB$ of the slit: that is, the uncertainty in the $x$-coordinate of the electron, $\Delta x$, is equal to the slit width, $\Delta x = AB$. By decreasing the slit width, it might seem that one can reduce the uncertainty $\Delta x$ in the $x$-coordinate to any arbitrary value and, as a consequence, improve the accuracy of one's knowledge of the $x$-coordinate of the electron to any arbitrary degree.

However, as follows from the foregoing, when the slit width becomes comparable with the de Broglie wavelength, the electrons should suffer diffraction on the slit and produce a diffraction pattern beyond it. On a fluorescent screen, $CD$, back of the slit, the observer will see the principal maximum arranged symmetrically about the $y$-axis, and the secondary maxima on each side of the principal maximum (the curve $MN$ in Fig. 68.7). It is important to note that while in front of the slit all electrons are moving along the $y$-axis and therefore have no appreciable $x$-component of momentum ($p_x = 0$), past the slit the electrons are deflected from their original path and acquire a momentum along the $x$-axis, $\Delta p_x$. From Fig. 68.7 it is seen that

$$\Delta p_x = p \sin \alpha = h \sin \alpha/\lambda \tag{70.7}$$

if one uses Eq. (69.2) for the momentum $p$.

Let us, for simplicity, take into account only the electrons that reach the screen $CD$ within the principal maximum, that is, within the angle $\alpha$ between the $y$-axis and the direction of the first diffraction minimum*. The position of this minimum on the screen is determined by the fact that the path length difference between the de Broglie waves diffracted from the upper and lower edges of the slit should be equal to the length of the wave:

$$\Delta x \sin \alpha = \lambda \tag{70.8}$$

Multiplying the left- and right-hand sides of Eq. (70.7) by the respective terms of Eq. (70.8) gives

$$\Delta x \, \Delta p_x = h$$

which is one of the Heisenberg indeterminancy relations.

* It can readily be shown that the minor diffraction maximum will not affect the reasoning and the results.

4. Strictly speaking, the indeterminancy relations are valid for megascopic bodies as well. However, the limitations they impose on coordinates and momentum in the classical sense are valid only when the wave-particle duality of megascopic objects is obvious. In all cases where the de Broglie wavelength becomes negligibly small (see Sec. 69.3), the limitations imposed by the indeterminancy principle on the description of a particle in terms of classical coordinates and velocity (or momentum) may be ignored.

In this connection, it is important to compare the basic law of motion in classical mechanics with the indeterminancy principle. This matter is discussed in detail in Sec. 14.3, and the reader is advised to go back to it again, especially to Para. 5, and to analyze the material in the light of the indeterminancy relations.

5. Apart from the indeterminancy relations for position and momentum, (70.4) through (70.6), there is one more relation of a similar kind for energy and time. It can be shown that if a particle is in a nonstationary state for a time interval $\Delta t$, the energy of this state can only be determined with an uncertainty $\Delta\mathcal{E}$. The uncertainty $\Delta\mathcal{E}$ in the energy of a particle is connected to the time interval $\Delta t$ by a relation similar to (70.4) through (70.6):

$$\Delta\mathcal{E}\,\Delta t \geqslant \hbar \tag{70.9}$$

Relation (70.9) can aiternatively be derived from Eq. (53.22) describing the nonmonochromaticity of a finite wave train*:

$$\Delta\omega\Delta t \geqslant 1 \tag{70.10}$$

where $\Delta t$ is the duration of the wave train and $\Delta\omega$ is the frequency range of the monochromatic waves making up the train. Re-writing Eq. (69.5) for incremental changes in energy and frequency, that is, $\Delta\mathcal{E}$ and $\Delta\omega$, gives

$$\Delta\mathcal{E} = \hbar\,\Delta\omega$$

Determining $\Delta\omega = \Delta\mathcal{E}/\hbar$ and substituting it in (70.10) will give the uncertainty relation of the form of (70.9). It will be shown later that this relation is of special importance to atomic and nuclear physics.

6. Sometimes it may be heard that the Heisenberg indeterminancy relations do not impose limitations on the application of the classical concepts of coordinates and momentum to quantum-mechanical particles, but only set a limit to the accuracy with which we can determine coordinates and momentum simultaneously at the present state of the art in physical experiment and theory. It is argued that ultimately, with further advances in quantum physics, the

* In a more rigorous derivation, the "$\approx$" sign should be replaced with the "$\geqslant$" sign.

coordinates and momentum of particles may be determined simultaneously with increasing accuracy.

This opinion is incorrect because the indeterminancy relations stem from the wave-corpuscular duality of particles, that is, because they express the natural limits imposed on the description of quantum-mechanical particles in terms of the classical concepts of coordinates and momentum.

7. In the light of the foregoing, it is legitimate to ask why the behaviour of quantum-mechanical objects must be described in terms of classical concepts such as position, momentum and the like, if they are not always applicable? This is done because any experiment made to obtain information about the behaviour and properties of quantum-mechanical objects is megascopic (the indication of a meter, the position of the light-spot on an oscilloscope, a photograph of a particle track, etc.). The action of any instrument used in quantum-mechanical research obeys the laws of classical mechanics and electrodynamics, and the information thus obtained is megascopic in nature, that is, it must be interpreted in terms of classical physics. It is, therefore inevitable that we must apply, at least partly, classical concepts to describe quantum-mechanical entities. Still, this applicability is limited, and the limits are described by the Heisenberg indeterminancy relations.

The interaction of a measuring device with the object under investigation is called *measurement*. It takes place in space and in time and is therefore an objective process. However, there is an important difference in the interaction of instruments with megascopic and microscopic objects. That between an instrument and a megascopic object is a megascopic process which can be described with sufficient accuracy in terms of classical physics. Furthermore, it may be thought that the instrument has no effect on the object that could not be precisely accounted for in terms of classical mechanics or electrodynamics or could not be made arbitrarily small.

The situation is different when an instrument interacts with microscopic particles. Because of the natural duality of quantum-mechanical particles, measuring, say, the position of a particle will lead to an uncertainty in its momentum, which cannot be reduced to zero, but can only be found in accordance with the Heisenberg indeterminancy relation, $\Delta p_x \geqslant \hbar/\Delta x$. This is why the effect of an instrument on the particle under investigation cannot be neglected. In fact, the instrument can change the state of the particle to a point where, say, its momentum can only be specified within a range limited by the indeterminancy relations.

8. The results of measurement are, in the final analysis, registered by an observer. Because of this, some physicists (including, above all, Heisenberg) emphasized the role played by the observer in quantum mechanics. In philosophy, this view is called subjective

idealism. As Heisenberg wrote, "While classical physics deals with objective events in space and in time to the existence of which observations are unimportant, quantum theory examines the processes which may be said to flash at the instant of observation and about which any physical statements during an interval between observations have no sense". It should be stressed that the indeterminancy relations offer no ground for these and similar idealistic conclusions.

9. One of the idealistic conclusions drawn from the indeterminancy relations is the assertion that the cause-and-effect principle cannot be applied to quantum-mechanical processes. At first glance it may appear that there is some ground for this view. Among other things, the cause-and-effect principle implies that if one knows the state of a system at some instant, one can accurately predict the state of the system at any other instant. In Newton's classical mechanics, from knowledge of the coordinates $x_0$, $y_0$ and $z_0$ and velocity components $v_{x0}$, $v_{y0}$ and $v_{z0}$ of a particle at an instant $t_0$, one can determine its position and velocity at any other instant $t$ by solving the equations of motion of that particle. This is known as *mechanical determinism*.

In contrast, since the position and velocity of quantum-mechanical particles can be determined simultaneously only with uncertainties given by the Heisenberg relations, one is prompted to conclude that the state of a system cannot be accurately specified at the initial instant, $t_0$, either, and so any succeeding stages cannot be predicted—that is, the cause-and-effect principle is no longer valid.

Actually, the state of a system in quantum mechanics is interpreted differently from classical physics, and so a different approach is necessary. The state of a quantum-mechanical entity is amply described by giving its wave function; giving the wave function for an instant $t_0$ will unambigously specify it for the instant $t > t_0$. In other words, in quantum mechanics, as in classical physics, the state of an entity specified for an instant $t_0$ uniquely defines its succeeding states in the precise agreement with the cause-and-effect principle. Thus the crux of the matter is that one cannot apply to quantum mechanics the cause-and-effect principle as it is used by classical physics and based on the "conventional" concepts of coordinates and momentum for the simple reason that the nature of microscopic objects is different.

## 70.3. MOTION OF A FREE PARTICLE

1. In this and subsequent sections we shall examine the motion of particles under conditions where their motion and energy are determined by their wave properties. It is important to differentiate

between the motion of a particle with no forces acting upon it (free motion) and the motion of a particle due to the action of forces (forced motion). In the latter case, the energy of the particle, $\mathscr{E}$, cannot take any values. If, in addition to its kinetic energy, $K$, a particle has a potential energy, $U$, its total energy, $\mathscr{E}$, is *quantized*. An observable quantity is said to be quantized when its magnitude is restricted to a discrete set of values.

2. To begin with, we shall examine a case where a particle having a mass $m$ is moving at a constant velocity $v$ in a direction chosen to be the $x$-axis; there are no forces acting on the particle, and so it is in free motion. The momentum of the particle is $p = mv$, and the de Broglie wave accompanying the moving particle has a wavelength $\lambda = h/p$ and a wave parameter $k = 2\pi/\lambda$. In Chapter 56 it is shown that a plane wave propagated in the $x$-direction and having a frequency $\omega$ and a wave parameter $k$ is described by an equation of the form (56.8):

$$s = A \cos(\omega t - kx) \qquad (70.11)$$

where $A$ is the amplitude of the wave. Eq. (70.11) may be applied to the de Broglie wave associated with a particle having a momentum $p$ and an energy $\mathscr{E}$, by expressing the frequency $\omega$ and the wave parameter $k$ in (70.11) in terms of the energy and momentum of the particle, $\omega = \mathscr{E}/\hbar$ and $k = p/\hbar$ (see Eq. (69.5) and (70.3)). Besides, the de Broglie wavelength for a free particle should be described by an equation of a more general form:

$$\psi = A \cos(\mathscr{E}t/\hbar - px/\hbar) \qquad (70.12)$$

Expression (70.12) shows that a free particle is accompanied by a plane de Broglie wave of a fixed amplitude $A$. The fact that the wave amplitude and, as a consequence, its square is constant indicates that the de Broglie wave must have a fixed, time-invariant intensity. According to the physical significance of de Broglie waves (see Sec. 69.3), this implies the same, fixed probability of finding the particle at any point on the $x$-axis. From the view-point of the indeterminancy relations, the free motion of a particle having a precisely specified momentum, $p$, means that its position on the $x$-axis is wholly indeterminate. The same follows from the equal probability of finding the particle at all points on the $x$-axis. The particle may move at any velocity $v$ corresponding to the energy $\mathscr{E} = mv^2/2$, which together with velocity can take any value.

Let us express the energy of the particle in terms of its de Broglie wavelength. According to (69.2), $\lambda = h/mv$; hence, $v = h/m\lambda$. Substituting in the expression for energy gives

$$\mathscr{E} = mv^2/2 = h^2/2m\lambda^2 \qquad (70.13)$$

Finally, noting that $\lambda = 2\pi/k$, the energy $\mathcal{E}$ may be expressed in terms of the wave parameter $k$ as:

$$\mathcal{E} = k^2h^2/8\pi^2m = k^2\hbar^2/2m \tag{70.14}$$

Referring to Fig. 70.1, the plot of the energy $\mathcal{E}$ of a free particle as a function of the wave parameter, $k$, of the associated de Broglie wave, that is, as a function of its velocity, $v$, yields a parabola.

## 70.4. THE PARTICLE IN A SQUARE POTENTIAL WELL

1. Now consider a particle in a limited motion along the $x$-axis. From the origin of coordinates, $x = 0$, to a point such that $x = L$ the particle is unimpeded in its motion. However, it is not free to leave the region $(0, L)$. In other words, at the boundary of the $(0, L)$ region, that is, at point $x = 0$ and $x = L$ the potential energy of the particle, $U$, goes to infinity (Fig. 70.2). The particle may be imagined moving over the bottom of a flat box with ideally reflecting, infinitely high walls. In such cases, the particle is said to be moving in an infinitely deep potential well and its motion is said to be limited by a potential barrier.

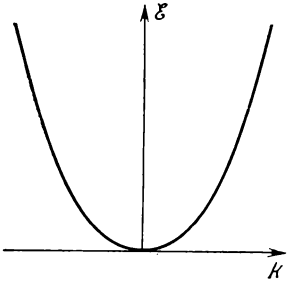


Fig. 70.1

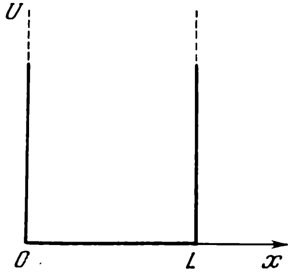


Fig. 70.2

Of course, such wells are non-existent. Yet, in investigating the electrical conductivity of metals we use this device of imagining that free (valence) electrons of the metal are confined in a flat-bottom potential box, with the potential barrier equal to the work function of the metal (see Sec. 44.9). Thus, what we have is a simplified model of a very important physical problem.

2. In this problem, the motion of the particle is constrained because it is confined in a square trap. The shape of the trap depends on the potential energy of the particle. In our case, the potential energy of the particle depends on the $x$-coordinate in a very simple way: if $x < 0$ or $x > L$, then $U = \infty$; if $0 \leqslant x \leqslant L$, then $U = 0$.

Now consider the behaviour of the de Broglie wave associated with the particle moving inside the square trap. The wave experiences reflection from the walls of the box, the incident and reflected components of the wave interfere and give rise to standing de Broglie waves. We have run into a similar situation in examining standing waves on a string fastened at both ends (see Sec. 57.2). As will be recalled, the string supported a whole number, $n$, of half-waves, Eq. (57.7):

$$n\lambda_n/2 = L \quad (n = 1,\ 2,\ 3,\ \ldots) \tag{70.15}$$

or

$$\lambda_n = 2L/n \tag{70.15$'$}$$

Thus, the length of a standing wave cannot take any value; it depends on the integers $n$, and is therefore designated $\lambda_n$. In other words, the magnitude of the wavelength is restricted to a discrete set of values.

It is obvious that the above reasoning is valid for the de Broglie wave associated with a particle moving in a square trap. The length of the potential well should hold a whole number of de Broglie half-wavelengths. Eq. (70.13) should be modified accordingly:

$$\mathscr{E}_n = mv_n^2/2 = h^2/2m\lambda_n^2 \tag{70.13$'$}$$

The subscript $n$ of the velocity $v$ and the energy $\mathscr{E}$ indicates that both cannot take arbitrary values. Thus, like the wavelength, $\lambda$, both the velocity and energy of the particle are quantized. Substituting (70.15′) in (70.13′) gives

$$\mathscr{E}_n = n^2h^2/8mL^2 \quad (n = 1, 2, \ldots) \tag{70.16}$$

Or in words, a particle confined in a square potential trap has a quantized energy directly proportional to the square of the integer $n$.

3. So far our discussion has been about any particle having wave properties and confined in a trap. To make the problem more specific, we now assume that the potential trap holds an electron which, as before, can have discrete values of energy, $\mathscr{E}_n$, called *energy levels*; the integers $n$ determining these energy levels of the electron are called its *quantum numbers*. Thus, the electron can occupy a certain energy level. Sometimes, it is stated that the electron is in a stationary quantum state, $n$. In this way, it is stressed that the state of an electron of energy $\mathscr{E}_n$ is independent of time and that in the absence of any external influences the electron can reside in that state for an arbitrary length of time.

4. Consider the effect of the linear dimensions of the trap on the quantization of energy. We shall show that the quantization of energy becomes important only if the linear dimensions of the potential box approach the size of an atom, $L = 10$ Å $= 10^{-9}$ m. To do this, we shall find the difference in energy, $\Delta\mathscr{E}$, between two adjacent

energy levels, $\mathscr{E}_{n+1}$ and $\mathscr{E}_n$. According to (70.16),

$$\Delta\mathscr{E} = \mathscr{E}_{n+1} - \mathscr{E}_n = (h^2/8mL^2)\,[(n+1)^2 - n^2] = (2n+1)\,h^2/8mL^2 \qquad (70.16')$$

Substituting in (70.16′) the numerical values $h = 6.62 \times 10^{-34}$ J s and $m = 9.8 \times 10^{-31}$ kg for an electron in a potential box with a linear dimension, $L = 10^{-9}$ m, comparable with the size of an atom, we get

$$\Delta\mathscr{E} = (2n+1)\,\frac{(6.62\times10^{-34})^2}{8\times9.8\times10^{-31}\times10^{-18}} = (2n+1)\times5.5\times10^{-20}\ \text{J}$$
$$= (2n+1)\times0.34\ \text{eV}$$

With increasing $n$, the spacing between adjacent energy levels increases in proportion to the series of odd numbers, $(2n + 1)$. It is important to note that the electron cannot have an energy not associated with one of the energy levels. The energies allowed to an electron in a square potential box are only those associated with the levels determined by Eq. (70.16).

A similar result can be obtained for a potential box of a megascopic size such that $L = 10^{-2}$ m;

$$\Delta\mathscr{E} = (2n+1)\,\frac{(6.62\times10^{-34})^2}{8\times9.8\times10^{-31}\times10^{-4}} = (2n+1)\times5.5\times10^{-34}\ \text{J}$$
$$= (2n+1)\times3.4\times10^{-15}\ \text{eV}$$

The energy levels, or states, are spaced so closely apart that they may be regarded as being practically continuous, or quasi-continuous (see Fig. 70.3). The quantization of energy of an electron in a trap of macroscopic size yields results only insignificantly differing from those given by classical physics which allows the electron to have any energy values, that is, which allows the electronic energy to vary continuously. It should be noted that at $L \to \infty$, the sequence of energy levels or states becomes strictly continuous, because $\Delta\mathscr{E} \to 0$.

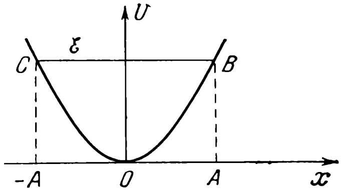


Fig. 70.3

5. It is useful to recall (see Sec. 14.3) that as regards the megascopic motion of a particle, the limitations imposed by the indeterminancy relations may be ignored and that the motion may well be described in terms of the trajectory of a particle. In contrast, it is not legitimate to apply the trajectory concept to the motion of an electron in an atom, where it is confined in a trap comparable in size with the atom.

It is thus seen that in a trap of a macroscopic size the energy of an electron behaves in a classical manner: it can take a continuity of values. The situation is different when an electron is confined in a trap of atomic size. Not only is the concept of trajectory no longer applicable, but also the energy of the electron, its key characteristic, becomes quantized. That is, it may only change by a fixed amount as the electron moves from one energy level or state to another. This point is fundamental to all of quantum mechanics, regardless of the shape of the potential well (trap) holding the electron or any other elementary particle.

6. To determine the effect of the quantum number, $n$, on the quantization of energy, we shall use Eq. (70.16′) for incremental energy $\Delta\mathscr{E}$ and write the relation $\Delta\mathscr{E}/\mathscr{E}_n$. Then

$$\Delta\mathscr{E}/\mathscr{E}_n = (2n+1)/n^2 \qquad (70.17)$$

With great quantum numbers, $2n+1 \approx 2n$, and (70.17) gives

$$\Delta\mathscr{E}/\mathscr{E}_n = 2/n \qquad (70.18)$$

As is seen, at $n \gg 1$, $\Delta\mathscr{E}/\mathscr{E}_n \ll 1$ or $\Delta\mathscr{E} \ll \mathscr{E}_n$, or in words, as the quantum number $n$ increases, the difference in energy between adjacent energy states increases at a slower rate than the energy of each state. This may be interpreted as saying that with increasing $n$ the energy levels should come closer together. At very high quantum numbers, the quantization of energy leads to results very close to those yielded by a classical treatment, that is, the levels become quasi-continuous. This is in agreement with the *correspondence principle* established by Bohr in 1922, which states that *in the limit of high quantum numbers the predictions of quantum theory agree with those of classical physics.*

In a broader statement, the correspondence principle asserts that any physical theory that has arisen from classical physics is tied in with the original classical theory—in certain limiting cases the predictions of the new theory should agree with those of the old one. As an example, in Chapters 12 and 13 we can see that the equations of kinetics and dynamics derived in the special theory of relativity reduce to those of Newton's classical mechanics at velocities $v \ll c$ such that

$$(v/c)^2 \to 0.$$

In Secs. 57.6 and 65.1 it has been shown that the predictions of wave optics agree with those of geometrical optics if the wavelength of light is negligible in comparison with any distances encountered in a given problem and also if it is assumed that $\lambda \to 0$. For quantum and classical mechanics this correspondence exists because we may neglect that $h$ is finite and set it approximately equal to zero, $h \approx 0$.

## 70.5. THE LINEAR HARMONIC OSCILLATOR IN QUANTUM MECHANICS

1. Let a particle of mass $m$ be moving along the $x$-axis under the action of a quasi-elastic force, $F = -kx$, proportional to the displacement of the particle from an equilibrium position. Here, $k$ is the elastic constant (see Sec. 8.4). This particle, called the linear harmonic oscillator, is a very fruitful model in optics and atomic physics. In examining the dispersion of light, we assumed the optical (valence) electrons of atoms (and molecules) to be vibrating under the action of the electric field associated with the light wave. In effect, we assumed them to be harmonic oscillators. The model of the atom as a harmonic oscillator, the atomic oscillator, has proved as fruitful in other applications. Furthermore, we ascribed thermal radiation to the fact that the vibrating atoms are sources of electromagnetic waves. According to Planck, each vibrating atom is a harmonic atomic oscillator whose energy may only be changed in discrete parcels (see Sec. 67.3). Planck's ideas have been verified and given further development in quantum mechanics.

2. To begin with, let us recall how the vibrations of a harmonic oscillator are treated by classical physics. Fig. 70.3 is a plot of the potential energy, $U$, of an oscillator, $U = kx^2/2$. The plot also gives the total energy, $\mathscr{E}$, of the particle. The maximum displacement of the particle from its equilibrium position is represented by points $B$ and $C$ where the velocity of the particle turns to zero and its total energy

$$\mathscr{E} = K + U(x) = mv^2/2 + U(x) \qquad (70.19)$$

becomes equal to the potential energy

$$\mathscr{E} = U(x) = kA^2/2 = mA^2\omega^2/2 \quad (\text{at } v = 0) \qquad (70.20)$$

The amplitude, $A$, of the oscillator is determined by the total energy, $\mathscr{E}$, stored:

$$A = \sqrt{2\mathscr{E}/k} = \sqrt{2\mathscr{E}/m\omega^2} = \frac{1}{2\pi\nu}\sqrt{2\mathscr{E}/m} \qquad (70.21)$$

The above expression utilizes the relation between the elastic constant $k$ and the natural frequency $\omega$ of elastic vibrations, that is, $k = m\omega^2$ (see Sec. 8.4).

From the view-point of classical physics, the vibrating particle cannot obviously leave the region $(-A, A)$. Otherwise the potential energy $U$ would exceed the total energy $\mathscr{E}$ of the particle, $U > \mathscr{E}$, and this would lead to an absurd result, namely, to a negative kinetic energy and, as a consequence, to an imaginary velocity: if $mv^2/2 < < 0$, then $v$ is an imaginary quantity.

3. Now consider a quantum-mechanical harmonic oscillator. In a quantum-mechanical approach, we should above all take into

account the wave properties of a particle confined within a parabolic potential trap (see Fig. 70.4). Because of the uncertainty principle, quantum theory yields a fundamentally different result: the total energy and amplitude of a harmonic oscillator cannot reduce to zero. For, if a particle is trapped in a range such that $\Delta x \approx A$, then, according to (70.4), $\Delta p_x \approx \hbar/A$, and the momentum $p$ of the particle cannot be zero. As is shown in Sec. 16.7

$$p \geqslant \Delta p_x \approx \hbar/A$$

On the other hand, the total energy satisfies the relation

$$\mathscr{E} \geqslant K = p^2/2m \geqslant \hbar^2/2mA^2 \tag{70.22}$$

Comparing (70.22) with (70.20) and dropping the amplitude $A$, we obtain

$$\mathscr{E}^2 \geqslant \hbar^2\omega^2/4,$$

or

$$\mathscr{E} \geqslant \hbar\omega/2$$

Thus, the energy of a harmonic oscillator cannot be lower than a minimum total energy

$$\mathscr{E}_0 = \hbar\omega/2 = h\nu/2 \tag{70.23}$$

which is called the *zero-point energy.*

4. The zero-point energy of an oscillator is solely determined by its natural frequency, ν. No cooling can remove it from the particle even at the absolute zero point of temperature, at which the oscillator still has one half quantum of energy and the motion corresponding thereto.

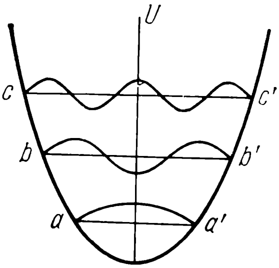


Fig. 70.4

The existence of zero-point energy has been verified by experiments on the scattering of light by crystals at extremely low temperatures. In crystals, light is scattered by the thermal vibrations of the atoms, molecules or ions located at the lattice points (see Sec. 62.8). According to classical physics, the intensity of the scattered light should tend to zero as the temperature falls to zero, because the light-scattering thermal vibrations of the lattice points should then cease. Observations have shown, however, that with decreasing temperature the intensity of the light scattered by crystals tends to a limit which remains unchanged with any further decrease in temperature. Even at the absolute zero point of temperature, the particles at the lattice points will keep vibrating ("zero-point vibrations") and scatter the incident light. Thus, the zero-point energy of atomic oscillators is associated with "zero-point" vibrations.

The zero-point energy is characteristic of any quantum-mechanical system of particles. At temperatures close to zero degrees absolute, any substance is in the crystalline state (except helium which remains a quantum-mechanical liquid down to near zero degrees absolute, provided the pressure does not exceed 22 atm), and its atoms (molecules or ions) behave as oscillators.

5. Now let us find the range of values for the total energy of a quantum-mechanical harmonic oscillator. The motion of the particle is limited by a parabolic potential curve, $U = m\omega^2x^2/2$ (Fig. 70.4). As with a particle confined within a square box, the energy of the particle in the parabolic potential trap is restricted to a discrete set of values. The quantized values of energy for the oscillator are determined by the odd number of de Broglie wavelengths that the effective length $2A_{eff} = aa'$, $bb'$ . . . can support.

Let us define the effective de Broglie wavelength:

$$\lambda_{eff} = h/p_{eff} = 2\pi\hbar/p_{eff} \tag{70.24}$$

where $p_{eff}$ is the effective momentum related to the energy as if there were no potential trap and the particle's motion were entirely free

$$\mathscr{E} = p_{eff}^2/2m = 4\pi^2\hbar^2/2m\lambda_{eff}^2 \tag{70.25}$$

Referring to Fig. 70.4, the effective amplitude $A_{eff}$ contains an odd number of effective de Broglie quarter-wavelengths:

$$A_{eff} = (2n+1)\,\lambda_{eff}/4 \tag{70.26}$$

Substituting (70.26) in (70.20) gives

$$\mathscr{E} = (2n+1)^2\, m\omega^2\lambda_{eff}^2/32 \tag{70.27}$$

Multiplying (70.25) by (70.27) eliminates $\lambda_{eff}$ and gives the discrete energy levels for a linear harmonic oscillator:

$$\mathscr{E}^2 = (2n+1)^2\,\pi^2\hbar^2\omega^2/16 \quad \text{or}$$

$$\mathscr{E} = (2n+1)\,\hbar\omega\pi/4 = \frac{\pi}{2}\left(n+\frac{1}{2}\right)h\nu$$

A rigorous quantum-mechanical approach based on the solution of the Schrödinger equation yields the following expression for the energies that a linear harmonic oscillator may have:

$$\mathscr{E}_n = h\nu\,(n+1/2) \quad (n = 1, 2, 3, \ldots) \tag{70.28}$$

Referring to Eq. (70.28), the energy levels of a harmonic oscillator make up a series of equidistant energy values (Fig. 70.5). Even the approximate calculation given above gives a correct relation between the energy and frequency $\nu$ of the linear harmonic oscillator, and also between the energy $\mathscr{E}_n$ and the quantum number $n$.

6. A rigorous quantum-mechanical treatment of the harmonic oscillator reveals one more important difference from the classical harmonic oscillator, namely that the particle may turn up outside the allowed range, $|x| \leqslant A$, that is, outside the points $C$ and $B$ in Fig. 70.3. In Para. 2 it is shown that the particle is then at a location where its total energy $\mathscr{E}$ is less than its potential energy. Owing to the wave properties of particles and the uncertainty principle, however, this is legitimate. Why this is so will be shown in greater detail in the next section.

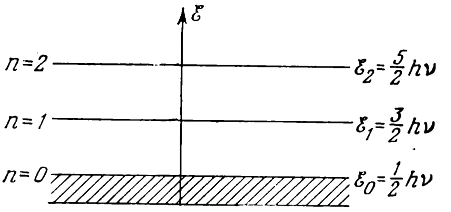

$\mathscr{E}$
$n=2$
$\mathscr{E}_2=\frac{5}{2}h\nu$
$n=1$
$\mathscr{E}_1=\frac{3}{2}h\nu$
$n=0$
$\mathscr{E}_0=\frac{1}{2}h\nu$

Fig. 70.5

## 70.6. TUNNELLING OF A PARTICLE THROUGH THE POTENTIAL BARRIER

1. In discussing a particle in a square potential well and a linear harmonic oscillator, we assumed that the de Broglie wave associated with the moving particle simply breaks off at the boundaries of the potential well. The actual process is much more elaborate. At the boundaries of a potential well, the de Broglie wave must behave as an electromagnetic wave does at the interface of two media differing in refractive index (see Sec. 63.1). As will be recalled, this wave is partly reflected from, and partly passes across, the interface. Even with total internal reflection, some of the radiation passes into the second medium (see Sec. 65.2). The de Broglie wave, too, is not only reflected from the boundary of the potential well; partly, it passes across the potential barrier. In other words, there is a probability of finding the particle in a region forbidden to it by classical physics.

2. From the view-point of classical physics, a particle is confined inside a potential box because the walls of the potential box act as a potential barrier which the particle cannot overcome. For the particle to overcome the potential barrier, it must be imparted an energy equal to or greater than the difference between the height of the barrier and its own energy, that is, its total energy should exceed the depth of the potential box.

3. Quantum mechanics leads to a basically new result, for it allows particles to *leak*, or *tunnel*, through potential barriers. Quite aptly, this is called *tunnelling*, or the *tunnel effect*.

The tunnel effect can readily be described in terms of the *penetration probability* of the potential barrier, $D$. By analogy with optics, the penetration probability of the potential barrier for the de Broglie waves may be defined as

$$D = I_t/I_i \tag{70.29}$$

where $I_i$ is the intensity of the de Broglie wave incident on the potential barrier and $I_t$ is the intensity of the de Broglie wave transmitted through the potential barrier. Thus, the *penetration probability* defines the probability of a particle passing or leaking through a potential barrier.

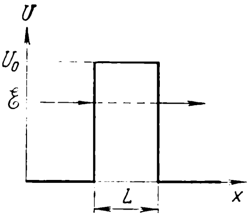


Fig. 70.6

Likewise by analogy with optics, it is legitimate to introduce the coefficient of reflection, $R$:

$$R = 1 - D$$

Quantum mechanics shows that the penetration probability depends on the shape and height of the potential barrier. For a square potential barrier of height $U_0$ and width (or thickness) $L$ (Fig. 70.6), the penetration probability, $D$, is given by

$$D = D_0 e^{-a} = D_0 e^{-2L\sqrt{2m(U_0-\mathscr{E})}/\hbar} \tag{70.30}$$

where $D_0$ is a constant close to unity, $m$ is the mass of the particle, and $\mathscr{E}$ is its total energy. The tunnel effect is a quantum-mechanical phenomenon which can take place if the penetration probability of the barrier is not too low.

For a square potential barrier, the parameter $a$ in (70.30) may be defined as

$$a = 2L\sqrt{2m(U_0 - \mathscr{E})}/\hbar$$

The particle is confined in the barrier with a linear dimension $L$, and its position is specified with an uncertainty $\Delta x \approx L$; therefore, the uncertainty in its momentum is $\Delta p \approx \hbar/L$, and the uncertainty in its energy is

$$\Delta\mathscr{E} \geqslant 3\, \frac{(p+\Delta p)^2}{2m} - \frac{p^2}{2m} = \frac{2p\Delta p}{2m} + \frac{\Delta p^2}{2m} \approx \Delta p^2/2m \approx 3\hbar^2/2mL^2$$

assuming that $p \geqslant \Delta p$ (see Sec. 16.7). For the particle to be able to penetrate the potential barrier, the uncertainty in its energy should be close to the difference between the barrier height and the particle energy:

$$U_0 - \mathscr{E} \approx \Delta\mathscr{E} \approx 3\hbar^2/2mL^2$$

Substituting this in the expression for $a$, we get

$$a \approx 2\sqrt{3} \approx 3.4$$

•

The above estimates are valid for cases where the linear dimensions of the potential barrier are comparable with the atomic size. As an example, in the macroscopic range, for $L = 10^{-2}$ m and $U_0 - \mathscr{E} \approx$ $\approx 10$ eV, the value of $a$ is about $10^8$, and the penetration probability is negligible, being $D = e^{-10^8}$. As the mass of the particle and the difference $U_0 - \mathscr{E}$ increase, the penetration probability rapidly falls off.

4. The tunnel effect has been found responsible for what is known differently as the *cold-cathode effect*, or *field emission*, or *auto-electronic emission*. Basically, this is the emission of electrons by cold conductors subjected to strong electric fields, although the field intensity is by a factor of several hundred less than would ordinarily be necessary for an electron to break its bond to the associated atom. The result may be explained in the following way. The accelerating field of intensity $E$ reduces the width $L$ of the barrier for electrons at the metal-vacuum interface, so that the electrons of an energy less in magnitude than the barrier height $U_0$ can pass through the barrier by the tunnel effect.

The tunnel effect is also behind *field ionization* which causes electrons to break loose from their atoms or molecules, thereby turning them into ions under the effect of an applied field. Here, too, the accelerating field appears to be weaker than classical physics requires it to be. A correct estimate of the accelerating field in agreement with experimental data is obtained by allowing for the passage of electrons through the potential barrier by the tunnel effect. Finally, the tunnel effect is a basic mechanism in the alpha-decay (see Sec. 81.9).

5. The tunnel effect, that is, the passage of a particle through the potential barrier, enables it to be found in a range not allowed by classical physics, that is, where its total energy $\mathscr{E}$ is less than the potential energy, $\mathscr{E} < U$, which constitutes the so-called tunnel-effect paradox. Actually, there is no paradox, and the reasoning that has led to the imaginary velocity of a particle in Para. 2 of Sec. 70.5 is invalid. The point is that the tunnel effect is a quantum-mechanical phenomenon. If we try, according to the correspondence principle stated in Para. 6 of Sec. 70.4, to describe it in classical terms by setting $\hbar \to 0$, the result will be $D \to 0$, and any talk of penetration through the barrier loses all sense. On the other hand, the quantum-mechanical description of the tunnel effect leads to a difficulty, unexpected from the view-point of classical physics, namely that one cannot represent the total energy, $\mathscr{E}$, of a particle as the sum of its kinetic energy, $mv^2/2$, and its potential energy, $U(x)$, that is,

as

$$\mathscr{E} = mv^2/2 + U(x) = p^2/2m + U(x)$$

In classical physics] this] representation is beyond any shadow of doubt and implies that both the kinetic energy and the potential energy of a particle are measurable *simultaneously* and to an arbitrary degree of accuracy. For the same reason the tunnel-effect paradox stemming from the representation of the total energy, $\mathscr{E}$, as the sum $p^2/2m + U(x)$, is non-existent in quantum theory. Specifying the position $x$ in the range of values $\Delta x$ for a particle and, as a consequence, its potential energy, $U(x)$, will immediately lead to an uncertainty in the momentum, such that $\Delta p \approx \hbar/\Delta x$. The kinetic energy of the particle, $p^2/2m$, will likewise change in an indeterminate manner.

It may be shown that the change in the kinetic energy of a particle, $\Delta K$, caused by specifying its position exceeds the difference between the barrier height $U_0$ and the total energy, $\mathscr{E}$, of the particle

$$\Delta K > U_0 - \mathscr{E}$$

In other words, $\Delta K$ exceeds the energy that a particle trapped in a potential well is lacking in order to escape from the well in a "classical" manner, that is, by *climbing over the barrier.*

# Chapter 71

# THE BOHR MODEL OF THE HYDROGEN ATOM

## 71.1. RUTHERFORD'S NUCLEAR MODEL OF THE ATOM

1. In 1911, Ernest Rutherford and his co-workers investigated the scattering of charged particles by interaction with atoms of a solid material. In his experiments, Rutherford used high-speed alpha-particles.

Alpha-particles are positive charged particles emitted from the nuclei of some heavy elements through radioactive decay. An alpha-particle has a charge $2e$ where $e$ is the electronic charge, and a mass about four times that of the hydrogen atom see Sec. 81.1). Alpha-particles emitted from radioactive elements are highly energetic. For example, alpha-particles from uranium have an energy of 4.05 MeV.

The experimental set-up used by Rutherford is shown in Fig. 71.1. Alpha-particles were emitted from a radon source, *1*, enclosed in a lead chamber so that all particles except those travelling in a narrow channel, *2*, were absorbed. The channel focused alpha-particles into a narrow beam to strike a thin gold foil, *3*, at right angles to its surface. Most of the alpha-particles passed through, and some of them were scattered. The scattered particles struck a fluorescent screen, *4*, thereby causing it to scintillate. A sufficient vacuum was maintained between the gold foil and the screen so as to avoid a further scattering of alpha-particles by air. With this arrangement, Rutherford could observe scattering at angles up to 150°.

2. It was found that, on passing through the gold foil, most alpha-particles kept moving in the same direction as before or were scattered through small angles. Yet, some were scattered through angles of the order of 135 to 150°. These large scattering angles could not be explained by the then existing concept of the atomic structure. As an explanation, Rutherford assumed that the positive charge and most of the atomic mass were concentrated in what was later called the *nucleus*, a very small central region in comparison with the remaining part of the atom which was, according to Rutherford, a cloud of negatively charged electrons totalling between them a charge equal to that of the positive charge of the nucleus. That was the inception of a *nuclear model of the atom* which has been so conducive to advances in present-day physics.

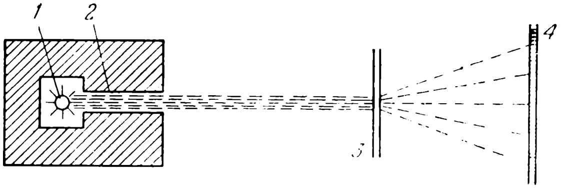


Fig. 71.1

3. The concept of the nuclear atom offers a very simple explanation for what Rutherford found in his experiments. When an alpha-particle passes through the electron shell of an atom, it deviates but little from its path because the mass of an electron is only a small fraction of that of an alpha-particle and because the negative charge of all electrons is distributed throughout the volume of the electron shell. This is why alpha-particles encountering the atomic electrons of gold pass through the foil without practically any scattering. In contrast, the alpha-particles passing near the nucleus are deviated from their original trajectory by a large amount because at short distances the forces of repulsion between a charged alpha-particle and the massive nucleus should be great. On the other hand, the probability that an alpha-particle will strike the small nucleus is low, and this is the reason why so few alpha-particles are scattered through large angles.

4. Rutherford deduced mathematical relations describing the motion of an alpha-particle in the Coulomb field of the nucleus

concentrated in a small volume containing a number $N$ of positively charged particles. His deduction may briefly be explained as follows. The Coulomb force of repulsion between an alpha-particle and a nucleus is given by

$$F = 2eNe/4\pi\varepsilon_0 r^2 \tag{71.1}$$

where $r$ is the separation between the alpha-particle and the nucleus, $\varepsilon_0$ is the dielectric constant in SI units, and $e = 1.6 \times 10^{-19}$ coulombs

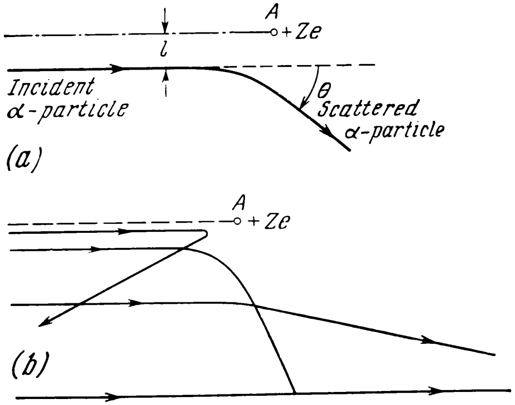


Fig. 71.2

is the elementary electronic charge. It may be shown that the repulsive force given by (71.1) will cause an alpha-particle approaching the nucleus to be deflected along a hyperbola, as shown in Fig. 71.2*a*. As is seen, an alpha-particle located at some finite distance $l$ is approaching the nucleus located at $A$. The force given by (71.1) scatters it through an angle $\theta$ and the particle traces out a hyperbola. Fig. 71.2*b* shows several hyperbolas for alpha-particles of the same energy for different values of the distance $l$.

Rutherford's formula connecting the number of alpha-particles scattered through an angle $\theta$ to the energy of the particles and the number $N$ of units of positive charge on the nucleus, can readily be verified by experiment, using alpha-particles of known energies. This was done in 1913 by Geiger and Marsden, who used thin gold and silver foil and counted the number of alpha-particles that produced scintillations on a fluorescent screen when scattered through definite angles. They concluded that the number $N$ of units of positive charge on the nucleus should be approximately equal to half the atomic weight. A little later (1932), Chadwick, an associate of Rutherford's, made more refined experiments. Using copper, silver and gold foil, he found that the number $N$ was very close to the atomic number $Z$ in Mendeleyev's Periodic Table, $N = Z$. Thus, Rutherford's nuclear model of the atom brought out the physical significance of the atomic number in the Periodic Table (see Sec. 73.2).

5. From knowledge of the nuclear charge, $Ze$, investigators determined the size of the region occupied by the atomic nucleus, that is, the upper limit of the nuclear "radius". The quotes stress the fact that it is not a precisely determinable quantity. The collision of an alpha-particle with the nucleus cannot be treated as one between two elastic spheres (see Sec. 17.3). Assuming that the nucleus and an alpha-particle are spherical in shape, the sum of their radii will be smaller than the minimum distance $d$ within which they can approach in the presence of repulsive force. Assume that an alpha-particle of a mass $m$ runs into the nucleus at a collision distance equal to zero. By the law of conservation of energy, at the distance, $d$, of closest approach the kinetic energy of the alpha-particle is wholly converted to the potential energy of electrostatic interaction, and the alpha-particle comes to a momentary stop:

$$mv^2/2 = (2e)\,(Ze)/4\pi\varepsilon_0 d \tag{71.2}$$

where $v$ is the initial velocity of the alpha-particle away from the nucleus. For alpha-particles emitted by radium-C, $v = 1.9 \times \times 10^7$ m/s. For gold ($Z = 79$), Eq. (71.2) gives $d$ as:

$$d = (2e)\,(Ze)/2\pi\varepsilon_0 mv^2 \approx 3.1 \times 10^{-14}\ \text{m}$$

The nucleus of the gold atom has linear dimensions smaller than this quantity. If we assume that an electron is a charged sphere (see Sec. 72.5), its classical radius should be of the same order of magnitude. This consideration along with other important factors examined in Chapter 80 has led to the prediction that electrons cannot exist in the nucleus.

## 71.2. CONFLICT BETWEEN CLASSICAL PHYSICS AND RUTHERFORD'S NUCLEAR MODEL OF THE ATOM

1. The nuclear model of the atom emerged from Rutherford's experiments on the scattering of alpha-particles by metal foil and mathematical studies. According to his model, all positive charge and practically all of the mass of the atom are concentrated in its central region, the *nucleus*, $10^{-15}$ or $10^{-14}$ m in diameter. The nucleus is surrounded by a region about $10^{-10}$ m in diameter where electrons with a mass only a small fraction of that of the nucleus are orbiting. As will be recalled, the mass of an electron is 1/1836.5th that of a proton, the nucleus of the hydrogen atom. Rutherford's nuclear atom looks very much like the solar system, with the nucleus occupying the centre of the system like the Sun, and with the electrons orbiting around like the planets. This is why this model of the atom is sometimes called the planetary atom.

2. The electrons of the nuclear atom cannot be at rest. If they stopped, Coulomb forces of attraction would immediately cause them to fall on the nucleus. The atom shows, however, an exceptional *stability*. For example the arrangement of the spectral lines in atomic spectra is the same for all atoms of a given chemical element. Yet, the stability of the atom cannot be adequately explained in terms of the classical laws of mechanics, electricity and optics.

As an example, we shall consider the nuclear model of the hydrogen atom, the simplest atom which consists of a single electron and a proton as the nucleus. The results will be applicable to any atom. For simplicity we assume that the electron revolves around the nucleus in a circular orbit. It should be noted that because of the wave properties of the electron and the uncertainty principle (see Sec. 14.3), the concept of orbit as a trajectory of motion is indefensible. The quantum theory defines the orbit as the locus of points where an electron in an atom has the highest probability of being found at a particular time (see Sec. 72.3). In our further discussion, we shall use the term "orbit" for an electron in an atom in its quantum-mechanical meaning.

3. We can determine the velocity of the electron revolving in a hydrogen atom in a circular orbit of a radius $r \approx 10^{-10}$ m from the centripetal force keeping the electron in orbit, or the Coulomb force of attraction:

$$mv^2/r = e^2/4\pi\varepsilon_0 r^2$$

Substituting the numerical values of electronic mass $m$, electronic charge $e$, and dielectric constant $\varepsilon_0$, we get $v \approx 10^6$ m/s, and the centripetal acceleration of the electron, $a = v^2/r = 10^{22}$ m/s$^2$.

4. It is seen that the velocity of the electron in the hydrogen atom is close to that of light, and its acceleration is such that the electron in the atom should act as an oscillator vibrating at a very high frequency. As will be recalled, such an oscillator should emit electromagnetic waves (see Sec. 59.5). Since electromagnetic waves are emitted continuously, the electron must inevitably give up its energy likewise continuously; at least, that is what follows from the application of classical laws to the nuclear atom. Then, by the same token, the atom cannot be stable, because an electron continuously giving out its energy cannot keep on its circular orbit; instead, it must approach the nucleus along a spiral path and land on it in $\tau \approx 10^{-10}$ s. Also, the frequency of the electron should likewise change continuously, and so should the frequency of the electromagnetic waves emitted by the electron. In other words, the hydrogen atom should have a continuous emission spectrum; that is, it cannot have a line spectrum.

5. Thus, the application of the classical laws of mechanics, electricity and optics to Rutherford's nuclear atom ends up in a com-

plete disagreement with observations. As we have seen above, if one used the classical theory, then: (a) the atom must be unstable because in emitting electromagnetic waves the electron must give up its energy continuously, and (b) there must be no spectral lines; instead, there may only be a continuous emission spectrum.

Actually, (a) the atom is exceptionally stable; (b) the atom emits electromagnetic waves under certain conditions; and (c) the atom emits light which has a line spectrum stemming from the structure and properties of the atom's electron shell.

The complete discord between the predictions based on the classical interpretation of the nuclear atom and observed data led to the abandonment of the classical theory and the inception of present-day quantum mechanics.

## 71.3. THE LINE SPECTRUM OF THE HYDROGEN ATOM

1. The radiation emitted by luminous gases has a line emission spectrum, that is, one consisting of separate spectral lines. When light is passed through such a gas, an absorption line spectrum is produced, for each atom absorbs the wavelengths that it can emit.

The first to be studied was the spectrum of the hydrogen atom. In 1885, Balmer found that the wavelengths of the nine lines in the hydrogen spectrum known at his time could be expressed very closely by the simple formula

$$\lambda = \lambda_0 n^2/(n^2 - 4) \tag{71.3}$$

where $\lambda_0 = 3646.13$ Å, and $n$ is a variable integer which takes on the successive values 3, 4, 5, . . ., 11 for, respectively, the first, second, third... line in the spectrum.

2. Rydberg proposed a different form for (71.3):

$$\nu^* = 1/\lambda = R\ (1/2^2 - 1/n^2) \tag{71.4}$$

where $R = 10\,967\,758\ \text{m}^{-1} = 109\,677.58\ \text{cm}^{-1}$ is called the *Rydberg constant*. The reciprocal of wavelength, $\nu^* = 1/\lambda$, is called the *wave number* and gives the number of wavelengths per unit length*. The integer $n$ can take values of 3, 4, 5, . . ., 11. The Balmer-Rydberg equation has brought out the importance of integers in spectral-series formulas and has proved extremely helpful in the studies of atomic structure.

3. At present, a far greater number of lines have been found in the spectrum of hydrogen, and all of them satisfy the Balmer-Rydberg equation to a high degree of accuracy.

As is seen from (71.4), the spectral lines associated with particular values of $n$ group together into a series of lines called the *Balmer*

* In Sec. 56.7, we gave a different definition for the wave number: $k = 2\pi/\lambda$, which gives the number of wavelengths contained within a length of $2\pi$ m (or cm). In optics, the quantity $\nu^*$ is mostly used. Obviously, $k = 2\pi\nu^*$.

*series*. With increasing $n$, the spectral lines crowd together. The limit for the Balmer series is the wavelength $\lambda_{lim}$ such that $n \to \infty$:

$$\lambda_{lim} = 4/R = 3645.981 \times 10^{-10} \text{ m} = 3645.981 \text{ Å}$$

Among other series discovered in the hydrogen spectrum are those in the invisible region, notably the *Paschen series* in the infrared:

$$\nu^* = R\,(1/3^2 - 1/n^2) \quad (n = 4, 5, 6, \ldots)$$

In the far infrared, three more special series have been discovered, namely the *Brackett series*

$$\nu^* = R\,(1/4^2 - 1/n^2) \quad (n = 5, 6, 7, \ldots)$$

the *Pfund series*:

$$\nu^* = R\,(1/5^2 - 1/n^2) \quad (n = 6, 7, 8, \ldots)$$

and the *Humphrey series*:

$$\nu^* = R\,(1/6^2 - 1/n^2) \quad (n = 7, 8, 9, \ldots)$$

In the far ultraviolet on the other side of the visible region, the *Lyman series* has been discovered, represented by

$$\nu^* = R\,(1/1^2 - 1/n^2) \quad (n = 2, 3, 4, \ldots)$$

With each of these series, as $n$ increases the spectral lines crowd together and the wave number approaches a "series limit" (that is, the frequency or wavelength limit), until the series "converges". Beyond that limit, no separate lines appear, but only a faint continuum.

The series found in the hydrogen spectrum are shown in Fig. 71.3. The wave numbers expressed as cm$^{-1}$ are given on the scale on the right of the diagram. The significance of the wave-number scale will be explained later (see Sec. 71.4).

4. The frequencies (or wave numbers) of all spectral lines of hydrogen can be expressed by a single formula

$$\nu^* = 1/\lambda = R\,(1/m^2 - 1/n^2) \tag{71.5}$$

where $m$ and $n$ are integers. For each particular series, $n = m + 1$, $m + 2$, etc. For the Lyman series $m = 1$, for the Balmer series $m = 2$, for the Paschen series $m = 3$, etc. As $n$ increases, the frequencies of all series converge to the respective limits. The limits for the wave numbers of the series in the hydrogen spectrum are given by

$$\nu^*_{lim} = R/m^2$$

5. Equation (71.5) has been verified experimentally to a high degree of spectroscopic accuracy. With particular clarity this equation reveals the importance of integers in the spectral relations,

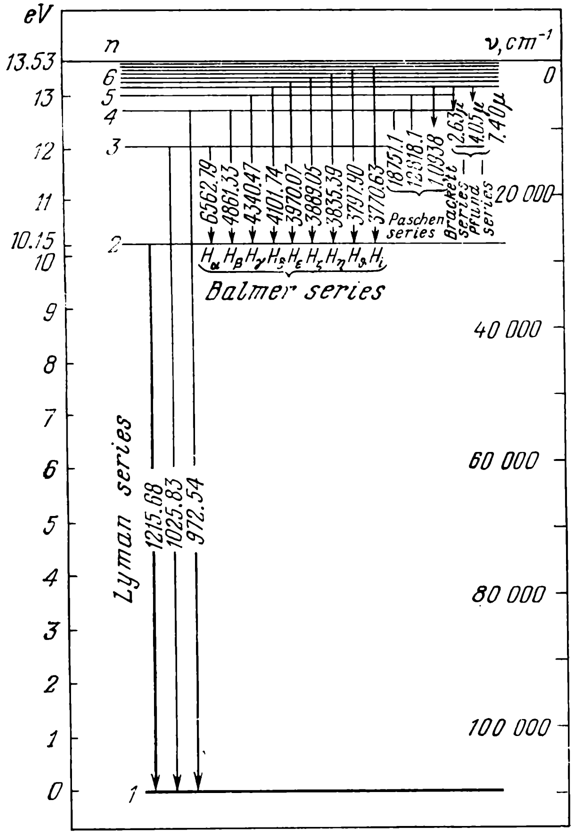


Fig. 71.3

which was fully realized only with the advent of quantum mechanics. We have seen in Chapter 70 that similar integers, the values of the quantum number $n$, define the discrete values of energy that electrons may have in a potential box or an oscillator. Jumping a little ahead, it may be noted that the numbers $m$ and $n$ in Eq. (71.5) are likewise quantum numbers defining the energy levels of the hydrogen atom.

The road from the discovery of spectral-series formulas for the hydrogen spectrum to a rigorous solution for the energy of the electron in the hydrogen atom on the basis of quantum theory has been one of outstanding discoveries and dramatism, although it spans a historically short time interval. As with all of physics in the early half of the 20th century, this road will for ever be identified with the name of the great physicist Niels Bohr.

## 71.4. BOHR'S THEORY OF ATOMIC SPECTRA

1. In 1913, Bohr came out with the first non-classical theory of the atom. The final goal of his theory was to provide a tie-in between the three predictions that existed at that time, namely:

(a) the empirical formulae for the line spectrum of the hydrogen atom discovered by Balmer and Rydberg;

(b) Rutherford's nuclear model of the atom which defied any explanation in classical terms;

(c) the quantum nature of the emission and absorption of light.

With this goal in view and maintaining a classical approach to the description of the electron behaviour in an atom, Bohr based his theory on three postulates, now known as the *Bohr postulates*. It should be stressed from the outset that the physical significance of these postulates could not be explained in terms of classical physics; in fact, they ran counter to the classical description of the electron behaviour in an atom. The true significance and meaning of the Bohr postulates were revealed later, after quantum mechanics was developed.

Bohr advanced his theory to explain the structure of the hydrogen atom and, generally, similar systems made up of a nucleus of charge $Ze$ and one electron in motion around this nucleus. Examples of such systems are singly ionized helium ($He^+$), doubly ionized lithium ($Li^{++}$), and other ions. Such systems are said to be *isoelectronic* with hydrogen. All spectral-series formulae for one-electron systems, notably Eq. (71.5), contain the product $RZ^2$ instead of $R$.

2. Bohr based his theory on three postulates. *Postulate 1* (the *postulate of stationary states*) asserts that an atom can exist only in certain stationary states unvarying with time unless influenced from without. While in one of these states, the atom does not emit electromagnetic waves. The stationary states are associated with stationary orbits in which electrons move. Although the electrons are in an accelerated motion, they do not emit electromagnetic waves. In this postulate, Bohr had abandoned the idea of energy emission by a charge in an accelerated motion supported by classical physics (see Sec. 59.6).

*Postulate 2* (the *rule of orbit quantization*) asserts that in a stationary state an electron moving in a circular orbit must have discrete, quantized values of momentum, given by the relation

$$L_n = mvr = n\hbar, \text{ where } n = 1, 2, 3, \ldots \qquad (71.6)$$

Here, $m$ is the mass of an electron, $v$ is its velocity, $r$ is the radius of the circular orbit, and $\hbar = h/2\pi$. Bohr's second postulate is simply interpreted in quantum mechanics. As with the potential box (see Sec. 70.4) and the harmonic oscillator (see Sec. 70.5), the number of wavelengths the circular orbit of length $2\pi r$ should contain must be an integral multiple of the de Broglie wavelength, $\lambda$:

$$2\pi r = n\lambda$$

Using Eq. (69.2) which defines the de Broglie wavelength, we get

$$2\pi r = nh/mv \quad \text{or} \quad mvr = nh/2\pi = n\hbar$$

which agrees with Bohr's second postulate.

*Postulate 3* (the *rule of frequencies*) asserts that as an atom jumps from one stationary state into another one quantum of radiant energy is emitted or absorbed. One quantum of radiant energy is emitted when an atom jumps from a higher into a lower energy state or, which is the same, when it moves from a more distant orbit to one closer to the nucleus. An atom absorbs energy during a transition from a lower to a higher energy state or, which is the same, from an orbit closer to the nucleus into one farther away. The emission or absorption of electromagnetic radiation by an atom brings about a change in the energy of the atom in proportion to the wavelength of this radiation. If $\Delta\mathscr{E}$ is the change in the energy of an atom due to the emission or absorption of electromagnetic radiation, $\mathscr{E}_n$ and $\mathscr{E}_m$ are the energies of the $n$th and $m$th stationary states, then the rule of frequencies may be written as

$$\Delta\mathscr{E} = \mathscr{E}_n - \mathscr{E}_m = h\nu \tag{71.7}$$

At $\mathscr{E}_n > \mathscr{E}_m$, a photon is emitted, while at $\mathscr{E}_n < \mathscr{E}_m$ a photon is absorbed. From Bohr's third postulate it follows that atoms can absorb only those spectral lines (frequencies) which they can emit. In optics, this process has, since Kirchhoff's time, been called the reversal of spectrum lines.

3. With his first and third postulates (which we shall put on a quantum-mechanical footing a little later), Bohr connected the three results listed in Para. 1. By 1913, they had been fully verified by experiment. The second postulate had remained a genius' guess until it was corroborated both experimentally and theoretically much later.

Let us compare equations (71.5) and (71.7). From this comparison we may draw a very important conclusion, namely that the energy $\mathscr{E}_n$ of a hydrogen atom in a certain stationary state is given by

$$\mathscr{E}_n = -Rch/n^2 \quad n = 1,\ 2,\ 3,\ \ldots \tag{71.8}$$

Thus, the integers that enter the spectral-series formula (71.5) determine the discrete, quantized values of energy that the hydrogen atom can have in the respective states. Thus, *the energy levels of a hydrogen atom are inversely proportional to the squares of the integers.** The

* It is assumed that the atomic nucleus is stationary and that the energy of a one-atom, or hydrogen-like, system is equal to that of the moving electron. The assumption that the nucleus is moving too will change the result but slightly.

integer $n$ defining the energy level of the hydrogen atom is called the *principal quantum number*. The energy state corresponding to $n = 1$ is called the *ground* or *normal* (*unexcited*) quantum state. All other states such that $n > 1$ are called *excited*.

The "—" sign in Eq. (71.8), implying negative energy states, is an indication that the electron experiences attraction by the nucleus or that it is bound to the nucleus by the Coulomb force of attraction (see Sec. 18.6). The absolute values of $\mathscr{E}_n$ in Eq. (71.8) give the *electron binding energy* for the $n$th state. The electron binding energy is the energy required to break an electron free from its nucleus, that is, to ionize the atom. Sometimes, it is called the *ionization potential* of that atom in a given state. It is obvious that the ionization potential is equal in magnitude to the electron binding energy for that state. For example, in the ground state ($n = 1$) the ionization potential $\mathscr{E}_{ion}$ of the hydrogen atom is 13.53 eV. The electron binding energy in the] ground state, $\mathscr{E}_1$, is $\mathscr{E}_1 = -13.53$ eV. On the left of Fig. 71.3 are the energy levels of the hydrogen atom in electron volts. The fact that the spacing between energy levels decreases with increasing $n$ implies that at $n \rightarrow \infty$, the energy $\mathscr{E}_n \rightarrow 0$. At $\mathscr{E}_\infty = 0$, the atom is ionized. The arrows in Fig. 71.3 indicate the transitions corresponding to the emission of various series of spectral lines.

## 71.5. ENERGY QUANTIZATION AND DETERMINATION OF RYDBERG'S CONSTANT ON THE BOHR THEORY

1. On the basis of his postulates, Bohr deduced the spectra of hydrogen and one-electron ions theoretically. The next objective was to deduce the relation expressed by Eq. (71.8) and Rydberg's constant already determined to a high degree of accuracy by experiment. Bohr believed that the electron in a hydrogen atom was moving in a circular orbit of radius $r$, and that in this orbit it was held by the Coulomb force of attraction equivalent to the centripetal force (see Sec. 7.1):

$$mv^2/r = (Ze)e/4\pi\varepsilon_0 r^2 \tag{71.9}$$

or, since $v = \omega r$, where $\omega$ is the angular velocity of the electron,

$$r^3 = Ze^2/4\pi\varepsilon_0 m\omega^2 \tag{71.9$'$}$$

Replacing $v$ in Eq. (71.6) with $\omega r$, and squaring the result, we get

$$m^2\omega^2 r^4 = n^2\hbar^2$$

Dividing the respective sides of the last two equations into each other, we obtain

$$r_n = n^2\hbar^2 4\pi\varepsilon_0/mZe^2 \tag{71.10}$$

or, in words, the radii of electron orbits in a hydrogen atom ($Z = 1$) are directly proportional to the square of the principal quantum number. Then the radius of the first orbit at $n = 1$, called the *first Bohr radius*, is

$$r_1 = a_0 = \hbar^2 4\pi\varepsilon_0/me^2 = 0.528 \times 10^{-10} \text{ m} = 0.528 \text{ Å} \qquad (71.10')$$

The first Bohr radius (or simply the Bohr radius) is a unit of length in atomic physics.

The energy of the electron in a hydrogen atom (or in a one-electron ion) is the sum of its kinetic energy $K$ and the potential energy $U$ of the electron in the presence of attraction by the nucleus:

$$\mathscr{E} = K + U = mv^2/2 - Ze^2/4\pi\varepsilon_0 r$$
$$= -Ze^2/8\pi\varepsilon_0 r \qquad (71.11)$$

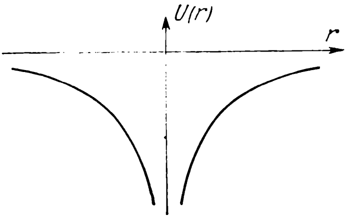


Fig. 71.4

In deriving the above expression, we have used Eq. (71.9) and also the fact that the potential energy of attraction is negative and given by

$$U = -Ze^2/4\pi\varepsilon_0 r$$

A plot of $U(r)$ for an electron in the field of a nucleus of charge $Ze$ is shown in Fig. 71.4. The nucleus is shown to reside at the origin of coordinates.

Substituting the expression for $r$ from (71.10) into (71.11) we get

$$\mathscr{E}_n = -Z^2 me^4/8h^2\varepsilon_0^2 n^2 \qquad (71.12)$$

Equating the respective terms in Eqs. (71.8) and (71.12), we get the following expression for Rydberg's constant

$$R = Z^2 me^4/8h^3\varepsilon_0^2 c \qquad (71.13)$$

It is an easy matter to see that, except the use of the Bohr postulates, Eq. (71.12) has been derived in a purely classical manner. The behaviour of an electron in an atom is described as if it were an ordinary classical particle. This is a manifestation of the inconsistency of Bohr's theory.

2. Rydberg's constant can be derived, using Bohr's principle of correspondence stated in Sec. 70.4. To begin with, let us find from Eq. (71.9′) the classical radius of the electron orbit in a hydrogen atom ($Z = 1$):

$$r = (e^2/4\pi\varepsilon_0 m)^{1/3}\,\omega^{-2/3}$$

Substituting it in (71.11), we obtain for the energy of the electron:

$$\mathscr{E} = -\frac{1}{2}\frac{(e^2)^{2/3}\, m^{1/3}\omega^{2/3}}{(4\pi\varepsilon_0)^{2/3}} \tag{71.11'}$$

Now consider the transition of the electron from the $n$th into the $(n-1)$th state at $n \gg 1$. According to the principle of correspondence, in the limit of high quantum numbers the results obtained by quantum theory must be in quantitative agreement with those obtained on the basis of classical theory. Notably, the expression for the energy of an electron should have the same form as (71.11′). As a proof, let us find from (71.5) the frequency for a transition between two adjacent levels:

$$\nu = c/\lambda = Rc\,[1/(n-1)^2 - 1/n^2]$$
$$= Rc\,(2n-1)/n^2\,(n-1)^2 \approx 2Rc/n^3$$

for $n \gg 1$ and the unity in the numerator and denominator may be neglected. Then,

$$n^3 = 2Rc/\nu = 4\pi Rc/\omega \quad \text{and} \quad n = (4\pi Rc)^{1/3}\,\omega^{-1/3}$$

Substituting the above expression for $n$ in Eq. (71.8), we get

$$\mathscr{E}_n = -Rch/n^2 = -\,(Rc)^{1/3}\, h\omega^{2/3}/(4\pi)^{2/3} \tag{71.11''}$$

By the principle of correspondence, the classical and quantum-theoretical expressions for the energy of an electron, (71.11′) and (71.11″) should agree. Equating them and raising to the third power we get the following expression for the Rydberg constant:

$$R = me^4/8h^3\varepsilon_0^2 c$$

As should be expected, the result is the same as Eq. (71.13) at $Z = 1$.

## 71.6. THE FRANCK-HERTZ EXPERIMENT

1. The Bohr theory was verified experimentally by Franck and Hertz in 1914, by passing a beam of electrons accelerated by an electric field through gases. At first, the electron beam was passed through mercury vapour. The experimental set-up is shown in Fig. 71.5. In this set-up, $K$ is the cathode, $S$ is the grid and $A$ is the anode, all arranged inside a glass envelope which also held an amount of mercury to produce mercury vapour under a pressure of about 0.1 mm Hg. The cathode was raised to incandescence and emitted electrons, and the anode was connected to a galvanometer, $G$. The difference of potential between the cathode and the grid set up an electric field which accelerated the electrons to an energy $e\varphi_1$ where $\varphi_1$ is the difference of potential between the cathode and the grid, and $e$ is the electronic charge. The electric field between

the grid and the anode was a weak retarding one, due to a potential difference, $\varphi_2$, of not more than 0.5 V.

2. On passing through mercury vapour, the electrons collide with the mercury atoms. These collisions may be of any one of two kinds. One kind is *elastic collisions* which do not change the velocity or energy of the electrons, and only change the direction of their velocities. The other kind is *inelastic collisions* in which the electrons transfer some of their energy to the mercury vapour.

Elastic collisions between the electrons and the mercury atoms cannot prevent the electrons from reaching the anode. Therefore, they cannot affect the value of anode current through the mercury-vapour tube, which is solely controlled by the accelerating field or, rather, by the potential difference between the cathode and the grid.

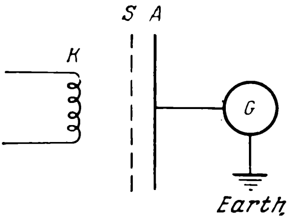


Fig. 71.5

In contrast, during inelastic collisions the electrons may lose so much of their energy that they will not be able to overcome the weak retarding field between the grid and the anode, and anode current may drop to practically zero.

According to Bohr's first postulate, the amount of energy that a mercury atom can receive from a colliding electron is just sufficient for the atom to undergo a transition to one of excited states. The excited state nearest to the ground or normal state for a mercury atom is separated by an energy gap of 4.86 eV. Until the electrons accelerated by the field acquire an energy of $e\varphi_1 = 4.86$ eV, they experience only elastic collisions in which they lose no energy and reach the anode so that anode current is rising. As soon as an electron gains an energy of 4.86 eV an inelastic collision may take place, the electron may transfer all of its energy to a mercury atom, and this may jump from the ground into the excited state. Obviously, the electron cannot now overcome the weak retarding field between the grid and the anode and will fail to reach the anode.

Thus, at a potential difference of 4.86 V between the cathode and the grid, anode current should show a sudden decrease. Similar events should happen when the potential difference is 2 × 4.86 V, 3 × × 4.86 V, etc., that is, when electrons may experience two, three or more inelastic collisions. This relation between anode current and the difference in potential, $\varphi_1$, between cathode and grid in the Franck-Hertz experiments is shown in the plot of Fig. 71.6. As is seen, anode current suddenly decreases at $\varphi_1$ equal to 4.86 V, 9.72 V and 14.58 V, in accordance with Bohr's first postulate.

3. The Franck-Hertz experiments came as a proof of Bohr's third postulate in the following way. The mercury vapour in the tube was

noted to emit ultraviolet light at a wavelength of 2537 Å. This emission is explained by the fact that the mercury atoms excited by their collisions with electrons can reside in an excited state for a very short time interval, about $10^{-8}$ s, and then jump back to their ground or normal state.* According to Bohr's third postulate, each transition of an atom to its ground state is accompanied by the emission of a quantum of radiant energy, or a photon of magnitude $\Delta\mathscr{E} = h\nu$. From $\Delta\mathscr{E} = 4.86$ eV $= 4.86e$ J, where $e = 1.6 \times 10^{-19}$ C is the electronic charge, the wavelength of emitted light is found to be

$$\lambda = c/\nu = hc/\Delta\mathscr{E} = 2537 \times 10^{-10}\ \text{m}$$
$$= 2537\ \text{Å}$$

This result fully checks with experiment, for mercury vapour emits primarily this wavelength.

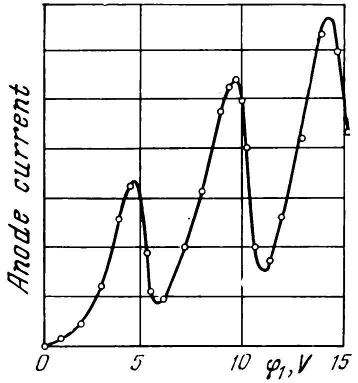


Fig. 71.6

4. Apart from an explanation of the line spectrum of hydrogen, Bohr's theory threw much light on the physical aspects of so-called characteristic X-rays (see Sec. 73.4), and some other phenomena lying outside the scope of this book. On the whole, Bohr's theory contributed immensely to the establishment of atomic physics. While it was in the making (between 1913 and 1925), very important discoveries were made, some of which are taken up in this book. These discoveries added to the wealth of present-day physics. Bohr's theory was especially stimulating to the advances of atomic and, partly, molecular spectroscopy. On the basis of Bohr's theory, a huge accumulation of experimental data on the spectra of atoms and molecules was presented in a systematic way and reduced to semi-empirical relations.

Yet, Bohr's theory showed some drawbacks from the outset. The most important among them was its inherent inconsistency because it was an attempt to reconcile classical physics and quantum-theoretical postulates. The gravest failure of Bohr's theory was its inadequacy as regards helium and generally any systems with a nucleus and more than one electron. Further developments showed that, successful in explaining some facts and incapable of interpreting others, Bohr's theory was a stepping stone to a consistent theory

* Actually, the valence electrons of an excited mercury atom jump from their normal or ground level to an excited level. The emission is just the reverse transition of the electrons.

of atomic and nuclear phenomena, quantum mechanics. Some of the basic principles of quantum mechanics have been discussed in Chapter 70, others will be taken up in the succeeding chapters.

# Chapter 72

# ONE-ELECTRON SYSTEMS IN QUANTUM MECHANICS

## 72.1. QUANTIZATION OF ELECTRON ENERGY OF THE HYDROGEN ATOM IN QUANTUM MECHANICS

1. The relations derived in the preceding chapter for the hydrogen atom on the basis of Bohr's postulates may be deduced by quantum mechanics without these postulates. Unfortunately, it is beyond the scope of this book to show how the motion of an electron in a Coulomb field of charge $e$ is treated in quantum mechanics. This involves the solution of the Schrödinger equation for an electron having a potential energy

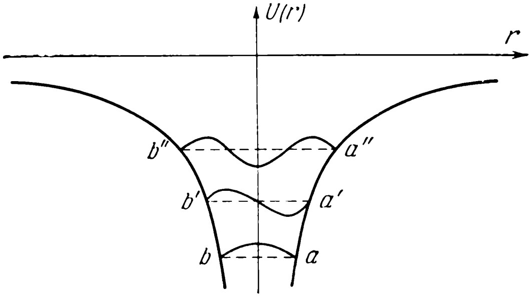


Fig. 72.1

$$U = -ee/4\pi\varepsilon_0 r$$

where $r$ is the spacing between the electron and the nucleus (Fig. 72.1). The most important result of this solution is that if the electron is in a hydrogen atom, is bound to it, and its total energy $\mathscr{E}$ is negative (see Eq. 71.11), then the electron should be in a periodic motion and its energy $\mathscr{E}$ is restricted to a discrete set of values given by Eq. (71.12) at $Z = 1$:

$$\mathscr{E}_n = -me^4/8h^2\varepsilon_0^2 n^2$$

where $n$ is the principal quantum number ($n = 1, 2, 3$, etc.).

2. Let us derive Eq. (71.12) qualitatively, using the same technique as in Sec. 70.5 for the harmonic oscillator. As the point of de-

parture we assume that the effective length within the range allowed by the potential trap (intercepts *ab*, *a'b'*, *a"b"*, etc. in Fig. 72.1) should be a multiple of the de Broglie half-wavelength (see Sec. 70.5). Since the effective length *l* is determined by the energy $\mathscr{E}$, the shape of the potential curve decides the quantization of energy. On the basis of equality between the potential and total energy of the electron, on the walls of the potential box, that is, at points *a*, *a'*, *a"*, *b*, *b'*, *b"*, etc. in Fig. 72.1, we have

$$\mathscr{E}_n = -e^2/4\pi\varepsilon_0 l_n \quad \text{or} \quad l_n = -e^2/4\pi\varepsilon_0\mathscr{E}_n \tag{72.1}$$

We can define the effective de Broglie wavelength, $\lambda_{eff}$, in a way similar to Eq. (70.13'). To begin with,

$$\overline{mv^2}/2 = h^2/2m\lambda_{eff}^2 \tag{72.2}$$

Then,

$$h^2/2m\lambda_{eff}^2 = \mathscr{E} - \overline{U} \tag{72.3}$$

and

$$n\lambda_{eff}/2 = 2l_n \tag{72.4}$$

where $n = 1, 2, \ldots$

The problem reduces to finding $\overline{U}$, which cannot be done by elementary methods. If, for simplicity, we assume that the electron has an equal probability of being found in any place inside the potential box, then we can determine that $\overline{U} = 3\mathscr{E}_n/2$ in a relatively simple (but not elementary) way. As regards the seeming violation of the law of conservation of energy (the potential energy appears to be greater than the total energy), we refer the reader to Sec. 70.6 where this matter has been discussed in detail. Now Eq. (72.3) gives

$$h^2/2m\lambda_{eff}^2 = -\mathscr{E}_n/2$$

whence

$$\lambda_{eff} = h/\sqrt{-m\mathscr{E}_n} \tag{72.3'}$$

Substituting this result and also Eq. (72.1) in (72.4), we find

$$nh/\sqrt{-m\mathscr{E}_n} = 4\,(-e^2/4\pi\varepsilon_0\mathscr{E}_n) \tag{72.4'}$$

Squaring both sides of equation (72.4'), we find $\mathscr{E}_n$ as

$$\mathscr{E}_n = -me^4/\pi^2h^2\varepsilon_0^2n^2 \tag{72.5}$$

Let us compare this equation with (71.12) for the electron energy levels in the hydrogen atom, deduced from the solution of the Schrödinger equation. As is seen, we have derived a correct expression for energy as a function of the principal quantum number and the universal constants *m*, *e* and *h*. The only difference from the exact equation is that instead of 8, the denominator has $\pi^2 \approx 9.98$. Of course, the above reasoning ought not to be taken as a derivation of an expression for the energy of the hydrogen atom. Its objective has been to illustrate the dependence of the energy $\mathcal{E}_n$ on the shape of the potential curve and to show that the energy of the electron displaying wave properties and moving in the Coulomb field of the nucleus in the hydrogen atom is restricted to a discrete set of values inversely proportional to the square of the principal quantum number.

## 72.2. QUANTIZATION OF ANGULAR MOMENTUM

1. Quantum mechanics has given a new significance to Bohr's second postulate. Thus, the angular momentum $L_1$ of an electron in any atom, and not only in the hydrogen atom, is restricted to a discrete set of values given by

$$L_l = \sqrt{l(l+1)}\,\hbar \tag{72.6}$$

where $l$ is the *orbital quantum number* whose limiting value is one less than that of the principal quantum number, or

$$l = 0, 1, 2, \ldots, (n-1) \tag{72.7}$$

From a comparison of Eq. (72.6) with Bohr's second postulate, Eq. (71.6), it is an easy matter to note that the result obtained by quantum mechanics differs from that of classical theory in the dependence of $L_l$ on the quantum numbers. Instead of the principal quantum number $n$ which enters the quantization rule for circular orbits, Eq. (71.6), the expression $\sqrt{l(l+1)}$ includes $l$, the orbital quantum number. It may be noted, though, that at $l \gg 1$, such that $l+1 \approx l$, Eq. (72.6) yields

$$L_l = l\hbar$$

which looks like Bohr's postulate

$$L = n\hbar$$

It is very important, however, that the limiting value of $l$ extends from zero to one less than that of the principal quantum number, $(n-1)$. Quantum mechanics allows an electron in any atom to be in a state in which the angular momentum, $L_l$, *is zero*. In Bohr's theory, this state is associated with the so-called "pendulum orbit" passing through the atomic nucleus. That is, Bohr's theory rules out the existence of states such that $l = 0$. It has been shown by experiment that the states in which an electron has no momentum associated with its orbital motion do exist (see Sec. 42.10).

2. The fact that the orbital quantum number may take on various values provides a basis in atomic physics for a classification of elec-

tronic states in atoms and molecules. The following notation has been adopted:

if $l = 0$, the electron is in the *s*-state;
if $l = 1$, the electron is in the *p*-state;
if $l = 2, 3$, etc., the electron is, respectively, in the *d*-, *f*-, etc. state, in the same order as the letters stand in alphabet.

## 72.3. PHYSICAL SIGNIFICANCE OF BOHR ORBITS IN QUANTUM MECHANICS

1. It has been shown in Sec. 14.3 that for an electron in an atom the Heisenberg uncertainty relations rule out the very idea of a trajectory. This brings us to ask what physical significance the Bohr orbit has in quantum mechanics? Consider this problem in detail, taking as an example the *s*-state of the electron in a hydrogen atom at $n = 1$, that is, the ground, unexcited quantum state. Computations based on the Schrödinger equation show that the probability of finding the electron in any point inside the atom depends only on the distance $r$ of the electron from the nucleus. That is, the electron has an equal probability of turning up at all points on the sphere of radius $r$ and with the nucleus as centre. In other words, the distribution of the probability of finding the electron in the atom is spherically symmetrical.

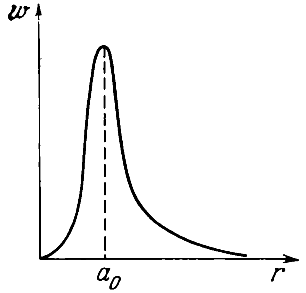


Fig. 72.2

2. This does not imply, however, that the electron has an equal probability of being found at any distance from the nucleus. Computations show that in quantum mechanics the probability $w$ $(r)$ of finding the electron at a given distance from the nucleus has the shape of the curve shown in Fig. 72.2. The probability is a maximum at a distance $r$ from the nucleus equal to the (first) Bohr radius $a_0$ Eq. (71.10′). Thus, the *Bohr orbits of the electron in an atom are loci of points where the probability of finding the electron is a maximum.* It is in this sense that the term electron orbits in an atom will be used in our further discussion.

## 72.4. SPACE QUANTIZATION OF ANGULAR MOMENTA

1. In Sec. 42.2 we examined the relation between the moment of momentum or the angular momentum $\mathbf{L}_l$ associated with the orbital motion of an electron, and its magnetic moment, $\mathbf{p}_m$. The orbital

angular momentum and magnetic moment of an electron are proportional to each other, oriented at right angles to the orbital plane of the electron and opposite in direction (see Sec. 40.6). The vectors $\mathbf{p}_m$ and $\mathbf{L}_l$ are related as follows

$$\mathbf{p}_m = e\mathbf{L}_l/2m_e \tag{72.8}$$

where $e$ is the electronic charge and $m_e$ is the electronic mass.* The quantity

$$g_l = e/2m_e \tag{72.9}$$

is called the *orbital gyromagnetic ratio.*

2. In quantum mechanics, the orientation of the vectors $\mathbf{p}_m$ and $\mathbf{L}_l$ relative to the plane of the electronic orbit cannot be specified for the reasons stemming from the physical significance of orbits in quantum mechanics. We may only choose a direction in space such that the angle $\alpha$ between this direction and the vector $\mathbf{L}_l$ decides the position of $\mathbf{L}_l$ in space. This direction may be the direction of the vector magnetic intensity $\mathbf{H}$ due to the field containing the atom and its electrons (Fig. 72.3). In the absence of an external magnetic field, the direction relative to which the orientation of $\mathbf{L}_l$ is specified may be that of the internal magnetic field set up by the nucleus and all electrons except that under consideration.

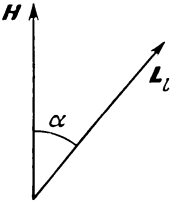


Fig. 72.3

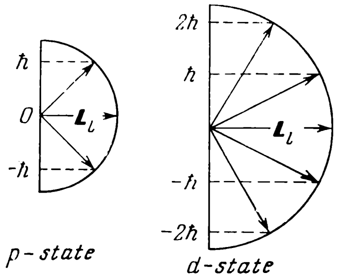


Fig. 72.4

3. The orientation of the vector magnetic moments of atoms and molecules in an external magnetic field is important to the magnetic properties of a material. According to classical physics, the vector $\mathbf{p}_m$ (or $\mathbf{L}_l$) may be oriented in an external magnetic field in any arbitrary manner. This is the basis of the classical theory of paramagnetism examined in Sec. 42.5.

4. This assumption has been proved to be in error. Quantum theory shows that the vectors $\mathbf{L}_l$ and $\mathbf{p}_m$ are *spatially quantized.* That is,

* In this and succeeding sections of the present chapter, the rest mass of an electron will be designated as $m_e$, and the letter $m$ will be reserved for another physical quantity.

*the vector angular momentum of an electron may be oriented in space only so that its component $L_{lz}$ in the z-direction of the external magnetic field is restricted to a discrete set of values which are multiples of $\hbar$:*

$$L_{lz} = m\hbar \tag{72.10}$$

where $m$ is the *magnetic quantum number* which has allowed values

$$m = 0, \pm 1, \pm 2, \pm 3, \ldots, \pm l \tag{72.11}$$

where $l$ is the orbital quantum number.

From Eq. (72.11) it is seen that the magnetic quantum number has a range of $(2l + 1)$ values. Accordingly, the vector $\mathbf{L}_l$ may have the number $(2l + 1)$ of orientations in space, that is, as many as there may be its components in the various directions of the external magnetic field. The allowed orientations of the vector $\mathbf{L}_l$ for the $p$- and $d$-electrons at $l = 1$ and $l = 2$ are shown in Fig. 72.4.

## 72.5. ELECTRON SPIN AGAIN

1. Space quantization was verified in the Stern-Gerlach experiment (see Sec. 42.10). Among other things, it unambiguously indicated that the magnetic moment (measured in the experiment) may have two orientations in an external magnetic field. If there were no space quantization and the magnetic moments $\mathbf{p}_m$ could be oriented arbitrarily, then instead of two sharply defined bands, the photographic plate would show a continuous distribution of atoms striking different points on the plate.

2. At first, Stern and Gerlach experimented with silver and other atoms in the first group of the Periodic Table of elements (see Sec. 73.2). These atoms have one outer valence electron. In the ground, unexcited state this electron is in the $s$-state, that is, it has an orbital quantum number of zero. As a consequence, the valence electrons of atoms in this group of elements and the atoms themselves have no orbital angular momentum ($L_0 = 0$)*. In their experiments with these atoms, Stern and Gerlach *could not possibly* detect any space quantization of the orbital angular momentum. Yet, as has been shown in Sec. 42.10, space quantization was in evidence unambiguously.

This contradiction, like many others to be discussed later, was not resolved until 1925 when Goudsmit and Uhlenbeck pointed out that certain features in atomic spectra could be explained if it was assumed that the electron "spins", that is, rotates about an axis through its centre of mass, and that it has both an *intrinsic angular momentum*, $\mathbf{L}_s$, and a *magnetic moment*, $\mathbf{p}_{ms}$ associated with this

* The orbital angular momenta of all electrons on inner shells (see Sec. 73.2) cancel out.

rotation or *spin*. From the Stern-Gerlach experiment it follows that the component of the spin magnetic moment of an electron in the direction of the external field is numerically equal to the Bohr magneton (see Sec. 42.10)

$$p_{msz} = \mu_B = e\hbar/2m_e \qquad (72.12)$$

3. The Stern-Gerlach experiments with the atoms of the first group of the Periodic Table were explained in a very simple manner when allowance for electron spin had been made. What was observed in the experiments was the space quantization of the *spin angular momentum*, $\mathbf{L}_s$. Like the orbital angular momentum $\mathbf{L}_l$ the spin angular momentum (or simply, spin), $\mathbf{L}_s$ and its component $\mathbf{L}_{sz}$ in the direction of the external magnetic field should be quantized. In quantum mechanics it is proved that the quantization of the spin angular momentum is given by

$$L_s = \sqrt{s(s+1)}\,\hbar \qquad (72.13)$$

Eq. (72.13) is not unlike Eq. (72.6), but instead of $l$ it contains the *spin quantum number*, $s$. This is not the end of the analogy between the orbital and spin angular momenta. It appears that the spin component $L_{sz}$ should be quantized so that the vector $\mathbf{L}_s$ may have the number $(2s + 1)$ of orientations. Since in the Stern-Gerlach experiment *two* orientations of spin were noted, it then follows that $2s + 1 = 2$, and $s = 1/2$. Thus the spin quantum number has only one allowed value and in this respect it differs from all other quantum numbers, namely the principal quantum number $n$, the orbital quantum number $l$, and the magnetic quantum number $m$. Furthermore, the spin quantum number $s$ is not an integer. The numerical value of the spin angular momentum may be found from Eq. (72.13):

$$L_s = \sqrt{(1/2)(1/2+1)}\,\hbar = \hbar\sqrt{3}/2 \qquad (72.13')$$

4. By analogy with the space quantization of the orbital angular momentum $L_l$ we may write the following expression for the space quantization of the spin angular momentum $L_s$:

$$L_{sz} = m_s\hbar$$

where $m_s$ is the *magnetic spin quantum number* which has two allowed values, namely $m_s = \pm 1/2$, because the spin component in the direction of the magnetic field likewise has two allowed values:

$$L_{sz} = \pm\hbar/2 \qquad (72.13'')$$

Physicists loosely say that the electron spin is equal to $\pm\hbar/2$ and may be oriented either with or against the direction of the magnetic field. What they actually have in mind is not the spin given by Eq. (72.13'), but its component $L_{sz}$. We, too, shall use spin in this meaning.

Eq. (72.12) is often interpreted to mean that the spin magnetic moment of an electron is equal to the Bohr magneton. This is a misnomer, too, because the absolute value of the spin magnetic moment component in the direction of the magnetic field is actually meant.

5. We shall use the results of the Stern-Gerlach experiment in the form of Eq. (72.12) and the space quantization of the spin angular momentum as expressed by Eq. (72.13″) in order to find the ratio of $p_{msz}$ to $L_{sz}$:

$$p_{msz}/L_{sz}=e/m_e=g_s \qquad (72.14)$$

The ratio of vector projections is equal to the ratio of the numerical values of the vectors themselves, that is:

$$p_{ms}/L_s=e/m_e=g_s \qquad (72.14')$$

The ratio $g_s = e/m_e$ is called the *spin gyromagnetic ratio*. From a comparison of (72.14′) and (72.9) it is seen that the spin gyromagnetic ratio is twice as great as the orbital gyromagnetic ratio. The value of the spin gyromagnetic ratio for ferromagnetics (see Sec. 42.9) was measured in the Einstein-de Haas experiment. As a result, a better insight was gained into the internal magnetic field of ferromagnetics and a quantum theory of ferromagnetism was advanced.

6. As a way of visualizing the electron spin, it is often said that the intrinsic angular momentum $L_s$ and the intrinsic magnetic moment $p_s$ of an electron are associated with its rotation about an axis through its centre of mass. It is sometimes stressed that this rotation makes the atomic structure look still more like the solar system in which the planets revolve around the Sun and rotate about their axes. This visualization, however, leads to major difficulties.

Imagine an electron as a ball of radius $r$, carrying charge $e$ on its surface (the structure of elementary particles is discussed in greater detail in Sec. 83.8). The radius $r$ can readily be found from the condition that the potential energy of a charged ball, $e^2/4\pi\varepsilon_0 r$, is its relativistic rest energy, $m_0c^2$ (see Secs. 18.8 and 16.1):

$$e^2/4\pi\varepsilon_0 r = m_0c^2$$

whence

$$r=e^2/4\pi\varepsilon_0 m_0c^2=2.8\times10^{-15}\text{ m}$$

on substituting the numeric values of all constants.

Now assume that this ball-shaped electron of classical radius is rotating about its axis and has an angular momentum $L_s$ of the amount $\hbar/2$. Let us find the linear velocity $v$ at which the points on the surface of the ball must move. Since

$$L_s = \hbar/2 = m_0vr$$

then

$$v = \hbar/2m_0r$$

Substituting the values of $\hbar$, $m_0$ and $r$ in this expression, we find that $v > c$. This result obviously runs counter to the main postulate of the special theory of relativity which asserts that nothing can move faster than light (see Sec. 12.6).

Thus, the visualization of electron spin as rotation is unsatisfactory. The spin of a particle, notably that of an electron, is just another property of particles, like its rest mass $m_0$ and electric charge $e$.

## 72.6. THE FINE STRUCTURE OF THE SPECTRUM

1. We have learned (in Sec. 71.3) that emitting atoms produce line spectra consisting of separate spectral lines. An example is the spectrum of the hydrogen atom examined in detail in Sec. 71.3. A closer examination of atomic spectra will show that some spectral lines are actually multiplets, that is, sets of several lines. For example, the spectrum of the sodium atom shows a bright yellow line, called the $D$-line. Present-day spectrometers unmistakably show this line to be a doublet, that is, one consisting of two spectral lines of wavelengths 5890.0 Å and 5895.9 Å. This is called the *fine structure*.

2. The occurrence of a spectral line as a multiplet, has been found to be due to the interaction between the orbital angular momentum and the spin angular momentum of the electrons in the emitting atoms. This interaction causes the energy levels of an excited electron to split up, and instead of a single spectral line it produces two, three, or more, as the case may be.

The part played by the spin angular momentum in the occurrence of fine spectrum structures may conveniently be studied, taking as an example an atom with one outer, optical electron.* Because of the spin magnetic moment, $p_{ms}$, the electron behaves like a magnetic dipole placed in the magnetic field set up by the orbital motion of that same electron. The behaviour of such a dipole has been examined in Sec. 41.10. As will be recalled, a dipole placed in an external field gains energy. This interaction causing a change in the energy of the electron and splitting up of its excited energy states is called the *spin-orbit effect* or *interaction*.

Depending on the orientation of the spin with or against the chosen direction of the magnetic field associated with the orbital motion of the electron, the energy gained by the electron due to its spin

* The electronic spectra of atoms are produced by changes in the state of their outer valence electrons. The electrons on the inner shells (see Sec. 73.2) do not contribute to the atomic spectra.

may be either smaller or greater than the energy the electron has due to its orbital motion alone.

Fig. 72.5*a* shows the relative positions of the ground and excited levels of an electron, leaving out its spin. In Fig. 72.5*b* it is shown that the spin-orbit effect causes the energy level of the electron to split into two. One of the new states has a lower energy and corresponds to the spin oriented with the field. The other has a greater energy and corresponds to the spin oriented against the field. This is a *doublet* energy level.

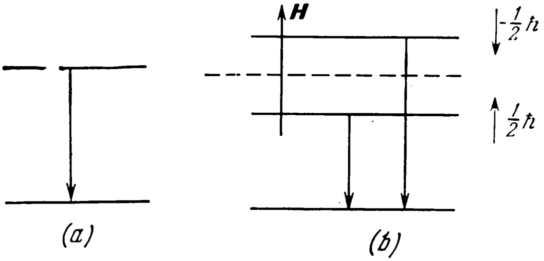


Fig. 72.5

3. In conclusion, it may be stated that, allowing for the spin of the electron, its state in the atom is specified by *four quantum numbers*, namely the principal quantum number $n$, the orbital quantum number $l$, the magnetic quantum number $m$, and the magnetic spin quantum number $m_s$ (or simply the spin quantum number $s$).

## 72.7. QUANTUM-THEORETICAL INTERPRETATION OF BOHR'S POSTULATES

1. Using Bohr's postulates as a basis, we examined in Sec. 71.4 the emission of spectral lines by an excited atom and also the absorption of light by atoms. The emission and absorption of radiation has been explained in full agreement with observations, and Bohr's postulates have been interpreted by quantum mechanics. We shall take up only some of the results obtained by quantum mechanics. Their deduction lies outside the scope of this book.

2. Consider the electron in a hydrogen atom (or in a one-electron ion)*, in a certain energy state. Let $n$ be the principal quantum number characterizing this state and determining the electron energy $\mathscr{E}_n$. As will be recalled (see Sec. 70.1), the state of an electron in quantum mechanics is uniquely specified by giving its wave function. Let $\psi_n$ $(x, y, z, t)$ be the wave function of an electron in a state of

* This limitation to one-electron systems does not subtract from the generality of the results obtained by quantum theory and is only used to simplify its comparison with Bohr's theory applicable solely to one-electron systems.

energy $\mathscr{E}_n$. The probability that the electron will be in a particular element of volume $\Delta V$ at a particular time is proportional to the square of its wave function, and so we get $|\psi_n|^2\Delta V$. The most important result obtained by quantum mechanics is that if an electron is in an energy state characterized by the principal quantum number $n$, *the probability that the electron will be found in a particular volume in the atom is independent of time*, that is, time-invariant. From the view-point of classical theory, an electron in such a state is not oscillating in the atom and cannot radiate, that is, its energy $\mathscr{E}_n$ remains unchanged. The state of an electron characterized by a particular energy $\mathscr{E}_n$ is called *stationary*, that is, time-invariant. This is Bohr's first postulate.

3. It is noted in Sec. 71.4 that Bohr's second postulate has likewise been explained by quantum mechanics in simple terms. This is also true of his third postulate (the rule of frequencies). If external influences affect the state of an electron so that it is forced to jump from state $n$ to state $m$, the probability that the electron will be in an element of volume $\Delta V$ in the atom is no longer given by $|\psi_n|^2\Delta V$. If the electron undergoes a quantum transition between states $n$ and $m$, it will reside part of the time in state $n$ and part of the time, in state $m$. Reasoning along the same lines as in Para. 2 above, we may say that the probability that the electron will be in a particular element of volume $\Delta V$ is now given by $\psi_n\psi_m\Delta V$. It is proved in quantum mechanics that under the circumstances the electron has a dipole electric moment which varies periodically with time. The frequency, $\omega$, at which the dipole moment changes is the same as the frequency of radiation emitted when the electron jumps from state $n$ to state $m$ and is given by

$$\omega = (\mathscr{E}_n - \mathscr{E}_m)/\hbar$$

in precise agreement with Bohr's third postulate.

We have seen that a correct theoretical explanation of Bohr's three postulates has only been given by quantum mechanics. This important result shows that Bohr's theory has only been a stepping stone to the present-day theory of atoms, molecules and their systems. Bohr's theory has been of paramount importance to all of physics. Yet one must clearly realize that it is only with the advent of quantum mechanics that a consistent theory of the structure and properties of microscopic particles came into being. The succeeding chapters will, as we hope, prove this point.

## 72.8. SPONTANEOUS EMISSION AND ABSORPTION OF LIGHT

1. According to quantum mechanics, an atom will remain in any stationary state for any arbitrary time, however long, provided there are no external factors that may change its energy. On the other

hand, observations show that an excited atom can jump to its ground or normal state of its own accord and emit light in the process. This is the *spontaneous emission of radiation.* An explanation of spontaneous transitions from higher to lower energy states called for further advances in quantum mechanics and has been accomplished by quantum electrodynamics which deals with the quantization, generation and collapse of the electromagnetic field in the most general terms. In this book we cannot give the reader even a sketchy outline of the consistent theory of the emission and absorption of light advanced by Dirac in 1927 (the Dirac electron theory).

2. In 1916, well before the advent of quantum mechanics and quantum electrodynamics, Einstein came out with a theory of radiation based on the conservation of energy during the interaction of atoms and molecules with the electromagnetic field. Consider some points of his theory.

If an atom is in a state $n$ and has an energy $\mathscr{E}_n$ at time $t$, then under internal influences whose mechanism cannot be accurately traced the atom can spontaneously jump to a state $m$ in which it will have a lower energy, $\mathscr{E}_m$. Ordinarily, an excited atom jumps to its ground or normal state where it has the lowest energy. The spontaneous transition from state $n$ to state $m$ can occur with a definite probability, because the atom can remain in state $n$ for some time. Einstein defined the probability that an atom will jump from state $n$ to state $m$ during one second as $A_{nm}$. The quantity $A_{nm}$ has come to be known as the *Einstein coefficient for spontaneous emission of radiation.* If $N_n$ atoms reside in the $n$th energy state at time $t$, then the number $\Delta N_n$ of atoms that jump to the $m$th level during an infinitesimal time interval $\Delta t$ will be proportional to the probability $A_{nm}$, the number $N_n$ of atoms in the $n$th energy level, and the time interval $\Delta t$:

$$-\Delta N_n = A_{nm} N_n \Delta t \qquad (72.15)$$

The "—" sign on the left-hand side of Eq. (72.15) indicates a reduction in the number of atoms in the $n$th level owing to their spontaneous transition to the $m$th level. Following about the same lines as in the derivation of an expression for the attenuation of waves in Sec. 55.4, we can write one for the time-dependent decrease in the number of atoms in the $n$th excited state as

$$N_n = N_{n0} \exp(-A_{nm} t) \qquad (72.16)$$

where $N_{n0}$ is the original number of atoms in the $n$th level at the initial instant of time $t = 0$. A plot of this relation appears in Fig. 72.6.

Each transition from the $n$th to the $m$th state is accompanied by the emission of a quantum of light (photon) of energy $\hbar\omega_{nm}$, where $\omega_{nm}$ is the cyclic frequency of the spectral line corresponding to the

$\rightarrow m$ transition. According to Bohr's third postulate (the rule of frequencies):

$$\mathscr{E}_n - \mathscr{E}_m = \hbar\omega_{nm}$$

3. In addition to the probability of spontaneous transitions, the spontaneous emission of radiation may be described in terms of the *mean lifetime*, $\tau_n$ *of an excited atom*, that is the mean time that an atom resides in an excited state. Stated differently, this is the time during which the number $N_{n0}$ of atoms in excited state $n$ reduces to $1/e$ of its original value

$$N_n = N_{n0}/e$$

From Eq. (76.16) it follows that at $t = \tau_n$, $N_{n0}/e = N_{n0} \exp(-A_{nm}\tau_n)$. Cancelling $N_{n0}$, we get

$$e^{-1} = \exp(-A_{nm}\tau_n)$$

or finally

$$A_{nm}\tau_n = 1$$

that is,

$$\tau_n = 1/A_{nm} \qquad (72.17)$$

Thus, the probability of spontaneous transitions, or the Einstein coefficient for spontaneous emission of radiation, is *the reciprocal of the mean life of an excited atom*.

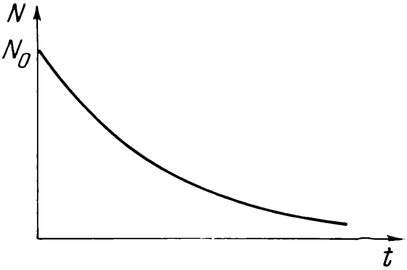


Fig. 72.6

4. The mean life of excited atoms can be measured experimentally, using Eq. (72.16) which may be re-written as

$$N_n = N_{n0} \exp(-t/\tau_n) \qquad (72.17')$$

In one of his experiments, Wien studied the radiation emitted by a beam of excited hydrogen ions travelling through a high vacuum. The experiment was arranged so as to prevent collisions and an exchange of energy between the particles. The excited ions jumped to their ground or normal state and emitted light in the process only because their mean lifetime $\tau_n$ was finite. In the experiment, the decrease in intensity was measured for each spectral line along the path of the particle beam travelling at velocity $v$ a distance $x$ such that $t = x/v$. The value of $\tau_n$ was found from Eq. (72.17'). Taking the logarithm of (72.17'), we readily get

$$\tau_n = t \ln(N_{n0}/N_n) = t \ln(J_0/J)^*$$

For the hydrogen line $H_\alpha$ ($\lambda = 6562$ Å), $\tau_n$ was found to be $1.5 \times 10^{-8}$ s, and for the mercury line ($\lambda = 2537$ Å), $\tau_n = 9.8 \times 10^{-8}$ s.

* It can be shown that the ratio $N_{n0}/N_n$ is equal to the ratio of intensities, $J_0/J$.

5. The value of $\tau_n \approx 10^{-8}$ s is typical of the lifetime of excited atoms. At the end of this time interval, they spontaneously jump to their ground state. The finite life, $\tau_n$, of an excited atom is responsible for the fact that the energy $\mathscr{E}_n$ of an excited atom can be determined only with an uncertainty $\Delta\mathscr{E}_n$ given by

$$\Delta\mathscr{E}_n \geqslant \hbar/\tau_n$$

on the basis of the Heisenberg uncertainty relations (see Eq. 70.9). The quantity $\Delta\mathscr{E}_n = \Gamma_n$ is called the *natural width* of the energy level $\mathscr{E}_n$, or, simply, the *level width.*

All excited levels have a level width other than zero because the mean life of the level $\mathscr{E}_n$ is finite, $\tau_n \approx 10^{-8}$ s. It is only the energy level corresponding to the normal or unexcited state of an atom that is infinitesimally narrow, because the atom can reside in that state for any length of time, however long, provided there are no external influences acting upon it. If $\tau_n$ tends to infinity, then $\Delta\mathscr{E}_n$ tends to zero. The quantities $\tau_n$ and $\Delta\mathscr{E}_n$ determine the so-called *natural line width*, $\Delta\omega_{nm}$, or the spread in the spectral line produced as the atom jumps from the $n$th to the $m$th level (see Sec. 61.4). According to Bohr's frequency rule:

$$\Delta\omega_{nm} = \Delta\mathscr{E}_n/\hbar \geqslant 1/\tau_n \qquad (72.18)$$

Eq. (72.18) asserts that infinitesimally narrow spectral lines cannot exist. The natural line width $\Delta\omega_{nm} \approx 10^8$ s$^{-1}$ corresponds to a wavelength interval of $\Delta\lambda \approx 10^{-4}$ Å.

In optics it is shown (see Sec. 61.4) that the chopping of wave trains which occurs because the lifetime of an excited state is finite renders any radiation non-monochromatic for fundamental reasons. We have just seen that Einstein's theory of radiation leads to similar results. It may be noted that the consistent quantum theory of light emission gives the same results.

## 72.9. INDUCED EMISSION OF LIGHT

1. According to Einstein, in a material acted upon by an electromagnetic field the atoms (or molecules) interact with the field in agreement with the laws of conservation of energy and momentum. As will be recalled (see Sec. 63.5), an electric dipole in the electromagnetic field of the incident light wave experiences forced vibrations. Depending on the phase relation between the dipole's own vibrations and those of the electric field intensity of the wave, the dipole can either absorb the field energy or give up energy through the *induced* (*stimulated*) *emission of radiation*. In the latter case, we have the so-called *negative absorption of light*, in contrast to the normal or positive absorption.

Einstein has shown that an atom in the electromagnetic field of a light wave has properties similar to those of an electric dipole. That is, in the presence of a field it must give up energy through the stimulated emission of radiation. In quantum-mechanical terms, an atom residing in an excited energy state $n$ may, *under the action of the field*, jump to a lower energy state $m$ with a certain probability. In a sense, the applied electromagnetic field "dumps" the atom from the excited level.

According to Einstein, an excited atom in level $n$ may jump into the lower level $m$ and emit one quantum, $h\nu$, either spontaneously or under the action of the field.

2. The induced emission of radiation predicted by Einstein in 1916 was experimentally discovered by V. A. Fabrikant in 1939. He observed the negative absorption of visible light in mercury vapour excited by an electric discharge. He also found that on passing through mercury vapour under conditions such that the upper excited levels had a larger population of atoms than the lower, less excited levels, light was amplified. On the basis of this experiments, Fabrikant formulated (in 1939-1940) the principle of light amplification in media capable of negative light absorption. In Sec. 79.4 we shall see that the stimulated emission of radiation has been verified experimentally and utilized in fundamentally new quantum-mechanical sources and amplifiers of light developed in recent decades.

3. In addition to transitions from an upper level $n$ to a lower level, $m$, atoms can also move in the reverse direction. On absorbing a photon of energy

$$\hbar\omega = \mathscr{E}_n - \mathscr{E}_m,$$

an atom may jump to a higher energy level, $n$.

Let a substance and the electromagnetic field be in a state of equilibrium at a certain constant temperature $T$. This means that all properties of the substance, notably its variables of state $p$, $v$ and $T$ (see Chap. 26) remain unchanged, and so do the variables of the field. For this equilibrium state to occur, a balance must exist between the emission and absorption of light, that is, the acts of light emission must be equal in number to acts of light absorption. This balance can be obtained in an isothermal enclosure, that is, a closed cavity with its walls held at constant temperature (see Sec. 67.1). The atoms in the cavity walls emit and absorb light in such a way that the condition of balance is satisfied. Einstein showed that given the condition of balance between, and the conservation of energy for, the emission and absorption of electromagnetic radiation by the atoms of an isothermal enclosure, one obtains the Planck radiation formula, Eq. (67.5).

# Chapter 73
# MANY-ELECTRON ATOMS

## 73.1. THE PAULI EXCLUSION PRINCIPLE

1. In the previous chapter it is shown that in radiative spontaneous and induced transitions atoms jump to the ground or a lower energy state. As regards one-electron atoms, this implies that the electron always tends to move into its normal state where it has the lowest energy. Obviously, this statement should also apply to many-electron atoms. As a result of quantum transitions, all electrons of a many-electron atom should tend to jump to the lowest energy level, that is, to occupy the most stable of all energy states. It would appear that all electrons of such an atom should "gather together" in the same (lowest) energy level.

2. According to a quantum-mechanical law enunciated by Pauli in 1925, no two particles of a given kind can however be in the same state; hence the name, *exclusion principle*. Today, the Pauli exclusion principle is formulated as follows: *In any system containing a multiplicity of electrons, no two electrons can occupy the same stationary state defined by a set of four quantum numbers, namely the principal quantum number n, the orbital quantum number l, the magnetic quantum number m, and the spin quantum number $m_s$.*

The Pauli exclusion principle has its basis in another quantum-mechanical concept of no less importance—particle indistinguishability. For example, no one can tell one electron in an atom or a molecule from another of the same kind, for they all have the same charge, mass and absolute value of spin. Interchanging two electrons in an atom will in no way affect its state. Particle indistinguishability leads to very important results, one of which is the Pauli exclusion principle. Unfortunately, a more detailed discussion of the relation between the Pauli exclusion principle and particle indistinguishability is outside the scope of this book. But we shall state more than once that electrons in atoms and other systems are indistinguishable (see Sec. 74.3). In quantum mechanics it is proved that apart from electrons the Pauli exclusion principle applies to all particles having half-integral spin, $\hbar/2$.

3. For the system of electrons in an atom, the Pauli exclusion principle may be written as

$$Z_1(n, l, m, m_s) = 0 \text{ or } 1 \qquad (73.1)$$

where $Z_1(n, l, m, m_s)$ is the number of electrons occupying the state described by a set of four quantum numbers $n$, $l$, $m$ and $m_s$.

We shall use the Pauli exclusion principle in order to find the maximum number of electrons in an atom that respectively have specified values of three quantum numbers ($n$, $l$, $m$), two quantum numbers ($n$, $l$) and, finally, one quantum number ($n$).

4. To begin with, we shall find the number $Z_2$ ($n$, $l$, $m$) of electrons occupying the states defined by a set of three quantum numbers $n$, $l$, and $m$. In such states, the electrons can only differ in spin orientation. Since the quantum number $m_s$ may only have two values, 1/2 and —1/2 (see Sec. 72.5), we write at once that

$$Z_2 (n, \ l, \ m) = 2 \qquad (73.2)$$

Now we shall find the maximum number of electrons, $Z_3$ ($n$, $l$), that occupy the states defined by two quantum numbers, $n$ and $l$, that is, which differ in the allowed values of the magnetic quantum number $m$. As will be recalled (see Sec. 72.4), the magnetic quantum number may have ($2l + 1$) values. Therefore, the maximum number $Z_3$ ($n$, $l$) of electrons is

$$Z_3 (n, \ l) = 2 \times (2l + 1) \qquad (73.3)$$

The values of $Z_3$ ($n$, $l$) for different values of $l$ are given in Table 73.1.

*Table 73.1*

| Values of orbital quantum number, $l$ | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| Symbol of the respective electronic state | *s* | *p* | *d* | *f* | *g* |
| Maximum number of electrons | 2 | 6 | 10 | 14 | 18 |

Finally, using the Pauli exclusion principle, we can find the maximum number $Z$ ($n$) of electrons occupying the states defined by the values of the principal quantum number, $n$. As has been shown (see Sec. 72.2), for a given value of $n$ the orbital quantum number may have values from 0 to ($n$ — 1). Therefore, the number $Z$ ($n$) can be obtained, if we sum Eq. (73.3) over values of $l$ from 0 to ($n$ — 1)*:

$$Z(n) = \sum_{l=0}^{l=n-1} Z_3(n, l) = \sum_{l=0}^{l=n-1} 2(2l+1) = [2(n-1)+2]\,n = 2n^2 \quad (73.4)$$

5. The electrons in an atom occupying a set of states having the same principal quantum number $n$ make up an *electron shell*. Each shell is designated according to the value of $n$. For example, the $K$-shell has an $n$-value of 1, the $L$-shell has an $n$-value of 2, the $M$-shell has an $n$-value of 3, the $N$-shell has an $n$-value of 4, the $O$-shell has an $n$-value of 5, etc. From Eq. (73.4) it follows that

* We leave it as an excercise for the reader to carry out the summation.

the maximum number of electrons is two for the $K$-shell and 8, 18, 32 and 50 for the $L$-, $M$-, $N$-, and $O$-shells, respectively.

*Table 73.2*

| $n$ | Electron shell | Numbers of electrons in | | | | | Maximum number of electrons |
|---|---|---|---|---|---|---|---|
| | | $s$-state ($l = 0$) | $p$-state ($l = 1$) | $d$-state ($l = 2$) | $f$-state ($l = 3$) | $g$-state ($l = 4$) | |
| 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 |

In each shell the electrons make up *sub-groups* or *sub-shells*, each corresponding to a particular value of the orbital quantum number, $l$. Table 73.2 gives the maximum numbers of electrons that occupy a particular shell for a given value of the orbital quantum number, that is, their number in the various sub-shells.

The Pauli exclusion principle has given a tremendous impetus to present-day atomic and nuclear physics. Without it, no theory of solids could have been formulated as we know it today (see Secs. 75 through 77). The Pauli exclusion principle has been the basis of Mendeleev's periodic law.

## 73.2. MENDELEEV'S PERIODIC LAW

1. In 1869, D. I. Mendeleev found that, once arranged in order of increasing atomic weight, the chemical elements of which sixty-four were known at that time would form columns, or *groups*, in which they had like physical and chemical properties that repeated themselves at regular intervals, called *periods*. With some elements, however, Mendeleev was forced to abandon this arrangement with increasing atomic weight because their actual properties dictated so.

2. As a way out, Mendeleev proposed the concept of *atomic number*, $Z$. When the elements were arranged in order of increasing atomic number, their physical and chemical properties were found to vary with atomic number in a systematic and predictable manner. There were several vacant spaces in Mendeleev's table which led him to predict the existence of six undiscovered elements which he called eka-boron (scandium), eka-aluminium (gallium), eka-silicon (germanium), eka-manganese (technetium), dvi-manganese (rhenium), and eka-tantalum (polonium), and to describe their chemical properties. The first three of these "missing" elements were discovered very soon after his predictions which were found to be extremely

accurate. On the basis of his periodic law, Mendeleev deduced the atomic weights and chemical properties of several elements, among them beryllium, titanium, caesium and uranium, which were far more accurate than those found empirically. That was a major success for Mendeleev's periodic law, and its acceptance as a fundamental property of nature became widespread. Until now it remains the basis of present-day chemistry, atomic and nuclear physics.

3. In physical terms, the atomic number $Z$ of an element in the Periodic Table was shown by Rutherford's model of the atom (see Sec. 71.1) to be identical with the number of protons, that is, positive elementary charges, on the nucleus (see Sec. 80.3). Every next element has an atomic number which is by one greater than that of the previous element.

4. The chemical, optical and some other physical properties of an element are governed by the behaviour of its outer electrons called *valence* or *optical electrons*. Among other things, the periodicity in the properties of elements is directly related to the periodicity in the arrangement of valence electrons in atoms.

5. The theoretical justification and explanation of the periodic law was given by Bohr's quantum theory in 1922, before the advent of quantum mechanics. The basic points of his theory may be stated as follows:

(a) the total number of electrons in an atom of a given chemical element is equal to the atomic number $Z$ of that element;

(b) the electronic state in an atom is defined by its four quantum numbers $n$, $l$, $m$, and $m_s$;

(c) the electrons in an atom are assigned to energy states so as to satisfy the principle of a minimum potential energy, that is, as the number of electrons increases, every next electron must occupy the lowest possible energy level;

(d) the energy levels in an atom must be occupied in accordance with the Pauli exclusion principle.*

6. The energy states in the various shells and also in the various sub-shells of a shell should be filled (closed or completed) in the same sequence as the energy levels are arranged according to the values of the quantum numbers $n$ and $l$. The first to be occupied are states with the lowest possible energy, then the states with a progressively increasing energy. For many atoms, this would mean that shells having the lowest values of $n$ should be closed before the next shell may get its complement of electrons. Within a shell, the first to be occupied are states with an $l$-value of 0, then states with greater $l$-values, up to $l = n - 1$. The periodic system arranged on this basis would have a structure and the number of elements per period

* Actually, Bohr gave a theoretical explanation of the periodic law before Pauli formulated the exclusion principle.

(the period length) according to Table 73.2. The actual Periodic Table is different.

7. Table 73.2 and the actual Periodic Table differ because each electron of an atom is in the electric field of a positively charged nucleus and in the field due to the other electrons interacting with the nucleus and with one another. Under the circumstances, the energy state of an electron moving in a complex field like this cannot be found accurately even in quantum mechanics.

To get an idea of how electrons are assigned to the energy states of an atom, the atom of every next element may be visualized as made up of that of the preceding element and one more proton added to its nucleus (and the requisite number of neutrons, see Sec. 80.3) and one more electron to its periphery. Then, according to Bohr, the electrons in the next element must be assigned to energy states in precisely the same manner as they are in the preceding one. However, interactions between the electrons in an atom bring about violations of this principle. Indeed, because of such interactions, at large values of $n$ the states with high $n$-values and low $l$-values may have a lower energy, that is, they may be more advantageous energetically than states with lower $n$-values but higher $l$-values. This is the reason why the actual Periodic Table differs from Table 73.2 as regards the occupancy of energy states.

8. Consider in greater detail the order in which electrons occupy their states in atoms in the ground, unexcited state.

In atomic physics, the electronic state of an atom is designated by the symbol $nl$ which gives the values of two quantum numbers. In the hydrogen atom ($Z = 1$) its only electron is in a $1s$ state with an $n$-value of unity, an $l$-value of zero, and an $m$-value of zero. Its spin component in the direction of the external field is defined by the spin quantum numbers $m_s = \pm 1/2$. In the helium atom ($Z = 2$) its second electron may likewise occupy a $1s$ state, that is, it may have the same values on $n$, $l$, and $m$, but its spin must be oriented in the opposite direction (so that the spin quantum number for one electron is 1/2 and for the other —1/2, or vice versa). The group of states such that $n = 1$, $l = 0$ and $m_s = \pm 1/2$ close the $K$-shell, and this completes the first period of the Periodic Table.

The next element is lithium ($Z = 3$), which has three electrons. According to the Pauli exclusion principle, the third lithium electron cannot occupy the already complete $K$-shell; therefore it must occupy the lowest possible state in a shell with $n = 2$ (the $L$-shell). This is a $2s$ state (with $n = 2$, $l = 0$, and $m = 0$). Lithium opens the second period in the Periodic Table.

The fourth element is beryllium ($Z = 4$), and its fourth electron likewise occupies a $2s$ state. The fifth electron of boron ($Z = 5$) must occupy a higher $2p$ state (with $n = 2$ and $l = 1$). Up to the element neon ($Z = 10$), the electrons in all atoms occupy a sub-shell such that $l = 1$ and $n = 2$. The neon atom has six such electrons, that is, the maximum number for this state. The $L$-shell of the neon atom is complete, and this closes the second period of the Periodic Table.

The eleventh electron of sodium ($Z = 11$) occupies the lowest $3s$ state in the $M$-shell ($n = 3$). This shell is consecutively filled until it is complete in argon ($Z = 18$). In the argon atom, all $3p$ states are full, and this completes the third period of the Periodic Table.

9. The nineteenth electron of potassium ($Z = 19$) must have occupied $3d$ state in the $M$-shell ($n = 3$, $l = 2$). However, the chemical and optical properties of potassium are similar to those of lithium and sodium in which the valence electron is in an $s$-state. Therefore in potassium, too, the valence (19th) electron must actually occupy an $s$-state, which it does in the $N$-shell ($n = 4$), and the state is a $4s$ state.

Starting with potassium, electrons begin to fill the $N$-shell, leaving the $3d$ sub-shell in the $M$-shell incomplete. This is because the energy $\mathscr{E}_{4,0}$ of an electron in a $4s$ state is lower than that of an electron in a $3d$ state as a result of interaction between electrons.* The chemical and optical properties of calcium ($Z = 20$) are such that its 20th electron must be in a $4s$ state of the $N$-shell. Starting with scandium ($Z = 21$), the $3d$ sub-shell is again filled in the normal manner, until it is complete in copper ($Z = 29$). Then the $N$-shell is filled in the normal way, too, until it is closed in krypton ($Z = 36$) which completes the fourth period of the Periodic Table.

The element rubidium ($Z = 37$) which follows krypton is similar to sodium and potassium in properties. Therefore its valence (37th) electron occupies a $5s$ state in the next $O$-shell ($n = 5$), although the $N$-shell is not complete yet. The same applies also to strontium ($Z = 38$) similar in properties to calcium. Starting with yttrium ($Z = 39$) and ending with palladium ($Z = 46$), a $4d$ sub-shell is filled. In silver ($Z = 47$) and cadmium ($Z = 48$), electrons again fill the sub-shells in the $N$-shell. Starting with indium ($Z = 49$) and ending with xenon ($Z = 54$), the $5p$ sub-shell is filled full, and xenon completes the fifth period of the Periodic Table.

Starting with caesium ($Z = 55$), electrons begin to fill the $P$-shell ($n = 6$).

10. The chemical elements from lanthanum ($Z = 57$) to lutecium ($Z = 71$) inclusive make up a separate group of rare earths called the *lanthanides* which show a similarity in chemical and some physical properties. This is again owing to the sequence in which the electrons occupy the various states in these elements. In lanthanum, the $5s$, $5p$ and $6s$ sub-shells are filled full, and the 57th lanthanum

* According to quantum theory, the energy of an electron in an atom is generally decided by both the principal and the orbital quantum numbers. It is only for the hydrogen atom that the energy depends solely on the principal quantum number (see Eq. (71.12)).

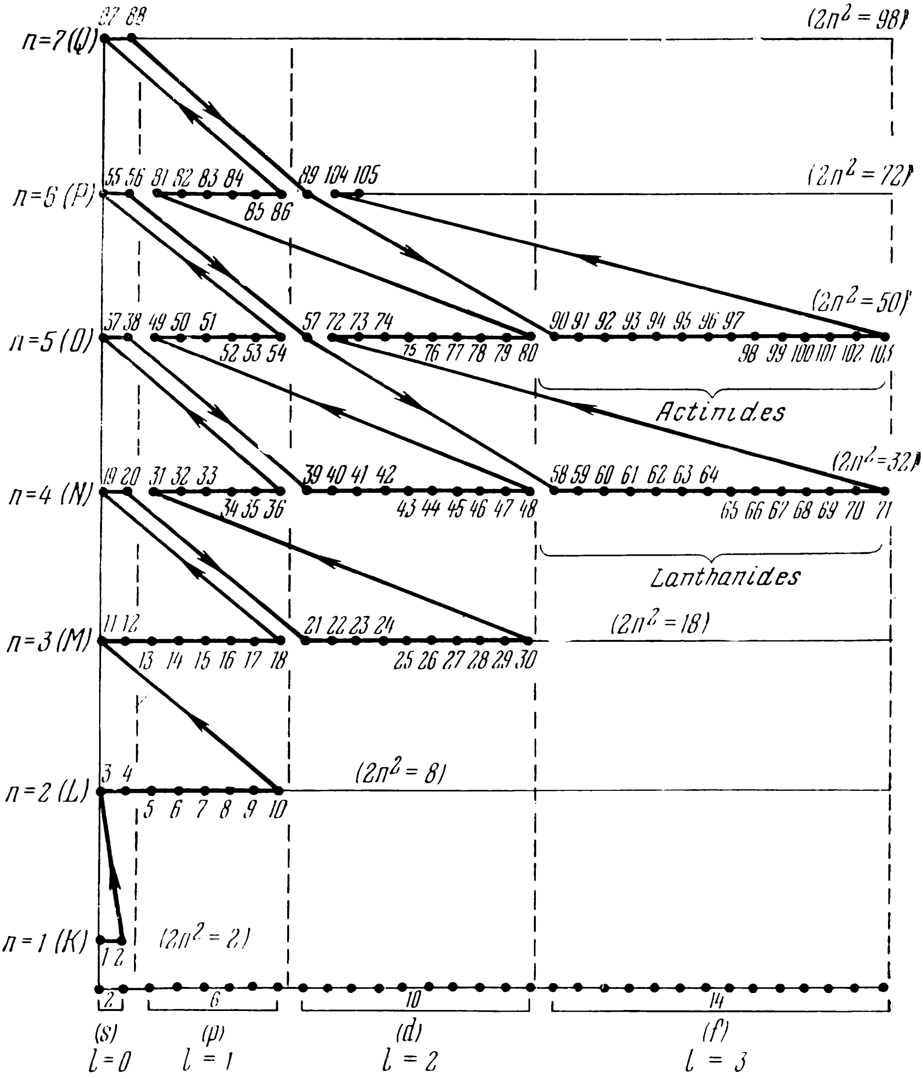


Fig. 73.1

electron occupies a $5d$ state while the inner $4f$ sub-shell remains incomplete. In the elements from caesium ($Z = 58$) to lutecium ($Z = 71$) this sub-shell is completed, but the outer $6s$ sub-shell remains unchanged. This is the reason why all lanthanides show a similarity in chemical properties.

Starting with hafnium ($Z = 72$), the $5d$ sub-shell is filled until it is complete in uni-valent gold ($Z = 79$). In mercury ($Z = 80$), the $6s$ sub-shell is complete, while the $6p$ sub-shell is filled full from thallium ($Z = 81$) to radon ($Z = 86$) which closes the sixth period of the Periodic Table. In francium ($Z = 87$) and radium ($Z = 88$), the $7s$ sub-shell of the $Q$-shell is filled full ($n = 7$).

11. The elements from actinium ($Z = 89$) to the element of atomic number 103 make up a second separate group, called the *actinides*. They comprise all the transuranic elements neptunium ($Z = = 93$), plutonium ($Z = 94$), americium ($Z = 95$), curium ($Z = 96$), etc. Two elements in this group have been named einsteinium ($Z = 99$) and fermium ($Z = 100$) after Einstein and Fermi, respectively. The element of atomic number 101 has been named mendelevium, after Mendeleev, the originator of the Periodic Table. In all actinides, the $5f$ sub-shell is complete, and their outer electrons occupy states similar to those in lanthanides.

The element of atomic number 104 was discovered at the Joint Institute of Nuclear Studies at Dubna (USSR) in 1964 and named kurchatovium after I. V. Kurchatov, an outstanding Soviet physicist. Kurchatovium has two isotopes, ${}_{104}\mathrm{Ku}^{260}$ and ${}_{104}\mathrm{Ku}^{264}$.

In diagrammatic form, the manner in which electrons occupy energy states in atoms is illustrated in Fig. 73.1. The full circles represent the elements, and the numerals their ordinal numbers. The electron structure of the atoms of some elements is shown in Table 73.3.

12. In 1969, the world celebrated the 100th anniversary of Mendeleev's periodic law. Its theoretical explanation has been a tremendous achievement for present-day physics. It has been possible on the basis of quantum mechanics, only.

## 73.3. BREMSSTRAHLUNG

1. A good deal of knowledge about the structure and properties of electron shells in complex atoms, molecules and, especially, the crystalline lattice of solids has been gained through the use of the rays discovered and called *X-rays* by Roentgen in 1895.

As has been found, $X$-rays are produced when fast electrons are braked by a material and their kinetic energy is converted to electromagnetic radiation. Thus, X-rays are electromagnetic waves with very short wavelengths extending from 0.01 to 800 Å. Since the shortest of violet rays visible to the eye have a wavelength of about 4000 Å, X-rays are invisible to the eye. The wave nature of X-rays has been verified by diffraction experiments such as described in Sec. 62.5.

2. X-rays are produced by X-ray tubes. An X-ray tube consists of a glass or metal envelope enclosing a cathode, an anode (or target)

*Table 73.3*

*ELECTRON STRUCTURE OF ATOMS (NORMAL STATE)*

| Z | Element | *K* | *L* | | *M* | | | *N* | | | | *O* | | | | *P* | | | | *Q* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1*s* | 2*s* | 2*p* | 3*s* | 3*p* | 3*d* | 4*s* | 4*p* | 4*d* | 4*f* | 5*s* | 5*p* | 5*d* | 5*f* | 6*s* | 6*p* | 6*d* | 6*f* | 7*s* |
| 1 | H | 1 | | | | | | | | | | | | | | | | | | |
| 2 | He | 2 | | | | | | | | | | | | | | | | | | |
| 3 | Li | 2 | 1 | | | | | | | | | | | | | | | | | |
| 4 | Be | 2 | 2 | | | | | | | | | | | | | | | | | |
| 5 | B | 2 | 2 | 1 | | | | | | | | | | | | | | | | |
| 6 | C | 2 | 2 | 2 | | | | | | | | | | | | | | | | |
| 7 | N | 2 | 2 | 3 | | | | | | | | | | | | | | | | |
| 8 | O | 2 | 2 | 4 | | | | | | | | | | | | | | | | |
| 9 | F | 2 | 2 | 5 | | | | | | | | | | | | | | | | |
| 10 | Ne | 2 | 2 | 6 | | | | | | | | | | | | | | | | |
| 11 | Na | 2 | 2 | 6 | 1 | | | | | | | | | | | | | | | |
| 12 | Mg | 2 | 2 | 6 | 2 | | | | | | | | | | | | | | | |
| 13 | Al | 2 | 2 | 6 | 2 | 1 | | | | | | | | | | | | | | |
| 14 | Si | 2 | 2 | 6 | 2 | 2 | | | | | | | | | | | | | | |
| 15 | P | 2 | 2 | 6 | 2 | 3 | | | | | | | | | | | | | | |
| 16 | S | 2 | 2 | 6 | 2 | 4 | | | | | | | | | | | | | | |
| 17 | Cl | 2 | 2 | 6 | 2 | 5 | | | | | | | | | | | | | | |
| 18 | Ar | 2 | 2 | 6 | 2 | 6 | | | | | | | | | | | | | | |
| 19 | K | 2 | 2 | 6 | 2 | 6 | | 1 | | | | | | | | | | | | |
| 20 | Ca | 2 | 2 | 6 | 2 | 6 | | 2 | | | | | | | | | | | | |
| 21 | Sc | 2 | 2 | 6 | 2 | 6 | 1 | 2 | | | | | | | | | | | | |
| 22 | Ti | 2 | 2 | 6 | 2 | 6 | 2 | 2 | | | | | | | | | | | | |
| 23 | V | 2 | 2 | 6 | 2 | 6 | 3 | 2 | | | | | | | | | | | | |
| 24 | Cr | 2 | 2 | 6 | 2 | 6 | 5 | 1 | | | | | | | | | | | | |
| 25 | Mn | 2 | 2 | 6 | 2 | 6 | 5 | 2 | | | | | | | | | | | | |
| 26 | Fe | 2 | 2 | 6 | 2 | 6 | 6 | 2 | | | | | | | | | | | | |
| 27 | Co | 2 | 2 | 6 | 2 | 6 | 7 | 2 | | | | | | | | | | | | |
| 28 | Ni | 2 | 2 | 6 | 2 | 6 | 8 | 2 | | | | | | | | | | | | |
| 29 | Cu | 2 | 2 | 6 | 2 | 6 | 10 | 1 | | | | | | | | | | | | |
| 30 | Zn | 2 | 2 | 6 | 2 | 6 | 10 | 2 | | | | | | | | | | | | |
| 31 | Ga | 2 | 2 | 6 | 2 | 6 | 10 | 2 | 1 | | | | | | | | | | | |
| 32 | Ge | 2 | 2 | 6 | 2 | 6 | 10 | 2 | 2 | | | | | | | | | | | |
| 33 | As | 2 | 2 | 6 | 2 | 6 | 10 | 2 | 3 | | | | | | | | | | | |
| 34 | Se | 2 | 2 | 6 | 2 | 6 | 10 | 2 | 4 | | | | | | | | | | | |
| 35 | Br | 2 | 2 | 6 | 2 | 6 | 10 | 2 | 5 | | | | | | | | | | | |
| 36 | Kr | 2 | 2 | 6 | 2 | 6 | 10 | 2 | 6 | | | | | | | | | | | |

and, sometimes, an anti-cathode spaced a certain distance apart and connected to a source of an extremely high tension (EHT) supply. The cathode acts as a source of electrons, and the anode (or anti-cathode) as a source of X-rays*. The field set up between the cathode and the anode accelerates the electrons to energies of $10^4$ or $10^5$ eV. In present-day particle accelerators (betatrons and synchrotrons), X-rays are produced by braking electrons with energies of the order of $10^2$ MeV and higher.

3. The invisible X-rays are detected by observing their effects. Among other things, X-rays produce a strong photochemical action which blackens photographic plates. They are also capable of ionizing gases and causing fluorescence in phosphors (see Sec. 79.2). For measurement purposes, use is mainly made of their photochemical and ionizing effects. In ionization chambers, the intensity of X-rays is determined by measuring the saturation current due to the ionization of the gas enclosed in the chamber, because the saturation current is proportional to the intensity of X-rays. The methods used for the detection of ionizing radiations will be discussed in the chapter on nuclear physics (see Sec. 81.8).

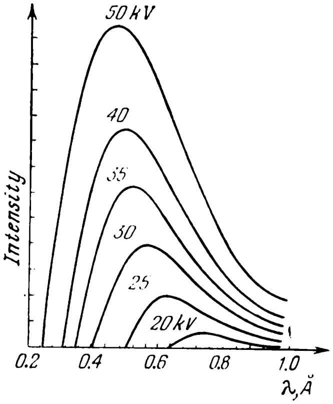


Fig. 73.2

4. Experiments have shown that there exist two kinds of X-rays, namely *continuous* and *characteristic.*

Continuous X-rays, as their name implies, have a continuous spectral distribution. They are produced when electrons accelerated in a vacuum strike a target and lose kinetic energy in passing through the strong electric field surrounding the target nuclei, thus giving rise to *bremsstrahlung* (the German for "braking radiation") and resulting in a continuous X-ray spectrum. As will be recalled (see Sec. 59.4), any charge in an accelerated (or retarded) motion emits electromagnetic radiation of continuous spectral distribution.

On the short wavelength side, the continuous X-ray spectrum is limited by a minimum wavelength, $\lambda_{\min}$, called the *continuous spectrum limit.* Continuous X-ray spectra for tungsten and various potential differences applied to the tube are shown in Fig. 73.2.

The continuous spectrum limit, $\lambda_{\min}$, cannot be explained by the classical wave theory of X-rays, for according to this theory the continuous X-ray spectrum produced by bremsstrahlung cannot be

* Apart from hot-cathode, high-vacuum X-ray tubes, there are cold-cathode, gas-filled tubes in which electrons are knocked out of the cathode by gas ions accelerated by a field.

bounded. Experiments have shown that the minimum wavelength $\lambda_{min}$ is inversely proportional to the kinetic energy $K$ of the electrons giving up bremsstrahlung. A very simple explanation for the existence of the minimum wavelength $\lambda_{min}$ is given by quantum theory.

It is obvious that the maximum energy, $h\nu_{max}$, of the X-ray quantum due to the energy of an electron cannot exceed its kinetic energy, $K$:

$$K = h\nu_{max} \tag{73.5}$$

On passing from frequency to wavelength in Eq. (73.5), we get

$$\lambda_{min} = c/\nu_{max} = ch/K \tag{73.6}$$

Eq. (73.6) agrees well with experimental data; at one time it provided a very precise tool for experimentally determining Planck's constant, $h$.

Eq. (73.5) is similar to Einstein's photoelectric equation, (68.3') (see Sec. 68.3) if we drop the work function from it. If the photoeffect is produced by X-rays, then the work function $A_0$ in Eq. (68.3') will be only a small fraction of a quantum of energy, $h\nu$, and it may well be neglected.

In fact, the photoeffect and bremsstrahlung are mutually reciprocal. Eq. (73.5) when read from right to left will give the kinetic energy of an electron in the photoeffect. When read from left to right, it will give the boundary frequency (or wavelength) that exists when the energy of a braked electron is fully converted to that of an X-ray quantum.

## 73.4. CHARACTERISTIC X-RAYS

1. These X-rays are called so because they are characteristic of the element in which they are produced. Characteristic X-rays have a line spectral distribution. A distinction of this distribution is that the atoms of each element, irrespective of the compound they form, produce a specific line spectrum of characteristic X-rays. In this, they strongly differ from the optical electron spectra of the same atoms, which vary according to whether the atoms are free or part of a chemical compound.

2. The optical spectra of atoms vary from compound to compound because they are determined by the behaviour of outer, or valence, electrons. As chemical bonds are formed, the states of valence electrons are changed, and so are their optical spectra. The nature of characteristic X-rays may be construed from the fact that X-ray line spectra characterize each element *individually* and irrespective of the compounds it may enter. Obviously, they are produced by processes taking place in the inner, fully complete electronic shells

of atoms, which remain unchanged in all chemical transformations.

3. The X-ray line spectra of atoms have been found to be similar from element to element, with homologous lines (spectral series) generally occurring at progressively shorter wavelengths as the atomic weight increases. In increasing order of wavelength, these groups of X-rays are called $K$-, $L$-, $M$-, $N$-, etc. series or radiation.

These series of X-rays owe their names to the fact that they are emitted by electrons in the respective atomic shells. As has been shown in Fig. 73.2, in atoms of large atomic numbers, $Z$, the $K$-, $L$-, $M$- and other shells closest to the nucleus are complete. If an electron is removed from a shell closer to the nucleus, its vacancy will be filled by an electron from a shell farther from the nucleus. This transition is accompanied by the emission of an X-ray quantum.

For example, if primary hard radiation or an incident electron removes an electron from the innermost ($K$-) shell, its vacancy can be filled with an electron from the $L$-, $M$-, $N$- or other shells. This transition will be accompanied by the emission of quanta of definite energies, and the resultant spectral lines will be those of the $K$-series. To remove an electron from the $K$-shell which is closest to the nucleus and where the electrons are attracted by the nucleus strongest of all, an input of energy is necessary, called the *excitation threshold* of the $K$-series. The energy of the incident electron or primary incident quantum should be at least equal to this threshold. As an example, for mercury ($Z = 80$), the excitation threshold of the $K$-series is about 82 keV.

4. As an electron from the $L$-shell jumps into the $K$-shell, it emits a quantum of the lowest energy corresponding to the $K_\alpha$-line which has the longest wavelength of all lines in the $K$-series of characteristic X-rays emitted by a given atom. The $K_\beta$-line corresponds to the transition of an electron from the $M$-shell to the $K$-shell. The $K_\gamma$-line corresponds to the transition of an electron from the $N$-shell to the $K$-shell. Together, the $\mathrm{K}_\alpha$-, $\mathrm{K}_\beta$-, $\mathrm{K}_\gamma$-, etc. lines make up the $K$-series.

The $L$-, $M$- etc. series of characteristic X-rays are emitted when an electron leaves a vacancy in the $L$-, $M$-, $N$-, etc. shells, respectively. For example, the transition of an electron from the $M$-shell to the $L$-shell produces the $L_\alpha$-line; that of an electron from the $N$-shell produces the $L_\beta$-line and so on. All transitions ending up on the $L$-shell produce the $L$-series of characteristic X-rays. In diagrammatic form, the production of several series of characteristic X-rays is illustrated in Fig. 73.3.

5. In 1913, Moseley found that the wavelengths of characteristic X-rays depend in a well defined manner on the atomic numbers of the elements that emit X-rays (the target element). This dependence,

known as the *Moseley law*, may be written as

$$\sqrt{\nu^*/R} = a\,(Z - \sigma) \qquad (73.7)$$

where $\nu^* = 1/\lambda$ is the wave number of the line (see Sec. 71.3), $R$ is the Rydberg constant in $\mathrm{m}^{-1}$ (or $\mathrm{cm}^{-1}$, see Sec. 71.3), $a$ is a proportionality constant, and $\sigma$ is the same for all the given series emitted by a particular target element. For the $K_\alpha$-lines, Moseley derived the following expression

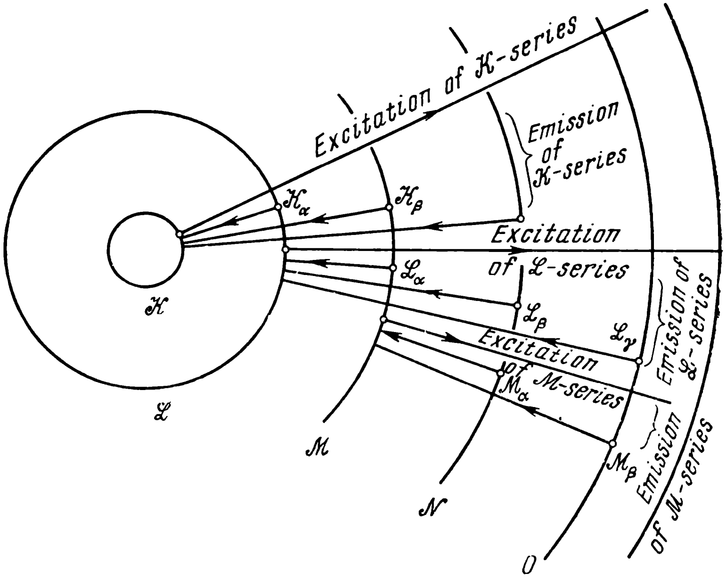


Fig. 73.3

$$\sqrt{\nu^*/R} = \sqrt{3/4}\,(Z - 1) \qquad (73.8)$$

From a comparison of (73.8) and (73.7) it is seen that for these lines $a = \sqrt{3/4}$ and $\sigma = 1$.

Eq. (73.8) may be re-written as follows:

$$\nu^*/R = (1 - 1/4)\,(Z - 1)^2 = (Z - 1)^2\,(1/1^2 - 1/2^2)$$

or

$$\nu^* = R\,(Z - 1)^2\,(1/1^2 - 1/2^2) \qquad (73.8')$$

In this form, Eq. (73.8') resembles the equation for the wavelength (or wave number) of the line in the Lyman series of the hydrogen atom (see Sec. 71.3). The difference is that the quantity $Z^2$ in the spectral series formulae for one-electron systems is less by $\sigma = 1$, called the *screening constant*. This constant reflects the fact that when, as a result of bombardment, one of the $K$ electrons is removed from a heavy atom containing the number $Z$ of electrons, there is one $K$ electron left near the nucleus. This electron "screens" the nucleus and makes its effective charge equal to $(Z - 1)e$ rather than to the entire charge $Ze$. Hence, the factor $(Z - 1)$.

The linear relation between $\sqrt{\nu^*/R}$ and the atomic number $Z$ for the $K_\alpha$-lines is illustrated by the *Moseley plot* of Fig. 73.4.

When first applied to the elements of the Periodic Table, the Moseley law confirmed that the nucleus charge should increase by one from element to element. This came as a proof of the validity of the nuclear atom and of the theory underlying the Periodic Table.

6. Because of the extremely short wavelengths and high "hardness", X-rays (both continuous and characteristic) are highly penetrating. Yet, in passing through a material X-rays are scattered and absorbed, and their intensity decreases. The attenuation of X-rays due to scattering is mainly due to the Compton effect (see Sec. 68.6). In the process, part of the energy of hard X-rays is transferred to the electrons of the material, while the scattered X-rays become softer and their wavelength increases. The absorption of X-rays is accompanied by conversion of the energy of X-rays quanta into the internal energy of the material. X-ray absorption is strongly dependent on the atomic number $Z$ of the material (more accurately, it is proportional to $Z^4$).

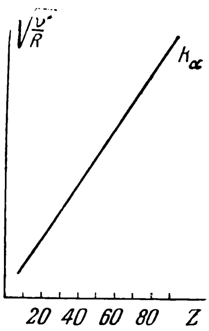


Fig. 73.4

The fact that X-rays are absorbed differently by different materials is widely utilized in medicine, science and technology (X-ray examination, flaw detection, etc.). For example, human bones consisting mainly of calcium phosphate absorb X-rays about 150 times as much as soft tissues in which they are chiefly absorbed by water. This is why bones stand out so distinctly on X-ray photographs. Industrial radiography likewise utilizes the fact that optically opaque solids absorb X-rays differently, according to their atomic numbers, $Z$. If the atomic number of a flaw or a foreign inclusion is greater than that of the host material, the area occupied by the flaw will appear as a lighter spot against a darker background. If the reverse is true, the flaw area will appear darker. X-rays are capable of determining not only the area, but also the thickness of a flaw. For this purpose, one measures the attenuation of X-rays inside and outside the flaw area.

# Chapter 74

# THE STRUCTURE AND SPECTRA OF MOLECULES

## 74.1. GENERAL CHARACTERISTIC OF CHEMICAL BONDS

1. A molecule is defined as the smallest particle of any substance that can exist free and still exhibit all of the chemical properties of the original substance.

Molecules may consist of atoms of one or several kinds, held together by *chemical bonds*. The stability of molecules is an indication that chemical bonds are caused by the forces arising from the interactions between atoms or groups of atoms. Observations show that some work need be done before a molecule can be broken up into its constituent atoms. Conversely, the formation of a molecule should be accompanied by the liberation of an amount of energy. For example, two atoms of hydrogen (H) in a free state have a greater energy than the same atoms making up a diatomic molecule ($H_2$). The energy given up as a molecule is formed is a measure of the forces that cause the bonding between the atoms in the molecule.

2. To understand why electrically neutral atoms can form a stable molecule, it will suffice to examine simple diatomic molecules in which the atoms may be of the same or different kinds. The forces that cause the bonding between atoms arise from the interaction between the outer, or valence, electrons of the atoms. This is confirmed by a marked change in the optical spectra of atoms as they form chemical compounds. As will be recalled, the line spectra of atoms are determined by the state of their outer or valence electrons. In contrast, the characteristic X-ray spectra governed by the inner electrons remain unchanged as a compound is formed (see Sec. 73.4). On the other hand, chemical bonds must be formed by electrons whose states may be changed through the expenditure of a relatively small amount of energy. Again, this applies to the outer, or valence, electrons, for their ionization potentials are much lower than those of the electrons in the inner complete shells (see Sec. 73.2).

3. Whatever the nature of the forces that cause the binding between atoms, they may be described in common terms. If atoms are a great distance apart, there is no interaction between them. As, however, they are brought closer together, that is, as the separation $r$ between them is decreased, a force of *mutual attraction* comes to work between the atoms. At short distances, $r$, comparable with (or even shorter than) the linear dimensions of atoms, a force of *mutual repulsion* arises, which keeps the electrons of one atom from going too deep into the electronic shells of the other.

Attractive and repulsive forces vary with the distance $r$ between atoms differently. In this respect they are not unlike the forces of intermolecular interaction examined in Sec. 31.4. Repulsive forces vary at a faster rate with changing $r$ than attractive forces do. As the distance $r$ between atoms increases, repulsive forces ebb away at a faster rate than attractive forces. Since the two forces—attractive and repulsive—are acting at the same time, at some distance $r$ between the atoms the two forces become equal, and their vectorial sum reduces to zero. This distance corresponds to the lowest potential energy, $U(r)$, of a diatomic molecule.

Fig. 74.1 shows three curves, one representing attractive force $F_2$, another repulsive force $F_1$, and the third their resultant, $F$, for the atoms in a diatomic molecule, as functions of the distance $r$ between the atoms. The repulsive force is assumed to be positive (see Sec. 31.4). A plot of the potential energy, $U(r)$ as a function of the distance $r$ between the atoms in a diatomic molecule appears in Fig. 74.2.

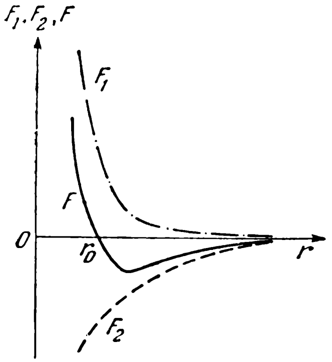


Fig. 74.1

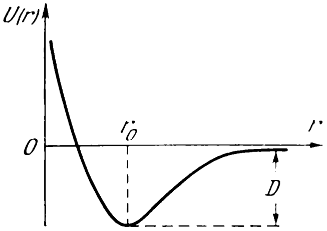


Fig. 74.2

4. The distance at which the forces of interaction between the atoms in a molecule are in equilibrium is called the *bond distance*, $r_0$. The quantity $D$ in the plot of Fig. 74.2 is called the *bond dissociation energy* or the *bond energy*. Numerically, it is equal to the work that should be done in order to break the chemical bond between the atoms in a molecule, that is, in order to dissociate the molecule into the constituent atoms (or ions, see Sec. 74.2) and move the atoms from one another to a distance where the interatomic forces are no longer effective. It is obvious that the bond dissociation energy is equal to the energy liberated when the molecule is formed, but is opposite in sign: the bond dissociation energy is negative, while the energy liberated as the molecule is formed is positive.

## 74.2. IONIC BONDS

1. In the simplest manner, the chemical bonds that hold the atoms of a molecule together may be attributed to an electric interaction. Such a molecule, however, would be stable only if the two participating atoms carry charges of opposite polarity. Then the attraction between the charges will ensure a strong chemical bond and a stable molecule.

This type of chemical bond does exist in some molecules. Typical examples are the molecules of alkali-metal halides, such as NaCl, RbBr and CsJ, that is, elements in the first and seventh groups of the Periodic Table. When two atoms interact one atom acquiring one or several more electrons comes by a negative charge and becomes a negative ion, while that giving up the respective number of electrons becomes a positive ion. These ions carrying unlike charge may then be drawn together into a molecule by electrostatic attraction.

This bond is called *ionic*. Sometimes, it is called the *heteropolar* bond (from the Greek "hetero" for "unlike"), or simply *polar*. Accordingly the molecules in which the bond is ionic are called *ionic* or *heteropolar*.

2. Consider the formation of a molecule of common salt, NaCl, in greater detail. Like all other atoms of the metals in the first group, the sodium atom has a relatively low ionization potential. It takes as little as 5.1 eV to remove its eleventh outer electron from the sodium atom. In contrast, the chlorine atom which has seven outer or valence electrons, like all other elements in the seventh group, has a high *electron affinity*, that is, gives up a greater amount of energy when an electron is attached to it (for chlorine, it is 3.8 eV). As an electron is passed from a sodium atom on to a chlorine atom, ions $Na^+$ and $Cl^-$ are formed. Each has a stable outer eight-electron shell (see Sec. 73.2), similar to that of the inert gases. The electrostatic attraction between the oppositely charged ions $Na^+$ and $Cl^-$ brings them closer together until, at very short distances, the attractive force gives way to a repulsive force which keeps the ions from moving any further. Finally, ions $Na^+$ and $Cl^-$ take up positions separated by the equilibrium distance $r_0$ in which the attractive and repulsive forces balance one another. Thus, a stable ionic molecule of NaCl is formed.

3. From the above example it is seen that the ionization energy of sodium exceeds the electronic affinity of chlorine by 5.1 eV — — 3.8 eV = 1.3 eV. In other words, the transfer of an electron from a sodium atom to a chlorine atom calls for the input of an amount of energy. On the other hand, it is a fact that an amount of energy is liberated each time a molecule is formed. Where does then the energy necessary to form a molecule come from? The point is that the energy given up as ions come closer together is due to their electrostatic interaction. Ions are formed and move closer together at

one and the same time, and then only after the atoms are close enough for the requisite energy to be liberated with the formation of ions.

As will be recalled, the potential electrostatic energy, $U(r)$ due to the interaction of two single-charge ions spaced a distance $r$ apart is given by

$$U(r) = e^2/4\pi\varepsilon_0 r$$

where $e$ is the ionic charge the same as the electronic charge, and $\varepsilon_0$ is the dielectric constant in SI units. It is an easy matter to find from this equation the distance $r$ at which the energy $U(r)$ makes up for the difference between the ionization energy and the electronic affinity (1.3 eV):

$$r = e^2/4\pi\varepsilon_0 U(r)$$

Substituting the numerical values, we get

$$r = 11 \times 10^{-10} \text{ m} = 11 \text{ Å}$$

Hence, the transfer of an electron from a sodium atom to a chlorine atom can take place only at $r \leqslant 11$ Å. According to X-ray diffraction analysis, however, the bond distance $r_0$ for NaCl is 1.4 Å. At this distance, the electrostatic energy $U(r_0)$ is 10.2 eV, which is 8.9 eV greater than is necessary to produce ions $\text{Na}^+$ and $\text{Cl}^-$ in one molecule. This works out to an energy of 49 200 J or 205 kcal liberated in the formation of one kilomole of NaCl. Experiments show that this value, although very accurate, is somewhat exaggerated, for we have not taken into account the fact that the repulsive force reduces the energy given up in the formation of ionic molecules.

## 74.3. COVALENT BONDS

1. Ionic bonds cannot arise in a molecule made up of atoms of the same kind, say in a hydrogen molecule, $H_2$, because ions carrying unlike charges cannot form in such a molecule. Instead, electrically neutral atoms are held together by what is known as *covalent* or *homopolar bonds* (from the Greek "homo" for "same"). Apart from diatomic molecules such as $H_2$, $O_2$ and $N_2$, covalent bonds are found in many other molecules, notably hydrogen fluoride HF, nitrogen oxide NO, ammonia $NH_3$, and methane $CH_4$, to name but a few.

2. Classical physics allows the existence of only gravitational forces (see Sec. 9.2) between electrically neutral particles or any other entities. However, they are too weak to explain the stability of a homopolar molecule. Furthermore, covalent bonds show the property of *saturation* which manifests itself in that atoms can only have a certain definite valence. That is, a hydrogen atom can form a bond with only one other hydrogen atom, and a carbon atom can form a

bond with only four hydrogen atoms, and no more. Like electric and magnetic forces, gravitational forces show no saturation. A single central body can attract an unlimited number of other bodies. In short, gravitational forces cannot account for the fact that neutral atoms can be linked together by a covalent bond.

3. The nature of covalent binding is readily explained in terms of quantum theory. To get an insight into the physical meaning of covalence, we shall use as an example a very simple molecule of this type, the hydrogen molecule $H_2$ consisting of two nuclei (protons) and two electrons. Spectroscopic studies have shown that the bond distance $r_0$ for $H_2$ is 0.74 Å, and the bond energy $D$ (see Sec. 74.1) is 4.718 eV or 103.24 kcal/kilomole. Quantum theory has proved the correctness of these experimental results.

4. At the basis of a quantum-theoretical explanation of covalence are the specific quantum-mechanical, and not classical properties of the valence electrons. This above all is the wave behaviour of electrons. For one thing, an electron has a certain probability of being found near the nucleus. In the simplest *s*-state (see Sec. 72.3), the probability distribution is spherically symmetrical, that is, the electron cloud is a sphere of a definite radius. For another, identical particles generally, and electrons in particular are *undistinguishable*. Indeed, the two electrons moving each around its "own" nucleus in a hydrogen molecule do not differ from each other: each has the same charge and rest mass and the same spin equal to $\hbar/2$. Hence, if the two electrons change places so that electron *1* formerly belonging to one nucleus takes the place of electron *2* formerly belonging to the other nucleus or vice versa, the state of the hydrogen molecule will not change in the least (Fig. 74.3). Of course, this exchange can take place only if the two nuclei of the hydrogen molecule approach so closely that their electronic charge clouds overlap.

Thus, the overlapping of the electronic charge clouds in the hydrogen molecule involves what is known as the *electron-exchange effect* or *interaction*. In other words, each electron in a molecule can alternately belong to each of the nuclei or, which is the same, the electrons are continually exchanging places. In a way, this may be likened to a continuous exchange of balls between two persons standing next to each other.

5. According to the Pauli exclusion principle, two electrons can be in the same state if they have opposite spins. Quantum-theoretical considerations show that if the overlapping charge clouds belong to electrons of opposite (anti-parallel) spin, the exchange energy is a minimum, and the atoms tend to form a stable molecule. When the overlapping cloud charges are due to electrons of the same (parallel) spin, the result is a repulsion between the atoms, and no molecule will form.

Plots of the exchange energy $U(r)$ for anti-parallel spins (curve *1*) and parallel spins (curve *2*) of the electrons in a hydrogen molecule appear in Fig. 74.4. The lowest point on curve *1* occurs when the bond distance is $r_0 = 0.83$ Å. The bond dissociation energy (or the bond energy), $D$, for the hydrogen molecule was deduced by Heitler and London to be 3.2 eV. It does not agree with the experimentally found value (4.72 eV) as well as the bond distance whose empirical value is 0.74 Å. The more accurate results obtained on quantum theory later showed a better agreement with experiment.

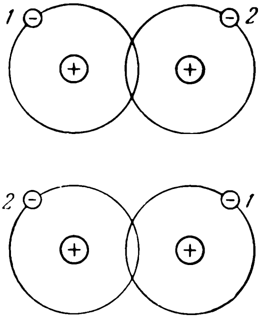


Fig. 74.3

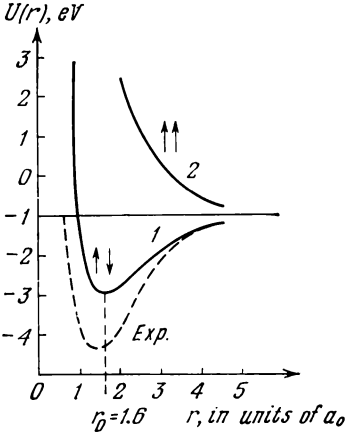


Fig. 74.4

Experiments show that the hydrogen molecule is diamagnetic (see Sec. 42.4), because it has no orbital magnetic moment (see Sec. 42.2), the spins are compensated, and the resultant magnetic moment is zero.

## 74.4. MOLECULAR SPECTRA

1. Molecular spectra markedly differ in appearance from the atomic spectra discussed in Sec. 71.3. Above all, they show wide bands made up of closely spaced spectral lines. Quite appropriately they are called *band spectra*. Bands of molecular spectra lines can be seen in the infrared, visible and ultraviolet regions. The closely spaced bands in molecular spectra make up groups of bands. In simple diatomic molecules, several groups of bands are usually present. As an example, Fig. 74.5 shows a photograph of a part of the spectrum of the iodine molecule. As molecules become more elaborate in structure, their spectra grow in complexity, too. Many-atom molecules of

a complex structure show spectra which have wide continuous absorption and emission bands in the visible and ultraviolet regions.

2. As with atomic spectra, each line in the spectrum of a molecule appears when the molecule undergoes a change in its energy state. The total energy of a molecule is the sum of five terms which, in a first approximation, may be thought of as being independent of one another (see Sec. 27.8 and 27.9). These terms are the translational energy $\mathscr{E}_t$ associated with the translational motion of the molecule's centre of inertia, the rotational energy $\mathscr{E}_r$ associated with the rotational motion of the molecule as a whole about an axis, the electronic energy $\mathscr{E}_e$ associated with the motion of the electrons in the atoms of the molecule, the vibrational energy $\mathscr{E}_v$ associated with the vibrational motion of the atomic nuclei and, finally, the nuclear energy $\mathscr{E}_N$ associated with the atomic nuclei. That is

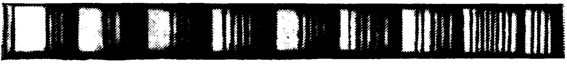


Fig. 74.5

$$\mathscr{E}=\mathscr{E}_t+\mathscr{E}_r+\mathscr{E}_e+\mathscr{E}_v+\mathscr{E}_N \tag{74.1}$$

The translational energy $\mathscr{E}_t$ may vary continuously as the conditions of the translational motion change, that is, the translational energy of a molecule is not quantized. Hence, a change in $\mathscr{E}_t$ cannot produce a line in the molecular spectrum.

If we ignore the optical processes due to nucleons, we may drop $\mathscr{E}_N$ from Eq. (74.1). Then, the energy $\mathscr{E}'$ determining the optical properties of a molecule is the sum of three terms

$$\mathscr{E}'=\mathscr{E}_e+\mathscr{E}_v+\mathscr{E}_r \tag{74.2}$$

Each term is quantized, that is, it is restricted to a set of discrete values. A change in each term, $\Delta\mathscr{E}_e$, $\Delta\mathscr{E}_v$, or $\Delta\mathscr{E}_r$, is likewise discrete, and so the energy $\mathscr{E}'$ of the molecule can only change by a discrete amount $\Delta\mathscr{E}'$ given by

$$\Delta\mathscr{E}'=\Delta\mathscr{E}_e+\Delta\mathscr{E}_v+\Delta\mathscr{E}_r \tag{74.3}$$

According to Bohr's third postulate (the frequency rule), the frequency $\nu$ of a quantum emitted by a molecule as its energy state is changed is

$$\nu=\Delta\mathscr{E}'/h=\Delta\mathscr{E}_e/h+\Delta\mathscr{E}_v/h+\Delta\mathscr{E}_r/h \tag{74.4}$$

Both experiment and theory show that the terms of Eq. (74.3) are such that

$$\Delta\mathscr{E}_r\ll\Delta\mathscr{E}_v\ll\Delta\mathscr{E}_e \tag{74.5}$$

3. Inequality (74.5) explains why molecular spectra occur in different regions of the electromagnetic spectrum and why the spectral lines make up bands.

As an example, let us examine the formation of the absorption spectrum of a molecule. Assume that a substance consisting of molecules is illuminated by an electromagnetic radiation of a low frequency $\nu$. This implies that the quantum, $h\nu$, of this radiation is small. Until the quantum becomes equal to the lowest possible difference between the nearest two energy levels in the molecule, no radiation will be absorbed and no absorption line will be produced. Absorption will be possible when the wavelength of the incident wave is 0.1 to 1 mm, that is in the far infrared. At these frequencies, a quantum of energy represents a change of $\Delta\mathscr{E}_r$ in the *rotational* energy of the molecule. On absorbing a quantum of energy, the molecule moves from one rotational energy level to a higher one, and this transition leads to the appearance of a line in the *rotation absorption spectrum*.* As the wavelength is reduced, more lines may occur in the rotation absorption spectrum. The complete set of lines gives an idea about the distribution of rotational energy states in the molecule.

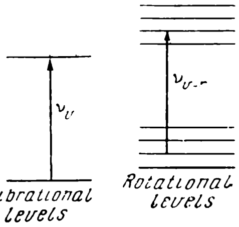


Fig. 74.6

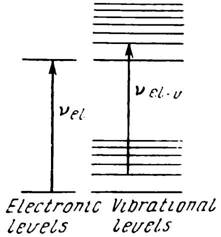


Fig. 74.7

4. When a substance absorbs electromagnetic radiation in the infrared region, that is, with wavelengths from 1 to 10 μm, transitions occur between the vibrational energy states of the molecule, and a *vibration molecular spectrum* is produced.

However, transitions between two vibrational energy states are accompanied by a change in the rotational energy states, and the molecule produces a *vibration-rotation spectrum*. In diagrammatic form, this is shown in Fig. 74.6. As is seen, each transition between two vibrational energy levels producing a line at frequency $\nu_v$ is accompanied by transitions between the respective rotational levels. The resultant vibration-rotation spectrum at frequencies $\nu_{v\text{-}r}$ consists of groups of closely spaced lines due to rotational transitions,

* The transition of a molecule from a higher to a lower rotational energy level produces a line in the rotation emission spectrum.

which make up a band corresponding to a particular vibrational transition.

5. The absorption of electromagnetic radiation in the visible and ultraviolet regions brings about transitions between *electronic* energy levels, and these transitions produce an *electronic molecular spectrum*. Each electronic energy level is associated with a particular distribution of constituent electrons in space, or a particular *electronic configuration* having a particular discrete value of energy. In turn, each electronic configuration is associated with a multiplicity of vibrational energy levels. A transition between two electronic levels is accompanied by many transitions between vibrational levels. This produces an *electronic-vibration spectrum* of the molecule consisting of a group of closely spaced lines that make up an electronic-vibration band, such as shown in Fig. 74.7. Furthermore, the system of rotational levels shown in the preceding figure is superimposed on each vibrational state. The total electronic-vibration spectrum in the visible and adjacent region is a system of several band groups often overlapping into a single wider band.

# Chapter 75

# THE PRESENT-DAY THEORY OF ELECTRICAL CONDUCTION IN METALS

## 75.1. LIMITATIONS OF THE CLASSICAL THEORY OF ELECTRICAL CONDUCTION IN METALS

1. The classical free electron theory of metals set forth in Sec. 44.4 was extremely simplified. According to it, the electrons in a metal behave like the particles of a classical gas. In fact, the original theory expounded by Drude assigned to them the same average velocity, $\bar{u}$, equal to that of the random thermal motion. Later, Lorentz, one of the originators of the classical free electron theory, proved that the electrons in metals obey Maxwell-Boltzmann statistics. That is, in the absence of an applied field, the velocity of the electrons that are responsible for electrical conduction in metals has the Maxwell distribution (see Sec. 25.2).

2. According to Lorentz, an applied voltage gives rise to an electric field which upsets the Maxwellian velocity distribution of electrons in the metal. As a result, the ordered motion of electrons due to the electric field is superimposed on the random thermal motion, and the average velocity of this ordered motion is proportional to the electric field intensity. Reasoning along the above lines, Lorentz

deduced Ohm's law in a form very similar to Eq. (44.13), while his expression for conductivity was not unlike Eq. (44.15):

$$\gamma = (2/3)\,(ne^2\bar{\lambda}/m)\,(\overline{1/u}) \tag{75.1}$$

Here, as in Sec. 44.4, $n$ is the free electron number density, $\bar{\lambda}$ is the electron mean free path, $e$ and $m$ are the electronic charge and rest mass, and $\overline{1/u}$ is the average reciprocal of the velocity of electron thermal motion found on the basis of the Maxwell velocity distribution. As is seen, Eqs. (75.1) and (44.15) differ only slightly. In the former, the conductivity is a function of the same physical characteristics of electrons in metals as it is in the latter.

Lorentz also derived an expression for the Wiedemann-Franz law which differs from Eq. (45.16) in that it has the factor $2k^2/e^2$ instead of $3k^2/e^2$:

$$K/\gamma = 2k^2T/e^2 \tag{75.2}$$

where $K$ is the coefficient of thermal conductivity of the electron gas, $k$ is Boltzmann's constant, and $T$ is the absolute temperature of the metal. From reference to Table 45.2, it may be seen that Eq. (75.2) agrees with experimental data not so well as Drude's expression, Eq. (45.16), because the value of $K/\gamma T$ is smaller than it is in Drude's theory.

Like the original theory of electrons in metals advanced by Drude, the classical free-electron theory formulated by Lorentz could not explain many things appearing in experiments. Although we mentioned them in Secs. 44.5 and 45.2, the limitations of the classical free-electron theory of metals deserve a more detailed discussion because of their fundamental importance.

3. In molecular physics it is proved that the average velocity of the thermal motion of particles is proportional to the square root of the absolute temperature, that is, $\bar{u} \backsim \sqrt{T}$. On the other hand, experiments have shown that the electrical conductivity of metals is inversely proportional to their absolute temperature ($\gamma \backsim 1/T$) over a wide range of temperatures. Drude's equation, (44.15), and Lorentz's equation, (75.1), could have explained why this is so, if it might be assumed that the product $n\bar{\lambda}$ is inversely proportional to $\sqrt{T}$. The expression for $\bar{\lambda}$ from the kinetic theory of gases (see Sec. 25.3), however, denies this possibility. So, the classical free-electron theory had no plausible explanation for the empirically found dependence of electrical conductivity (or resistivity) on temperature.

4. Using Eq. (44.15) or (75.1), it is an easy matter to determine the mean free paths, $\bar{\lambda}$, of an electron that correspond to the values of electrical conductivity $\gamma$ (or resistivity, $\rho = 1/\gamma$) observed experimentally. The first step is to determine $\bar{u}$ from the equations deri-

ved in the kinetic theory of gases. Then the electron number density can be found from, say, the Hall effect (see Sec. 44.2). It is found then that at room temperature ($T \approx 300$ K) $\bar{\lambda}$ is about $10^{-9}$ to $10^{-8}$ m, which is tens or even hundreds of times the lattice constant of metals, while, according to the Drude-Lorentz theory, it must be comparable with the ionic radius of metals.

5. On the basis of the experimental data collected in Table 44.1, it is shown in Sec. 45.2 that the electron gas does not contribute to the specific heat of metals. This is at variance with the classical free-electron theory which asserts that the electrons in a metal must behave like a monatomic gas and must have a molar heat capacity of 3 kcal/kmole·K. This contradiction as regards the specific heats of metals is important inasmuch as it implies a violation of the law of conservation of energy. The limitations and contradictions of the classical free-electron theory of metals have been removed by the quantum theory of metals.

## 75.2. QUANTIZATION OF ELECTRON ENERGY IN METALS

1. Advances in quantum mechanics have led to a quantum theory of solids giving a deeper and more consistent insight into the electrical, optical and other properties of metals, crystalline dielectrics and semiconductors. In turn, a better knowledge of solids has promoted their wider use in science and technology. This and succeeding chapters will examine some of the ideas and applications of the present-day quantum theory of solids. Above all, we shall discuss the present-day concepts of electrical conduction in metals, taking into account all aspects in the behaviour of electrons in metals and recalling that the electrons in atoms, molecules and crystals obey the laws of quantum mechanics.

2. We shall assume that free electrons in metals make up an electron gas whose particles are moving as if the positive lattice ions set up no electric field. Then the motion of electrons may be described, using a model of a potential box with a flat bottom (see Sec. 70.4). If we assume that the potential energy of an electron outside the metal is zero, then inside the metal it is $A_0$, where $A_0$ is the positive work function of the metal. In other words, in a metal free electrons are inside a potential box with vertical walls of a finite depth.

In Sec. 70.4 it is shown that in such a potential box the energy of electrons is restricted to a set of discrete values. It should be noted, though, that the energy $\mathscr{E}_n$ of an electron in a metal cannot be found by Eq. (70.16) because the walls of the potential box in a metal have a finite depth. At present, however, we are not concerned with the precise dependence of the electron energy on the size of the box and the value of the principal quantum number $n$. What is important is that *the electrons in a metal may occupy only definite energy* *levels as in an atom*. There is a material difference, however, between the arrangement of electronic energy levels in metals and in isolated atoms: in atoms the difference in energy between the electrons at two adjacent levels is much greater than it is in crystals.

## 75.3. FERMI LEVEL FOR ELECTRONS IN METALS

1. The classical free-electron theory of metals asserted that the electrons in metals obey the classical Maxwell-Boltzmann statistics. It was on its basis that Lorentz deduced his expression for electrical conductivity, Eq. (75.1). However, electrons and other microscopic particles have properties which are ignored in the statistical description of particle systems on the basis of classical theory. Among other things, classical physics could not allow for the dual wave-particle nature of particles as this remained unknown until 1924 when the first steps were made in the formulation of quantum theory. Nor could classical statistics give its due to the indistinguishability of microscopic particles (see Secs. 73.1 and 74.3). Finally, electrons and other microscopic particles with half-integral spin ($\hbar/2$) obey the Pauli exclusion principle (see Sec. 73.1) which limits their energy distribution. In 1926, Fermi and Dirac formulated a quantum statistics for an assembly of particles which allowed for all the properties of electrons and other particles with half-integral spin.

2. The statistics of an assembly of particles seeks to determine the velocity and energy distributions of the particles. For example, in Secs. 25.2, 26.10 and 26.11 we have examined the velocity distribution of gas molecules developed by Maxwell and the energy distribution of gas molecules in a gravitational field.

The statistics of free electrons in a metal seeks to determine the number of electrons out of their total at a given temperature $T$ that has velocities in the range from $v$ to $v + \Delta v$ or, respectively, energies in the range from $\mathscr{E}$ to $\mathscr{E} + \Delta\mathscr{E}$. That is, if we have a number $n$ of electrons per unit volume of a metal, we seek to find their proportion, $\Delta n$, that has energies within the narrow interval $\Delta\mathscr{E}$.

3. In seeking a solution to this problem, it is important to take into account all the basic properties of electrons in a metal listed in Para. 1 of this section.

The point of departure is that the electrons in a metal may only have certain definite energies or, which is the same, may occupy only some allowed energy states. It is obvious that all electrons tend to occupy the lowest energy levels as they are most stable of all. However, the electrons obey the Pauli exclusion principle which limits the number of electrons that may be in the same state (see Sec. 73.1). As applied to electrons in metals, the Pauli exclusion principle somewhat differs from its statement in Sec. 73.1: *of all* *the electrons in a metal, there may be not more than two electrons in the same state, and the spins of these electrons must be anti-parallel.*

4. According to the Pauli exclusion principle, pairs of electrons occupy allowed energy levels, starting from the lowest one. The horizontal lines in Fig. 75.1 represent the energy levels occupied by electrons. As is seen, *the work function $A_0$ of the metal should be measured from the topmost of all occupied levels.* This topmost occupied energy level is pivotal to all of the quantum theory of solids. It is called the *Fermi level*, after Enrico Fermi, an outstanding physicist of our time, who has made a major contribution to present-day physics. The energy of an electron in the Fermi level is designated $\mathscr{E}_\mathrm{F}$, while its velocity and momentum as $v_\mathrm{F}$ and $p_\mathrm{F}$, respectively. The manner in which these quantities are determined will be shown later.

## 75.4. MOMENTUM SPACE OF ELECTRONS IN A METAL

1. If we neglect the potential energy of electrons in a metal due to the electric field of the lattice ions and the energy of interaction between the electrons, the total energy $\mathscr{E}$ of an electron will be a function of its velocity $v$ and its momentum $p$:

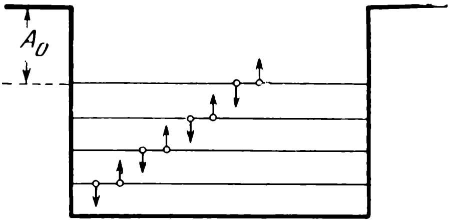


Fig. 75.1

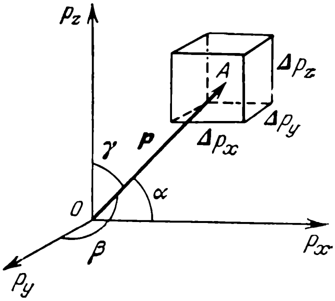


Fig. 75.2

$$\mathscr{E} = mv^2/2 = p^2/2m \qquad (75.3)$$

Notably, on the Fermi level

$$\mathscr{E}_\mathrm{F} = p_\mathrm{F}^2/2m \text{ and } p_\mathrm{F} = \sqrt{2m\mathscr{E}_\mathrm{F}} \qquad (75.4)$$

Let us lay off the rectangular components of the electron momentum, $p_x$, $p_y$ and $p_z$ along the respective axes of a Cartesian coordinate system (Fig. 75.2). Any point $A$ in this three-dimensional "space" represents the momentum of an electron as a vector, that is, as a quantity possessing both magnitude and direction. Thus, by connecting the point $A$ to the origin of coordinates, we obtain a vector

**p** whose magnitude is equal to the numerical value of the momentum of the electron:

$$p = \sqrt{p_x^2 + p_y^2 + p_z^2} \quad \text{or} \quad p^2 = p_x^2 + p_y^2 + p_z^2 \qquad (75.5)$$

The angles $\alpha$, $\beta$, and $\gamma$ made by the vector **p** with the Cartesian axes define the direction of the momentum of the electron:

$$\cos\alpha = n_x/p, \quad \cos\beta = p_y/p, \quad \cos\gamma = p_z/p \qquad (75.6)$$

The space we are discussing is called the *momentum space*. It may be recalled that in an ordinary space the position of a point, $B$, is decided by a vector **r** drawn from the origin of coordinates to this point (Fig. 75.3).

2. The momentum space may be used to determine the energy state of the electrons at the Fermi level. In this momentum space, the constant electron energy given by Eq. (75.3) is represented by a simple geometrical figure, namely the surface of a sphere of radius $p$. Consider an electron residing at the Fermi level and having a maximum energy $\mathscr{E}_F$. Its energy is related to its momentum by Eq. (75.4). Thus, in the momentum space, a free electron of energy $\mathscr{E}_F$ is represented by a *Fermi surface* which is a sphere of radius $p_F$ (Fig. 75.4). Another name for the Fermi surface is the *maximum energy surface*.

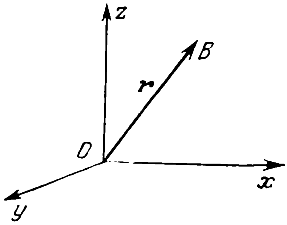


Fig. 75.3

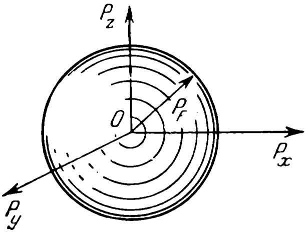


Fig. 75.4

For a bound electron moving in the composite electric field due to the ions of the crystalline lattice, the Fermi surface has a very elaborate shape.

3. Let us connect the Fermi energy, $\mathscr{E}_F$, to the number $n$ of free electrons per unit volume of a metal. We allow two of them to be at a level of energy $\mathscr{E}_F$. Then the momenta of all other electrons should lie inside a sphere of radius $p_F = \sqrt{2m\mathscr{E}_F}$. Now we divide the entire momentum space into elementary cells and try to determine their size. We arrange them so that the tip of the momentum vector of any electron, point $A$ in Fig. 75.2, will always be inside any one of these elementary cells. According to the quantum theory of metals,

*each elementary cell is a quantum state having a definite energy. No more than two electrons with opposite spins may be in the same cell.*

In order to determine the size of an element of the momentum space, we shall use the Heisenberg uncertainty principle (see Sec. 70.2). Imagine in a metal a small cube with an edge $l$ such that $l^3 = V$. If an electron is in free motion in this volume, its position in space may be determined accurate to the length of the cube edge, that is, $\Delta x \approx \Delta y \approx \Delta z \approx l$. Then from relations (70.4) through (70.6) it follows that the rectangular components of the electron momentum may be determined accurate to:

$$\Delta p_x \approx \Delta p_y \approx \Delta p_z \approx \hbar/l$$

The elementary volume of the momentum space discussed above (Fig. 75.2) is given by

$$\Delta\omega = \Delta p_x \Delta p_y \Delta p_z = \hbar^3/l^3 = \hbar^3/V \tag{75.7}$$

Rigorous calculations give

$$\Delta\omega = 8\pi^3\hbar^3/V = h^3/V \tag{75.8}$$

4. Now we shall divide the entire space bounded by the Fermi surface into spheres of radius $p_F$, thus obtaining elementary volumes $\Delta\omega$. If the volume $V$ of a metal contains the number $N$ of electrons, then, according to the foregoing, each quantum state, that is, the elementary volume $\Delta\omega$, will be occupied by two electrons. The total number of occupied elementary volumes will be $N/2$ and their overall volume will be $(N/2)\,(h^3/V)$*. On the other hand, this is the volume of the Fermi sphere in the momentum space of radius $p_F$. Hence

$$4\pi p_F^3/3 = (N/2)\,(h^3/V) = nh^3/2 \tag{75.9}$$

where $n = N/V$ is the number of electrons per unit volume in the metal. From Eq. (75.9), $p_F$ and $\mathscr{E}_F$ are related to $n$ as follows:

$$p_F = h\,(3n/8\pi)^{1/3} \tag{75.10}$$

$$\mathscr{E}_F = p_F^2/2m = (h^2/2m)\,(3n/8\pi)^{2/3} \tag{75.11}$$

It follows from the last two equations that the momentum and energy of an electron at the Fermi level are decided solely by the number of electrons per unit volume. For example, with $n \approx 10^{29}\ \text{m}^{-3}$, substituting in Eq. (75.10) the numerical values $h = 6.62 \times 10^{-34}$ J s and $m = 9 \times 10^{-31}$ kg, we get $p_F = 2 \times 10^{-24}$ kg m/s. The velocity $v_F$ corresponding to this momentum is

$$v_F = p_F/m \approx 2 \times 10^{-24}/9 \times 10^{-31} \approx 2 \times 10^6\ \text{m/s}$$

* In our reasoning, we have assumed that all elementary volumes inside the Fermi surface have the same probability of being occupied.

According to (75.11), $\mathscr{E}_F \approx 1.6 \times 10^{-18}$ J $\approx 10$ eV. It will be shown in Sec. 75.6 that the above estimates hold if the electrons in a metal have an absolute temperature of zero, $T = 0$ K.

5. Now we shall compare the energy of an electron at the Fermi level with the average energy, $kT_{cl}$, of an electron given by classical theory. From the condition $\mathscr{E}_F = kT_{cl}$ we find that a particle of the classical electron gas would have the energy $\mathscr{E}_F$ at a temperature $T_{cl}$ given by

$$T_{cl} = \mathscr{E}_F/k = 1.6 \times 10^{-18}/1.38 \times 10^{-23} \approx 10^5 \text{ K}$$

which is the temperature at which no metal can remain solid. It is obvious that the electron gas in metals does not obey the classical Maxwell-Boltzmann statistics. In other words, the properties and behaviour of electrons in metals can be correctly visualized only if it is remembered that the electron gas in metals is a "quantum" gas, that is, one obeying quantum statistics.

### 75.5. DEGENERACY OF ELECTRONS IN A METAL

1. The quantum statistics of electrons and other assemblies of particles having half-integral spin, $\hbar/2$, takes into account the fact that identical particles are indistinguishable, that they obey the Pauli exclusion principle and that the energy of an electron in a metal is restricted to a set of discrete allowed values. It is seen from the numerical estimates given in Paras. 4 and 5 of the preceding section that the properties of the quantum electron gas differ markedly from those of the classical electron gas. This departure is called *degeneracy*. The temperature below which a gas behaves as a degenerate one is called the *degeneracy temperature*, $T_d$. This point has already been discussed in Sec. 26.8.

2. We shall show that the electron gas is degenerate always. To do this, we shall first determine the degeneracy temperature $T_d$ of an assembly of quantum-mechanical particles (notably, electrons in a metal). According to Sec. 26.8, the degeneracy temperature is given by

$$T_d = \hbar^2 n^{2/3}/3km \tag{75.12}$$

If we neglect the fact that the Planck constant is finite and set $\hbar \approx 0$, then $T_d \rightarrow 0$; that is, the degenerate gas is a quantum-mechanical system. For the electron gas in metals, $n \approx 10^{29}$ m$^{-3}$ and $m = 9 \times 10^{-31}$ kg. According to (75.12), $T_d \approx 1.84 \times 10^4$ K. Thus, the electron gas in metals is practically always degenerate owing to the small mass and the high density of electrons. It is only at a temperature of tens of thousands of degrees that the electrons in a metal would obey the classical Maxwell-Boltzmann statistics. However, no metal can exist in a condensed state at such temperatures.

In semiconductors, the concentration of the electron gas is much lower than it is in metals, being $10^{18}$ m$^{-3}$ in some cases. Under the circumstances, the degeneracy temperature is negligibly low ($T_d \approx 10^{-4}$ K), and the electron gas in semiconductors is non-degenerate, that is, one obeying classical statistics.

Another example of a degenerate gas is the photon gas. As a proof, imagine a closed cavity whose walls are at a common temperature $T$ and which contains an electromagnetic field. We have already used such an isothermal enclosure in connection with thermal radiation (see Ch. 67). Treating this radiation as the photon gas and noting that the rest mass of a photon is zero ($m = 0$), we find that for the photon gas the degeneracy temperature is $T_d = \infty$. That is, the photon gas will remain degenerate at any finite temperature.

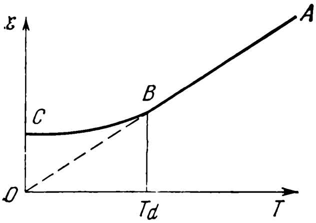


Fig. 75.5]

A plot of the energy of the electron gas in metals as a function of temperature is shown in Fig. 75.5. As is seen, at $T > T_d$ in the region $BA$ the gas is non-degenerate, and its energy is proportional to temperature as with an ordinary gas (see Sec. 26.5). Below $T_d$, that is, at $T \leqslant T_d$, the electron gas becomes degenerate, and the energy and velocity of the electrons within this temperature region (region $CB$ on the curve) are practically independent of temperature. Thus, the classical definition of temperature as a physical quantity proportional to the average translational kinetic energy of the molecules of an ideal gas (see Sec. 26.5) remains valid only above the degeneracy temperature. As regards ordinary, molecular gases, the classical definition of temperature holds at practically any temperature (see Sec. 26.8).

3. The marked difference in properties between the degenerate electron gas and ordinary classical gases may be illustrated as follows. As will be recalled (see Sec. 26.3), an ideal gas is defined as one whose molecules do not exert forces on one another and are in a free motion in which they only collide. As regards an ordinary gas, we may say that it approaches an ideal gas as the potential energy of interaction between its molecules decreases in comparison with their kinetic energy. Also, as a gas becomes more rarefied, its density decreases, and it approaches an ideal gas in properties. The opposite is true of the degenerate electron gas in metals. It approaches the ideal gas in properties *as its density increases* or, which is the same, *as the separation between the electrons decreases*.

This can be proved as follows. The potential interaction energy $U$ of electrons is proportional to $e^2/a$, where $e$ is the electronic charge and $a$ is the average spacing between electrons equal in magnitude to $n^{-1/3}$*. Thus, $U \approx e^2n^{1/3}$. For a region in which the degeneracy of electrons is considerable, the kinetic energy $K$ of an electron in a metal is given by

$$K \approx 3kT_d/2 \approx 3\hbar^2n^{2/3}/2m$$

It is seen that *the kinetic energy of electrons increases with increasing electron number density n at a faster rate than their potential interaction energy does.* This is the reason why the electron gas comes to the ideal gas very closely.

## 75.6. ENERGY DISTRIBUTION OF ELECTRONS IN METALS AT ABSOLUTE ZERO

1. From the preceding sections of this chapter it is clear that a degenerate electron gas must obey statistics markedly differing from the classical one (see Secs. 25.2, 26.10 and 26.11).

Consider some of the results obtained for the electrons in metals on the basis of Fermi-Dirac quantum statistics. The difference between the classical and quantum statistics is most striking at extremely low temperatures. Fig. 75.6 illustrates the distribution of electrons in a metal between the allowed energy levels at $T = 0$ K. The number of energy levels, $N$, counted from the bottom of the potential box is laid off as abscissa, and the number of electrons at a given level as ordinate. Since the levels are restricted to a set of discrete values of energy, the distribution at energies below $\mathscr{E}_{\rm F}$ is represented by a large number of dots on a straight line (Fig. 75.6 simply shows this line, *2*, drawn through the points). According to the Pauli exclusion principle, pairs of electrons occupy all energy levels from the bottom of the potential box up to and including the Fermi level. The top Fermi level is numbered $n/2$, where $n$ is the number density of electrons in the metal.**

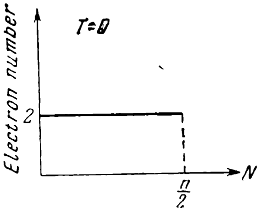


Fig. 75.6

2. The average separation between two adjacent energy levels for electrons in a metal is

$$\Delta\mathscr{E} = \mathscr{E}_{\rm F}/(n/2) = 2\mathscr{E}_{\rm F}/n$$

Setting $n = 10^{29}$ m$^{-3}$ and $\mathscr{E}_{\rm F} \approx 10$ eV (see Sec. 75.4), we obtain

$$\Delta\mathscr{E} \approx 10^{-22} \text{ eV}$$

that is, the energy levels are arranged so closely as to form a dense, nearly continuous sequence.

* If a volume of 1 m$^3$ contains a number $n$ of particles, and the average separation between them is $\Delta x$, then, obviously, $\Delta x^3 \cdot n = 1$, whence, $\Delta x = n^{-1/3}$.

** We refer to the energy level distribution of electrons per unit volume.

However, the plot of Fig. 75.6 does not give the energy distribution of electrons at $T = 0$ K. To describe this distribution, we should recall that, according to theoretical calculations, the number of levels with energies from $\mathscr{E}$ to $\mathscr{E} + \Delta\mathscr{E}$ is directly proportional to the product $\sqrt{\mathscr{E}} \cdot \Delta\mathscr{E}$. From this consideration we can readily find the number $\Delta n$ of electrons out of the total number $n$ that have energies from $\mathscr{E}$ to $\mathscr{E} + \Delta\mathscr{E}$ at $T = 0$ K. As is seen from the *energy distribution curve for electrons in a metal*, representing the ratio $\Delta n/\Delta\mathscr{E}$ as a function of $\mathscr{E}$ at $T = 0$ K in Fig. 75.7, no electrons have an energy exceeding $\mathscr{E}_F$. The curves of Figs. 75.6 and 75.7 agree with each other and show that the Fermi energy, $\mathscr{E}_F$, is the highest energy that electrons can have in metal at $T = 0$ K.

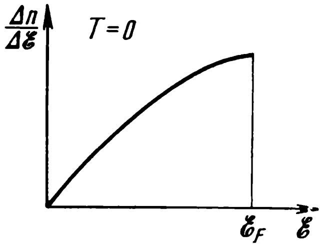


Fig. 75.7

## 75.7. THE EFFECT OF TEMPERATURE ON THE ENERGY DISTRIBUTION OF ELECTRONS

1. According to quantum statistics, the most important property of electrons in metals is that their number density and energy change but little with temperature. Let us see how the energy distribution curve shown in Fig. 75.7 changes with temperature. On heating, the electrons are thermally excited to higher energy levels. This should inevitably change the energy distribution that exists at $T = 0$ K. To find this change, we recall that at the absolute zero point of temperature the Fermi energy, $\mathscr{E}_F$, is about 10 eV and that average energy transferred to an electron by heating is about the same as the average energy of thermal motion, $kT$. At room temperature ($T \approx 300$ K), the value of $kT$ is 0.025 eV, that is, the condition

$$kT \ll \mathscr{E}_F \tag{75.13}$$

is satisfied.

2. Inequality (75.13) shows that at room temperature thermal excitation is only possible for the electrons occupying energy levels near the Fermi level, that is, the topmost filled level at $T = 0$ K. These levels make up a narrow band of wid.h $kT$, directly adjacent to the Fermi level.

In Fig. 75.8, these bands are shown shaded. It is especially important that the electrons in the inner levels remain practically unaffected, because the energy they gain is insufficient to excite them beyond the Fermi level, while the higher levels are all occupied. Upon heating, some electrons having an energy a little below $\mathscr{E}_F$ jump into levels with an energy somewhat above $\mathscr{E}_F$, and a new energy distribution of electrons is obtained. Figs. 75.9 and 75.10 show curves similar to those of Figs. 75.6 and 75.7 and satisfying the condition (75.13). From a comparison of the curves at $T = 0$ K and $T \neq 0$ K it is seen that they only differ in the rate of fall near the level numbered $n/2$ and near the level of energy $\mathscr{E}_F$. At $T = 0$ K, there is a sudden downward jump, while at $T \neq 0$ K, the change is gradual, with the curve approaching the $x$-axis. The curves of Figs. 75.9 and 75.10 illustrate the fact that heating only excites the electrons occupying the levels closest to the Fermi level, while the energy distribution of electrons occupying the inner levels remains the same as at $T = 0$ K.

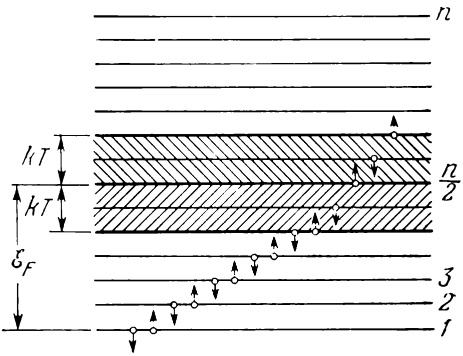


Fig. 75.8

3. Let us determine, at least approximately, the number of electrons, $\Delta n$, which occupy the band of levels $kT$ wide below the Fermi level. The number of levels in this band is $kT/\Delta\mathscr{E}$, where $\Delta\mathscr{E} = 2\mathscr{E}_F/n$ is the level spacing (see Sec. 75.6). The number of electrons that can occupy these levels is given by

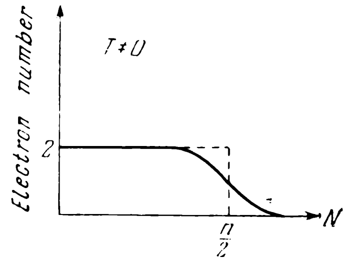


Fig. 75.9

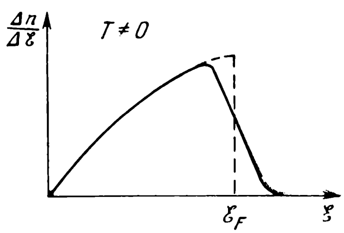


Fig. 75.10

$$2kT/\Delta\mathscr{E} = 2kTn/2\mathscr{E}_F = kTn/\mathscr{E}_F$$

The factor 2 appears because, according to the Pauli exclusion principle, two electrons with opposite spins may occupy the same level. If we assume that half the electrons occupying the band of width $kT$ jump beyond the Fermi level, then the number $\Delta n$ of thermally excited electrons will be given by

$$\Delta n \approx kTn/2\mathscr{E}_F \tag{75.14}$$

At room temperature and at $\mathscr{E}_F \approx 5$ to 10 eV, we get that $\Delta n/n < < 0.01$. That is, only a small proportion of electrons (less than 1%) is excited by heating. In other words, over the entire range of temperatures in which the electron gas is degenerate, the energy distribution of its electrons differs but little from that existing at the absolute zero point of temperature. But, as is shown in Sec. 75.5, the electron gas is degenerate at any temperature. Therefore, the energy distribution of electrons in metals differs but little from that at absolute zero. Hence, we may state that *the number density and thermal velocity of electrons in metals are independent of temperature.*

4. From the previous discussion it might appear that the position of the Fermi level remains unchanged with temperature. Actually, we have used the Fermi energy, $\mathscr{E}_F$ existing at the absolute zero point of temperature. It is proved in quantum statistics that *the Fermi level is brought somewhat down as the temperature rises.*

The fact that the Fermi energy depends but little on temperature is responsible for the thermoelectric effects which occur when the surfaces of two unlike metals are brought in contact (Sec. 44.8).

5. In conclusion, it should be noted that a regular relation exists between the energy distribution of particles given by quantum statistics (Fig. 75.10) and the energy distribution of gas molecules according to the classical Maxwell law. At elevated temperatures, when the gas is no longer degenerate, the energy distribution curve of quantum-mechanical particles is the same as the Maxwell distribution curve (see Sec. 25.2).

### 75.8. THE SPECIFIC HEAT OF THE DEGENERATE ELECTRON GAS

1. As we have already seen (see Secs. 45.2 and 75.1), the classical free electron theory of metals fails to explain why the electrons do not practically contribute to the specific heat of metals. The heat capacity of the electron gas is extremely low in comparison with that of the crystalline ion lattice. The explanation was found only after the properties of the degenerate gas in metals had become known.

2. Let us find the energy $\Delta U_m$ absorbed by thermally excited electrons. So that we can then conveniently pass on to the specific heat of metals, we assume that thermal excitation affects a number $\Delta N$ of electrons out of the total number $N$ of free electrons per kilomole of a metal. It is obvious that

$$\Delta n/n = \Delta N/N$$

where $\Delta n$ and $n$ refer to unit volume of a metal. When thermally excited, each electron absorbs an energy of the same order of magnitude as $kT$. All electrons, $\Delta N$, which are excited to higher energy levels gain an energy given by

$$\Delta U_m = kT \cdot \Delta N = NkT\,(kT/2\mathscr{E}_F)$$

In deriving the above expression we have used Eq. (75.14).

As will be recalled, there is one free electron per atom in univalent metals (see Sec. 44.2, Table 44.1). Then the number $N$ of free electrons per kilomole of a metal is the same as Avogadro's number, $N_A$, and the product $N_A k = R$ is the universal gas constant (see Sec. 26.9). Thus,

$$\Delta U_m = RTkT/2\mathscr{E}_F$$

The molar specific heat for the electron gas can be obtained from Eq. (45.4):

$$C_m = \Delta U_m/T = RkT/2\mathscr{E}_F \qquad (75.15)$$

3. Let us compare Eq. (75.15) with the specific heat $C_{cl}$ of a non-degenerate monatomic gas obeying the classical kinetic theory of gases. As will be recalled (see Eq. 27.14):

$$C_{cl} = 3R/2 \qquad (75.16)$$

Using (75.15) and (75.16), we readily find that

$$C_m/C_{cl} = kT/3\mathscr{E}_F \qquad (75.17)$$

As is seen, the specific heat of the degenerate electron gas in metals is markedly lower than that of a non-degenerate monatomic gas. More specifically, at room temperature $kT \approx 0.025$ eV, $\mathscr{E}_F \approx 3$ to $10$ eV, and, according to Eq. (75.13), $kT/\mathscr{E}_F \approx 0.01$ and so

$$C_m \sim 0.01 C_{cl}$$

This result agrees well with observations and explains why the electron gas has a low specific heat, thereby resolving a major limitation of the classical free electron theory of metals.

## 75.9. QUANTUM THEORY OF ELECTRICAL CONDUCTION IN METALS

1. The quantum theory of metals has brought with it marked changes in the classical concept of electrical conduction in metals. It is a well known fact that the electric current in metals is due to an ordered motion of electrons. This motion is initiated by the electric field set up in the metal by a current source. Before electrons can take part in the ordered motion under the action of the electric field, their energy state must be changed, that is, the electrons should "accept" an energy from the current source*. Under the usual applied

* We leave it as an exercise for the reader to determine the energy an electron acquires over its mean free path under the action of the electric field due to an ordinary cell. Use the material presented in Secs. 39.3, 44.2, and 44.6.

voltages, an electron can pick up a small amount of energy only if there are unoccupied energy levels nearby. Then, on gaining energy, the electrons will jump into these empty levels, giving rise to an electric current, that is, the motion of electrons in a direction opposite to that of the applied electric field.

2. The theory of electrical conduction in metals based on the Fermi-Dirac quantum statistics owes its origin to Sommerfeld. Among other things, it has led to an expression for Ohm's law of current density similar to Eq. (44.13). According to quantum theory, the resistivity is given by an equation outwardly similar to Eq. (44.15):

$$\begin{aligned}\gamma_q &= ne^2\bar{\lambda}(F)/m\bar{u}(F)\\ \gamma_{cl} &= ne^2\bar{\lambda}/2m\bar{u}\end{aligned} \qquad (75.18)$$

In essence, however, this result differs greatly from that given by classical theory. In Eq. (75.18), $\bar{\lambda}(F)$ is the mean free path of an electron *at the Fermi level*, $\bar{u}$ $(F)$ is the mean thermal velocity of an electron at the topmost filled energy level. The meaning of the quantities entering the expression for $\gamma_{cl}$ has been explained in Secs. 44.4 and 44.5.

3. The mean velocity $\bar{u}$ $(F)$ is practically independent of temperature because the Fermi level remains unchanged as it varies. According to the classical free electron theory of metals, $\bar{u} \approx \sqrt{T}$, and because of this it was inadequate to explain why $\gamma$ depends on temperature (see Sec. 75.1). The expressions for $\gamma_{cl}$ and $\gamma_q$ differ most of all because the classical and quantum theories of metals drastically differ in interpreting the mean free path of electrons, $\lambda$. On the classical theory, free electrons are particles of a classical electron gas which collide with the positive ions of the crystalline lattice, which accounts for the resistance that a metal offers to the flow of electric current. Quantum theory treats the motion of electrons through the crystalline lattice as the propagation of de Broglie waves. The interaction of these waves with the ions of the crystalline lattice is qualitatively different from a simple collision between an electron and an ion. *The de Broglie waves are scattered by the ions of the crystalline lattice.*

4. In quantum theory, the free mean path of an electron is replaced with the free mean path of a de Broglie wave, which is defined as the average distance that a de Broglie wave travels between successive scattering by the sites of the crystalline lattice. The lattice points are not an unsurmountable obstacle to the propagation of de Broglie waves. In fact, the waves can move round the lattice sites and travel considerable distances without scattering. This is why the free mean path $\bar{\lambda}$ is not directly related to the lattice spacing and may be as great as hundreds of them. If the mean free path of de

Broglie waves is sufficiently long, the probability of finding an electron that has travelled hundreds of lattice spacings will likewise be other than zero. In other words, an electron can travel large distances in a crystal. In classical terms, this amounts to stating that an electron has a mean free path of hundreds of lattice spacings.

5. This new view on the interaction of electrons with the metal lattice has led to a different interpretation of the resistance that metal conductors present to the flow of current and of its dependence on temperature.

It is known from optics (see Sec. 62.8) that the intensity of a radiant flux passing through a turbid medium (a fog, colloidal solutions, etc.) is reduced because some of the flux is scattered. The particles that scatter the incident flux should be spaced a distance $d$ apart, comparable with the wavelength $\lambda$ of light. If $d \ll \lambda$, light will not be scattered, and the medium will be optically uniform. At $d \ll \lambda$, the light-scattering irregularities of the medium scatter no light, and the radiant flux passes through, as if the medium is perfectly transparent. In Sec. 45.3 it is shown that about the same is true of sound waves scattered in solids. Electron waves, on passing through the metal lattice, experience a similar scattering.

6. A perfectly regular crystalline lattice having stationary ions at its points would not scatter electron waves, for there would be no scattering centres such as lattice imperfections exceeding the wavelength of de Broglie waves. A stream of free electrons would pass through such a lattice of stationary ions unobstructedly with no resistance to the flow of electric current.

In all solids, however, the particles occupying lattice sites do oscillate already at room temperature in a random manner and produce fluctuations in density which are as random (see Sec. 28.10). This may be explained as follows. The random thermal vibrations of particles at lattice points cause the spacings between material particles and, as a consequence, the density of the material in adjacent elementary volumes inside the metal to change continually, and this leads to the appearance of local density discontinuities. The linear dimensions of these discontinuities are much greater than the de Broglie wavelength, and the free electrons passing through the metal lattice are scattered by these thermal vibrations. This is why even pure metals offer an opposition to the flow of electric current. In vibrating, the lattice points cause the propagation of progressive waves in the acoustical mode (see Sec. 45.3), or quasi-particles called *phonons*. Thus, the electrical resistance of metals is due to the scattering of conduction electrons by phonons.

7. Since the attractive and repulsive forces between particles depend on their separation differently, the vibrations of particles in the crystalline lattice of a solid *are not harmonic*. The fact that the vibrations are not harmonic is essential to the understanding of

what causes pure metals to oppose the flow of electric current. The point is that the departure of these vibrations from harmonic ones upsets the lattice periodicity and, as a consequence, gives rise to discontinuities in density, and these serve as scattering centres for electron waves. If the lattice periodicity were not disturbed, the lattice ions would not scatter the electron waves, and the resistance of metals would be zero at any temperature. It is because the *lattice is no longer periodic* that electrons are scattered, there is an opposition to the flow of electric current, and heat is produced in any conductor.

8. As the temperature of a conductor rises, the scattering of electron waves by the thermal vibrations of the lattice increases, and the mean free path of electrons decreases. At room temperature, $\bar{\lambda}(F)$ is inversely proportional to temperature, $\bar{\lambda}(F) \backsim 1/T$. Because of this, the electrical conductivity, $\gamma_q$ is inversely proportional to temperature, too, $\gamma_q \backsim 1/T$, and this is confirmed by observations. At very low temperatures, the mean free path is inversely proportional to the fifth power of temperature, $\bar{\lambda}(F) \backsim 1/T^5$. Therefore, in accordance with observations, the resistivity of pure metals at extremely low temperatures is directly proportional to the fifth power of temperature, $\rho = 1/\gamma$, $\gamma \backsim 1/T^5$ and $\rho \backsim T^5$.

## 75.10. SUPERCONDUCTIVITY

1. It has been found that at $T_c \approx 4.2$ K, the resistance of purified mercury suddenly drops to zero (Fig. 75.11). At this temperature, the electric current induced in a mercury conductor has been noted to remain unchanged for any time interval, however long. This property has come to be known as *superconductivity*.

The temperature, $T_c$, at which a substance becomes a superconductor is called the *characteristic transition temperature*. It ranges from less than 1 K to about 20 K for different metals and alloys. The values of $T_c$ for some superconducting elements and compounds are tabulated in Table 75.1.

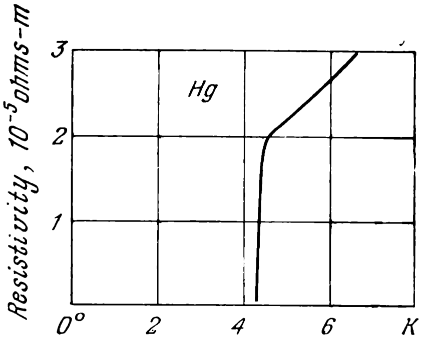


Fig. 75.11

The elements having the property of superconductivity occupy the central part of the Periodic Table. At room temperature, they usually have a lower electrical conductivity than non-superconducting metals.

*Table 75.1*

| Superconducting element or compound | $T_c$, K | Superconducting element or compound | $T_c$, K |
|---|---|---|---|
| Ir | 0.14 | MoC | 8 |
| Ti | 0.39 | Tc | 9.3 |
| Cd | 0.56 | $Mo_3Re$ | 9.8 |
| Zn | 0.85 | $Nb_3Zn$ | 10.5 |
| Ga | 1.10 | NbC | 11.1 |
| Tl | 2.39 | MoTc | 14 |
| Ag | 4.16 | NbN | 14.7 |
| Pb | 7.22 | $V_3Si$ | 17.0 |
| Nb | 8.9 | $Nb_3Sn$ | 18.2 |

2. The characteristic transition temperature of elements depends on their isotopic make-up (see Sec. 80.1). As a rule, $T_c$ decreases with increasing average atomic weight of the element having several isotopes. Impurities added to a pure metal or lattice imperfections only slow down the transition to superconductivity, but do not stop it altogether. This is an indication that during a superconducting transition the electrons cease to interact with the crystalline lattice of the metal.

3. Below $T_c$, superconductivity may be destroyed by a sufficiently strong magnetic field at $T$=constant. The field may be external or set up in the superconductor by the current traversing it. At any temperature such that $T < T_c$, there is a minimum magnetic field strength, $H_c$, which is sufficient to destroy superconductivity. This threshold value of magnetic field strength is called *critical*. It increases with decreasing temperature of the superconductor. A plot of the field strength $H$ as a function of temperature for white tin is shown in Fig. 75.12.

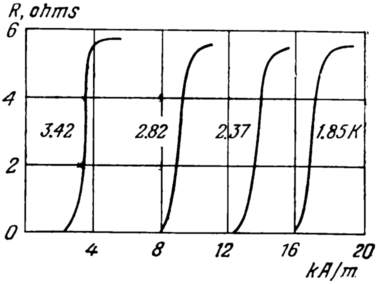


Fig. 75.12

4. At $T \leqslant T_c$, a superconductor placed in a magnetic field pushes out the magnetic flux. This is known as the *Meissner effect* after its discoverer. Fig. 75.13*a* shows the magnetic flux threading an ordinary conductor. Fig. 75.13*b* shows the lines of magnetic induction pushed out of a superconductor. The magnetic induction $B$ inside a superconductor is zero (provided the field strength is below its critical value), and the superconductor behaves as an ideal diamagnetic

of magnetic susceptibility $\varkappa_m = -1$ (see Sec. 42.4). At this value of $\varkappa_m$ the permeability $\mu = 1 + \varkappa_m = 0$ and $B = \mu_0 \mu H = 0$.

In practical semiconductors, the magnetic field does penetrate for some depth. This depth depends on the temperature and shape of the specimen. At temperatures which are 1 or 2 K below $T_c$, the flux penetrates for about $10^{-5}$ cm. This is why very thin superconducting films are not perfect diamagnetics; in them, the magnetic field is never zero.

5. The transition to the superconductive state changes the thermal properties of the substance. Among other things, in the absence of a magnetic field and at $T_c$ the heat capacity of a metal changes suddenly. In the presence of a field, the isothermal transition from the superconductive to the normal state is accompanied by a sudden change in the heat capacity and by the absorption of heat while the reverse transition is accompanied by the liberation of heat. The thermal conductivity of the material changes as suddenly.

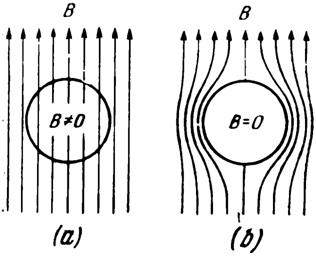


Fig. 75.13

The experimental facts listed above served as a basis for the *thermodynamic theory of superconductivity* to which major contributions were made by L. D. Landau and V. L. Ginzburg. According to this theory, the superconductive and normal states of a substance are two different phases of that substance. Given certain values of the variables of state, namely temperature $T$ and field strength $H$, these phases may change into each other.

6. Although it gave a formal explanation for the basic experimental facts, the thermodynamic theory of superconductivity could not of course reveal the true nature of the phenomenon. This has been done only recently on the basis of the quantum theory of solids.|

In 1950 it was suggested that superconductivity arises from interactions between conduction electrons and vibrations of the crystalline lattice, or phonons (see Sec. 45.3). The studies that followed led to the theory published in 1957 by J. Bardeen, L. N. Cooper and J. R. Schrieffer. According to this theory conduction electrons in a superconductor tend to interact in pairs being scattered by phonons, with the result that exchange forces come to act between the pairs (see Sec. 74.3). The exchange interaction is quantum-mechanical in nature and consists in *the mutual attraction of electrons*. Mathematically, this problem was solved by Bogoliubov of the Soviet Union in 1957.

It has been found that exchange interaction is especially strong for pairs of electrons with opposite spins and momenta. Given certain conditions, the attraction between these electrons may exceed their electrostatic repulsion by a wide margin. Because of this strong interaction, *conduction electrons in metals form a bound system which cannot give up energy in small portions*. This is why collisions with lattice ions do not affect the energy states of conduction electrons, and the metal behaves as an ideal superconductor of zero resistivity. In order to break up the bond between any electron and the remaining electrons of the bound system, an energy must be expended, corresponding to the average energy of thermal lattice vibrations at the transition temperature, $T_c$. This is why at $T > T_c$ superconductivity cannot exist.

These ideas have finally led to a consistent theory of superconductivity which explains all the properties of superconductors, including magnetic and thermal. The dependence of temperature on the isotopic make-up of metals and that of threshold field strength on temperature deduced theoretically agree well with observations. This theory has also given criteria for the existence of bound states in a system of interacting electrons and of superconductivity.

7. Superconductivity has found many applications in science and technology. The fact that heavy currents can be obtained from a low-voltage source with no Joule's heat liberated in the process is utilized in instrumentation. For example, a galvanometer whose coil is made from a superconductor has a sensitivity of about $10^{-12}$ V. The fact that a magnetic field can change a conductor into a superconductor and back is utilized to amplify extremely low direct currents and voltages. In such an amplifier, the weak d.c. signal to be amplified is fed into a superconductor placed in an alternating magnetic field. The field strength is chosen such that the conductor alternates between the normal and superconductive states. The resultant alternating current has the frequency of the magnetic field. After that, it is amplified in the usual manner. Superconducting materials may be used to make high-$Q$ resonant cavities (see Sec. 51.6), that is, cavity resonators having a very low impedance and a low attenuation.

Superconductivity can be utilized to obtain very strong magnetic fields. This can be done with electromagnets whose windings are wound with a wire fabricated from a superconductive alloy with a high value of $H_c$. Such windings can build up extremely high current densities and, as a result, a very strong field. Solenoids wound with wire made from superconducting alloys set up fields with a magnetic induction of over 10 T. Such solenoids do not dissipate power, while ordinary solenoids carrying copper windings dissipate a huge amount of heat at a magnetic induction of 10 T.

The fact that superconductivity can be destroyed by a magnetic field is utilized in devices known as cryotrons. The switching time of a miniature film cryotron is $10^{-9}$ to $10^{-10}$ s.

Computer memory devices utilize superconductivity, too. For their operation, such devices depend on the fact that the current induced in a superconducting core can be stored without impairment for a very long time.

A major limitation to the utilization of superconductivity is the need for extremely low temperatures. It is hoped, therefore, that the development of superconductors having a high (room) transition temperature will open up new vistas for the use of superconductivity in science and technology.

# Chapter 76
# THE BAND THEORY OF SOLIDS

## 76.1. AN OUTLINE OF THE BAND THEORY OF SOLIDS

1. When we discussed electrical conduction in metals in the preceding chapter we assumed that the electrons in metals were moving free inside a flat-bottomed potential box, or, which is the same, we assumed that the electrons in a metal had the same potential energy everywhere. Actually, this is not so. For, in addition to free electrons there are positive ions to reckon with. In a metal these ions set up an electric field which affects the motion of free electrons.

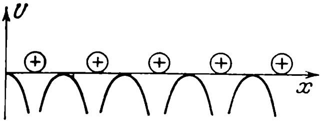


Fig. 76.1

2. The ions occupy the lattice sites and make up an ordered array. Accordingly, the distances between the ions along the $x$-, $y$- and $z$-axes are equal to the lattice spacings in these directions and the electric field due to the ions is a periodic function of the coordinates $x$, $y$, $z$. The potential energy of electrons in a metal is likewise a periodic function of their position. Changes in the potential energy of an electron along the $x$-axis are shown in Fig. 76.1. The $x$-axis is drawn through the lattice sites. At the location of ions, there appear "potential funnels"—minima of energy. In approaching an ion, the electron seems to fall in such a funnel.

3. The regular arrangement of atoms or molecules at the lattice points is characteristic of all solids, and not only of metals. The spacing between the lattice points is comparable in magnitude with the linear dimensions of the atoms and molecules themselves. Under the circumstances, the atoms (or molecules) cannot be treated as

isolated entities; one has to take into account their interactions. Nor ought one to think that the atoms (or molecules) of a crystal do not set up an internal electric field. As will be recalled, even an electric dipole, this simplest and electrically neutral system, gives rise to an electric field (see Sec. 10.4). This is also true of atoms (or molecules) which carry both positive and negative charge arranged in a specified manner. Thus, *in any crystalline solid, there is a periodic electric field established by the particles occupying the lattice sites.* Accordingly, the statement that the potential energy of electrons moving in a metal varies periodically applies to all solids.

4. A further step in the development of the quantum theory of metals (and of all types of solids) has been the *band theory of solids.*

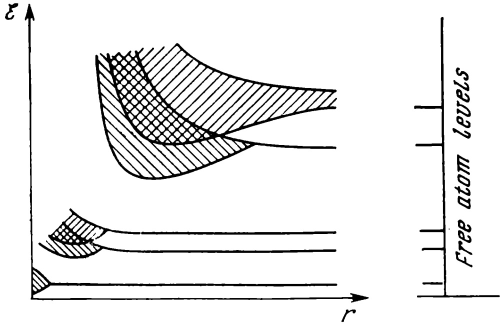


Fig. 76.2

The band theory treats a solid as an ordered array of particles taking up positions at the lattice sites and setting up a periodic electric field such that each electron is moving in a field due to all other particles. The term "band theory" will be clear from the subsequent discussion.

The central problem of the band theory of solids is to trace changes in the discrete energy levels of electrons in an isolated atom as the atoms are brought together to make a solid.

It has been shown (see Sec. 71.4) that the energies which electrons are allowed to have in an atom are separated by wide regions of forbidden energies. As the atoms are brought together to make a solid, the periodic electric field of the lattice and the interactions between the atoms markedly affect the energy levels of electrons in the solid. As a result, the electronic energy levels are split up. Instead of one energy level the same for all $N$ isolated atoms, there appear, in the solid, a number $N$ of closely spaced, but separate levels which fall into groups, or *bands.*

5. The manner in which the energy levels of isolated atoms split up as the atoms are brought together and the energy bands form in a solid is illustrated in Fig. 76.2. Level splitting is shown as a function

of the interatomic distance $r$. As is seen, not all levels split in the same way. Interactions between the atoms of the solid affect most of all the energy levels of the outer valence electrons which are bound loosely to their nuclei and have the greatest energy. The energy levels of the inner electrons split much less, and then only at distances substantially shorter than the lattice constant.

A solid may be looked upon as a huge molecule made up of a multiplicity of atoms. The energy levels of the inner electrons are practically the same as they are in isolated atoms. In contrast, the outer electrons are common to, or shared by, all the atoms of the molecule (or the solid). The outer electrons may have energies falling within the shaded areas in Fig. 76.2; these areas represent *allowed energy bands.*

## 76.2. SPLITTING OF THE ENERGY LEVELS OF OUTER AND INNER ELECTRONS IN THE ATOMS OF SOLIDS

1. Why the energy levels of electrons in a solid fall into allowed bands and why the energy levels of the outer and inner electrons split differently can be shown on the basis of quantum theory. For this, we should turn again to the indeterminancy relations for energy and time, Eq. (70.9).

It has been shown (see Sec. 72.8) that the mean life of an excited electron in an isolated atom is $\tau \approx 10^{-8}$ s. Accordingly, the natural energy band width $\Delta\mathscr{E}$ is

$$\Delta\mathscr{E} \geqslant \hbar/\tau \approx 10^{-7}\ \mathrm{eV}$$

and this decides the natural spectral line width (see Sec. 72.8). In making a solid, the independent atoms are brought together within a distance of about the same order of magnitude as the lattice constant. Now the valence electrons more loosely bound to their nuclei than the inner electrons can move from atom to atom by tunnelling through the potential barrier (see Sec. 70.6) which separates the atoms in a crystal. Let us show that this causes the energy levels of such electrons to spread until they turn into allowed energy bands, and find the range of values, $\Delta\mathscr{E}$, that a valence electron may have in a solid.

2. For simplicity, we assume that the atoms in a crystal are separated by a rectangular barrier such as shown in Fig. 76.3. The penetration probability of this barrier is given by (70.30):

$$D \approx \exp\left[-2L\sqrt{2m(U_0-\mathscr{E})}/\hbar\right]$$

where $U_0 - \mathscr{E}$ is the height of the barrier for an electron of energy $\mathscr{E}$, which in our case is about 10 eV (which is of the same order of magnitude as the ionization energy of an atom) and $L$ is the barrier width. For the interacting atoms in a crystal, it is reasonable to set $L \approx 10^{-10}$ m which is comparable with the lattice constant. Now we find the average time $\tau$ during which the valence electron stays with a given atom, that is, the time during which the electron remains inside a rectangular potential well whose linear dimensions are $d \approx 10^{-10}$ m (the linear dimensions of an atom). If an electron in an atom (the potential well) is moving with a velocity of $v = 10^6$ m/s, then the electron will strike the barrier at a rate of $v/d$ per second. The frequency at which the electron tunnels through the barrier is given by

$$\nu = vD/d = (v/d) \exp\left[-2L\sqrt{2m(U_0 - \mathscr{E})}/\hbar\right]$$

Then, the mean life time of an electron in a given atom is

$$\tau = 1/\nu = (d/v) \exp\left[2L\sqrt{2m(U_0 - \mathscr{E})}/\hbar\right]$$

On substituting the numerical values of all terms, we get $\tau \approx 10^{-15}$ s. Thus, the time interval during which a valence electron stays with a given atom is one-ten millionth of the mean life of an excited atom in an independent atom.* Obviously, this results in a sudden spread in the energy level of the valence electron in an atom which is part of a crystal. From the indeterminancy relation we get

$$\Delta\mathscr{E} \geqslant \hbar/\tau \approx 1 \text{ eV}$$

Instead of a level width of $\Delta\mathscr{E} \approx \approx 10^{-7}$ eV that an electron has in an independent atom, in a crystal there appears a band of allowed energy values about 1 eV wide, which is $10^7$ times as great.

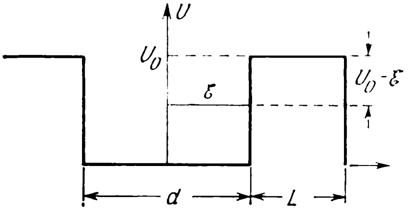


Fig. 76.3

3. The estimates quoted above apply only to valence electrons. The inner electrons, that is, those occupying complete shells in atoms, have a very negligible probability of tunnelling through the barrier and jumping to another atom. For the inner electrons, the potential barrier height increases considerably, being $U_0 - \mathscr{E} \approx \approx 10^3$ eV. A relatively small increase in the barrier width ($L \approx \approx 3 \times 10^{-10}$ m) brings about an entirely different result: on the average, an electron stays with a given atom for $\tau \approx 10^{20}$ years. It is obvious that the spread in the energy levels of atoms occupying the inner shells in the atoms of a crystal is in no way comparable with the spread in the excited levels of the valence electrons in an independent atom.

* This estimate shows that in crystals having loosely bound valence electrons, that is, in metals, an electron gas is formed.

## 76.3. ARRANGEMENT OF ENERGY BANDS IN SOLIDS. BAND-TO-BAND AND INTRABAND ELECTRON TRANSITIONS

1. The allowed energy band arising from one level existing in an independent atom is made up of $N$ closely spaced levels, where $N$ is the total number of atoms in a solid. It is of the order of $10^{28}$ or $10^{29}$ to the cubic metre, and so an allowed energy band contains as many levels. Since the difference in energy between adjacent levels is about $10^{-22}$ eV (see Sec. 75.6), then the total width of an allowed energy band is a few electron-volts. Much as in an isolated atom the discrete energy levels are separated by energy gaps, so the allowed energy bands in a solid are separated by forbidden bands. The forbidden bands are about the same in width as the allowed bands. Wider allowed energy bands arise from the more excited levels in an independent atom. As energy increases, however, the allowed bands grow wider, and the forbidden bands narrower. An energy band diagram for a solid is shown in Fig. 76.4.

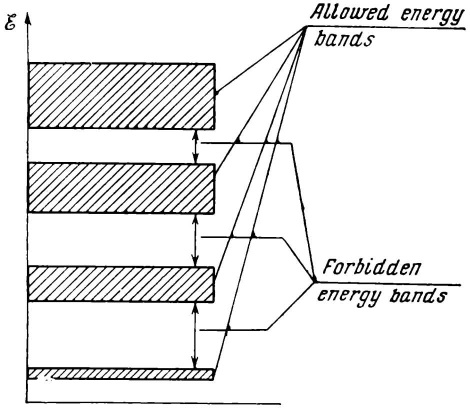


Fig. 76.4

2. In an independent atom, the quantized allowed energy levels may either be filled with electrons or remain empty. In a solid, the electrons form a system of particles obeying the Pauli exclusion principle. Let $2N$ electrons occupy certain levels in $N$ isolated atoms. Then, in a crystal these $2N$ electrons will make up pairs that will occupy $N$ levels in the respective band, the spins of the electrons in each pair being in opposite directions. In a solid, some energy bands may be filled full, others only partly filled, and still others empty.

3. Much as they do in an independent atoms, the electrons in a crystal may move from one allowed band to another (band-to-band or interband transitions) or between levels in a given band (intraband transitions). For an electron to move from a lower allowed band

to an adjacent higher band, it must be imparted an amount of energy equal to the width of the forbidden band separating the two allowed bands. This energy is a few electron volts. The energy input required for an electron to undergo a transition between levels in the same band is as low as about $10^{-22}$ eV, which is the difference in energy between adjacent energy levels.

4. Consider some of the factors that may cause electrons to undergo transitions between levels in a band. The energy that an electron gains under the action of the electric field due to a current source usually is $10^{-4}$ to $10^{-8}$ eV,* which is sufficient to initiate an intraband transition. The average energy of thermal vibrations of atoms in solids at room temperature is $kT \approx 0.025$ eV. If imparted to electrons, this energy will also be sufficient to cause transitions between levels inside an allowed band.

5. Interband transitions call for an energy input of a few electron volts. At the potential differences that can usually be applied to a solid, the energy that electrons can gain over their mean free path is insufficient for an electron to climb to a higher allowed band. The energy sufficient for an upward transition to take place can be imparted by heating the crystal. Thus, the thermal excitation of electrons may lead to transitions both inside a band and between bands.

6. The foregoing applies to all solids as well as to metals. For the energy levels of electrons split up and energy bands form because the atoms or any other particles in the lattice sites interact and set up a periodic field inside the solid. It has been shown in Sec. 76.1 that even if the particles at the lattice sites are neutral atoms or molecules, their nuclei and electrons establish an electric field which is especially strong near the lattice sites. Thus, in any solid made up of particles (ions, atoms or molecules) arranged in a regular array at the lattice points, there is an energy band structure and an energy band spectrum.

## 76.4. METALS AND DIELECTRICS IN THE LIGHT OF THE BAND THEORY OF SOLIDS

1. In terms of the band theory of solids, the difference in electrical properties between metals (which are conductors of electric current) and dielectrics (which do not conduct current) is decided by two factors. One is the energy band scheme, or more specifically, the forbidden band width. The other is the difference in electron population between the allowed bands. For a solid to be a conductor, it is important that there should be free energy levels to which an applied electric field can excite electrons.

* This figure applies to the energy that an electron gains over its free mean path.

2. It has been shown in Sec. 76.3 that conventional current sources are capable of only causing intraband transitions. Let us establish the condition necessary for electrical conduction in metals, using metallic sodium as an example. As will be recalled (see Sec. 73.2) an independent sodium atom has two complete shells, one containing two and the other eight electrons. With the eleventh, valence, electron, the upper energy level of the sodium atom is only half-full in accordance with the Pauli exclusion principle. A crystal of metallic sodium has two completely filled energy bands identified with these complete shells of the independent sodium atoms. They will not interest us, because an applied electric field cannot bring about intraband transitions in them. In making a crystal, the valence electrons form a *valence band* which is only half-full with electrons and is therefore a *conduction band* (Fig. 76.5). In other words, the electrons in this band can take part in the transport of electric current because an applied electric field can impart them an amount of energy sufficient for an upward transition to free energy levels, and the electrons will constitute a motion of charges, that is, an electric current.

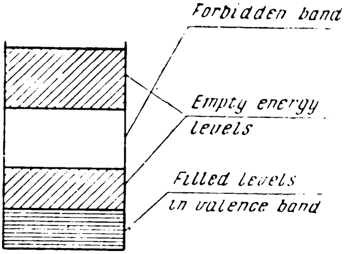


Fig. 76.5

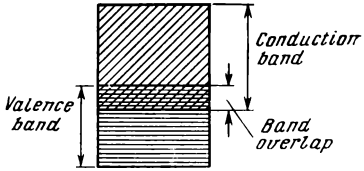


Fig. 76.6

Thus, if the valence band is not filled full, the solid will always be an electric conductor. Examples are the metals in the first group of the Periodic Table (lithium, sodium, potassium, rubidium and caesium).

3. A solid may also be a conductor when its conduction band overlaps with the band formed due to the splitting of energy levels of the valence electrons. Examples are the crystals of the elements in the second group of the Periodic Table (beryllium, cadmium, magnesium, zinc, etc.). In this case, a wide "hybrid" band is formed which is filled by valence electrons only partly. Above and close to the filled levels there are empty levels and, as we have already learned, the solid will be a conductor (Fig. 76.6).

4. The band theory of solids has explained why the electrical conductivity of metals does not rise with increasing valence. From

a classical point of view, an increase in the valence of a metal, that is, an increase in the number of "free" valence electrons per atom, should have resulted in a greater electrical conductivity. For example, aluminium is tri-valent because it has three valence electrons per atom, and copper is uni-valent because it has one valence electron per atom. It would seem that the electrical conductivity of aluminium should be greater than that of copper. Actually, the reverse is true. As can be seen from Table 44.2, the resistivity of copper at 0 °C is about half as great as that of aluminium; that is, its conductivity is twice as great.

Now it is known that the electrical conductivity of a metal does not depend on the number of valence electrons per atom. It does depend on the number of electrons for which the upper conduction band has a sufficient number of free energy states.

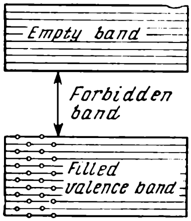


Fig. 76.7

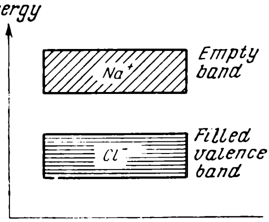


Fig. 76.8

In bivalent alkali-earth metals, the valence electrons in the lattice atoms occupy energy levels in the hybrid band in such a manner that some of them are free for occupancy. However, bivalent metals have fewer electrons that an applied electric field can cause to jump into these free states, than uni-valent metals, and so their electrical conductivity is lower than that of the latter. This reasoning applies still more to tri-valent metals.

5. In some solids, the bands do not overlap, while the valence band which contains the outer electrons of the atoms or ions is filled full and the higher bands are empty (Fig. 76.7). Such a solid is a dielectric (an insulator), that is, a substance which does not conduct electric current. An example is common salt, NaCl, whose molecules have ionic bonds. As will be recalled (see Sec. 74.2), in a molecule of NaCl the outer, eleventh, electron of the Na atom moves into the outer shell of the Cl atom. As a result, ions $Na^+$ and $Cl^-$ are formed, in which the outer electron shells are complete. In a crystal of NaCl, there is a complete valence band due to $Cl^-$, above which, 6 eV higher, is the energy band of the ion $Na^+$, empty of electrons (Fig. 76.8). The electric field due to a current source cannot force electrons from the complete band of $Cl^-$ to the empty conduction band of $Na^+$, and crystals of NaCl behave as dielectrics.

# Chapter 77

# ELECTRICAL PROPERTIES OF SEMICONDUCTORS

## 77.1. INTRINSIC ELECTRON CONDUCTION IN SEMICONDUCTORS

1. Butting in between metals with a resistivity of $10^{-6}$ to $10^{-8}$ ohms-m, and dielectrics with a resistivity of $10^8$ to $10^{13}$ ohms-m are a great number of materials classed as *semiconductors*. Their resistivity ranges from $10^{-5}$ to $10^8$ ohms-m. Almost everything around us in nature consists of semiconducting materials. Oxides, sulphides and tellurides of many metals also show semiconducting properties. In the Periodic Table, semiconductors make up a group of elements shown in Fig. 77.1. On the bottom left of the semiconductors are metals. On the top right are elements which are dielectrics in the solid state.

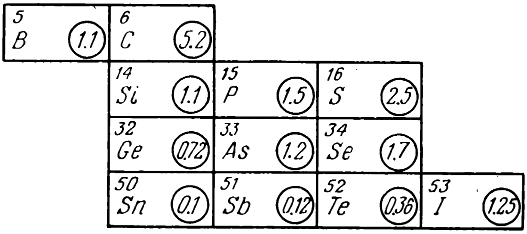


Fig. 77.1

Typical representatives of semiconductors, widely used in technology, are germanium, silicon and tellurium.

2. Germanium, Ge, is used most of all semiconductors. It belongs to group IV, Period 4, of the Periodic Table. An isolated germanium atom has four electrons in its outer shell. The remaining 28 electrons are distributed between the inner complete shells. A germanium crystal has an atomic lattice. The four valence electrons of each atom form covalent bonds with the electrons of neighbouring atoms in the lattice so that no free electrons are left in pure germanium (Fig. 77.2). Because of this, a crystal of pure germanium must be a good insulator.

Silicon, Si, is widely used in present-day semiconductor technology. The fourteen electrons of an isolated silicon atom are arranged around its nuclei so that four of them are in the outer shell as in germanium. In silicon crystals, as in germanium, the outer valence electrons form covalent bonds with their neighbours, and silicon must likewise be an insulator.

3. The covalent bonds in germanium and silicon may readily be broken already at relatively low temperatures. As a result, free electrons will be produced, giving rise to what is called *intrinsic conduction*, that is, the conduction characteristic of a pure, ideal crystal.

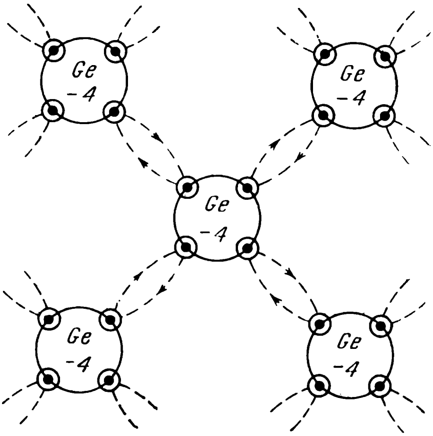


Fig. 77.2

Experiments show that the electrical conductivity of pure semiconductors increases with rising temperature, which constitutes a major distinction of semiconductors from metals in which (see Sec. 44.5) the conductivity decreases with rising temperature.

4. This change in the resistance (or conductance) of pure semiconductors with temperature is adequately explained by the band theory of solids. If the upper filled band is separated from the nearest allowed band by a narrow forbidden band, the solid can be a dielectric only at very low temperatures. With rising temperature, thermal excitation can force the electrons from the upper boundary of the filled band into the upper band having free energy levels. For this to happen, the energy input should be equal to at least the width of the forbidden band $\Delta\mathscr{E}_0$. The relative position of energy bands in a semiconductor and a dielectric is shown in Fig. 77.3. The quantity $\Delta\mathscr{E}_0$ is called the *activation energy* of intrinsic conduction. It is a very important characteristic of semiconductors. In Fig. 77.1, the activation energies (in electron volts) of semiconductors are given inside the circles.

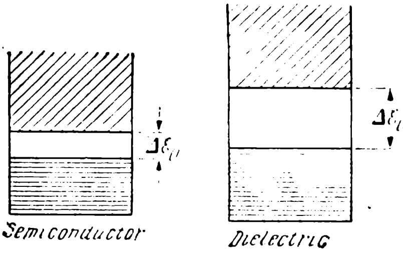


Fig. 77.3

As the temperature of a pure semiconductor rises, the number of electrons that are forced by thermal excitation into a free energy band and contribute to electrical conduction increases. This is why the conductivity of semiconductors increases with rising temperature.

## 77.2. INTRINSIC HOLE CONDUCTION IN SEMICONDUCTORS

1. The electron-free energy levels in the filled band have a decisive effect on the electrical properties of semiconductors. An electron gaining an energy as a result of thermal excitation breaks its covalent bonds with the other electrons of the atom and leaves its site. As an example, assume that in an electrically neutral substance one of its electrons leaves its site for another one, say, in an adjacent ion. Then an excess positive charge will be left at its former site, because in the normal state the substance is electrically neutral. This excess positive charge appearing at the vacant electron site is called a *hole*. It behaves as a positive charge equal in magnitude to an electronic charge. An adjacent electron may move into this vacancy or, which is the same, the hole may be said to move to another site left by an adjacent electron. In a similar manner, the vacancy left in a file by a soldier will immediately be filled by one next behind, leaving a new vacancy one place farther, to be filled by still another soldier, and it will appear as if the vacancy moves down the file, in the direction opposite to the movement of soldiers.

2. In an applied electric field the electrons move against the electric field intensity, while the holes move with the field, that is, in the direction in which the field would cause a positive charge to move. The electrical conduction of a semiconductor due to the motion of holes is called the *intrinsic hole conduction*. The holes move with and the electrons against the field throughout the bulk of the semiconductor specimen. Thus, in a semiconductor conduction is by both electrons and holes.

From the view-point of the band theory of solids, hole conduction is due to the fact that thermal (or any other) excitation of electrons from the filled band leaves vacant levels in that band. These levels may be taken over by electrons in the same band. As shown in Sec. 76.3, intraband transitions under the action of an electric field give rise to conduction in a solid. It is important to stress that hole conduction is maintained by the ordinary electrons located in the shells around the nuclei, and not by the positive charges of the atoms (nuclei). Yet, the conduction mechanism involves the motion of electrons from particle to particle in a manner equivalent to the motion of a positive charge. This is why, in the physics of semiconductors, it is customary to speak of hole conduction and of a hole current. The concept of hole current is a very convenient tool in practical applications (see Chapter 78).

3. From the foregoing it is seen that the carriers of charge in semiconductors may be electrons and holes. However, the above discussion applies only to chemically pure semiconductors. Unfortunately, crystals of pure semiconductors are not all easy to grow. Most semiconductor devices (see Chap. 78) use materials in which electrical conduction is due to the presence of impurities.

## 77.3. IMPURITY OR EXTRINSIC ELECTRONIC (*N*-TYPE) SEMICONDUCTORS

1. An impurity added to a semiconductor strongly affects its electrical properties. An impurity may be atoms or ions of other chemical elements which occupy sites in the lattice of the host material. Or it may be a lattice imperfection, such as a vacancy, a dislocation and the like. The impurities bring about changes in the periodic field of the crystal and affect the behaviour and energy levels of electrons. When impurity atoms are added to the lattice of the host semiconductor, their valence electrons have energy levels which cannot be accommodated in the allowed energy bands of the host. Because of this, *impurity energy levels* appear in the forbidden band.

Impurities affect the number of charge carriers in a semiconductor differently. Some may serve as suppliers of excess electrons in the crystal, others may act as centres to which the electrons already present in the crystal may attach themselves.

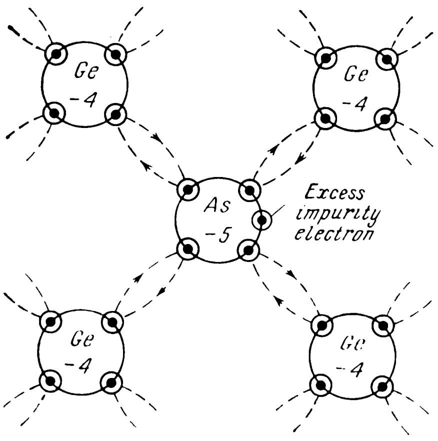


Fig. 77.4

2. As an example, let us see what happens when an atom in the germanium lattice is substituted by an impurity atom which has five valence electrons (this may be phosphorus, arsenic, or antimony) (Fig. 77.4). Four electrons of the impurity atom form covalent bonds with the electrons of adjacent germanium atoms, but the fifth electron cannot do so. This excess electron is only loosely bound to its atom and can be moved into the conduction band of the semiconductor with relative ease.

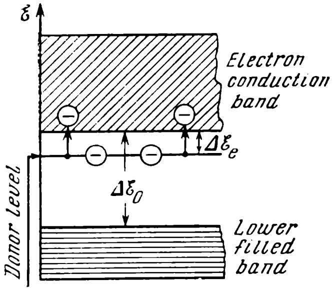


Fig. 77.5

The energy of such excess impurity electrons is somewhat lower than the energy corresponding to the bottom of the conduction band. Therefore the energy levels of the impurity electrons occur somewhat below the bottom of the conduction band. These levels are filled with electrons and are called *donor levels*, and the impurity atoms supplying these excess electrons to the host crystal are called *donor atoms.* A small amount of energy, $\Delta\mathscr{E}_c$ is required to force the electrons from the donor levels into the conduction band, this energy can be supplied by heating. For example, for silicon $\Delta\mathscr{E}_e$ is 0.054 eV. if the impurity atom is arsenic.

3. With electrons from donor levels transferred to the conduction band, the semiconductor shows *extrinsic* or *impurity electronic* (*N-type*) *conduction.* Accordingly, we have an *N*-type semiconductor (with *N* for "negative", because the conduction is primarily by negative electrons).

The relative positions of the energy bands and impurity levels in an *N*-type semiconductor are shown in the energy diagram of Fig. 77.5.

## 77.4. IMPURITY OR EXTRINSIC HOLE (*P*-TYPE) SEMICONDUCTORS

1. Now let an atom of a tri-valent element, such as boron, aluminium or indium, be present in the germanium lattice (Fig. 77.6). This impurity atom has one electron too few than is required to satisfy all the covalent bonds in germanium (see Fig. 77.2). However, this impurity atom can satisfy the binding arrangement by borrowing an electron from the nearest germanium atom. Then a hole will appear where the borrowed electron has been, and it can in turn be filled by an electron from another adjacent germanium atom, and so on. As has been shown in Sec. 77.2, this amounts to the motion of a hole in the semiconductor, and this hole becomes a charge carrier. Since the impurity atoms borrow or, rather, accept electrons from the host material, they are called *acceptor impurities*, and the impurity energy levels free from electrons that they form in the forbidden band are called *acceptor levels.*

2. The electron energy levels of acceptor atoms are somewhat higher than the top of the filled band in the host semiconductor (Fig. 77.7). With tri-valent boron as impurity and a silicon crystal as host, the acceptor levels are $\Delta\mathscr{E}_h = 0.08$ eV above the top of the filled band. The energy $\Delta\mathscr{E}_h$ is only a fraction of the total width of the forbidden band.

In such a semiconductor, conduction may be by impurity holes. This is because the electrons residing at the top of the filled band may readily be transferred by heating to the acceptor levels until these levels are filled full. Once they are in the acceptor levels, the applied field cannot change the energy states of these electrons—they are trapped there. As a result, the bottom band will have electron vacancies, and these will behave as holes. Thus, the bottom band becomes a *hole conduction band.* The applied field will cause the electrons in the bottom band to move from level to level inside the band and to fill holes progressively. This amounts to the motion of holes against the flow of electrons. This is *P-type conduction*, with *P* for "positive", because the conduction is mainly by positive holes. Accordingly, the materials showing it are called *P-type* semiconductors. The arrangement of impurity levels in a *P*-type semiconductor is shown in the diagram of Fig. 77.7.

3. In conclusion, it is important to stress the difference in behaviour between the carriers in semiconductors and electrons in metals.

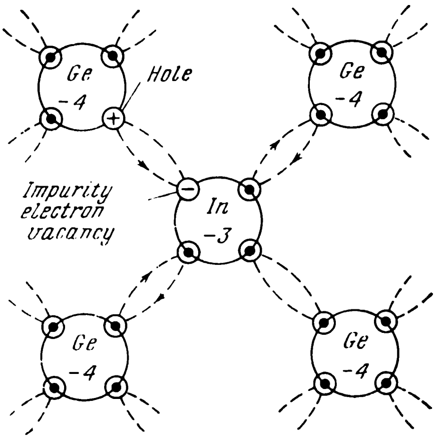


Fig. 77.6

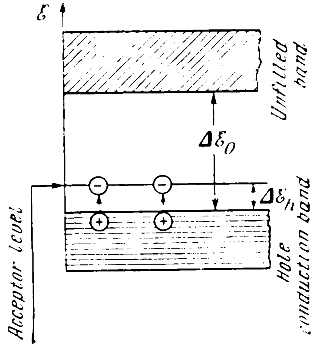


Fig. 77.7

The quantum theory of metals has shown that the number density and energy of electrons in a metal are practically independent of temperature. In other words, the number of carriers in a metal cannot practically be controlled, and this has been corroborated by observations. In a semiconductor of any type of conduction, the number of carriers is considerably smaller than that in metals. However, the number density and energy of carriers in semiconductors depend strongly on temperature, increasing with rising temperature. This offers a means for controlling the number and energy of carriers in semiconductors—a feature widely utilized in semiconductor devices.

# Chapter 78

# PHYSICAL PROCESSES IN SEMICONDUCTOR DEVICES

## 78.1. CONTACT PHENOMENA IN METALS

1. It has been known for a long time that a metal-to-semiconductor junction can rectify alternating currents at very high radio frequencies. This property was at the basis of the early crystal detectors and

rectifiers and, later, of crystadyne amplifiers. Rectification at a metal-semiconductor junction, or unilateral conduction through such a junction, is attributed to the existence of a *barrier* or *blocking layer* on the contact surface, passing current in one direction and effectively blocking it in the opposite direction.

2. The barrier layer is a thin layer near the junction with the potential energy of electrons changing considerably across the layer. The barrier layer is formed as the carriers of opposite signs gather on each side of the junction. There forms the so-called *electrical double* or *dipole layer* which interacts with the carriers that pass through the junction.

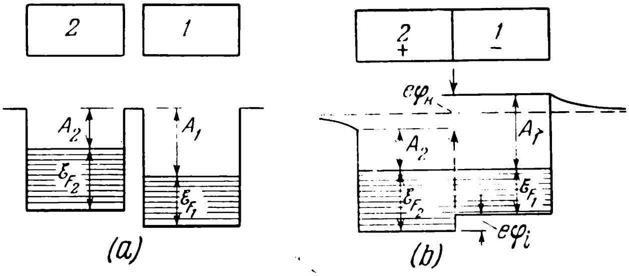


Fig. 78.1

The electrical double layer sets up an electric field in which a stream of electrons is moving. In their motion, the electrons run into a potential barrier. If a voltage is applied across the junction in a direction such that the potential barrier is reduced, the junction is said to be *forward-biassed*, the resistance in this direction (the *forward resistance*) is brought down, and a current is allowed to cross the junction. With a voltage applied in the reverse direction, the junction is said to be *reverse-biassed*, its potential barrier for electrons is increased, as is the resistance in this direction (*reverse resistance*), and no current is allowed to flow across the junction.

3. Consider the formation and function of the dipole layer, taking as an example two metals, *1* and *2*, differing in work function, $A_1$ and $A_2$, that is, in the height of the top electron-filled Fermi level (Fig. 78.1*a*).

Immediately after the two metals are brought in contact, the flow of electrons is predominantly from metal *2* to metal *1*, for the former has a lower work function than the latter. This goes on until the tops of the filled energy levels in the two metals are separated the same distance from the junction, and a dipole layer forms, consisting of a positive layer in metal *2* due to the positive charge of the ionic lattice, and a negative layer in metal *1* into which the electrons from metal *2* are swept (Fig. 78.1*b*). At equilibrium, the electric field due to the dipole layer puts an end to the flow of electrons.

4. Thus, equilibrium between two or more metals in contact is possible only when the metals have the same work function or the same Fermi level. This can be proved from the general thermodynamic conditions for an equilibrium of the electron gas in the metals in contact, with their volume and temperature held constant. Qualitatively, everything reduces to the fact that the pressure of the electron gas in the metals that are brought in contact becomes the same—a condition essential to equilibrium—as the energy levels of the electrons become the same, not unlike the liquid level in communicating vessels.

As metal *2* is charged positively to a certain potential $U_2$, all energy levels are brought down by $eU_2$, while in metal *1* charged negatively to a potential $-U_1$, all energy levels are raised by an amount $eU_1$ above their positions in the uncharged metal.

Why this is so may be explained as follows. The transition of an electron from a level in an uncharged state to the same level after the metal has been charged to a negative potential $-U_1$ requires an energy input numerically equal to $eU_1$. In other words, the potential energy of the electron increases by $eU_1$, which implies that all levels go up by this amount. The same reasoning applies to the levels in the positively charged metal, except that they go down, in comparison with their position in the uncharged metal.

5. As soon as the Fermi levels in metals *1* and *2* become the same, the factor causing the predominant flow of electrons from metal *2* to metal *1* ceases, and a dynamic equilibrium sets in between the two metals, in which the electric double layer has a certain definite width $l$. However, the electrons in direct vicinity to the metal surfaces differ in potential energy, and a difference of potential is established between the metals, given by

$$\varphi_k = (A_1 - A_2)/e$$

usually referred to as the *extrinsic contact potential difference*. As is seen, it exists due to the difference in work function between the metals in contact. For different pairs of metals, the value of $\varphi_k$ ranges from a few tenths of a volt to several volts (see Sec. 44.9) and depends on the general condition and cleanliness of the contacting surfaces.

6. There is also a potential difference between points within the contacting surfaces. It exists because the potential energy of electrons in metals *2* and *1* is not the same, despite the fact that the Fermi levels are the same. From Fig. 78.1*b* it is seen that the potential energy of electrons in metal *2* is lower than that of electrons in metal *1*, the difference being $\mathscr{E}_{F2} - \mathscr{E}_{F1}$. Accordingly, the potentials within

the metals differ by an amount

$$\varphi_i = (\mathscr{E}_{F2} - \mathscr{E}_{F1})/e$$

which is called the *internal* or *intrinsic contact potential difference.*

7. The difference between the potentials inside the conductors gives rise to an additional diffusion current from the second to the first metal. Using Eq. (75.11) for the Fermi energy at $T = 0$ K, we may write

$$\mathscr{E}_F = (h^2/2m)(3n/8\pi)^{2/3}$$

where $n$ is the number of electrons per unit volume of metals. Substituting this expression for $\mathscr{E}_F$ in the equation for $\varphi_i$, we get

$$\varphi_i = (h^2/2me)(3/8\pi)^{2/3}(n_2^{2/3} - n_1^{2/3})$$

As is seen, the internal contact potential difference stems from the difference in electron gas concentration between the metals in contact. It causes electrons to diffuse so as to bring down the higher concentration. Qualitatively, this agrees with the results obtained in Sec. 44.7 on the basis of the classical free-electron theory of metals. However, there is a marked quantitative difference between the classical and quantum-mechanical results.

8. At a temperature other than zero, the Fermi energy is temperature-dependent. Because of this, the internal contact potential difference depends on the temperature at the contact. This dependence explains the thermoelectric effects discussed in Sec. 44.8. Without going into further details, it may be pointed out that Eq. (44.23) for the thermo-emf is valid approximately, but the factor $\alpha$ is given by a far more elaborate expression and is generally a function of temperature.

## 78.2. RECTIFICATION AT A METAL-SEMICONDUCTOR JUNCTION

1. Consider a metal-semiconductor junction, choosing an $N$-type semiconductor to make the problem more definite.

Let the metal have a greater work function than the $N$-type semiconductor. Then the electrons at the donor impurity levels will move from the semiconductor into the metal, and the semiconductor layer near the junction will be depleted of free electrons to become positively charged. On collecting excess electrons, the metal will charge negatively, and a double electric layer will form between them.

2. The magnitude of the contact potential difference at the metal-semiconductor junction is governed by the fact that the semiconductor has a relatively low carrier concentration. The equilibrium thickness of the double layer is considerable; that is the Fermi levels of

the metal and semiconductor become the same with the double layer extending for a considerable distance into the semiconductor.

The double layer drastically differs in properties from the bulk of the semiconductor. Above all, the electrons in this layer have a greater potential energy than the bulk electrons, and the energy levels of electrons in the energy bands of the semiconductor near the junction go up. On the side in contact with the metal, a layer is formed in the semiconductor, depleted of electrons. This is a *depletion*, *barrier* or *blocking* layer.

3. The most important property of the depletion layer is the difference in opposition which it offers to a flow of current in different directions. Precisely this property is at the basis of rectification at a metal-semiconductor junction. When the junction is biassed in the forward direction (that is, the "—" side of the source is applied to the semiconductor), the resistance of the depletion layer is brought down, and the current is allowed to pass with no or little resistance. This happens because the width of the double layer is reduced, the potential barrier for electrons is brought down, and electrons find it easier to move from the semiconductor into the metal. With a sufficiently high applied voltage across the junction, the resistance of the double layer becomes the same as that of the bulk of the semiconductor. When the junction is reverse-biassed, the applied voltage leads to a further depletion of the semiconductor layer adjacent to the metal. The double electrical layer increases in width, its opposition to the flow of current rises, and little or no current can flow across the junction.

4. The action of the barrier (blocking or depletion) layer at the junction between a metal and a semiconductor is to a marked degree dependent on the area of the contact surface. More specifically, a depletion layer appears only when the contact surface is small in area. This is because the impurity concentration is not the same throughout the bulk of the host material. Some small areas have them, and others have not, and a barrier layer forms where there is a concentration of impurities. In areas devoid of impurities, nothing exists to bring about a sudden change in resistance, and its value is low. On the other hand, where the contact surface is large in area, it will include regions high in electrical conductance, and these will shunt the barrier layer, thereby reducing its rectifying properties.

## 78.3. RECTIFICATION AT A *P-N* JUNCTION

1. The region in a single-crystal semiconductor where the *N*-type of conduction (by electrons) changes into the *P*-type of conduction (by holes) is called the transition region or, most commonly, the *P-N junction*. It is produced by doping a single crystal with appro-

priate impurities so as to produce regions differing in type of conduction (an $N$ and a $P$ region). Alternate $P$-$N$ regions can be obtained by adding a donor to the melt, then pulling the crystal a short distance, then adding an acceptor and pulling the crystal again, and so on. A device with one $P$-$N$ junction is called a *crystal diode*.

The impurities added to the host material bring about stable changes in its properties and produce the desired type of conduction.

2. The thin transition region is the basis of rectification, or unilateral conduction, by semiconductor devices.

When two semiconductors differing in type of conduction are brought in contact, electrons from the $N$-type semiconductor diffuse into the $P$-type. Because of this, the $N$-type semiconductor is depleted of its electrons near the junction, and an excess positive charge

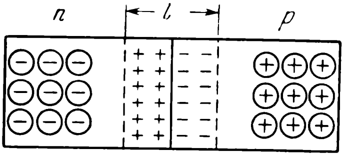


Fig. 78.2

appears in the $N$-type conductor. Holes from the $P$-type semiconductor diffuse in the opposite direction, leaving behind them an excess negative charge in the $P$-type semiconductor. As a result, a double electrical layer of width $l$ is formed (Fig. 78.2) which prevents any further migration of electrons and holes across the junction. For both types of carriers, this double electrical layer acts as a potential barrier with a height of a few tenths of a volt. Electrons and holes can overcome this obstacle only at a temperature of several thousand degrees, and so the contact layer effectively blocks the flow of carriers.

3. The resistance offered by this blocking layer may be greatly affected by an applied electrical field. Let the $N$-region of a semiconductor device be connected to the "—" side of a power supply whose "+" side is connected to the $P$-region of the same device (Fig. 78.3). The applied electrical field will cause electrons in the $N$-region to move towards the transition region. The same field will cause the holes in the $P$-region to move towards the junction, too, that is, against the flow of electrons.

This type of connection is called *forward biassing*. With it, the width of the blocking layer is progressively decreased, because the electrons swept across the junction by the field will fill the holes, that is, the electrons and the holes will recombine. Thus, the $P$-$N$ junction offers no resistance to the flow of current brought about by the applied voltage. This voltage only maintains the flow of electrons and holes towards one another.

4. If we reverse the polarity of the applied voltage, the holes in the *P*-region and the electrons in the *N*-region will be forced to move away from the junction in opposite directions (Fig. 78.4). Now the blocking or depletion layer on each side of the junction will increase in width—that is, there will be an increase in the width of the regions depleted of their electrons or holes, respectively, and the double electrical layer will gain in size, too. The *P*-region will acquire an excess negative charge, and the *N*-region an excess positive charge.

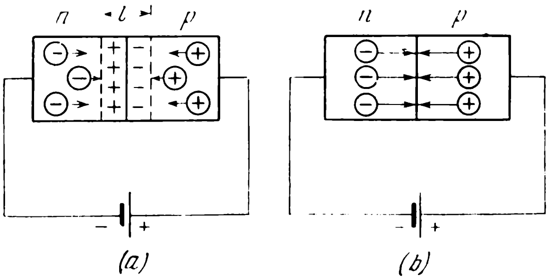


Fig. 78.3

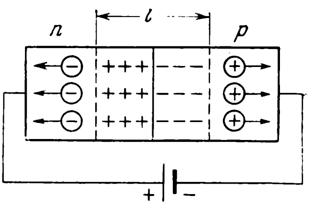


Fig. 78.4

With an increase in the applied voltage there will be a continuous growth in the depletion layer and, as a consequence, in its resistance. This is *reverse biassing*. With a sufficiently high reverse-bias voltage, the depletion layer will be practically an insulator devoid of mobile carriers. If the device is again biassed in the forward direction, the recombination of electrons and holes will bring down the potential barrier at the *P-N* junction and the junction resistance will be drastically reduced. Thus, with an alternating applied voltage, the device conducts only in one direction, that is, the alternating current is rectified.

## 78.4. TRANSISTORS

1. While in their effect germanium or silicon diodes are similar to vacuum or gas-filled diode valves, transistors which have two *P-N* junctions are similar to triode valves. To-day almost any circuit, however complex, may well be constructed with crystal diodes and transistors alone, without any vacuum or gas-filled valves. Among the advantages of such circuits are small size, no need for filament or heater supplies, and a long service life.

Most often, transistors are fabricated from germanium and silicon. This is because charge carriers in these semiconductors have the highest mobility of all. No less important, germanium and silicon are strong mechanically, show good chemical stability and a relatively low rate of recombination—the carriers of opposite sign have enough time to travel distances of the order of 0.01 to 0.1 mm before they recombine.

2. A major limitation of transistors is that they can normally operate within a relatively narrow temperature range. For germanium, the intrinsic temperature is about 100 °C. Near this temperature, the concentration of free carriers drastically increases, and it becomes difficult to keep their number just sufficient for normal operation. This is why the upper temperature limit for germanium transistors does not exceed 55° to 75 °C. At low temperature, the energy of thermal motion is insufficient to liberate the requisite number of carriers in the bulk of the material. As a result, the resistance of the device is increased, and its operation is upset.

### 78.5. PHOTORESISTORS AND PHOTODIODES

1. Carriers in a semiconductor may be liberated due to the absorption of light or irradiation with fast electrons, alpha or any other particles. If the energy of the absorbed photon is greater than the activation energy of intrinsic or extrinsic conduction, an electron (or a hole) will pass into the conduction band of the semiconductor and contribute to its conductivity. The conductivity resulting from the generation of such electrons and holes is called *photoconductivity*.

Not all of the carriers liberated by radiation contribute to conduction. Some of them join the impurity atoms and take over the sites left by other carriers. Yet, the net result of illuminating the semiconductor with light of sufficient intensity is an increase in the concentration of free carriers and in electrical conductivity. It has been found that the polarity, mobility and other properties of "light" carriers are usually the same as those of ordinary "dark" carriers. Among other things, this has been borne out by the measurements of the Hall constant (see Sec. 44.2).

For many semiconductors, the energy of a photon of visible light is sufficient to drive carriers into the conduction band where they contribute to the electrical conductivity of the material. For some semiconductors, an increase in electrical conductivity can be obtained only if they are illuminated with light at very low frequencies, that is, one in the far infrared region of the spectrum. It has been shown (see Sec. 67.1) that this radiation is emitted by hot bodies. Thus, the presence of hot bodies may be detected at great distances by noting the effect of their radiation: there will be an increase in electrical conductivity in a circuit containing a light-sensitive semiconductor. This increase in current may be boosted by amplifiers to a value sufficient to generate a warning that a hot radiating body has been detected.

2. Photoconduction, that is, the increase in electrical conduction brought about by light, results in a rapid decrease in resistance, for, as we have learned (see Sec. 68.2), the photoeffect is practically free from time lag.

The photoconductive or inner photoelectric effect is the sole basis of operation for the class of *photoconductive cells* or *photoresistors*. The illumination of a semiconductor with light at a sufficiently high frequency usually produces the outer photoelectric or photoemissive effect—the liberation of electrons from their bonds to the lattice.

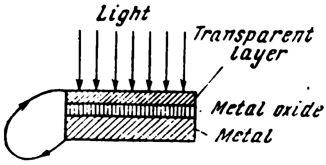


Fig. 78.5

An elementary photoresistor is a glass plate to which a thin layer of semiconductor is applied and electrodes are made to the deposit. The entire assembly is then given a coat of transparent varnish. Photoresistors typically show a non-linear relationship between the photocurrent and the incident radiant flux at high values of illumination. A major limitation of photoresistors is that their properties are highly sensitive to temperature.

Photoresistors intended for operation in the visible region of the spectrum are made of cadmium sulphide and thallium sulphide, and those for use in the infrared are made of lead selenide and lead telluride.

3. Photoresistors are used in sound motion pictures, as alarms, in television, automatic and remote control. In automatic or remote control, any disturbance in the normal run of a process will change the radiant flux incident on a photocell, and this will generate a signal to warn the operator or to stop the process.

Photoresistors are used to sort mass-produced articles according to size and colour. The articles to be sorted are moved by a belt conveyor past a photocell so that they intercept the light beam reaching the detector. The colour or size of the object vary the amount of radiation reaching the detector and, as a consequence the resultant photocurrent. It may be arranged so that the articles will be dumped into one box or another according to the current they produce.

4. A practically important example of utilizing photoconductivity is offered by photovoltaic or barrier-layer cells. In sketch form, a photovoltaic cell is shown in Fig. 78.5. It is made up of a metal and its oxide (semiconductor) brought in contact and given a transparent coat of metal. The junction between the metal and the semiconductor acts as a rectifier, that is, it allows electrons to pass only from the oxide to the metal (from, say, cuprous oxide to copper). Light

incident on the photovoltaic cell maintains the stream of electrons from the semiconductor to the metal, and no applied voltage is required to control the stream. A photovoltaic cell is a direct converter of light into electric energy.

A photovoltaic cell faithfully follows all changes in the intensity and spectral make-up of the light source, provided light acts on the entire bulk of the semiconductor material. Ordinarily, photovoltaic devices show some time lag, and the photocurrent takes some time to reach a maximum after the light source is turned on, and to drop to its "dark" value when the source is turned off.

5. There are purely semiconductor photosensitive devices. They offer a number of advantages over vacuum photocells (mechanical strength, freedom from noise, high sensitivity to various regions of the spectrum). Among them are *photodiodes* and *phototransistors.* A photodiode is a *P-N* junction device, and a phototransistor is a *P-N-P* (or *N-P-N*) device. In both, light controls the number of free carriers generated and reduces the resistance of the *P-N* junction, thereby bringing about an increase in the current flowing through the junction due to a potential difference.

# Chapter 79

# SOME OPTICAL PROPERTIES OF SOLIDS

## 79.1. RAMAN SCATTERING

1. In 1927, Academicians L. I. Mandelshtam and G. S. Landsberg of the Soviet Union investigated the spectral make-up of the light scattered by quartz crystals. They noticed that, in addition to the frequency $\nu_0$ equal to that of the incident light, the scattered radiation had a number of other frequencies above or below $\nu_0$. They called it the *combination scattering of light.* At about the same time, the phenomenon was discovered by Raman and Krishnan of India, who investigated the scattering of light by the molecules of transparent gases, liquids and solids. Raman published his report in the magazine *Nature* before the Soviet scientists, and so the phenomenon has come to be known as the *Raman effect* or *Raman scattering* and the associated spectra as the *Raman spectra.*

2. As already noted, a Raman spectrum has two extra groups of spectral lines. The lines at frequencies $\nu_0—\nu_1$, $\nu_0—\nu_2$, etc., that is, lower than the frequency of the incident light, are called *red companions*, for the reason that they are shifted in the direction of the red region of the electromagnetic spectrum. The lines having frequencies $\nu_0 + \nu_1$, $\nu_0 + \nu_2$, etc., that is, higher than the frequency

of the incident light, are called *violet companions* because they are shifted toward the violet region of the spectrum.

It has been found that these extra frequencies are independent of the frequency of the incident light and are characteristic of the scattering substance. Observations have shown that the violet companions have a lower intensity than the red companions. Besides, a rise in temperature has been found to raise the intensity of the violet companions while nothing has been noted in the intensity of the red companions. Their intensity is practically independent of temperature.

The quantum physics of atoms and molecules and the quantum theory of light have offered a simple explanation of Raman scattering.*

3. Let the photon incident on a substance have an energy $h\nu_0$, where $\nu_0$ is the frequency of the incident light. As will be recalled (see Sec. 74.4), the molecules of a substance can occupy different vibrationary energy levels. Let the energies of these levels respectively be $\mathscr{E}_1$, $\mathscr{E}_2$, etc., such that $\mathscr{E}_1 < \mathscr{E}_2$, etc. The level of energy $\mathscr{E}_1$ is the ground, or unexcited, level, and $\Delta\mathscr{E} = \mathscr{E}_2 - \mathscr{E}_1$ is the difference in energy between two adjacent levels. The incident photon can impart a molecule in the state $\mathscr{E}_1$ an energy $\Delta\mathscr{E}$ necessary for the molecule to jump to an excited level of energy $\mathscr{E}_2$. In the process, the energy of the photon will decrease by $\Delta\mathscr{E}$ to become equal to:

$$h\nu = h\nu_0 - \Delta\mathscr{E} \tag{79.1}$$

A frequency $\nu$ equal to

$$\nu = \nu_0 - \Delta\mathscr{E}/h = \nu_0 - \nu_1, \quad \text{where} \quad \nu_1 = \Delta\mathscr{E}/h \tag{79.2}$$

will appear in the scattered light. This is the frequency of a red companion.

A molecule may be excited to higher levels with energies $\mathscr{E}_3$, $\mathscr{E}_4$, and so on. The requisite energy can likewise be supplied by the incident light, with the result that further red companions will appear.

4. To understand why violet companions appear, it should be recalled that the molecules of a substance may for some time be in excited states (see Sec. 72.8). Let a molecule be in an excited state with energy $\mathscr{E}_2$, $\mathscr{E}_3$, or any other. A quantum of energy $h\nu_0$ may cause it to jump to the ground, unexcited vibrational state with energy $\mathscr{E}_1$. As a result, the energy of the photon will increase by $\Delta\mathscr{E} = \mathscr{E}_2 - \mathscr{E}_1$, to become equal to

$$h\nu = h\nu_0 + \Delta\mathscr{E} \tag{79.3}$$

* There is also a classical explanation of the Raman effect.

and a frequency corresponding to one of the violet companions will appear in the scattered light:

$$\nu = \nu_0 + \Delta\mathscr{E}/h \tag{79.4}$$

Transitions from the more excited states with energies $\mathscr{E}_3$, $\mathscr{E}_4$ and so on, to the ground state will be accompanied by the appearance of other violet companions.

It should be stressed that the shift in the frequency $\nu_0$ of the incident light towards the red or violet region is determined by the frequencies of the transitions between the vibrational energy levels of the molecules. Because of this, the frequency shifts displayed by the extra spectrum lines in the Raman spectra coincide with the frequencies of molecular vibrations in the infrared region.

5. The number of molecules occupying the ground vibrational energy state, $\mathscr{E}_1$, is always greater than that of molecules in the excited states. This is why the scattering of a photon accompanied by an upward transition of molecules to excited states is more probable than the scattering by excited molecules, accompanied by a downward transition. In other words, the intensity of the red companions must be greater than that of the violet companions. With heating, the number of molecules in the excited states increases, and so does the number of downward transitions to the ground state. Thus, an increase in temperature entails an increase in the intensity of the violet companions. On contrast, heating can affect the number of molecules in the ground vibrational energy state only slightly, and so the intensity of the red companions remains practically unaffected by rising temperature.

6. Raman scattering is widely used in the study of natural frequencies of vibrations in polyatomic molecules and, through them, molecular structure and behaviour. It is also useful in the quantitative analysis of organic compounds and their mixtures.

## 79.2. LUMINESCENCE

1. In Chapter 67 we have discussed the radiation given up by hot bodies, or thermal radiation. There is also the emission of light due to other cause than high temperature, the so-called *luminescence* (from the Latin *luminis* for light). According to the action producing it, we have cathodoluminescence, electroluminescence, photoluminescence, chemiluminescence, and some other types of luminescence. In the case of cathodoluminescence, by bombarding a metal with electrons or other charged particles, small amounts of the metal are vaporized in an excited state and emit radiation. Electroluminescence is luminescence excited by electric fields or currents. Photoluminescence is the emission of light as a result of irradiation with

visible light, X-rays or gamma-rays (see Sec. 81.10). Chemiluminescence is luminescence produced by chemical action.

2. In all of these cases, the emission of light is due to the stimulation of the substance to an excited state. This stimulation or excitation involves the input of energy from a source of suitable radiation. In solids, centres of luminescence are atoms, ions or groups of ions located near a lattice imperfection. Such imperfections in the lattice may be produced by introducing an activator, that is an impurity atom, or by producing a vacancy. The transition of an excited centre of luminescence to the ground state is accompanied by the emission of light. This is luminescence.

The time during which luminescence can be observed depends on the time interval between the acts of excitation and emission.

3. As a rule, the life time of an excited state is about $10^{-8}$ s (see Sec. 72.8). When the light is emitted only as long as the exciting emission is maintained, this is called *fluorescence* (from the Latin *fluor* for flow). For fluorescence, the time interval between the acts of excitation and emission is $10^{-8}$ to $10^{-9}$ s. It is observed when some liquids or gases are illuminated with light. For example, when illuminated by day light, kerosene emits a faint bluish fluorescence. Ultraviolet rays emitted by a mercury-vapour lamp cause solutions of some dyes to give up visible light.

Sometimes the time interval between the acts of excitation and emission may be as long as $10^{-4}$ s to a few minutes. Accordingly, luminescence lasts longer. Luminescence which is delayed for a long time (about $10^{-8}$ s) after excitation is called *phosphorescence* (from Greek *phos* for light and *phoros* for carrying). Phosphorescence is displayed by some solids, for example, crystalline zinc sulphide. A coat of zinc sulphide applied to a sheet of cardboard makes a fluorescent screen which will emit a faint glow for minutes after the exciting emission is removed.

It is relevant to note that the delineation of luminescence into fluorescence and phosphorescence is an arbitrary one, because no exact time line can be drawn between them.

4. As a rule, photoluminescence is excited by ultraviolet rays or their neighbours in the spectrum. Stokes found that *when luminescence is excited by radiation, the frequency of the luminescence is usually less than that of the incident radiation.* Or in terms of wavelengths, the wavelength of luminescence excited by radiation is always greater than that of the exciting radiation. This is known as the *Stokes law*. Quantum optics offers a simple explanation of the Stokes law by the law of conservation of energy. The point is that if the light incident on a substance has a frequency $\nu$, the energy of a photon is $h\nu$. Each photon spends some of its energy to excite luminescence. The remainder is taken up by various non-optical processes. If $\nu_{lum}$ is the frequency of the luminescence emitted by the substance,

then by the law of conservation of energy

$$h\nu = h\nu_{lum} + \Delta\mathscr{E} \quad \text{or} \quad \nu = \nu_{lum} + (\Delta\mathscr{E}/h) \qquad (79.5)$$

where $\Delta\mathscr{E}$ is the part of the photon's energy taken up by non-optical processes. As a rule, $\Delta\mathscr{E} > 0$ and $\nu > \nu_{lum}$, that is, $\lambda_{lum} > \lambda$. This is the Stokes law.

Sometimes the Stokes law is violated, and the luminescence has a shorter wavelength than that of the exciting radiation (*anti-Stokes lines*). This happens when the energy associated with the thermal motion of particles is converted to radiation. Then an amount of energy, $\Delta\mathscr{E}_1$, due to the internal energy of the substance is added to the energy of a photon, $h\nu$, of exciting radiation:

$$h\nu_{lum} = h\nu + \Delta\mathscr{E}_1,$$

and

$$\lambda_{lum} < \lambda$$

5. For practical applications of luminescence, it is important to know the proportion of exciting energy that is converted to the luminescence of a substance. In the Soviet Union, this problem was investigated by Academician S. I. Vavilov. The ratio of the energy emitted by photoluminescence to that absorbed during the act of excitation is called the *photoluminescence efficiency*. According to Vavilov, the photoluminescence efficiency, $D$, is a function of the wavelength of exciting radiation (Fig. 79.1).

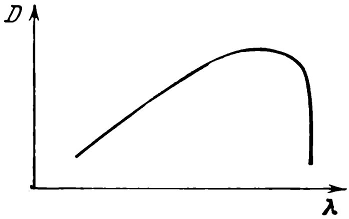


Fig. 79.1

The Vavilov law can be explained by introducing the concept of the *quantum efficiency of photoluminescence*, which is defined as the ratio of the number of photons of luminescence to the number of quanta absorbed. As the wavelength of the exciting radiation increases, that is, as its frequency decreases, the number of photons absorbed increases. On the other hand, each photon absorbed may give rise to a quantum $h\nu_{lum}$ emitted. Therefore, the quantum efficiency of luminescence increases with increasing wavelength. At some wavelength, the quantum efficiency shows a sudden slump. This is because photons of energy $h\nu$ cannot excite the centres of luminescence any longer at that wavelength.

6. Luminescence is at the basis of qualitative and quantitative *fluorescence analysis*. In this type of analysis, even the minutest quantities of impurities (down to $10^{-11}$ g) may be detected in the substance under investigation from the intensity of their fluores-

cence lines. Fluorescence analysis is widely used in industry, medicine, biology, and some other fields.

7. Luminescence is utilized in *fluorescent lamps*. The light efficiency of conventional incandescent lamp is very low, being of the order of 12 to 20 lumens per watt; only a few per cent of the energy input appears as visible light. In contrast, fluorescent lamps are more efficient, require no filaments to heat, and emit in a very narrow region of the spectrum. For example, sodium vapour lamps where luminescence is excited by an electric discharge (see Sec. 48.6) have an efficiency of about 60 lumens per watt and emit yellow light close to the 598-nm yellow line of sodium, which corresponds to the maximum sensitivity of the human eye.

In mercury-vapour lamps, available in a great number of designs, the internal surfaces of the envelopes are coated with *phosphors*, that is, substances that exhibit luminescence. In these lamps, the mercury vapour emits ultra violet rays which are absorbed by the phosphor, and this emits luminescence. The spectrum make-up of this light is very close to day light. Mercury-vapour and fluorescent lamps are widely used for street lighting, at factories, and in the homes.

## 79.3. NEGATIVE ABSORPTION OF LIGHT

1. The mid-50s saw rapid advances in quantum electronics. In 1954, Academicians N. G. Basov and A. M. Prokhorov of the USSR published articles describing a microwave quantum-mechanical oscillator, now known as the *maser* (for Microwave Amplification by Stimulated Emission of Radiation). In 1960, optical masers were developed, more commonly known as lasers (for Light Amplification by Stimulated Emission of Radiation). Both masers and lasers depend for their operation on the induced or stimulated emission of radiation mentioned in Sec. 72.9. It has been stated that this effect is due to the interaction between electromagnetic waves and the atoms of the substance through which the waves pass. Since the behaviour of atoms is described in terms of quantum laws, the names of both types of devices contain the word "quantum", that is, a quantum-mechanical oscillator or a quantum-mechanical amplifier.

2. In Sec. 72.9 we have learned that induced (or stimulated) emission of radiation may result in the negative absorption of light. Since this is the basis of lasers, a more detailed examination is in order.

A medium in which the passing light beam gains in intensity is called *active*. The existence of such media was postulated by Einstein in his explanation of stimulated emission of radiation. For a better understanding of the material presented in the sections that follow, it is important to dwell on the properties of induced radiation in greater detail.

According to Einstein, induced radiation must be precisely identical in properties with the exciting radiation. Every new photon appearing because an atom (or a molecule) of the substance jumps from a higher to a lower energy state under the action of light has the same energy and moves in the same direction as the exciting photon. In terms of wave theory, the effect of the induced emission of radiation consists in an increase in the amplitude of the transmitted wave without any changes in its frequency, direction of propagation, phase and polarization. In other words, the *induced radiation is coherent with the exciting radiation.*

3. Because of stimulated emission, the light passing through the active medium is amplified. It should be remembered, however, that in addition to the stimulated emission of radiation there is also the absorption of light. By absorbing a photon, an atom in a state with energy $\mathscr{E}_1$ will move to a higher energy level with energy $\mathscr{E}_2$ (Fig. 79.2*a*), and the intensity of the light beam passing through the medium will be reduced. Thus, two competing processes are taking place at the same time. Through acts of stimulated emission, a photon of energy $h\nu$ "dumps" an atom from level $\mathscr{E}_2$ to level $\mathscr{E}_1$ and now two photons instead of one move on (Fig. 79.2*b*). In contrast, the acts of absorption reduce the number of photons passing through the medium. The net effect of the medium will depend on which of the two processes is predominant. If absorption of photons predominates, the medium will be attenuating, and not amplifying for the light passing through it. If, on the other hand, acts of stimulated emission are predominant, this will be a light amplifying medium.

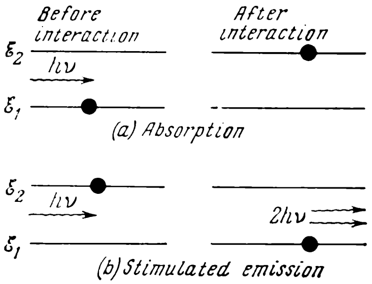


Fig. 79.2

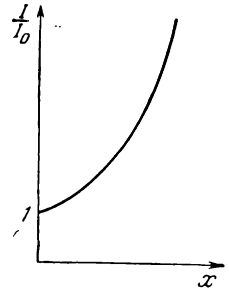


**Fig.** 79.3

4. The absorption of light by a substance obeys the Bouguer law (see Sec. 55.4):

$$I = I_0 \exp(-\alpha x) \tag{79.6}$$

where $\alpha$ is the positive coefficient of absorption, $x$ is the thickness of the absorbing layer, $I_0$ is the intensity of the incident light, and $I$ is the intensity of the transmitted light.

For the first time, the behaviour of a medium showing the negative absorption of light were investigated by V. A. Fabrikant of the Soviet Union (see Sec. 72.9). According to him, for such a medium Eq. (79.6) has a similar form, but the absorption coefficient is a *negative* quantity, so that instead of attenuation, the light beam passing through it experiences amplification. At $\alpha < 0$, Eq. (79.6) points to a steep increase in light intensity with increasing thickness of the amplifying medium (Fig. 79.3). That is, in such a medium the number of photons increases in an avalanche-like manner because acts of stimulated emission predominate. Two photons produced by an act of stimulated emission (Fig. 79.2*b*) collide with atoms in an excited state, so that after the atoms are forced to a lower state, four photons continue their travel. These events repeat themselves over and over again (Fig. 79.4). From the view-point of wave theory, the amplitude of an electromagnetic wave and its square which is proportional to light intensity will increase due to the energy contributed by excited atoms.

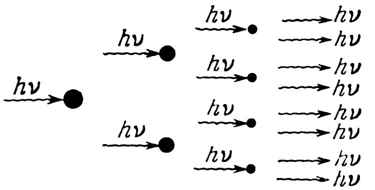


Fig. 79.4

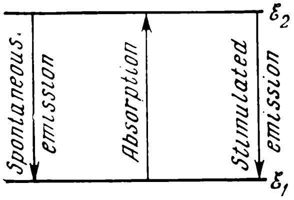


Fig. 79.5

5. Let us determine the absorption coefficient, $\alpha$, of a medium without presuming that it must be an amplifying one.

Under conditions where spontaneous emission of radiation is negligible, the absorption coefficient $\alpha$ should be determined by two conflicting processes, absorption and stimulated emission.

Consider two energy levels, *1* and *2*, that the atoms (molecules) of the medium can occupy and between which, according to Einstein, three types of optical processes can take place, namely spontaneous emission, absorption and stimulated emission (Fig. 79.5). For simplicity we shall assume that spontaneous emission accompanied by the spontaneous transitions of atoms to the normal state may be neglected. Later we shall learn when this is possible. The number of acts of photon absorption is proportional to the number $N_1$ per unit volume of atoms in the excited state with energy $\mathscr{E}_1$. The number of acts of stimulated emission is proportional to the number $N_2$ per unit volume of atoms in the upper energy level with energy $\mathscr{E}_2$. It is rigorously proved that the proportionality factor is the same in both cases. The absorption coefficient $\alpha$ in Eq. (79.6) is proportional to the difference in number between the acts of absorption and stimulated emission:

$$\alpha = k (N_1 - N_2) \tag{79.7}$$

where $k > 0$ is a proportionality factor.

### 79.4. LASERS

1. Under ordinary conditions, that is, at thermal equilibrium, the number $N_2$ of atoms in excited level *2* is smaller than the number $N_1$ of atoms at a lower level, *1*, that is, $N_2/N_1 < 1$. Therefore, at equilibrium $\alpha > 0$. In other words, the acts of ordinary (positive) absorption exceed in number the transitions accompanied by negative absorption, that is, stimulated emission. However, from Eq. (79.7) it follows that there may be media for which $\alpha$ is negative ($\alpha < 0$). To obtain a medium having negative absorption, it is necessary to bring the system out of balance, that is, to arrange so that the number $N_2$ of atoms in the excited level will be greater than the number $N_1$ of atoms in the normal state, that is, $N_2/N_1 > 1$. This is the *inversion of the normal population distribution* or simply *population inversion*. Here, the inversion means that the upper energy level has a greater concentration of atoms than the lower one has.

2. Population inversion is obtained by what is called *pumping*. The most natural way of doing this is to use optical pumping. In simple terms, the active material is illuminated with light at a frequency $\nu$ such that $h\nu = \mathscr{E}_2 - \mathscr{E}_1$, with the result that the atoms in the lower energy state absorb photons and jump to the higher energy state. If the active material is a gas, population inversion can be obtained as a result of inelastic collisions between atoms and electrons in a gas discharge (electrical pumping). However, with a two-level scheme, these methods of pumping fail to produce the desired population inversion, because the excited atoms residing in the excited state for a very short time interval (see Sec. 72.8) lose their energy through spontaneous radiation and through collisions with electrons and drop to the lower level.

According to Fabrikant (see Sec. 72.9), this difficulty can be circumvented by selectively destroying some of the lower energy levels with the use of suitable molecular impurities so that the upper energy levels will have a greater population. In the maser developed by N. G. Basov and A. M. Prokhorov and, independently of them, by

Townes, the upper state molecules were separated from the lower state ones by an electrostatic field. Still another method is to obtain population inversion with an auxiliary radiation and a three-level scheme.

3. The three-level scheme was for the first time proposed by N. G. Basov and A. M. Prokhorov in 1955. One of the first operating lasers using the three-level scheme of population inversion and a solid-state laser element was developed in 1960. The laser rod was a ruby crystal which chemically is aluminium oxide, $Al_2O_3$, with an addition of 0.03 to 0.05% chromium oxide, $Cr_2O_3$, in which ions $Cr^{3+}$ replace some of the Al atoms in the crystal lattice. In such a lattice, induced transitions occur in the chromium ions, $Cr^{3+}$. As is seen from the energy level diagram (Fig. 79.6), a $Cr^{3+}$ ion has two wide energy bands *A* very close to the ground level *C*, and also a double level *B*; transitions from the latter to the ground level correspond to the emission of red light, at wavelengths of 6927 Å and 6943 Å.

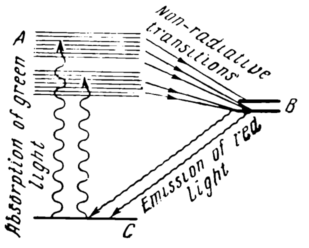


Fig. 79.6

When the ruby is illuminated by a strong green light from a flash tube filled with a mixture of neon and krypton (the pumping lamp or tube), chromium ions move to the levels within the wide band *A* from which it is most probable that a non-radiative transition of ions to the double level *B* will take place, with the excess energy being transferred to the ruby lattice. In this way, the double level *B* may be brought up to a state in which its ion population will exceed that of the ground level *C*. In other words, the populations of these two levels will be inverted. Such a device can emit at 6927 A and 6943 Å. Population inversion is promoted still more by the low probability of spontaneous transitions for chromium ions from level *B* to level *C*.

In one of the gas lasers, the lasing medium is the plasma of an r.f. gas discharge produced in a mixture of helium and neon. Due to collisions with electrons, the helium atoms are excited to an upper level with energy $\mathscr{E}_3$. As the excited helium atoms collide with neon atoms, the latter are likewise excited to one of the upper levels for neon, located close to the respective level for helium. A transition of neon atoms from this level to one of the lower levels with energy $\mathscr{E}_2$ is accompanied by the emission of radiation. A simplified diagram of a three-level laser is shown in Fig. 79.7.

The cumulative increase in light intensity in the active element signifies that this element acts as an amplifier for electromagnetic waves. The principle of this amplification was formulated by V. A. Fabrikant, M. M. Vudynsky and F. A. Butayeva in 1951.

4. The amplification of light by stimulated emission of radiation may be further augmented by allowing the light being amplified to traverse the same active element many times before the beam is coupled out. Until that moment, the light beam is contained within the laser optical cavity formed by two plane or concave mirrors between which the laser rod or cell is placed. In schematic form, a laser is shown in Fig. 79.8. Any photon given up spontaneously by optically pumped atoms in the laser rod may serve as a "starter" for the lasing action.

Consider a photon moving parallel to the axis of the gas cell or the ruby crystal. It gives rise to an avalanche of photons moving all in the same direction (Fig. 79.9*a*). Some of this avalanche will pass outside through semi-transparent mirror *3* and part will be reflected to be built up in the active element, *1* (Fig. 79.9*b*). On reaching the 100% reflecting mirror, *2*, some of the photons will be absorbed, but the greater proportion will be reflected, and the amplified beam will traverse the same path through the laser rod as the "starting" photon (Fig. 79.9*c*). Thus, the mirrors provide a feedback mechanism essential to sustain oscillations (see Sec. 52.1). Multiply reflected, amplified and coupled out of the laser cavity through the semi-transparent mirror, the laser beam is highly collimated and intense.

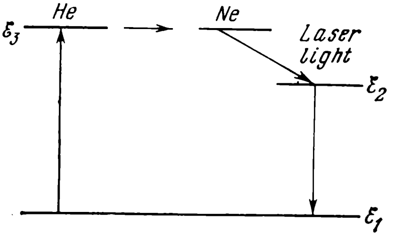


Fig. 79.7

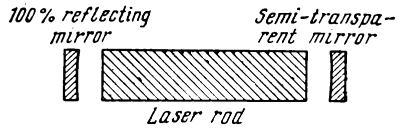


Fig. 79.8

5. For the proper build-up of oscillations in a laser, it is essential that the amplification between two consecutive reflections of the beam from mirror *2* can at least make good the losses that occur due to reflections from the mirrors. A measure of amplification for light in a laser over a path length $L$ is given by

$$K = I/I_0 = \exp(\beta L)^*$$

where $L$ is the optical path length between the two mirrors. The distance travelled by photons between two consecutive reflections is $2L$, and so the amplification is given by $\exp(2\beta L)$. To allow for the losses of photons at the mirrors, we designate the reflection coef-

* In Eq. (79.6), we set $\beta = -\alpha$.

ficients of mirrors *2* and *3* (Fig. 79.9) as $r_2$ and $r_3$. The total loss of photons reflected consecutively from both mirrors is proportional to the product $r_2r_3$. With allowance for the losses at the mirrors, the gain of a laser may be written as

$$K' = I/I_0 = r_2r_3 \exp(2\beta L) \qquad (79.8)$$

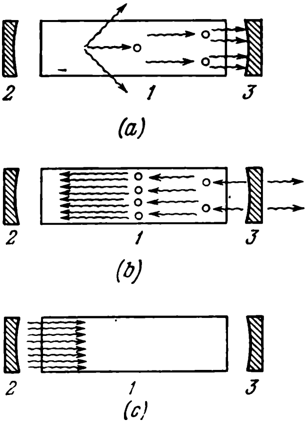
From Eq. (79.8) it is an easy matter to derive the condition for the losses at the mirrors to be made good by the amplification in the laser rod, and so $I = I_0$, that is, $K' = 1$:

$$r_2r_3 \exp(2\beta L) = 1 \qquad (79.9)$$

Taking a logarithm of Eq. (79.9), we find the negative absorption coefficient $\beta$ of the laser

$$\beta = -\ln(r_2r_3)/2L \qquad (79.10)$$

Eq. (79.10) is used to determine the threshold value of pumping energy necessary for lasing action. It is obvious that if the pumping energy is raised to a point where the losses at the mirrors are made good, the avalanche of photons will build up, and the laser output beam will gain in brightness. However, light amplification in a laser cannot increase without bound. As it increases, there is a companion increase in spontaneous emission of radiation by the atoms in the upper energy levels. As a result, the population inversion is reduced, the number of induced transitions decreases, the amplification goes down, and the accumulation of photons slows down. This is known as the *saturation of laser gain.*



(a)

(b)

(c)

Fig. 79.9

6. So far we have ignored the fact that the laser radiation is coherent with the exciting one. The wave properties of light lead therefore to some additional conditions important for a laser to sustain oscillations. In terms of wave theory, light amplification by a laser implies a continuous and marked increase in the amplitude of the light wave. For this to occur, however, it is important that the wave returning to some point in the active element have the same phase as that of the original wave with any number of reflections from the mirrors. This imposes a certain constraint on the relationship between the wavelength $\lambda$ and the length $L$ of the laser rod. The optical path length travelled by a wave between two consecutive reflections should be a multiple of the wavelength:

$$2L = n\lambda \quad \text{or} \quad L = n\lambda/2 \quad n = 1, 2, \ldots \qquad (79.11)$$

Then the primary and all secondary waves will interfere constructively, and the amplitude of the resultant wave will increase appreciably. If condition (79.11) is satisfied, the waves coupled out of the laser through mirror *3* (Fig. 79.9) with each reflection, will be coherent with one another. The difference in phase between two consecutive emerging waves is $\Delta\varphi = 2\pi\ (2L/\lambda)$ and is determined by the optical path length difference $2L$ (see Sec. 57.5). The light coupled out of the laser is produced by the constructive interference of many coherent waves differing in phase by some multiple of $2\pi$. This is why the laser beam has a very high resultant amplitude and an extremely high intensity. As will be recalled (see Secs. 57.6 and 61.7), the constructive interference of many coherent waves produces very narrow and sharply defined intensity maxima. If condition (79.11) is not satisfied, the waves will no longer be coherent, and no constructive interference will bo possible.

7. Eq. (79.11) expresses the conditions for phase relationships which is as important to the lasing action as that for loss compensation, Eq. (79.9). From Eq. (79.11) it follows that if the space bounded by the two end mirrors in a laser is treated as a mirror cavity resonator, then the length $L$ of the resonator should accommodate a whole number $n$ of standing half-waves (see Sec. 57.2). Thus, Eq. (79.11) expresses at the same time the condition of resonance between the electromagnetic waves and the mirror cavity.

The frequencies generated by a laser can be found from Eq. (79.11). Noting that $\lambda = c/\nu$ and substituting it in (79.11), we get

$$\nu_n = nc/2L \tag{79.12}$$

For each value of $n$ there is a certain definite frequency, $\nu_n$. Besides, the frequencies generated by a laser should satisfy Bohr's frequency rule (see Sec. 71.4) which connects frequency to the difference in energy between the atoms of the laser rod. At first sight it may appear that the need to satisfy both Eq. (79.12) and Bohr's rule complicates the construction of a practical laser out of all proportions. Indeed, the distance $L$ must be maintained to an extremely high degree of accuracy, or else the resultant waves will not be coherent. Actually, however, the situation is not so hopeless. For one thing, Bohr's frequency rule may be satisfied accurate to the finite width of the atomic energy level (see Sec. 72.8). For another, a spreading of radiation frequencies inevitably occurs due to several causes, above all due to the Doppler effect.

8. A detailed discussion of the width of the spectral lines emitted by a laser is outside the scope of this book. It may be shown that the laser spectrum consists of several very narrow lines whose frequencies, as is seen from Eq. (79.12), spaced $\Delta\nu = c/2L$ apart. For a gas laser with $L = 10^2$ cm, $\Delta\nu$ is 150 MHz. It is remarkable that the spectral line width for a gas laser is only a fraction of the intrinsic line width due to the finite lifetime of excited atoms (see Sec. 72.8).

With the advent of lasers, much headway has been made towards the generation of truly monochromatic light. The high monochromaticity of laser light implies that the duration of the continuous wave trains emitted by a laser is by several orders of magnitude greater than that obtainable with conventional optical means. Hence, the length of a continuous wave train from a laser is many times that of a wave train in conventional optics. This removes a very important limitation usually imposed in optics on interference experiments, namely the requirement that the optical path length difference be kept to a minimum. With a laser beam, interference experiments may be carried out over distances from tens of metres upwards.

9. A remarkable feature of the laser beam is its negligible angular divergence which stems from the very mechanism of stimulated emission. The point is that the "starting" photon necessary for the lasing action to begin must travel parallel to the rod axis. Any photon travelling at an angle to the rod axis will build up an avalanche of photons but this will disappear from the rod between the mirrors rapidly without contributing to light amplification (Fig. 79.9*a*). Thus, it is only the photons travelling parallel to the rod axis that contribute to both oscillation and amplification. This accounts for the extreme parallelism of the output laser beam. Yet, there is some angular spread left in the laser beam because of the wave properties of light. The lowest angular divergence, $\theta_{\min}$, for the laser beam is set by the diffraction of light. The angular divergence of a beam cannot be smaller than the diffraction angle on a circular screen of diameter $D$:

$$\theta_{\min} \geqslant \lambda/D \tag{79.13}$$

where $D$ is the diameter of the mirror in a laser. $\theta_{\min}$ is of the order of $10^{-5}$ to $10^{-6}$ radians. For example, it applies to gas lasers.

10. Owing to high coherence and negligible angular divergence, laser beams can advantageously be used in communications, ranging and detection of targets, generation of high heat within a limited volume, and the like. With a bandwidth of 1 Å and a wavelength of 1 μm, as many as 10 000 radio programs can be transmitted over the same laser beam, and do so over distances of astronomical magnitude. A laser beam can pierce a minute hole in the hardest of all materials, such as diamond, weld together microscopic components, or heal exfoliation of the retina in the human eye.

The capabilities and performance of existing lasers are still far away from what they can do ultimately. For example, a laser can, at least in principle, build up a pressure equivalent to millions of atmospheres. The prospects for advances in laser technology are unlimited, indeed.

# PART EIGHT THE BASIC PHYSICS OF THE NUCLEUS AND ELEMENTARY PARTICLES

## Chapter 80

## MAIN PROPERTIES AND STRUCTURE OF THE ATOMIC NUCLEUS

### 80.1. CHARGE AND MASS OF ATOMIC NUCLEI

1. In Sec. 71.1 we have discussed the experiments that had led to the formulation of the theory by which the nuclear atom was accepted. In the closing chapters of this book we shall dwell on the structure of the atomic nucleus and examine some of the problems facing the physics of fundamental particles.

The most important characteristics of a nucleus are its charge $Z$ and its mass $M$. The charge on the atomic nucleus is determined by the number of positive charges it contains. The carrier of an elementary charge, $e = 1.6021 \times 10^{-19}$ C, on the nucleus is a proton. Since an atom as a whole is electrically neutral, the nuclear charge simultaneously determines the number of electrons around the nucleus. The arrangement of electrons in the shells and sub-shells of an atom depends to a great extent on their number in the atom (Sec. 73.2). This is why the charge on the nucleus determines the energy distribution of the electrons and the position of each element in the Periodic Table. In other words, chemical elements are identified by their nuclear charge or, which is the same, by their *atomic numbers*.

2. The mass of an atomic nucleus is practically the same as that of the entire atom because the mass of the electrons in an atom is negligible. As will be recalled, the mass of an electron is 1/1836th that of a proton. It is customary to measure the mass of an atom in physical *atomic mass units*, abbreviated *amu*. The physical amu is defined as one-sixteenth of the mass of the oxygen atom of mass number 16, symbolized ${}_8O^{16}$ (the symbol will be explained later in this section), or

$$1 \text{ amu} = 1.65976 \times 10^{-27} \text{ kg}$$

Apart from the amu, use is made of the unified amu symbolized $u$, equal to one-twelfth of the mass of the neutral ${}_6C^{12}$ atom, the carbon isotope (see Sec. 26.9):

$$1\ \mathrm{u} = 1.6603 \times 10^{-27}\ \mathrm{kg}$$

The masses of atoms have been measured to a high degree of accuracy (see Sec. 41.8). These measurements have revealed the existence of *isotopes*, two or more nuclides having the same atomic number, hence constituting the same element, but differing in mass. The masses of isotopes when expressed in atomic mass units are always close to integers. These numbers are called *isotopic masses*.

In most cases, each element is a mixture of isotopes and its atomic mass is the average of the masses of all of its isotopes. This is the reason why the comparative (of ten called chemical) atomic masses of such elements are never whole numbers. For example, the atomic mass of boron is 10.82; neon, 20.183; magnesium, 24.32; chlorine, 35.457; iron, 56.85; cobalt, 58.71; nickel, also 58.71; copper, 63.54; zinc, 65.38; germanium, 70.60; and that of krypton 83.80.

With the discovery of isotopes, it was suggested that the nucleus should be made up of particles with an atomic mass close to unity. Since the nucleus of the hydrogen atom, the proton, accurately fits this condition (its atomic mass is 1.008 amu accurate to the third decimal place, and it carries a positive charge), it appeared obvious that the atomic nucleus must include protons.

3. For convenience, use is made of *mass numbers*, the whole numbers closest in value to the atomic masses when those quantities are expressed in atomic mass units. The symbol is $A$. The mass number is commonly written as a superscript after the symbol of the atom, such as ${}_Z\mathrm{X}^A$ or $\mathrm{X}_Z^A$, where X is the symbol of the atom in the Periodic Table, having the nuclear charge (or atomic number) $Z$.

4. Two or more nuclides having the same number of nucleons in their nuclei, and having therefore identical mass numbers and about the same atomic mass are called *nuclear isobars*. They are mostly encountered among heavy nuclei and usually in diades and triades. Examples of long-lived isobaric diades are ${}_{16}\mathrm{S}^{36}$ and ${}_{18}\mathrm{Ar}^{36}$, ${}_{44}\mathrm{Ru}^{104}$ and ${}_{46}\mathrm{Pd}^{104}$. An example of an isobaric triade is ${}_{40}\mathrm{Zr}^{96}$, ${}_{42}\mathrm{Mo}^{96}$ and ${}_{44}\mathrm{Ru}^{96}$.

## 80.2. SPIN AND MAGNETIC MOMENT OF THE NUCLEUS

1. In Sec. 72.6 we have seen that the spin of an electron results in the fine structure of the spectrum. For atoms having one valence electron the relative orientation of the orbital and spin moments of the electron leads to the splitting of all energy levels (except the $s$-level) and as a result, to the splitting of spectral lines. With further improvement in the resolving power of spectroscopic instruments, investigators were able to investigate such lines. In 1928, A. N. Terenin and L. N. Dobretsov found that each of the two $D$-lines of sodium

was in turn a doublet, that is, consisting of two very closely spaced spectral lines. For the $D_2$ line ($\lambda = 5890$ Å) the spacing was found to be 0.021 Å, and for the $D_1$ line ($\lambda = 5896$ Å), 0.023 Å. This is the *hyperfine structure* of the spectrum.

Pauli suggested that the hyperfine structure might be due to an occurrence of angular momentum in the atomic nucleus. The total angular momentum, or nuclear spin, along with nuclear charge and nuclear mass, is the most important characteristic of the nucleus.

The nucleus is made up of protons and neutrons each of which has spin $\hbar/2$. The nuclear spin is the vector sum of the spin angular momenta of all the component particles. A nucleus made up of an even number of nucleons has integral spin (in units of $\hbar$) or zero spin. A nucleus made up of an odd number of nucleons has half-integral spin (in units of $\hbar$). If a nucleus has a spin $I$ (in units of $\hbar$), then the total spin of an atom having one valence electron may be either $I + 1/2$ or $I - 1/2$, because an electron always has half-integral spin (in units of $\hbar$)*. An optical transition of an electron in the sodium atom from the upper to the lower level, split up because of nuclear spin, produces the hyperfine structure of the $D_2$ line. The splitting-up of spectral lines is illustrated on the left of Fig. 80.1, and the hyperfine structure of the $D_2$ line is explained on the right. The spin $I$ can be determined from the observed intensities of the lines in the hyperfine structure. For the sodium atom, it is $3/2\hbar$.

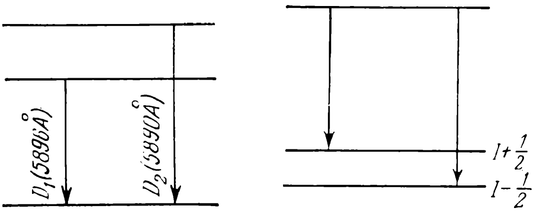


Fig. 80.1

In Sec. 42.10 we have learned that electronic spin was measured in the experiments conducted by Stern and Gerlach from the intensities of the rotational spectra of the hydrogen molecule, $H_2$ (for the origin of rotational spectra see Ch. 74). The relative intensities of spectral lines in the rotational spectra of hydrogen have shown that the hydrogen nucleus, ${}_1H^1$, has half-integral spin. That is, the proton has half-integral spin, $\hbar/2$.

The spin of the neutron is determined from that of the deutron, ${}_1D^2$, the nucleus of heavy hydrogen. In the ground state, the deutron

* The spins of the remaining electrons add together in pairs to give a total spin of zero.

has been found to have integral spin of magnitude $\hbar$. Since the proton has half-integral spin, $\hbar/2$, the neutron might have a spin equal to either $\hbar/2$ or $3\hbar/2$. The latter assumption would lead to a disagreement between theory and observation and has been discarded. Thus, the neutron has half-integral spin, $\hbar/2$.

2. In addition to nuclear spin, the nucleus has a *magnetic moment*. Thus, all atomic particles (the nucleus and electrons) have a magnetic moment.

The magnetic moment of a nucleus is determined by those of its component particles. By analogy with the Bohr magneton (see Sec. 42.2), the magnetic moments of nuclei are expressed in terms of the so-called *nuclear magneton* defined as

$$\mu_N = e\hbar/2m_p \tag{80.1}$$

where $m_p$ is the mass of a proton used instead of the electronic mass in the expression for the Bohr magneton, because of which the nuclear magneton is 1/1836th of the electronic Bohr magneton ($m_p/m_e \approx 1836.5$) : $\mu_N = (5.05038 \pm 0.00018) \times 10^{-27}$ J/T $= (5.05038 \pm 0.00018) \times 10^{-24}$ erg/G.

The relation between the nuclear spin $I$ in units of $\hbar$ and the magnetic moment $p_{mN}$ in nuclear magnetons is expressed by an equation similar to that for electronic moments (Eq. 72.14′):

$$p_{mN} = g_N I \tag{80.2}$$

where $g_N$ is the *nuclear gyromagnetic ratio*.

3. Experiments whose description lies outside the scope of this book have shown that the neutron has a negative magnetic moment of magnitude $-(1.91314 \pm 0.00005)\ \mu_N$.

For the first time, the magnetic moment of the proton was measured from the deflection of a beam of hydrogen molecules in a nonhomogeneous magnetic field. In principle, this technique did not differ from that used by Stern and Gerlach (see Sec. 42.10). A major difficulty in the experiment was the need to make up for the electronic magnetic moment which is nearly 2000 times as great as the nuclear magneton, $\mu_N$. The value found by precise measurement was then surprising, being $(2.79275 \pm 0.00003)\ \mu_N$ instead of one nuclear magneton. The "+" sign of the proton magnetic moment indicates that the magnetic moment and spin of the proton have the same orientation. From a classical point of view this implies that the proton owes its magnetic moment to the rotation of a positive charge. In contrast, the spin and magnetic moment of the neutron are in opposite directions. The anomalous value of magnetic moment for the proton and the negative magnetic moment of the neutron are explained by the internal structure of these particles (see Sec. 83.8).

### 80.3. CONSTITUTION OF THE NUCLEUS

1. When the nuclear model of the atom was advanced the composition of the nucleus became a crucial problem of nuclear physics. An answer to this question could only be given after the discovery of various properties of the nucleus, notably nuclear charge $Z$, nuclear mass, and nuclear spin. The nuclear charge was found to be defined by the sum of the positive charges it contains. Since an elementary positive charge is associated with the proton, the presence of protons in the nucleus appeared to be beyond any doubt from the outset (see Sec. 80.1). Two more facts were also established, namely:

(a) The masses of all isotopes (except ordinary hydrogen), expressed in proton mass units, were found to be numerically greater than their nuclear charges expressed in elementary charge units, this difference growing with increase in $Z$. For the elements in the middle of the Periodic Table the isotopic masses (in amu) are about twice as great as the nuclear charge. The ratio is still greater for the heavier nuclei. Hence one was forced to think that the protons were not the only particles that make up the nucleus.

(b) The masses of the isotopic nuclei of all chemical elements suggested two possibilities, either the particles making up the nucleus had about the same mass, or the nucleus contained particles differing in mass to a point where the mass of some was negligible in comparison with that of the others, that is, their mass did not contribute to the isotopic mass to any considerable degree.

2. The latter possibility appeared especially attractive because it fitted nicely with the proton-electron model of the nucleus.

That the nucleus might contain electrons seemed to follow from the fact that natural beta-decay is accompanied by the emission of electrons (see Sec. 81.12). The proton-electron model also explained the fact why the isotopic atomic weights were nearly integers. According to this model, the mass of the nucleus should be practically equal to the masses of the protons that make it up, because the electronic mass is about 1/2000th that of the proton. The number of electrons in the nucleus must be such that the total charge due to the positive protons and the negative electrons is the true positive charge of the nucleus.

For all its simplicity and logic, the proton-electron model was refuted by advances in nuclear physics. In fact, it ran counter to the most important properties of the nucleus.

3. If the nucleus contained electrons, the nuclear magnetic moment would be of the same order of magnitude as the electronic Bohr magneton (see Sec. 42.2). It has been shown in Sec. 80.2 that the nuclear magnetic moment is defined by the nuclear magneton which is about 1/2000th the electronic magneton.

Data on nuclear spin also witnessed against the proton-electron model. For example, according to this model the beryllium nucleus,

${}_4\mathrm{Be}^9$, would contain nine protons and five electrons so that the total charge would be equal to four elementary positive charges. The proton and the electron have each a half-integral spin, $\hbar/2$. The total spin of the nucleus made up of 14 particles (nine protons and five electrons) would have to be integral. Actually, the beryllium nucleus, ${}_4\mathrm{Be}^9$, has half-integral spin of magnitude $3\hbar/2$. Many more examples might be cited.

Last but not least, the proton-electron model conflicted with the Heisenberg uncertainty principle. If the nucleus contained electrons, then the uncertainty in the electron position, $\Delta x$, would be comparable with the linear dimensions of the nucleus, that is, $10^{-14}$ or $10^{-15}$ m. Let us choose the greater value, $\Delta x = 10^{-14}$ m. From the Heisenberg uncertainty relation for the electron momentum we have

$$\Delta p \approx \hbar/\Delta x \approx 10^{-33} \div 10^{-14} = 10^{-19} \text{ kg m/s}$$

The momentum $p$ is directly related to its uncertainty, that is $\Delta p : p \approx \approx \Delta p$ (see Sec. 16.7). Once the momentum of the electron is known, one can readily find its energy. Since in the above example $p > > m_e c = 10^{-30}$ kg $\times$ 3 $\times$ $10^8$ m/s, one should use the relativistic relation for energy and momentum (see Sec. 16.3):

$$\mathscr{E}^2 = c^2p^2 + m_e^2c^4$$

Then we get

$$\mathscr{E} = c\sqrt{p^2 + m_e c^2} = 3 \times 10^8 \sqrt{10^{-38} + (10^{-30} \times 3 \times 10^8)^2}$$
$$\approx 2 \times 10^8 \text{ eV} = 200 \text{ MeV}$$

This figure is greatly in excess of that (7-8 MeV) found for the total binding energy by experiment (see Sec. 80.4) and is many times the energy of electrons emitted in beta-decay. If, on the other hand, the electrons in the nucleus were assumed to have the energy comparable with that associated with the particles emitted in beta-decay (usually a few MeV), then the region where the electrons must be localized, that is, the size of the nucleus as found from the uncertainty relations would be much greater than that found by observation.

4. A way out was found when in 1932 Chadwick discovered a new fundamental particle. From an analysis of the paths followed by the particles produced in some nuclear reactions and applying the law of conservation of energy and momentum, Chadwick concluded that these paths could only be followed by a particle with a mass slightly greater than that of the proton and with a charge of zero. Accordingly, the new particle was called the *neutron*.

Soon after Chadwick's discovery, in 1934, D. D. Iwanenko of the Soviet Union came out with a hypothesis that the *atomic nucleus should consist solely of protons and neutrons*. A similar hypothesis was advanced by Heisenberg.

5. Before long these views met with general acceptance and served as a basis for the present-day theory of the atomic nucleus. According to the latest views, the mass number $A$ represents the total number of protons and neutrons in the nucleus. The charge number $Z$, a multiple of the protonic charge, defines the number of protons, so that the difference $A - Z = N$ defines the number of neutrons in the nucleus of a given isotope.

If we take as a guide the number of protons $Z$ and of neutrons $A - Z$ in the nuclei of the various elements in the Periodic Table, we shall discover that in the elements as far as the middle of the table the nuclei have about as many neutrons as protons, so that $(A - Z)/Z \approx 1$. With increase in mass number, that is, in the heavier nuclei the number of neutrons exceeds that of protons, so that at the end of the Periodic Table the ratio $(A - Z)/Z \approx 1.6$.

In nuclear physics it is said that the proton and the neutron are two charge states of the same particle, the nucleon. The proton is the protonic state of the nucleon with a charge $+e$, and the neutron is its neutronic state with zero charge. According to the latest data, the rest mass of a proton and of a neutron respectively is

$$m_p = 1.0075975 \pm 0.000001 \text{ amu} = (1836.09 \pm 0.01)\, m_e$$

$$m_n = 1.008982 \pm 0.000003 \text{ amu} = (1838.63 \pm 0.01)\, m_e$$

The proton and the neutron have the same mass number equal to unity.

In the nucleus, the nucleons are in states substantially differing from their free states. This is because in all nuclei, except that of ordinary hydrogen, there are at least two nucleons between which a special *nuclear interaction* or *coupling* exists.

6. The proton-neutron model of the nucleus accounts for both the observed values of isotopic masses and, the magnetic moments of the nuclei. For, since the magnetic moments of the proton and the neutron are of the same order of magnitude as the nuclear magneton (Sec. 80.2), it follows that a nucleus built up of nucleons should have a magnetic moment of the same order as the nuclear magneton (Sec. 80.2). Therefore, with protons and neutrons as the building blocks of nuclei, the magnetic moment should be of the same order of magnitude. Observations have confirmed this.

Also with protons and neutrons as the constituents of nuclei, the uncertainty principle leads to reasonable values of energy for these particles in a nucleus, in full agreement with the observed energies per particle (Sec. 80.4).

Finally, with the assumption that nuclei are composed of neutrons and protons, the difficulty arising from nuclear spin has likewise been resolved. For if a nucleus contains an even number of nucleons (an even mass number $A$), it has integral spin (in units of $\hbar$). With an odd number of nucleons (an odd mass number $A$), its spin will be half-integral (in units of $\hbar$).

## 80.4. BINDING ENERGY OF THE NUCLEUS. MASS DEFECT

1. Nuclei containing positive protons and uncharged neutrons make up stable systems despite the fact that the protons experience Coulomb repulsion.

The stability of nuclei is an indication that there must be some kind of binding force between the nucleons. It would appear that for a proper insight into this binding force one must precisely know how the attraction between the nucleons depends on distance. However, the binding force can be investigated on the energy basis alone, without invoking any considerations concerning the nature and properties of nuclear forces.

An idea about the strength of a system can be gleaned from the effort required to break it up—the greater the effort, the stronger the system. As regards the nucleus, to destroy it means to break the bonds between the nucleons or, in other words, to do work against the binding. This approach based on the law of conservation of energy leads to several important facts about the forces that hold the nucleons in a nucleus together.

The energy required to remove any nucleon from the nucleus is called the *binding* (or *separation*) *energy of that nucleon* in the nucleus. It is equal to the work that must be done in order to remove the nucleon from the nucleus without imparting it any kinetic energy. The *total binding energy* of a nucleus is defined as the amount of work that must be done in order to break up the nucleus into its constituent nucleons. From the law of conservation of energy it follows that in forming a nucleus, the same amount of energy must be released as is put in to break it up.

2. The magnitude of the binding energy of nuclei may be estimated from the following considerations. The rest mass of any permanently stable nucleus has been found to be less than the sum of the rest masses of the nucleons that it contains. It appears as if in "packing up" to form a nucleus the protons and neutrons lose some their masses.

An explanation of this phenomenon is given by the special theory of relativity (see Sec. 20.1). This fact is accounted for by the conversion of part of the mass energy of the particles into binding energy. In Sec. 20.5 we have taken up the changes in the internal energy of a body during nuclear processes. Now we shall discuss the point in greater detail.

The rest energy of a body, $\mathscr{E}_0$, is related to its rest mass, $m_0$, by the expression

$$\mathscr{E}_0 = m_0c^2$$

where $c$ is the velocity of light in a vacuum. Designating the energy given up in the formation of a nucleus as $\Delta\mathscr{E}_b$, then the mass equivalent of the total binding energy

$$\Delta m_0 = \Delta\mathscr{E}_b/c^2 \tag{80.3}$$

is the decrease in the rest mass as the nucleons combine to make up the nucleus. If a nucleus of mass $M$ is composed of a number $Z$ of protons with a mass $m_{\text{p}}$ and of a number $A - Z$ of neutrons with a mass $m_{\text{n}}$, the quantity $\Delta m_0$ is given by

$$\Delta m_0 = Zm_{\text{p}} + (A - Z)\, m_{\text{n}} - M \tag{80.3$'$}$$

The quantity $\Delta m_0$ gives a measure of the binding energy, for it follows from (80.3) and (80.3′) that

$$\Delta\mathscr{E}_b = \Delta m_0 c^2 = [Zm_{\text{p}} + (A - Z)\, m_{\text{n}} - M]\, c^2 \tag{80.4}$$

3. In nuclear physics, energies are expressed in atomic energy units (aeu) corresponding to atomic mass units:

$$1 \text{ aeu} = c^2 \times 1 \text{ amu} = 9 \times 10^{16} \text{ m}^2/\text{s}^2 \times 1.660 \text{ kg} = 1.491 \times 10^{-10} \text{ J}$$
$$= 931.1 \text{ MeV}$$

Thus, in order to find the binding energy of a nucleus in MeV, one should use the equation

$$\Delta\mathscr{E}_b = [Zm_{\text{p}} + (A - Z)\, m_{\text{n}} - M]\, 931.1 \tag{80.5}$$

where the masses of the nucleons and the mass of the nucleus are expressed in atomic mass units.

On the average, the binding energy per nucleon is about 8 MeV, which is a fairly large amount.

4. In practical calculations, in addition to the binding energy one often uses the so-called *mass defect* (or *mass decrement*), which is the difference between the exact atomic mass, $M_a$, of an isotopically pure specimen, expressed in atomic mass units, and the mass number $A$:

$$\Delta_X = M_a - A \tag{80.6}$$

where $X$ is the symbol of an element. The term mass defect also applies to the quantity $\Delta m_0$ defined by Eq. (80.3′).

The mass defects of free protons and neutrons having mass numbers of unity respectively are:

$$\Delta_{\text{p}} = 0.0075957 \text{ amu}, \quad \Delta_{\text{n}} = 0.008982 \text{ amu}$$

For oxygen, ${}_8\text{O}^{16}$, the mass defect is zero:

$$\Delta_{\text{O}} = 0 \text{ amu}$$

because the isotopic mass of ${}_8\text{O}^{16}$ is the mass number exactly.

For most nuclei, the mass defect is a negative quantity. The accuracy with which mass defects can be determined depends on that of the isotopic masses. For heavy nuclei, the mass defects are found with errors ranging between wide limits. This is why it is usual to employ the mass defect per nucleon. This quantity is called the *packing fraction*, $f$:

$$f = \Delta_X /A = (M_a - A)/A = M_a/A - 1 \qquad (80.7)$$

The packing fraction can be determined with the same accuracy for both heavy and light nuclei.

A plot of the packing fraction as a function of mass number appears in Fig. 80.2. It has the highest value for neutrons and protons, being equal to the mass defects of these particles. As $A$ increases, the packing fraction rapidly decreases at first, then, on passing through zero for the ${}_8O^{16}$ nucleus, it becomes negative.

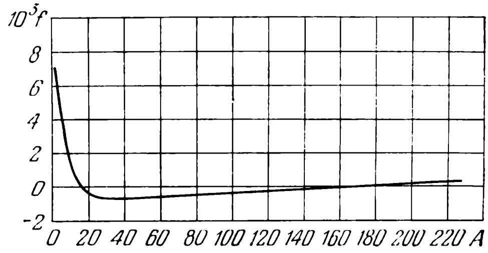


Fig. 80.2

5. Theory asserts that the nuclear binding energy $\Delta\mathscr{E}_b$ mainly depends on the total number of particles in the nucleus and less on the proton/neutron ratio. This is in agreement with observations which show that in a first approximation the nuclear binding energy rises linearly with an increase in the mass number. Physically, this means that each nucleon added to a nucleus causes the liberation of about the same amount of energy. The mean binding energy per nucleon is given by

$$\Delta\varepsilon_b = \Delta\mathscr{E}_b/A \qquad (80.8)$$

A plot of the binding energy per nucleon as a function of mass number $A$ is shown in Fig. 80.3. As is seen from the plot, the strength of binding varies with the mass number of the nuclei. The binding is at its strongest in the middle of the Periodic Table, in the range $28 < A < 138$, that is, from ${}_{14}Si^{28}$ to ${}_{56}Ba^{138}$. In these nuclei, the binding energy is very close to 8.7 MeV. With further increase in the number of nucleons in the nucleus, the binding energy per nucleon decreases. For the nuclei at the end of the Periodic Table (for example, uranium), $\Delta\varepsilon_b$ is about 7.6 MeV.

In the region of small mass numbers, the binding energy per nucleon shows characteristic maxima and minima. Minima in the binding energy per nucleon are shown by nuclei containing an odd number of protons and neutrons, such as ${}_3Li^6$, ${}_5B^{10}$ and ${}_7N^{14}$. Maxima in the binding energy per nucleon are associated with nuclei having an even number of protons and neutrons, such as ${}_2He^4$, ${}_6C^{12}$ and ${}_8O^{16}$.

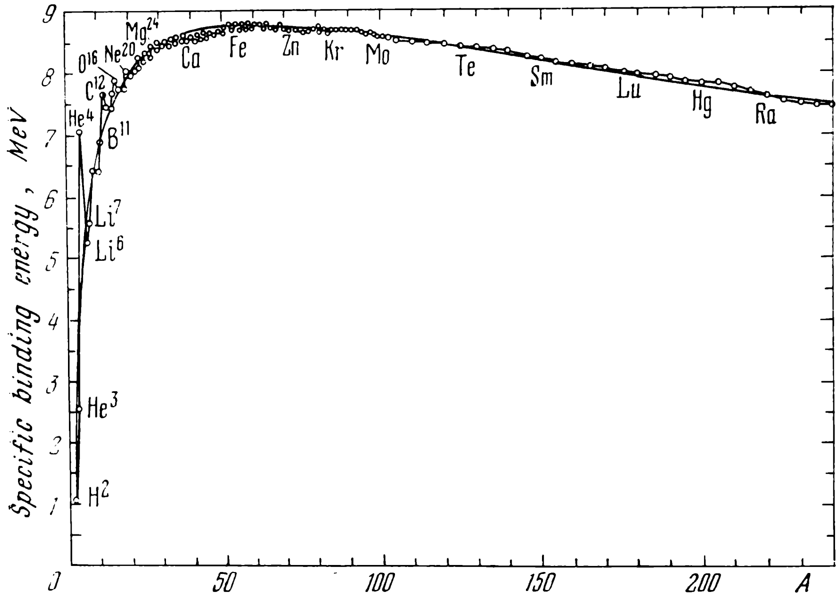


Fig. 80.3

The general course of the curve gives a clue to the mechanisms by which nuclear energy is released. We find that nuclear energy can be released either by the fission of heavy nuclei and the fusion of light nuclei from still lighter ones. It is clear from general considerations that energy will be released in nuclear reactions for which the binding energy per nucleon in the end products exceeds the binding energy per nucleon in the original nuclei. This general condition can be satisfied either by the fission of heavy nuclei into constituents in the middle of the Periodic Table or by the fusion of lighter nuclei at the beginning of the Periodic Table. These points will be raised again in greater detail in Ch. 82.

## 80.5. NUCLEAR FORCES

1. The forces operating between nucleons in a nucleus are called *nuclear forces*. An idea about these forces can be gained from general considerations. The stability of nuclei and the release of energy as a nucleus is formed from nucleons are indications that up to a certain distance between the nucleons, nuclear forces are those of attraction. Nuclear attraction is a good deal stronger than the electrostatic repulsion of the protons and accounts for the high binding energy and stability of the nuclei.

Nuclear forces cannot be ordinary electrostatic forces, for then a stable nucleus composed of a proton and a neutron would be inconceivable. Yet, such a nucleus does exist as the deuteron, the nucleus of heavy hydrogen or deuterium, ${}_1D^2$. The deuteron is a stable system with a binding energy of 2.2 MeV.

It is an easy matter to prove by substituting the mass and charge of a proton in the expressions for Coulomb's law and the universal law of gravitation that the gravitational attraction between protons is $1/10^{36}$ of the electrostatic repulsion between them. This is why gravitational forces play practically no role in the nucleus, despite the negligible separation between its nucleons.

As to magnetic interaction between protons in motion in the nucleus, it has been shown in Sec. 40.3 that the magnetic interaction of moving electrical charges is $1/(v/c)^2$ of their electrostatic interaction, where $v$ is the velocity of each charge relative to the chosen reference system, and $c$ is the velocity of light in vacuum. For protons, $v < c$, and their magnetic interaction is weaker than the electrostatic one. Apart from this magnetic interaction, however, all nucleons, both protons and neutrons, experience the interaction of their intrinsic magnetic moments (see Sec. 80.2). Calculations show that the energy of this interaction between a proton and a neutron is about $10^5$ eV, which is considerably less than the binding energy in the deuteron.

To sum up, nuclear forces are a special kind different from all forces known before their discovery.

2. The nucleus occupies a finite element of space (Sec. 80.6), and within this element the nucleons must be a definite distances apart. Obviously at a certain distance, attractive force gives way to repulsive force. The distance at which this transition occurs is expressed in terms of *fermis* (fm). The fermi is defined as

$$1 \text{ fm} = 10^{-15} \text{ m} = 10^{-13} \text{ cm}$$

The fermi is not unlike the unit of the first Bohr radius in the hydrogen atom used in the measurement of distances in atomic physics (Sec. 71.5).

Observations and theory have revealed some other properties of nuclear forces.

3. Above all, they have been found to be *short-range* forces. They drop to a negligible value at a distance as short as $4.2 \times 10^{-15}$ m = = 4.2 fm. The distance of 2.2 fm has come to be known as the *range of nuclear forces.*

4. Nuclear forces are *charge-independent.* That is, interactions between two nucleons are independent of whether one or both nucleons have electric charge. In other words, neutron-neutron, neutron-proton and proton-proton interactions are almost identical in character. Thus, as regards specifically nuclear interactions, protons and neutrons are identical particles. The charge independence of nuclear forces has been established from experiments on the scattering of protons by deuterons and of neutrons by protons. We cannot discuss this point in greater detail. It will suffice to note that experiments on the scattering of neutrons by protons have solved the important problem of neutron spin (see Sec. 80.2).

The charge independence of nuclear forces can be confirmed by analysing the difference in binding energy between the so-called mirror nuclei. Nucleus $B$ is a mirror image of nucleus $A$ if the number of protons in $B$ is equal to the number of neutrons in nucleus $A$, while the number of neutrons in nucleus $B$ is equal to the number of protons in nucleus $A$. Or, which is the same, each member of a proton-neutron pair can be transformed into the other by exchanging all neutrons for protons and vice versa.

As an example of mirror nuclides, we shall consider the nucleus of the super-heavy isotope of hydrogen called tritium, ${}_1H^3$ (or ${}_1T^3$) and the nucleus of the light isotope of helium, ${}_2H^3$. The former contains one proton and two neutrons, and the latter two protons and one neutron. The binding energies of these nuclei respectively are 8.49 MeV and 7.72 MeV. Each nucleus has three nucleons, but they are bound stronger in tritium than in helium. Assuming the charge independence of nuclear forces, the difference in binding energy equal to 0.77 MeV should be attributed to Coulomb repulsion between the two protons in helium, because tritium has only one proton. Obviously, the mutual repulsion of protons reduces their binding energy and, as a consequence, that of the entire helium nucleus. If we assume that 0.77 MeV is the potential energy due to Coulomb repulsion between the protons, $U$, then using the expression $U = q_1q_2/4\pi\varepsilon_0 r = e^2/4\pi\varepsilon_0 r$ from electrostatics, we can find the distance between the protons at which their Coulomb energy is equal to this value. It is 1.9 fm, or of the same order as the range of nuclear forces.

5. Nuclear forces are *noncentral*, or *tensor*, forces, that is, those whose direction depends in part on the spin orientation of the nucleons, which may be parallel or anti-parallel. This has clearly been shown by experiments on the scattering of neutrons by the molecules of parahydrogen and orthohydrogen. A molecule of parahydrogen differs from that of orthohydrogen in that in the former the protons

have anti-parallel spin orientation, and in the latter, parallel spin orientation. If the interaction between nucleons were independent of spin orientation, neutrons would be scattered identically by orthohydrogen and parahydrogen. Observations have testified to the opposite, that is, nuclear forces are dependent on spin orientation.

6. Nuclear forces have the property of *saturation*, similar to that of chemical bonds between valence electrons in the atom. This saturation consists in that a nucleon interacts can only form bonds with a certain definite number of its neighbours and no more, even though they may be within the effective range of nuclear forces. The saturation of nuclear forces stems from the dependence of the total binding energy on the mass number $A$. If there were no saturation and each of the number $A$ of nucleons could interact with every other one, $(A - 1)$, the total binding energy would be proportional to the number of nucleon pairs in the nucleus, that is, to the number of arrangements of $A$ particles taken two at a time. In algebra it is proved that this number is equal to $A\ (A - 1)/2$. Therefore, if attractive forces existed between all such pairs, the total binding energy would be roughly proportional to $A^2$ rather than to $A$. However, data on the binding energy and mass defect show that the dependence of binding energy on mass number $A$ nearly linear. Thus, the forces between nucleons must saturate.

Much as the saturation of chemical bonds results in the formation of stable groups of atoms, or molecules, so the saturation of nuclear forces is responsible for the extremely high stability of some nuclides.

The saturation of nuclear forces is practically complete with alpha-particles which are stable formations of two protons and two neutrons. The saturation of nuclear forces may be explained by their short range if we assume that beyond the range of attractive force the repulsive force is such that any more particles are prevented from getting within the range of attractive force.

7. The short range of internucleon forces has been successfully explained on the basis of the hypothesis that they have an *exchange character*. The idea that interaction between two particles might be effected through an exchange of a third particle was for the first time advanced by I. E. Tamm and D. D. Iwanenko of the Soviet Union in 1934.

The present-day quantum field theory asserts that the field is quantized in much the same manner as the most important characteristics of quantum-mechanical particles, say, their energies. An exchange interaction arises from the continuous exchange of field quanta between the particles. Accordingly, electromagnetic interaction is treated as an exchange of electromagnetic-field quanta, or photons (see Ch. 68), and gravitation as an exchange of gravitational-field quanta, or gravitons.

Gravitons have not yet been found, but research is going on. The fundamental difficulty in the experimental observation of gravitons consists in that the gravitational waves emitted by the prospective sources are of negligible intensity.

In the field picture of interaction, the interaction of nucleons should likewise be treated as an exchange of quanta of a special *nuclear field*. It took a relatively long time to elucidate what quanta of this nucler field might be like. At first, it was believed that they might be electrons. The emission of electrons in beta-decay seemed to support the idea. It was argued that although there are no electrons in the nucleus (see Sec. 80.4), they might be produced by some processes occurring in the nucleus and they might serve as the agents of interaction between the nucleons. It has been shown theoretically by Tamm that electrons cannot be quanta of the nuclear field because this would not fit the observed short range of nuclear forces and high total binding energy.

In 1935, the Japanese physicist Yukawa published a theory of nuclear forces in which he suggested that the "quanta" of the nuclear field must be particles with a rest mass about 200 electron masses. Such particles have come to be known as *mesons* (from the Greek *mesos* for *middle*) for the reason that their rest mass must be intermediate between the mass of the electron and the mass of the proton. Before long, Yukawa's mesons were found experimentally and called *pi-mesons*, or *pions*. There are also other kinds of mesons having different rest masses (see Sec. 83.3).

For an insight into Yukawa's theory and in order to estimate the mass of the pi-meson at least roughly, we may reason as follows. On forming near a nucleon, the meson travels towards the other nucleon which absorbs it. The time, $\Delta t$, it takes the meson to travel from one nucleon to the other is their interaction time. During this time interval, the energy of the nucleon which gives up a meson is decreased, and that of the nucleon which absorbs the meson is increased. It may be said that during the time interval $\Delta t$ there is an uncertainty about the energy of each of the interacting nucleons, that is, we cannot say for certain when the meson leaves one nucleon and when it joins the other. According to the Heisenberg uncertainty principle (Sec. 70.2), the uncertainty in energy, $\Delta\mathscr{E}$, is connected to the time during which it exists by the relation $\Delta\mathscr{E}\Delta t \approx \hbar$. Since the uncertainty in energy is associated with the loss or gain of a meson, it cannot be smaller than the energy of the meson. For simplicity, we may put $\Delta\mathscr{E}$ to be equal to the rest energy of the meson, that is:

$$\Delta\mathscr{E} = \mathscr{E}_0 = m_\pi c^2$$

where $m_\pi$ is the rest mass of a pi-meson. As a consequence,

$$\Delta t \approx \hbar/\Delta\mathscr{E} \approx \hbar/m_\pi c^2$$

Let the mesons in the nucleus be relativistic particles, that is, ones travelling at a velocity $v$ very close to that of light in vacuum, $c$. To simplify the matter, we may put $v = c$. Since the interaction effected by the meson is one between nucleons, the distance $R_\pi$ which the meson covers is equal to the range of nuclear forces which may be put equal to $R_\pi = 1.5$ fm. Then,

$$R_\pi = v\Delta t = c\Delta t \approx c\hbar/m_\pi c^2 = \hbar/m_\pi c \tag{80.9}$$

whence

$$m_\pi = \hbar/R_\pi c \tag{80.10}$$

Substituting the numerical values in this expression we obtain the mass of the meson predicted by Yukawa, $m_\pi \approx 250 m_e$, where $m_e$ is the rest mass of an electron. According to present-day data, the rest mass of a pi-meson is 264 or 273 electron masses, according to whether it is charged or uncharged (Sec. 83.4).

The above reasoning and estimates do not apply to high energies. In other respects, rough estimates are quite satisfactory. Eq. (80.9) defines the range of nuclear forces transmitted by particles having a definite rest mass. From Eq. (80.9) it is seen that this range is equal to the Compton wavelength of a particle having the respective rest mass (see Sec. 68.6).

For a given rest mass of the pi-meson, Eq. (80.9) defines the range of nuclear forces. This relation may also be interpreted in a different way. Over the time $\Delta t = \hbar/m_\pi c^2 \approx 10^{-23}$ s, there is a continuous emission of pi-mesons which enter the picture only as a way of describing the interaction of nucleons by means of a meson field. These are *virtual pi-mesons*. If, on colliding with another nucleon, a given nucleon gains energy, the nucleon will emit a real, and not a virtual, pi-meson or pion (Sec. 83.4).

If the field quanta are photons, then $R_{ph} = \infty$, because their rest mass is zero. In other words, the range of electromagnetic forces is infinite: in effecting the interaction, photons can travel any distance, however long. This is in agreement with the observation that electromagnetic forces (electrical and magnetic, alike) gradually decrease with increasing separation between the interacting charges or currents, until they turn to zero at infinity.

## 80.6. THE NUCLEAR SIZE

1. In connection with Rutherford's experiments on the scattering of alpha-particles by nuclei, mention was made of a method for determining the size (or radius) of the nucleus. Obtained by this method, the nuclear radius appears to be dependent not only on the properties of the nucleus and the energy of the alpha-particles used for bombardment. That is, the nuclear radius thus determined is

decided by Coulomb forces having a far greater range than the specifically nuclear forces.

2. This is why we ought to define the *specifically nuclear radius*, that is, one decided by internucleonic interactions. It must be remembered, however, that a nucleus is a system of particles obeying quantum mechanics and, among other things, the Heisenberg uncertainty relations. Because of this, the radius of the spherical volume containing the nucleons can be specified only within the limits fixed by these relations. In other words, the boundary of the spherical volume defined by the nuclear radius has of necessity a spread. This fully applies to the spherical volume occupied by the electrons in an atom, that is, to the size of the atom as a whole.

3. Experimentally, the nuclear radius can be determined from the scattering of uncharged particles having a sufficiently high energy and of highly energetic electrons by nuclei. Experiments on the scattering of neutrons by nuclei have shown that the nuclear radius increases with increase of mass number according to the relation:

$$R = R_0 \sqrt[3]{A} \tag{80.11}$$

where $R_0 \approx 1.4$ or $1.5$ fm.

Eq. (80.11) may be interpreted as follows. The nucleus is a collection of particles having about the same size and spaced about the same distance apart so that each particle occupies about the same "effective" volume. Then the volume of the nucleus is proportional to the number of nucleons it contains, which fact is stated in Eq. (80.11). For, if $R$ is the nuclear radius, and $R_0$ is the radius of a single nucleon, then,

$$4\pi R^3/3 = 4\pi R_0^3 A/3$$

whence we obtain Eq. (80.11).

The heaviest nuclei, for example uranium, measure about $3 \times 10^{-28}$ m$^2$ in cross section and their nuclear radii are of the order of $10^{-14}$ m.

4. Using Eq. (80.11), we can find the mean density, $\rho$, of nuclear matter. Assuming that the nucleus is a sphere of radius $R$, we have

$$\rho = M_N/(4\pi R^3/3) \tag{80.12}$$

where $M_N$ is the mass of the nucleus. If we put $M_N = m_n A$, where $m_n$ is the mass of a neutron, then

$$\rho = \frac{1.674 \times 10^{-24}}{4/3\pi\,(1.5 \times 10^{-15})^3} \approx 1.3 \times 10^{14}\ \text{g/cm}^3 = 1.3 \times 10^{17}\ \text{kg/m}^3$$

As is seen, the density of nuclear matter is tremendous, exceeding by an extremely wide margin that of ordinary substances made up of the atoms of elements or their compounds. It may be added that the density of nuclear matter is independent of the mass number $A$.

## 80.7. THE LIQUID-DROP MODEL OF THE NUCLEUS

1. We still lack accurate data about nuclear forces, and so no theory of the atomic nucleus may be exhaustive. This is why different models of the nucleus have been devised in an attempt to describe and calculate the various quantities representing the properties of the nucleus and the processes occurring in it. In a way, these models are not unlike photographs of the same scene taken from different angles. Each gives only a partial picture.

We shall only discuss the simplest model of the nucleus, the *liquid-drop model*. Apart from its simplicity, it enables us to describe not only some of the properties of the nuclei, but also their fission. The latter point is of special importance for an insight into the physical principles of nuclear power generation (Ch. 82).

2. For the first time, a liquid-drop model of the nucleus was proposed in 1936 by Ya. I. Frenkel and elaborated by Bohr and von Weizsäcker. This model is based on an outer analogy between the atomic nucleus and a charged drop of a liquid. By way of comparison, the interaction forces in a liquid drop operate over a short range like the nuclear forces. As is with the nucleus, the forces at work in the liquid drop have the property of saturation. Much as the nucleus has an approximately constant binding energy per nucleon and a fixed density independent of the number of nucleons, so the liquid drop has a fixed density (at specified ambient temperature and pressure) independent of the number of particles in the drop. Finally, the constituent molecules of the drop have a certain mobility, like that of nucleons.

However, the nucleus is charged and obeys the laws of quantum mechanics. In this, it differs from a liquid drop.

A point of special importance in the development of the liquid-drop model of the nucleus was that the ratio $(A - Z)/Z$ gradually increases from unity at the beginning of the Periodic Table to 1.6 at its end with increase in $Z$ and $A$. Of all nuclei, the most stable ones are those for which the "concentration" of neutrons and protons is the same. The increase in the ratio $(A - Z)/Z$ with increase in $A$ means an increase in the concentration of neutrons in the "liquid-drop" nucleus. The increase in the Coulomb energy of repulsion between protons, which is proportional to $Z^2$, entails a decrease in the concentration of protons and, accordingly, an increase in the concentration of neutrons.

3. The liquid-drop model of the nucleus was put on a theoretical footing by a semiempirical formula that von Weizsäcker had deduced for the overall energy of the atomic nucleus. In this formula, the overall energy of the nucleus is represented as a sum of six terms:

$$\mathscr{E} = \mathscr{E}_1 + \mathscr{E}_2 + \mathscr{E}_3 + \mathscr{E}_4 + \mathscr{E}_5 + \mathscr{E}_6 \qquad (80.13)$$

The first term, $\mathscr{E}_1$, is the *rest energy* of the nucleus, connected to the rest mass of the constituent nucleons (Secs. 80.3 and 80.4) by the following relation:

$$\mathscr{E}_1 = [Zm_p + (A - Z)\, m_n]\, c^2$$

The second term, $\mathscr{E}_2$, represents the energy released as the nucleons form a bond to make up the nucleus, that is, the *binding energy* of the nucleus. This term is negative because the binding energy is released when a nucleus is formed. It has been shown in Sec. 80.4 that in a first approximation the binding energy per nucleon may be taken to be the same for all nuclei, and so the total binding energy is proportional to the number $A$ of nucleons in the nucleus:

$$\mathscr{E}_2 = -\alpha_1 A$$

where $\alpha_1$ is a coefficient which is to be found by experiment.

The third term, $\mathscr{E}_3$, represents the so-called *surface energy of the nucleus*, which is not unlike the surface energy of a liquid (see Sec. 34.7). Like the molecules situated in the surface region of a liquid drop, the outer nucleons at the boundary of the nucleus experience one-sided attraction directed inside the nucleus. This adds to the potential energy of the outer nucleons and, as a consequence, to that of the entire nucleus. As with a liquid drop, the surface energy of a nucleus is proportional to its surface area:

$$\mathscr{E}_3 = \sigma 4\pi R^2$$

where $\sigma$ is the coefficient of "surface tension" for the nucleus (see Sec. 34.7). Since according to (80.11), the nuclear radius is proportional to $A^{1/3}$, the expression for $\mathscr{E}_3$ may be re-cast as follows:

$$\mathscr{E}_3 = \alpha_2 A^{2/3}$$

where $\alpha_2$ is a further constant to be determined empirically.

The fourth term, $\mathscr{E}_4$, represents the Coulomb energy due to the electrostatic repulsion between protons in the nucleus:

$$\mathscr{E}_4 \approx q^2/4\pi\varepsilon_0 R \approx (Ze)^2/4\pi\varepsilon_0 R$$

where $\varepsilon_0$ is the dielectric constant in SI units. Replacing $R$ in accordance with (80.11), we get

$$\mathscr{E}_4 = \alpha_3 Z^2 A^{-1/3}$$

where $\alpha_3$ can be deduced theoretically.

The term, $\mathscr{E}_5$, called the *assymmetry* (or *neutron-excess*) *energy* explains the difference in stability between nuclei containing unequal numbers of protons and neutrons. The stability of a nucleus is directly related to its energy, and, as usual, that system is most stable of all which has the lowest energy. Therefore, we find the highest stability and the highest binding energy per nucleon, in nuclei containing equal numbers of protons and neutrons:

$$Z = A - Z$$

that is,

$$A = 2Z, \quad A - 2Z = 0$$

The expression $(A - 2Z)/A$ is called the neutron excess ratio. Thus, the asymmetry energy is a function of this ratio. Its exact form is unknown. Assuming that it is independent of the sign of the argument $(A - 2Z)/A$, that is, of whether protons are more numerous or vice versa, this function may be quadratic

$$\mathscr{E}_5/A = \alpha_4 \, [(A - 2Z)/A]^2$$

or

$$\mathscr{E}_5 = \alpha_4 \, (A - 2Z)^2/A$$

where $\alpha_4$ is to be found by experiment.

Finally, the last term, $\mathscr{E}_6$, called the *pairing energy*, allows for the fact that the interaction of nucleons depends on their relative spin orientation. Because of this, the total binding energy is a maximum for nuclei containing an even number of protons and neutrons ("even-even", or e-e nuclei), and a minimum for nuclei containing an odd number of protons and neutrons ("odd-odd", or o-o nuclei). Most often, it is connected to $A$ by the relation:

$$\mathscr{E}_6 = \pm \alpha_5 A^{-3/4}$$

where $\alpha_5$ is an empirical constant. The "+" sign applies to nuclei with an even number of protons and neutrons, and the "—" sign to nuclei with an odd number of these nucleons. For other nuclei, the term $\mathscr{E}_6$ is dropped.

By definition, (80.5), the total binding energy of a nucleus is the difference between the rest energy of its constituent nucleons and its total energy

$$\Delta\mathscr{E}_b = \mathscr{E}_1 - \mathscr{E} \tag{80.14}$$

At present, the following numerical values have been found for the various coefficients in the Weizsäcker formula: $\alpha_1 = 15.75$ MeV, $\alpha_2 = 17.8$ MeV, $\alpha_3 = 0.71$ MeV, $\alpha_4 = 23.7$ MeV, and $\alpha_5 = 34$ MeV. Substituting them in Eq. (80.13), we obtain the following formula for the total binding energy of nucleus:

$$\Delta\mathscr{E}_b = 15.75A - 17.8A^{2/3} - 0.71Z^2A^{-1/3} - 23.7\,(A - 2Z)^2\,A^{-1} \pm 34A^{-3/4} \tag{80.15}$$

4. So far we have used no criterion for the stability of atomic nuclei. Let a nucleus be stable if its composition remains unchanged with time.

The relation between charge number $Z$ and mass number $A$ at which nuclei are most stable can be found from the semiempirical formula (80.15). Consider nuclei having a specified value of $A$ ($A =$ $=$ constant), only differing in the value of $Z$, that is, nuclear isobars. For such nuclides the total binding energy $\Delta\mathscr{E}_b$ given by Eq. (80.15) will be a function of only $Z$. Under the circumstances, a criterion for the stability of isobars can be deduced from general considerations. It is obvious that for a specified value of $A$ those nuclei will be most stable which have values of $Z$ corresponding to the lowest total energy $\mathscr{E}$ of the nucleus. These values of $Z_{st}$ can be found from the expression for the minimum energy of the nucleus. Calculations which we omit* yield the following result:

$$Z_{st} = A/(1.98 + 0.015A^{2/3}) \tag{80.16}$$

It is usual to choose for $Z_{st}$ the integral value closest to that yielded by Eq. (80.16). This equation agrees with observations. Incidentally for not too heavy nuclides, it gives $Z_{st} \approx A/2$, or in words, the number of protons is equal to that of neutrons in the nucleus.

## Chapter 81
## NATURAL RADIOACTIVITY

### 81.1. RADIOACTIVITY DEFINED

1. The term 'natural radioactivity' applies to the spontaneous transformation of one nuclear species into another with the emission of some particles (such as alpha, beta, antineutrinos, and neutrinos) or electromagnetic radiations (gamma-rays). As a rule, natural radioactivity is displayed by the heavy nuclei at the end of the Periodic Table, beyond lead. There are also naturally radioactive light nuclei, such as the potassium isotope $_{19}K^{40}$, the carbon isotope $_6C^{14}$, and the rubidium isotope $_{37}Rb^{87}$, to name but a few.

Natural radioactivity was discovered by Henri Becquerel in 1896. At that time, he was busy investigating fluorescence associated with X-rays. As part of his program, Becquerel used a sample of the double sulphate of potassium and uranium and placed it on a photographic plate wrapped in black paper, intending to expose it to sunlight. After awaiting in vain for sunshine for some days, he de-

* In deriving Eq. (80.16), use the condition for the minimum of a function in differential calculus.

veloped the plate as it was not suitable for his experiments. Much to his surprise, the image of the crystals stood out clearly, far more clearly than in any of his previous tests. Further experiments showed that the blackening of the plate was quite independent of exposure to light, and even compounds which exhibited no fluorescence could produce the effect so long as they contained uranium, especially in metallic form. These observations led Becquerel to the conclusion that uranium emitted special kinds of rays.

It was soon found that the rays from uranium could pass through thin metal films and ionize the gas through which they travelled. A very remarkable feature about this emission was that it occurred *spontaneously*, at a constant rate and independently of changes in ambient illumination, pressure and temperature.

Pierre and Marie Curie found that the radiation from pitchblende was four times as strong as from uranium. This led to an intensive search for the source of this stronger radiation. Finally, in 1898, the Curies succeeded in discovering two new substances which they named polonium, $_{84}Po^{210}$, and radium, $_{88}Ra^{226}$.

The substances emitting the newly discovered radiation were called *radioactive*, and the newly discovered property was named *radioactivity* by Mme M. Curie.

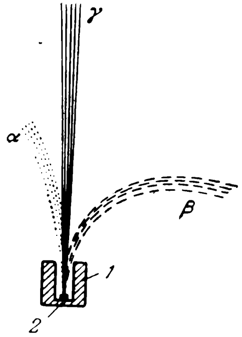


Fig. 81.1

2. It was soon found that the rays from these radioactive substances were of three kinds, called alpha-rays, beta-rays, and gamma-rays. Their existence was established from the deflection of radioactive emission in a magnetic field. The arrangement used to separate alpha-, beta- and gamma-rays is shown in Fig. 81.1, where the magnetic field is directed downwards at right angles to the plane of the drawings. At *1* is a thickwalled lead container holding a sample of radium, *2*. From the deflection of the radiation in the magnetic field it is seen that alpha-rays are positive, beta-rays are negative, and gamma-rays are uncharged.

Further investigations showed that alpha-rays were helium nuclei. For their identification, Rutherford carried out the following experiment. A glass vial holding a sample of radon, a radioactive gas ($_{86}Rn^{222}$) was placed in a glass vessel from which practically all air had been evacuated. The alpha-particles emitted by the radon sample were absorbed by the walls of the vessel, each captured two electrons, and turned to helium atoms. These were driven from the walls of the vessel by heating.

The spectrum of the gas in the vessel was found to be identical with the emission spectrum of helium, and this confirmed that the alpha-particles emitted by the radon sample turned to helium. Applying the methods of magnetic and electrostatic deflection, Rutherford determined the specific charge, $q/m_\alpha$, of alpha-particles (where $m_\alpha$ is the mass of an alpha-particle) and found that their charge was $2e$ and the mass the same as that of the nucleus of the helium isotope, ${}_2He^4$.

3. Beta-rays are streams of very fast electrons whose velocity exceeds that of ordinary cathode (electron) rays and approaches that of light in a vacuum. Their energy is 10 MeV. The character of beta-rays has been confirmed by measuring their specific charge, $q/m_\beta$, where $m_\beta$ is the mass of a beta-particle.

4. Gamma-rays are a hard electromagnetic radiation much more penetrating of all radioactive rays. The properties of gamma-rays have been investigated by the same methods as those of X-rays (see Sec. 62.5), mostly from their absorption and scattering by substances. It has been found that they cause a weak ionization in the material they traverse. Since they have higher frequencies (that is, shorter wavelengths) than X-rays, their quantum-mechanical properties stand out with special clarity.

5. Experiments have shown that all radioactive radiations cause chemical effects, blacken photographic plates, ionize gases and, sometimes, condensed materials on passing through them, and cause some solids and liquids to fluoresce. These properties are at the basis of experimental techniques for the detection and investigation of radioactive rays (see Sec. 81.8).

## 81.2. TRANSITION RULES FOR RADIOACTIVE DECAY

1. In his experiments on the identification of alpha-particles, Rutherford found that the amount of radioactive radon decreased with time exponentially as $\exp(-bt)$, where $b$ is the decay constant independent of the environments and the concentration of radioactive atoms. The disintegration of radium in $RaCl_2$ and $RaBr_2$ has been found to be dependent solely on the number of radium atoms in the compounds, that is, the rate of the disintegration is independent of whether the sample is a pure element or a compound. These facts have led to the conclusion that radioactive transformations are the property of nuclei which can undergo these transformations spontaneously.

2. The nuclear transformations accompanied by the emission of alpha- and beta-particles are called alpha- and beta-decay, respectively. Gamma-decay is non-existent. The nucleus that undergoes a decay is called the *parent*, the intermediate products are called *daughters*, and the final stable element is called the *end product*.

Experimental studies into radioactive disintegrations have led to the formulation of transition rules:

for alpha-decay: $_ZX^A \xrightarrow[\alpha]{} {}_{Z-2}Y^{A-4} + {}_2He^4$

for beta-decay: $_ZX^A \xrightarrow[\beta-]{} {}_{Z+1}Y^A + {}_{-1}e^0$

where X is the chemical symbol of the parent nucleus, Y is that of a daughter nucleus, $_2He^4$ is the helium nucleus (the end product), and $_{-1}e^0$ is the electron of charge —1 (in units of elementary charge $e$) and of mass number zero, since the electronic mass is 1/1836 the protonic mass.

The transition rules are based on the conservation of charge and of mass number: the sum of charges (and of mass numbers) of the daughter nuclei and end products is equal to the charge (mass number) of the parent nucleus. This is exemplified by the decay scheme of radium with the emission of radon and an alpha-particle:

$$_{88}Ra^{226} \xrightarrow[\alpha]{} {}_{86}Rn^{222} + {}_2He^4$$

Thus, the alpha-transformation removes four units of mass and two units of charge, producing an element two steps down in the Periodic Table. The beta-disintegration removes one negative charge and essentially no mass, producing an element one step higher in the Periodic Table.

3. The daughter nucleus produced by radioactive decay is, as a rule, capable of further decay, and so is the next daughter produced by the decay of the first. This process continues until finally a stable substance is formed. Thus we have a *radioactive series* or chain. Each member of a radioactive series is a radioactive isotope (radioisotope) of the element occupying the respective square in the Periodic Table.

The naturally radioactive nuclei form three *radioactive series*, namely the *uranium series*, the *thorium series* and the *actinium series*, thus called after the respective parents, $_{92}U^{238}$, $_{90}Th^{232}$, and $_{89}Ac^{235}$. There is one more radioactive series produced artificially and starting with neptunium, $_{93}Np^{237}$, a transuranic element. In each radioactive series, each nuclide transforms into the next through a chain of alpha- and beta disintegrations, each chain terminating in a stable isotopic nucleus. The thorium series terminates in the $_{82}Pb^{208}$ nucleus, the uranium series in the $_{82}Pb^{206}$ nucleus, and the actinium series in the $_{82}Pb^{207}$ nucleus. The neptunium series terminates in the $_{83}Bi^{209}$ (bismuth) nucleus.

Even though we might not know which member of a given series undergoes radioactive decay by the emission of alpha- or beta-particles, we are in a position to state precisely how many alpha- and beta-transitions should take place before the parent turns into a specified product nucleus. As an example, we shall take up the transformation of the uranium nucleus into the lead nucleus: ${}_{92}U^{238} \rightarrow \ldots \rightarrow {}_{82}Pb^{206}$.

The number $n_\alpha$ of alpha transitions can be found at once by dividing the difference in mass number between the parent and the end product by four, because each alpha transition removes four units of mass. In our example, $n_\alpha = (A_1 - A_2)/4 = 8$.

To find the number of beta transitions, we first determine the decrease in charge number: 92 — 82 = 10 units. However, it should be recalled that each alpha transition removes two units of charge, while each beta transition adds one unit of charge. Thus, the number of beta-transitions is given by the equation:

$$Z_1 - Z_2 = 2n_\alpha - n_\beta,$$
$$2n_\alpha - n_\beta = 10$$

From the value of $n_\alpha$, we find that $n_\beta = 6$. Thus, the uranium nucleus undergoes eight alpha transitions and six beta transitions before it transforms to the lead nucleus.

## 81.3. THE FUNDAMENTAL LAW OF RADIOACTIVE DECAY

1. With time, the number of parent nuclei decreases because of radioactive decay. This decrease obeys a certain law which we seek to find. Let at the initial instant of time, $t = 0$, there be $N_0$ nuclei of a radioactive element X and let us find the number of nuclei of the same element that will remain untransformed by an arbitrary time $t$. Since we are dealing with spontaneous transformations, it is natural to assume that a greater number of nuclei will decay over a longer interval of time. Furthermore, the number of nuclei undergoing decay per unit of time (say, a minute) will be greater with a larger original sample. These two points underly the fundamental law of radioactive decay. If we have $N$ untransformed nuclei present at time $t$, and $N - \Delta N$ untransformed nuclei existing at time $t + \Delta t$, then the change in the number of untransformed nuclei that is the number of nuclei decaying in time $\Delta t$ will be proportional to $N$, that is:

$$\Delta N \sim N\Delta t; \quad \text{or} \quad \Delta N = -\lambda N \Delta t \qquad (81.1)$$

where $\lambda$ is a positive proportionality factor called the *decay constant*; it has a definite value for each nuclear species. The minus sign on the right-hand side of Eq. (81.1) indicates that $\Delta N$ decreases as time increases, that is, their final number is less than the initial number. From Eq. (81.1) it follows that the decay constant is the fractional decrease in the number of nuclei decaying per unit time:

$$\lambda = (-\Delta N/N)/\Delta t \qquad (81.2)$$

In other words, the decay constant represents the proportion of nuclei decaying per unit time, or the *decay* rate. The decay constant is independent of ambient conditions and is solely determined by the internal properties of the nucleus. It has dimensions of $|\lambda| = T^{-1}$.

2. Using Eq. (81.1) we can readily solve the problem stated in Para. 1, namely find the time dependence for radioactive decay. Reasoning along the lines similar to those in Sec. 55.4 where we derived the relation describing the attenuation of waves, we can show that the number of atoms of the original kind remaining after time $t$ is

$$N = N_0 \exp(-\lambda t) \qquad (81.3)$$

where $N_0$ is the initial number of radioactive nuclei existing at $t = 0$ and $N$ is the number of radioactive nuclei present at $t$. A plot of $\ln(N/N_0)$ as a function of time appears in Fig. 81.2. The decay constant $\lambda$ can be found from the slope of the curve, for it can be proved that $\tan\alpha = \lambda K$ (see page 221).

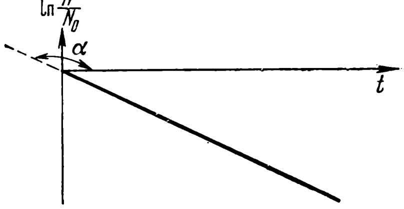


Fig. 81.2

3. In practice the stability of radioactive nuclei against decay and the decay rate are most often estimated in terms of the *half-life*, $T$, rather than the decay constant $\lambda$. The half-life is defined as *the time at which half of the original nuclei have decayed.* Stated somewhat differently, the half-life is the time after which one-half the original number of nuclei remains untransformed. Thus, $t = T$, if $N(T) = N_0/2$. By this definition and on the basis of the exponential decay law, Eq. (81.3), $T$ and $\lambda$ are related as

$$N_0/2 = N_0 \exp(-\lambda T)$$

Cancelling $N_0$ and taking a logarithm, we obtain

$$T = \ln 2/\lambda = 0.693/\lambda$$

or

$$1/\lambda = T/0.693 = 1.44T \qquad (81.4)$$

The half-lives of naturally radioactive elements range between wide limits. For uranium it is 4 500 million years, for radium 1590 years, for protactinium 32 000 years, for radon 3.825 days, and for radium-C (an isotope of polonium) it is $1.5 \times 10^{-4}$ s. For some induced radioactive elements the half-life is a few millionths or even hundred-millionths of a second.

4. The constancy of $T$ (or $\lambda$) for a given radioactive element implies that these quantities represent huge numbers of atomic nuclei. Thus, radioactive decay is a statistical process (see Sec. 81.6).

The above definition of the half-life is sometimes incorrectly construed as implying that the total number of nuclei in a sample will decay in a time equal to $2T$. This is not so because if the number of nuclei remaining after the time $T$ is $N_0/2$, then after the time $2T$ this number will be half the number $N_0/2$, or one-quarter of $N_0$, and in the time $3T$ this number will be half of $N_0/4$, that is, $N_0/8$, and so on.

### 81.4. ACTIVITY AND ITS MEASUREMENT

1. It is natural to ask how one can measure a very long and a very short half-life. It is obvious that Eq. (81.3) cannot be used for this purpose directly. Help comes from the fact that the members of a radioactive series are radioactive, too. Generally, the number of daughter nuclei is changing with time as well. This will continue until the decay rate of a radioactive product (daughter nuclei) becomes just equal to its rate of formation from the previous member of the chain (the parent nuclei). This condition is called *ideal equilibrium*. Thus, at ideal equilibrium

$$-\Delta N_p/\Delta t = -\Delta N_d/\Delta t$$

or, in words, the number of parent nuclei decaying in time $\Delta t$ is equal to that of daughter nuclei disintegrating in the same time. But, according to (81.1)

$$-\Delta N/\Delta t = \lambda N \qquad (81.5)$$

and so, at equilibrium the following relation holds

$$\lambda_p N_p = \lambda_d N_d$$

or

$$N_p/N_d = \lambda_d/\lambda_p = T_p/T_d \qquad (81.6)$$

*At ideal equilibrium, the numbers of parent and daughter nuclei are proportional to their half-lives.* This relation is used in cases where the half-life of a nuclear species is either too short or too long for direct determination from Eq. (83.1).

2. By definition, the total number of decays per unit time, $Q = -\Delta N/\Delta t$, is called the *activity* of a given radioactive sample. On the basis of (81.5), *the activity of a radioactive source is equal to the product of decay constant by the number of unchanged nuclei present in that source.*

Since the number of unchanged nuclei is decreasing continually, the activity of the remaining source is decreasing too. This decrease is immaterial if the source has a long half-life (as is with uranium or radium). It has to be taken into account, however, if the half-life is a few years or, the more so, a few days. For example, the activity of radon baths used for medical purposes is halved in less than two days, because the half-life of radon is 3.825 days. This is why such a preparation has to be renewed rather frequently.

3. In the International System (SI) of units, activity is expressed in $s^{-1}$. A source is said to have one unit of activity if it undergoes one decay every second.

Activity is often expressed in curies. One curie (Ci) is the activity of 1 g of radium, that is, the number of decays per second occurring in one gram of radium. Let us find this number.

To begin with, we shall express the half-life of radium equal to $T = 1590$ years in seconds and find its decay constant $\lambda$. Then, we shall determine the number of nuclei in one gram of radium, that is, the number of radium atoms per gram. It is equal to Avogadro's number, $N_A$, divided by the mass of one kilomole, $M$:

$$N = N_A/M = \frac{6.023 \times 10^{26} \times 1/\text{kmole}}{226 \text{ kg/kmole}} = 2.67 \times 10^{24} \text{ kg}^{-1}$$
$$= 2.67 \times 10^{21} \text{ g}^{-1}$$

Then the activity of one gram of radium will be

$$Q = \lambda N = 0.693N/T = \frac{0.693}{1590 \times 365 \times 24 \times 3600} \times 2.67 \times 10^{21}$$
$$= 3.7 \times 10^{10} \text{ s}^{-1}$$

That is, the number of decays per second in one gram of radium is 37 000 million.

The definition of the curie used at present reads as follows: The curie is a unit of radioactivity defined as the quantity of any radioactive nuclide in which the number of decays per second is $3.7 \times 10^{10}$.

The curie is a very large unit, because radium is a very active element, and the mass of one gram is a fairly large amount for any practical preparation.* This is why in practice use is made of submultiples of the curie, namely the millicurie (mCi) and the microcurie (μCi)

$$1 \text{ mCi} = 10^{-3} \text{ Ci}$$

$$1 \text{ μCi} = 10^{-6} \text{ Ci}$$

An alternative unit is the rutherford (Rd), a unit of radioactivity equal to $10^6$ decays per second, 1 Rd = $10^6$ $s^{-1}$. Obviously, 1 Ci = $= 3.7 \times 10^4$ Rd.

## 81.5. USE OF THE EXPONENTIAL DECAY LAW

1. Equations (81.1) through (81.5) come in handy in many problems involving radioactivity. For example, Eq. (81.1) may be used to find the number of nuclei (or mass) of a given radioactive element decay-

* The total amount of radium available in the world is a few kilograms.

ing or remaining untransformed over a given time interval. It should be remembered, however, that Eq. (81.1) may be used only if the time of observation is a small fraction of the half-life, that is, if $\Delta t \ll T$. Otherwise, one might end up with a number of transformed nuclei exceeding their original number. The following example will stress the point. Using Eqs. (81.1) and (81.5), we find that the proportion of the original number of nuclei decaying in time $\Delta t$ is

$$| \Delta N/N | = \lambda \Delta t = 0.693 \Delta t/T$$

Hence, at $\Delta t/T > 1/0.693 = 1.44$, that is, at $\Delta t > 1.44T$, we have that $\Delta N/N > 1$. That is, the number of transformed nuclei exceeds the original number. With nuclei having a short half-life, this trap may be passed unnoticed. Thus, it appears natural at first sight to apply Eq. (81.1) to the following problem: Find the proportion of radon disintegrating in six days. However, the answer will then be preposterous: $\Delta N/N > 1$, that is, $\Delta N > N$.

2. A rigorous solution for such problemsi is obtained with Eq. (81.3). The first step is to find the number $N(t)$ of nuclei remaining unchanged at time $t$. Then the number of nuclei disintegrating in time from $t = 0$ to $t$ will obviously be equal to $N_0 - N(t)$. This approach avoids errors, although it involves a greater amount of computation. Let us find when Eq. (81.1) may be used instead of Eq. (81.3). Substituting the relation between $\lambda$ and $T$ according to (81.4) in Eq. (81.1), we get:

$$\lambda N(t) \Delta t = -0.693 N(t) \Delta t/T$$

As is seen, the accuracy of the answer given by Eq. (81.1) improves with decrease in the ratio $\Delta t/T$, that is, with increase in the inequality $\Delta t \ll T$.

## 81.6. RADIOACTIVE DECAY AS A STATISTICAL PROCESS

1.The law of radioactive decay, Eq. (81.3), has been derived on the assumption that radioactive decay occurs spontaneously, and we cannot pin down the nucleus that will decay in a given time interval $\Delta t$. The point is that all nuclei of a given chemical element are *undistinguishable*. The best we can do is to find an average number of nuclei decaying in the time interval from $t$ to $t + \Delta t$. Thus, what we have is a statistical process, that is, the decay of a given nucleus is a random event having a certain probability of occurrence.

The decay probability per unit time per nucleus may be derived as follows. If we have $N$ original nuclei and the number decaying in a time $\Delta t$ is $\Delta N$*, then the relative decrease, $-\Delta N/N$, in the number of nuclei per unit time, that is, the quantity $-(\Delta N/N)/\Delta t$ gives the decay probability per unit time per nucleus.

2. This definition agrees with the meaning of the decay constant, $\lambda$, given by Eq. (81.2). By definition, the decay constant is the decay probability per unit time per nucleus.

It may further be assumed that the decay constant is independent of time. Physically, this means that for a spontaneous decay of a nucleus it is immaterial how "old" this nucleus is and that $\lambda$ is characteristic of all nuclei of a given kind. It can be shown that this definition of the decay constant leads to the exponential radioactive decay law, Eq. (81.3), which is thus a *statistical law.*

## 81.7. RADIOACTIVE DATING IN GEOLOGY AND ARCHAEOLOGY

1. The decrease in the number of radioactive nuclei according to Eq. (81.3) may be used as a means for measuring the time that passes between the instant when the number of radioactive nuclei is $N_0$ and the instant when their number is $N$. In other words, radioactivity provides a kind of time scale. According to Eqs. (81.3) and (81.4), the time interval between the instants when the number of radioactive nuclei is $N_0$ and $N$ is

$$t = (1/\lambda) \ln N_0/N = 1.44T \ln N_0/N \tag{81.7}$$

As a rule, $N$ represents the number of unchanged nuclei at the present time, so that Eq. (81.7) gives the age of given species of radioactive nuclei. Practically, a different radioactive time scale is required for each application. In determining the age of rocks that make up the Earth's crust, one should use a sufficiently slow radioactive time scale, that is, radioactive decays with a half-life of the same order of magnitude as geological epochs, running into hundreds of millions or even millions of millions of years. This condition is satisfied by the half-live of uranium-238 and uranium-235. Naturally occurring uranium is actually a mixture of both. Their half-lives are 4500 million and 900 million years, respectively.

At present, chemically pure and naturally occurring uranium contains 99.28% ${}_{92}\mathrm{U}^{238}$, 0.714% ${}_{92}\mathrm{U}^{235}$, and 0.006% ${}_{92}\mathrm{U}^{234}$, the latter being the decay product of uranium-238. Since its content is very small, uranium-234 may be neglected. Each of the uranium-238 and uranium-235 isotopes is the parent of a radioactive series of its own, both of which terminate in lead isotopes (see Sec. 81.2). Thus, lead nuclei are the end products of the radioactive decay of uranium nuclei. Using the ratio between uranium and the lead derived from

* Presumably, $\Delta N$ is sufficiently large, since otherwise $\Delta N/N$ cannot be used for the determination of decay probability.

it in natural uranium, one can readily determine the time interval during which this amount of lead has accumulated.

2. In archaeology, radioactivity is used to date the objects found in excavations. In such applications, the uranium time scale is unsuitable for at least two reasons. For one thing, artifacts have never contained uranium. For another, the uranium time scale clock is too slow for human history where time is usually measured in centuries or millenia. In other words, archaeological dating needs a radioactive time scale with a half-life of a few centuries or millenia. Nature has provided such time scales.

3. The particles that make up the so-called primary cosmic rays are extremely energetic and, colliding with the nuclei of the elements that form the Earth's atmosphere, break them up into fragments (see Sec. 83.2). These fragments are highly energetic, too, and form the so-called secondary cosmic rays. The interaction of cosmic rays with the nuclei of atmospheric nitrogen turns them into the nuclei of carbon with mass number 14, instead of 12, as with ordinary carbon. This carbon isotope, ${}_6C^{14}$, is radioactive, and its half-life is about 5570 years, which fits archaeologists well. Moreover, because the intensity of primary cosmic rays (Sec. 83.2) remains practically constant, there is an unvarying supply of radioactive carbon in the atmosphere. Radioactive carbon produces radioactive carbon dioxide which is assimilated by plants in precisely the same manner as stable $CO_2$. With plants, radioactive carbon finds its way into animals and becomes part of their organs and tissues.

In a living plant or animal, the per cent content of radioactive carbon in comparison with the ordinary carbon does not change with time, because any losses are made good by food. If, however, a plant or an animal dies, food cannot replenish the loss of radioactive carbon any longer. Thus, death starts the radioactive clock: the radioactive carbon content of the organism or of an artifact made of organic materials decreases according to the radioactive decay law, Eq. (81.3). Thus, one can determine the time passing since the death of the organism or the age of an artifact made of an organic material.

4. Using a charged particle counter (Sec. 81.8), it has been found (${}_6C^{14}$ decays by emission of beta-particles) that one gram of radioactive carbon contained in the cellulose of a living or a recently cut tree emits an average of 17.5 particles per minute. That is, the activity of the radioactive isotope is 17.5 decays per minute. Converting $T = 5570$ years into minutes, we find the number of ${}_6C^{14}$ nuclei that have this magnitude of activity:

$$N = (1/\lambda)(\Delta N/\Delta t) = 1.44T\,(\Delta N/\Delta t) = 1.44 \times 5570 \times 365 \times 24 \times 60 \times 1.75 \approx 7.5 \times 10^{10}$$

Thus, one gram of carbon in the cellulose of a living or a recently cut tree contains 75 000 million nuclei of radioactive carbon. This number progressively decreases because of radioactive decay. If the number lost through decay is not replenished (and this happens when the tree is cut), the original number will decrease with time according to Eq. (81.3). That is, the activity of the remaining radioactive carbon will decrease progressively. If we compare its activity at a time $t$ with that existing at $t = 0$, that is, when the tree was just cut down, we can determine the time interval between these two instants. According to (81.5), we may write the following relations for $t = 0$ and for the present time, $t$:

$$Q_0 = |(\Delta N/\Delta t)_0| = \lambda N_0$$

$$Q_t = |(\Delta N/\Delta t)_t| = \lambda N(t)$$

According to the radioactive decay law, however,

$$N(t) = N_0 \exp(-\lambda t)$$

Dividing the first relation by the second termwise, we get

$$Q_0/Q_t = \exp(\lambda t)$$

whence the sought time interval is

$$t = (1/\lambda) \ln Q_0/Q_t = 1.44\, T \ln Q_0/Q_t$$

5. When this technique is applied to wooden artifacts usually found in archaeological excavations, one actually finds the time at which a tree was cut. This gives the age of the artifacts made from it.

As an example, let in a wooden artifact found in an excavation one gram of carbon emit 350 particles registered by a counter over 40 min. Then the activity is $350 \div 40 = 8.75$ decays per minute, and the age of the artifact is

$$t = 1.44T \ln(17.5/8.75) = (1/\ln 2)T \ln 2 = T = 5570 \text{ year}$$

## 81.8. DETECTION OF RADIOACTIVE RADIATIONS AND PARTICLES

1. Present-day nuclear physics uses a variety of techniques for detecting radioactive radiations (alpha and beta particles and gamma-quanta) and fundamental particles. At the basis of these techniques are the ionizing and photochemical effects of the particles under study. We shall examine some of them.

Sir William Crookes observed that a screen coated with small crystals of phosphorescent zinc sulphide displayed a brilliant luminosity when bombarded by alpha-particles. When the surface of the screen was examined with a magnifier, the light from the screen was found to consist of individual flashes of light, or *scintillations*. Experiments proved that each scintillation was caused by a single alpha-particle striking the screen,and this was used to count alpha-particles.

A visual observation technique consisting of a phosphorescent screen and a microscope was used in early *scintillation counters*. However, it was extremely tiresome. In the modern scintillation counters that have appeared since the 1940s, the microscope and human eye combination has been replaced with a photomultiplier tube. The scintillations from the fluorescent screen are converted into electrical pulses at the output of the photomultiplier tube. These pulses are boosted by an amplifier and then drive an electromechanical pulse counter. Additionally, the pulses may drive an oscilloscope which displays waveforms representing the intensity of each particle. Thus, this technique determines both the number of particles and their energy distribution.

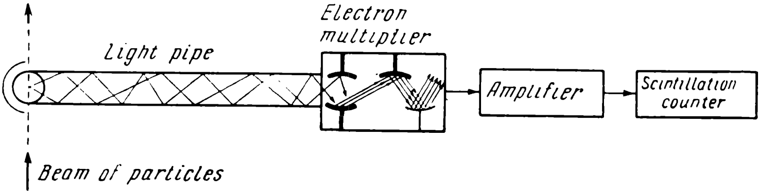


Fig. 81.3

In sketch form, a scintillation counter of this type is shown in Fig. 81.3. In order that the greater proportion of light emitted by each scintillation can reach the photocathode, the scintillator crystal is optically coupled to the photomultiplier tube by an acrylate "light pipe". Inside this "light pipe", the light is propagated by continuous total internal reflection, without practically any losses.

2. Charged particles travelling through a transparent material at a velocity exceeding the phase velocity of light in that material and also gamma-quanta (from the secondary electrons they produce) can be counted with Cerenkov radiation counters or detectors. In sketch form, a Cerenkov counter or detector is shown in Fig. 81.4. The charged particles to be counted are made to pass through an acrylic-plastic block with a refractive index of $n = 1.5$. Cerenkov radiation (see Sec. 59.7) is emitted at an angle (the Cerenkov angle) $\theta$ such that $\cos\theta = 1/(v/c)n$ and focused by the spherical surface of the plastic block, *1*, and reflected by mirrors, *2*, onto the photocathodes, *3*, of two photomultipliers placed outside the path of the moving particles. At $n = 1.5$, the Cerenkov counter can detect electrons with energies in excess of 0.18 MeV, protons with energies upwards of 320 MeV, and gamma-quanta producing secondary electrons of sufficiently high energies.

Since Cerenkov counters determine the angle at which the radiation is emitted, they can also determine the direction in which the particles are travelling. From the Cerenkov angles, it is possible to determine the velocities and energies of the moving particles. At present, Cerenkov counters are installed on space probes as a means of studying cosmic radiations. They have played a major role in the discovery of the anti-proton and the anti-neutron (Sec. 83.7).

3. The passage of a charged particle through a gaseous medium causes ionization of the latter. This effect is the basis of ionization chambers used for the observation of charged particles (Sec. 48.2). If the energy of the particles passing through a medium exceeds its ionization energy, such particles are capable of producing primary or secondary ions of both polarities. Primary ions are directly produced by alpha-and beta-rays, and secondary ions by X-rays or gamma-rays. In the latter case, the incident radiation first produces secondary electrons (photoionization), and these then produce the ionization of the molecules or atoms of the gas. Neutrons are detected by their interactions with nuclei.

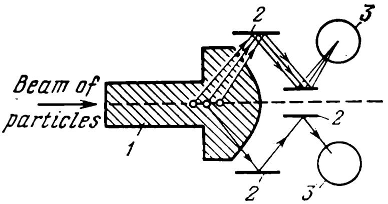


Fig. 81.4

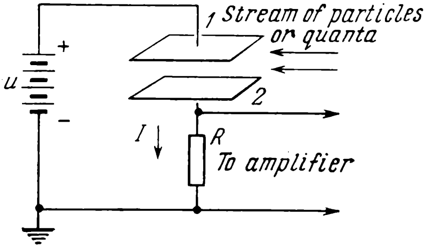


Fig. 81.5

The number of ionic pairs formea in the medium per unit time gives a measure of the intensity of the particles or quanta causing the ionization. This number can be determined by electrical detection methods in which the ions produced by the passage of a charged particle are separated by an electrostatic field. The ions move towards the electrodes where they are detected as a mean ionization current proportional to the number of ionic pairs produced per second, that is, to the intensity of the particle stream causing the ionization. This proportionality is observed only in the *saturation current region* (Sec. 48.2), when all ions pass on to the electrodes rather than recombine or diffuse through the walls. Devices operating in the saturation current region are called *ionization chambers*. Schematically, an ionization chamber is shown in Fig. 81.5. The electrodes, *1* and *2*, may be flat, cylindrical or spherical. A voltage, $u$, of several hundred volts is applied to electrode *1*. Electrode *2*, called the *collector*, is connected to an amplifier. The mean ionization current is derived from the voltage drop across a high-value resistor, $R$.

4. Some charge-detecting devices operate in the gas-multiplication (or gas amplification) mode (see Sec. 48.4). Of special importance to their operation is a gas-discharge region called the *Geiger* region (after Geiger who discovered it in cooperation with Mueller in 1928). The Geiger region is characterized by a strong discharge due to collisions and, by the marked effect of ultraviolet light of the discharge which knocks photoelectrons out of the molecules and atoms of the gaseous medium and out of the chamber walls. In the Geiger region, the ionization current is independent of the number of primary ions formed by each primary ionizing particle entering the counter. The type of counter using this mechanism is called the Geiger (Geiger-Mueller or G-M) counter (for its description, read Sec. 48.4).

5. A major contribution to the study of fundamental particles was the *cloud chamber* invented by C.T.R. Wilson in 1912. Its operating principle and design have been described in Sec. 36.9. As a further refinement, proposed by D. V. Skobeltsyn of the Soviet Union, it is usually placed in a strong and uniform magnetic field (the Wilson-Skobeltsyn technique).

In this field, charged particles are subject to Lorentz forces, and their paths are curved (Sec. 41.2). Once the radius of curvature and velocity of a particle are known, it is an easy matter to determine its specific charge. Conversely, if the specific charge is known, the radius of curvature makes it possible to find the velocity and energy of the particle.

6. The path of an ionizing particle may be made visible by letting it pass through a superheated liquid (see Sec. 36.10) which boils as its pressure is suddenly reducedbyletting it expand.The ions forming along the path of the charged particle act as centres on which a string of bubbles can grow. This principle is utilized in the *bubble chamber* first proposed by Glaser in 1952. The liquids that may be used in bubble chambers are liquid hydrogen, propane ($C_3H_8$) and other low-boiling-point liquids (mainly Freons). A major advantage of the bubble chamber over the Wilson cloud chamber is that the liquid used is about $10^3$ times denser, and highly energetic particles are retarded over distances a few thousandths of those in the Wilson cloud chamber. That is, the track of a particle photographed in a bubble chamber corresponds to a path thousands of times that in the cloud chamber. Practically, it means that a bubble chamber can register a complete path of a very fast particle where a cloud chamber can only cover a small fraction thereof.

7. One more method used for the detection of charged particles is the *nuclear emulsion technique*. In the Soviet Union, it has been developed by L. V. Mysovsky and A. P. Zhdanov. This technique utilizes the blackening of photographic emulsion caused by the passage of fast charged particles. Nuclear emulsions are used in thicknesses

from 0.5 to 1 mm, so that one can investigate the paths of highly energetic particles. For example, a particle with an energy of about 10 MeV produces a track about 0.1 mm long within the emulsion.

With very energetic particles that would travel many tens of metres in air, it is possible to form a block of any desired thickness by using emulsion strips separated by thin sheets of paper and arranged at an angle to the path of the particle. Then the consecutive segments of the particle track can be investigated from the blackened areas on the emulsion strips following one after another.

Apart from its simplicity, the nuclear emulsion technique offers the advantage of producing permanent records of any nuclear event occurring within the block. This is why the nuclear emulsion technique has been widely used in the study of new fundamental particles and in space exploration (blocks of emulsion have been flown in sounding rockets, artificial Earth satellites, and space probes). With additions of boron or lithium, blocks of emulsion can be used to investigate the trajectories of neutrons which interact with the boron or lithium nuclei and produce alpha-particles that cause the emulsion to blacken. From these tracks left by alpha-particles it is possible to determine the velocities and energies of the primary neutrons.

## 81.9. THEORY OF RADIOACTIVE DECAY

1. A good deal of light was thrown on the mechanism of alpha-decay by Rutherford's experiments on the scattering of alpha-particles by the uranium nuclei. It was found that alpha-particles with an energy of 8.8 MeV were repelled from the nucleus according to the Coulomb law at any distance from the nucleus up to 30 fm. Hence, the height of the Coulomb potential barrier in the uranium nucleus must be at least 8.8 MeV. On the other hand, the alpha-particles emitted by the uranium nucleus were found to have an energy of 4 MeV, or less than the height of the potential barrier. It was concluded, therefore, that the release of an alpha-particle by the nucleus in the case of alpha-decay is due to the tunnelling effect (Sec. 70.6), that is, the quantum-mechanical penetration through the potential barrier by the particle. With this discovery, it has become possible to explain some experimental facts about alpha-decay. Among them is the empirical *Geiger-Nuttall relation* which states that the alpha-particle range increases with the decay constant. Quantitatively, this relation is expressed as follows:

$$\ln \lambda = A + B \ln R \tag{81.8}$$

where $R$ is the alpha-particle range in air, at 0°C, $\lambda$ is the decay constant, $A$ and $B$ are empirical constants having the same values for the members of a given natural radioactive series.

2. The theory of alpha-decay explains why the alpha-particles emitted by elements occupying adjacent places in the Periodic Table differ but little in energy and very markedly in half-life. For example, the initial velocities of the alpha-particles emitted by the polonium isotopes, ${}_{84}\mathrm{Po}^{218}$ and ${}_{84}\mathrm{Po}^{214}$, are 1.68 × $10^7$ m/s and 1.92 × $10^7$ m/s, that is, the difference is small, while their half-lives are 3.05 min and $10^{-6}$ s, or differ greatly.

3. According to quantum mechanics (Sec. 70.6), no matter how high the potential barrier may be, there is a non-zero probability of tunnelling through the barrier at energies lower than the barrier height. As applied to alpha-decay, this means that a radioactive nucleus may emit alpha-particles with energy of a few MeV, which is lower than the barrier height. For a quantitative consideration of the tunnel effect, it is important to know the shape of the potential barrier, that is, the dependence of the potential energy of the alpha-particle on its radial distance from the centre of the nucleus. In approximate form, this dependence is shown in Fig. 81.6. The outer wall of the barrier due to the Coulomb repulsion of the alpha-particle from the nucleus is a hyperbola. The shape of the inner wall is determined by the dependence of nuclear force on distance. Since nuclear forces have a far shorter range than the Coulomb forces, the inner wall of the potential barrier is much steeper than the outer one. In Fig. 81.6, it is represented by a vertical line.

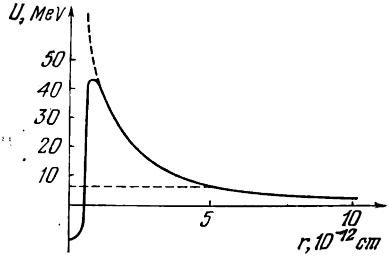


Fig. 81.6

As already stated in Sec. 70.6, the transparency of the barrier strongly depends on its shape. But even for a fairly simple barrier such as shown in Fig. 81.6, a mathematical solution of the problem of alpha-decay is not at all simple. It is instructive, therefore, to consider a rectangular barrier, least complicated of all potential barriers, a plot for which is shown in Fig. 70.6. The transparency of this barrier is given by Eq. (70.30):

$$D \approx \exp\left[-2L\sqrt{2m(U_0-\mathscr{E})}/\hbar\right] \qquad (81.9)$$

The notation used is explained in Para. 3, Sec. 70.6.

The transparency of the barrier can readily be connected to the decay constant, λ. On approaching the barrier, an alpha-particle contained in the potential well may either be reflected from the barrier and remain in the well or "penetrate" the barrier and leave the nucleus. The transparency is the probability of escape of an alpha-particle from the nucleus as a result of a single attempt to penetrate the barrier. By definition, the decay constant, λ, is the decay probability per unit time per nucleus (see Sec. 81.3). Thus, to find the decay constant, the transparency must be multiplied by the number of times that the alpha-particle attempts to penetrate the barrier wall, or the escape repetition rate, $n$:

$$\lambda = Dn \tag{81.10}$$

The escape repetition rate $n$ is the reciprocal of the time $\tau$ during which an alpha-particle traverses the nucleus, that is, the distance $L$ equal to the nuclear diameter, $L = 2r_0$, where $r_0$ is the nuclear radius. Designating the velocity of an alpha-particle as $v$, we find that the average time for traversal of the nucleus or the time interval between successive attempts to escape is

$$\tau = 2r_0/v$$

and so the escape repetition rate is

$$n = 1/\tau = v/2r_0 = \sqrt{\mathscr{E}/2mr_0^2} \tag{81.11}$$

where the velocity of an alpha-particle, $v$, is expressed in terms of its energy $\mathscr{E}$ according to the relativistic equation $\mathscr{E} = mv^2/2$ to simplify calculations. Substituting (81.11) and (81.9) in (81.10), we finally get:

$$\lambda = \sqrt{\mathscr{E}/2mr_0^2}\,\exp\left[-2\sqrt{2m(U_0-\mathscr{E})}\,2r_0/\hbar\right] \tag{81.12}$$

4. Equation (81.12) relates the decay constant to the energy of the emitted particles. It should include the Geiger-Nuttall relation. Taking a logarithm of (81.12), we get

$$\ln\lambda = \ln\sqrt{\mathscr{E}/2mr_0^2} - (2/\hbar)\sqrt{2m(U_0-\mathscr{E})}\,2r_0 \tag{81.13}$$

The first term remains practically constant within each radioactive series. Therefore, Eq. (81.13) may be given a form not unlike the Geiger-Nuttall relation:

$$\ln\lambda = A + Bf(\mathscr{E}) \tag{81.8'}$$

The second term of (81.8′) is not a logarithmic function of the alpha-particle range, as it is in the Geiger-Nuttall relation, but this discrepancy can readily be explained. The point is that the Geiger-Nuttall relation is anything but precise, as is the theory set forth above. So, no precise agreement ought to be expected between observation and theory. There is, however, a qualitative agreement, because according to (81.12) or (81.13) the decay constant increases with increase in the energy of the released alpha-particles and, as a consequence, in their range. This might be visualized as follows: as the energy of an alpha-particle in the nucleus increases, it becomes increasingly more difficult for the nucleus to hold the particle, and the probability that the nucleus will decay by alpha-emission increases.

5. Equation (81.12) explains why, despite the small difference in velocity between the emitted alpha-particles, the emitting nuclei may differ so greatly in their half-lives. This is because the energy of an alpha-particle, $\mathscr{E}$, is in the exponent of Eq. (81.9), and exponential functions are strongly sensitive to changes in the exponent.

A more rigorous theory of alpha-decay, allowing for the shape of the potential barrier shown in Fig. 81.6, yields a far more elaborate equation which connects the decay constant to the energy of the alpha-particle and the atomic number of the element in the Periodic Table. This equation fits observations and is a theoretical improvement upon the empirical Geiger-Nuttall relation.

6. Observations have shown that each radioactive element emits *groups of monoenergetic alpha-particles*, that is, all particles in a particular group have nearly the same energies. Therefore, it may be said that the emitted alpha-particles have a *line energy spectrum*. This suggests an analogy between the electronic shells of the atom and its nucleus. Much as the line spectrum of the photons emitted by the atom stems from the quantization of the energy of the atom, so the line energy spectrum of the alpha-particles emitted by the nuclei of a given element is an indication that the energy of the nucleus is quantized, that is, *it is restricted to a discrete set of values.*

### 81.10. GAMMA-RAYS

1. Observations have shown that the emission of gamma-rays by nuclei is not, as a rule, an independent kind of radioactivity. They are companions to alpha- and beta-decays. Let us see how the nucleus emits gamma-rays.

On emitting an alpha-particle, the parent nucleus turns into a daughter nucleus. As a rule, a daughter nucleus is in the excited state. Jumping to the normal or a lower excited state, the daughter nucleus emits a gamma-photon, much as an excited atom emits a photon of optical radiation or of X-rays in jumping to the normal state. The nucleus emits gamma-rays by the same mechanism as the atom emits photons. A very important difference is that gamma-photons have far higher energies than optical photons. This is because the difference in energy between the various levels in the nucleus is far greater than it is between the levels in the atom. While in atoms, the spacing between the energy levels is of the order of one electron volt, in the nuclei this spacing, as measured from the energies of gamma-photons, is about 0.1 MeV.

These estimates for the energies of electronic and nucleonic transitions can readily be obtained by using the equation

$$\Delta\mathscr{E} = \hbar^2/2ma^2$$

(see Eq. (16.24)) and assuming that the electron and the nucleon are trapped in an atomic region with $a \approx 10^{-10}$ m and in a nuclear region with $a \approx 10^{-14}$ m, respectively. Then

$$\Delta\mathscr{E}_e \approx \frac{10^{-68}}{2\times10^{-30}\times10^{-20}\times1.6\times10^{-10}} \approx 1 \text{ eV}$$

$$\Delta\mathscr{E}_N = \frac{10^{-68}}{2\times1.6\times10^{-27}\times10^{-28}\times1.6\times10^{-19}} \approx 0.1 \text{ MeV}$$

Thus, gamma-rays are an electromagnetic radiation of short wavelengths not exceeding $10^{-11}$ m, or 0.1 A.

2. *The gamma-rays accompanying the alpha-decay of a parent nucleus are emitted by the daughter nucleus.*

To prove this statement, we shall examine the alpha-decay of radium to radon. If a radium nucleus emits an alpha-particle with a maximum energy, the daughter nucleus will have a minimum energy, that is, it will be in the normal state. Conversely, if a radium nucleus emits an alpha-particle with a lower energy, the daughter nucleus will be in an excited state. The difference in energy between the excited and normal states of the daughter nucleus must be equal to the energy of the released gamma-photon.

Observations have fully corroborated the above reasoning.

3. From studies into the energy spectra of alpha-particles and the energies of gamma-photons, it is safe to conclude that *gamma-rays have a line spectrum*, not unlike that of alpha-particles. Incidentally, the energy spectra of alpha-particles and gamma-rays offer a powerful tool for the study of the energy levels of atomic nuclei.

4. The high energy of gamma-photons makes them highly penetrating. In fact, their penetrating ability is even greater than that of X-rays.

The high penetrating ability of gamma-rays is utilized in isotope radiography, a non-destructive method of testing materials widely used in metallurgy, shipbuilding and other industries. The intensity of the gamma-rays reaching a detector located on the far side of the workpiece varies according to the composition, thickness, density and other properties of the material. Since any flaws in, say, welded joints, castings and the like are actually departures from the normal properties, they are duly revealed by the emergent beam of gamma-rays as cracks, blowholes, lack of fusion in welded joints, etc. From a gamma-ray radiograph one can readily identify the location, size and shape of such flaws.

5. Like X-rays, gamma-rays are absorbed by the medium through which they pass and thus may affect its properties. As with other kinds of ionizing radiation, the effect of gamma-rays on substances is estimated in terms of the absorbed *dose of radiation*, $D$.

It is defined as the ratio of the energy delivered with radiation to the mass of the body absorbing the radiation. The unit is *one joule* *per kilogram*, J/kg, that is, a dose such that a mass of 1 kg receives an energy of 1 Joule with ionizing radiation.

An off-system unit is the *rad* (radiation absorbed dose) defined as

$$1\ \text{rad} = 10^{-2}\ \text{J/kg}$$

The absorbed dose per unit time gives the *dose rate*, $N$:

$$N = D/t$$

Its unit is one watt per kilogram, W/kg.

6. The *exposure dose*, $D_e$, is measured in *coulombs per kilogram*, C/kg, and expresses the energy of radiation as estimated from the ionization of dry atmospheric air by photons, X-rays or gamma-rays. More specifically, it is a dose for which the sum of ions of the same sign produced by the electrons released by irradiated air with a mass of 1 kg at the full utilization of the ionizing capacity is 1 coulomb.

An off-system unit of the exposure dose is the *roentgen*, R, defined as

$$1\ \text{R} = 2.58 \times 10^{-4}\ \text{C/kg}$$

One roentgen represents an exposure dose such that the sum of charges due to ions of the same sign in one cubic centimetre of dry air under normal atmospheric pressure is equal to one absolute electrostatic unit of charge.

The exposure dose per unit time gives the *exposure dose rate*, $N_e = D_e/t$, measured in *amperes per kilogram*, A/kg. One unit of exposure dose rate represents an exposure dose of a photon radiation such that the exposure dose increases by one coulomb per kilogram in one second.

The off-system units of the exposure dose rate are:

$$1\ \text{R/s} = 2.58 \times 10^{-4}\ \text{A/kg}$$

$$1\ \text{R/min} = 4.30 \times 10^{-6}\ \text{A/kg}$$

$$1\ \text{R/h} = 7.17 \times 10^{-8}\ \text{A/kg}$$

7. The biological effect produced by an ionizing radiation is measured in terms of an equivalent dose. The unit is one joule per kilogram, J/kg.

An off-system unit is the *roentgen-equivalent-man* (rem). This is the dose which produces the same biological effect as that produced by one roentgen of high-voltage X-radiation:

$$1\ \text{rem} = 10^{-2}\ \text{J/kg}$$

There are equations which relate the dose rate to the activity of a radiation source having a definite geometry and located a definite distance from the detector.

To man, the safe dose rate is about 250 times that produced by cosmic rays and the radioactive radiations from inside the Earth. The lethal dose for man is in excess of 500 R absorbed singly.

## 81.11. THE MOESSBAUER EFFECT

1. Although we say that atoms and gamma-rays have line spectra, that is, spectra composed of individual monochromatic lines, it must be remembered that a purely monochromatic radiation is nonexistent. This is because an atom or a nucleus can remain in an excited state for a finite length of time. It is only in the ground or normal state in which its energy is a minimum that an atom or a nucleus can reside for any time, however long. All excited states of a nucleus have energies which can be determined accurate to $\Delta\mathscr{E}$, according to the uncertainty relation (see Sec. 70.2):

$$\Delta\mathscr{E} \approx \hbar/\Delta t \tag{81.14}$$

where $\Delta t$ is the life-time of the nucleus in an excited state. For the ground state of a stable nucleus, $\Delta t = \infty$ and $\Delta\mathscr{E} = 0$. The shorter the time range $\Delta t$, the greater the uncertainty in the energy of the excited state, $\Delta\mathscr{E}$.

The uncertainty in the energy of an excited state leads to the uncertainty in the frequency of the gamma-photons released by an excited nucleus as it disintegrates. Hence, the gamma-radiation of nuclei cannot be monochromatic. Let us estimate this lack of monochromaticity.

2. As an example, let us consider the iridium (${}_{77}\mathrm{Ir}^{191}$) nucleus. In the excited state, it has an energy of $\mathscr{E} = 129$ keV. The uncertainty in energy in this state is $\Delta\mathscr{E} \approx \hbar/\Delta t$, where $\Delta t$ is the length of time during which the nucleus remains in the excited state. For the purpose of an estimate, we let $\Delta t$ be equal to the decay half-life of the nucleus. For the iridium isotope which we have chosen as an example, this is $\Delta t = T = 10^{-10}$ s. Then, on the basis of (81.14), the uncertainty in the energy state with an energy of 129 keV will be $5 \times 10^{-6}$ eV. This quantity is called the *natural energy level width*, $\Gamma$, and the quantity $\Delta\nu = \Delta\mathscr{E}/h \approx 1/\Delta t$ is called the *natural line width* for gamma-radiation of frequency $\nu = \mathscr{E}/h$ (see Sec. 72.8). In our case, the partial energy level (or spectral line) width is $\Gamma/\mathscr{E} = 4 \times 10^{-11}$ The ratio $\Gamma/\mathscr{E}$ describes the lack of monochromaticity in gamma-radiation.

3. A very important problem in nuclear physics has been to find methods for measuring minute changes in energy, comparable with the natural level width, $\Gamma$. With such methods, one can measure the energy of levels in nuclei to a very high degree of accuracy in $\Gamma/\mathscr{E}$.

One such method utilizes the *resonance absorption* of gamma-rays by nuclei. If a nucleus absorbs gamma-rays at frequency $\nu$, we may say that the difference in energy between its excited and ground states is $\Delta\mathscr{E} = h\nu$. If the absorbed gamma-rays are of just the right energy to raise the absorbing nucleus to the excited state, we have what is known as *resonance absorption*. This happens when the nucleus absorbs gamma-rays of the same frequencies as it can emit. A similar situation has been discussed in Sec. 71.4 in the optical region: an atom absorbs light only at the frequencies that it can emit.

Until 1958, however, this method could not be put to practical use because the energy of the emitted or absorbed gamma-rays was found to differ from the energy required for nucleus to undergo a transition from level to level. This difference is due to the recoil of the nucleus as it emits or absorbs a gamma-photon, which is necessary to conserve momentum. In other words, the emitting nucleus receives a momentum equal in magnitude to that of the emitted photon, but in the opposite direction. The recoil energy, $\mathscr{E}_R$, is given by:

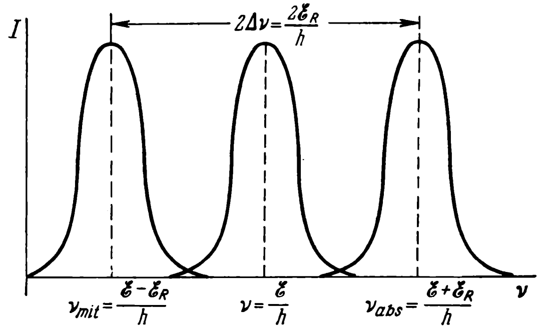


Fig. 81.7

$$\mathscr{E}_R = p_R^2/2m_R = p_{ph}^2/2m_R = (h\nu/c)^2(1/2m_R)$$
$$= (h\nu)^2/2m_Rc^2 = \mathscr{E}^2/2m_Rc^2 \qquad (81.15)$$

where $p_R$ the momentum imparted to the nucleus due to recoil, numerically equal to the photon momentum, $p_{ph} = h\nu/c$ (see Sec. 68.5), and $m_N$ is the nuclear mass. Substituting the numerical values, we find that $\mathscr{E}_R \approx 0.05$ eV. This is much greater than $\Gamma$, the natural energy level width. Thus, the recoil reduces the energy of the emitted gamma-photon by $\mathscr{E}_R$, and its frequency by $\Delta\nu = \mathscr{E}_R/h$. Accordingly, $h\nu_{emit} = \mathscr{E} - \mathscr{E}_R$.

Therefore, in order that a nucleus with an excitation energy $\mathscr{E}$ can absorb a photon, the energy of the photon should exceed $\mathscr{E}$ by the recoil energy of the absorbing nucleus. The energy of the absorbed photon should be sufficient not only to excite the nucleus, but also to supply the recoil energy to the absorbing nucleus (Fig. 81.7):

$$h\nu_{abs} = \mathscr{E} + \mathscr{E}_R$$

Thus, the frequency shift between the gamma-photons that a nucleus can emit and absorb is $2\Delta\nu = 2\mathcal{E}_k/h$, and the absorption is not resonant. This reduces the value of the method as a whole. Even if we observe the absorption of gamma-rays at a definite frequency, we cannot determine the energy levels of the nucleus without further analysis.

4. In 1958, Moessbauer developed a method by which the recoil energy loss can be reduced practically to zero, thereby making the observation of resonance absorption possible. Accordingly, the phenomenon of *recoil-free gamma-ray resonance absorption* has come to be known as the *Moessbauer effect*. Moessbauer's idea can be understood from reference to Eq. (81.15): the recoil energy of a nucleus decreases with an increase in its mass. Of course, the rest mass of a nucleus cannot be increased. Instead, Moessbauer discovered that the effect of an increase in the mass of the nucleus could be obtained if the radioactive nucleus were firmly held in a crystal lattice, so that the recoil momentum of the gamma-ray emission could be taken up by the entire crystal rather than by the radioactive nucleus alone. Under such conditions, the recoil energy loss may be taken equal to zero, and the gamma-rays emitted from the source will have just the right energy (frequency) to be resonantly absorbed. The lines in the emission and absorption spectra have a width of about the same order as the natural width, so that minute differences in energy or frequency can be measured accurate to within the natural level or line width. The value of $\Gamma/\mathcal{E}$ for gamma-transitions in the $Fe^{57}$ nucleus with a nuclear transition energy of $\mathcal{E} = 14.4$ keV has been found to be $3 \times 10^{-13}$, and for gamma-transitions in the $Zn^{87}$ nucleus with a nuclear transition energy of $\mathcal{E} = 93$ keV, it is $5 \times 1\ ^{-16}$.

5. Owing to its capability to measure extremely small changes in energy to a high degree of accuracy, the Moessbauer effect has found uses as a means for determining very subtle effects in present-day physics. For example, it was used in 1960 to measure the effect of gravitation on source energy at different heights under laboratory conditions (the so-called "red shift"). This effect was mentioned in Sec. 24.6 in connection with Einstein's theory of gravitation. Now we shall discuss it in greater detail.

When a photon moving in a gravitational field travels from a point with a gravitational potential $\varphi_1$ to a point with a gravitational potential $\varphi_2$, its energy changes by $\Delta\mathcal{E} = -m(\varphi_2 - \varphi_1) = -m\Delta\varphi$. In a homogeneous gravitational field, this change will be $\Delta\mathcal{E} = -mg\Delta h$ (see Sec. 18.4). Much as an increase in the potential energy of a body moving in a gravitational field brings about a proportional decrease in its kinetic energy, so the increase in the gravitational energy of a photon entails a decrease in its "intrinsic" energy, $\mathcal{E} = h\nu$. This explains why there is a minus sign in the relation

$$\Delta\mathscr{E} = -m\Delta\varphi \text{ or } h\Delta\nu = -m\Delta\varphi \qquad (81.16)$$

The mass of a photon is related to its energy and frequency by the expression

$$m = \mathscr{E}/c^2 = h\nu/c^2$$

The fractional change in frequency as the photon moves through a gravitational potential difference $\Delta\varphi$ is

$$\Delta\nu/\nu = -\Delta\varphi/c^2 \qquad (81.17)$$

This relation manifests itself as follows. Suppose that an observer on the Earth has registered a particle of solar radiation. The potential of the Sun's gravitational field on the Earth's surface is greater than it is on the Sun's surface ($\Delta\varphi > 0$) and, according to (81.17), $\Delta\nu/\nu$ is negative. As a result, all spectral frequencies of the Sun and stars observed on the Earth will be reduced, that is, shifted towards the red end of the spectrum. This is why the effect is called the *gravitational red shift*. For solar radiation, the gravitational red shift is

$$\Delta\nu/\nu = -\Delta\varphi/c^2 \approx 10^{-6}$$

In order to see whether this is a small or a large quantity, we shall replace frequency with wavelength. As has been shown in Sec. 61.4, the fractional change in wavelength is equal to the fractional change in frequency, that is, $\Delta\nu/\nu = -\Delta\lambda/\lambda$. The minus sign indicates that the wavelength increases as the respective frequency decreases. On the average, the wavelength of the spectral lines coming from the Sun is $\lambda = 5000$ Å and its gravitational red shift is $\Delta\lambda = 10^{-6}\,\lambda = 5 \times 10^{-3}$ Å. That is, the red shift affects the third decimal place. If the spectral line of radiation coming from the Sun has a wavelength of 5000 Å, on the Earth it will be 5000.005 Å. This is an extremely fine effect even for present-day physics, although the Sun's gravitational field is far stronger than the Earth's.

6. Using the Moessbauer effect, the gravitational red shift has been measured under laboratory conditions from the motion of a photon in the Earth's weak gravitational field over short distances. The idea of the experiment is very simple. If a gamma-photon having a frequency $\nu$ at the laboratory's floor be directed upwards, its frequency at the laboratory's ceiling will be reduced by the gravitational red shift. On measuring this shift as accurately as practicable and comparing it with that deduced theoretically, one can test the validity of the prediction made by the general theory of relativity about the red shift. Let us determine the requirements that the experimental apparatus should satisfy.

On moving vertically in the Earth's gravitational field to a height of 10 m, the gravitational red shift will be

$$|\Delta\nu/\nu| = \Delta\varphi/c^2 = g\Delta h/c^2 \approx 10 \times 10 \div (10 \times 10^{16}) = 10^{-15}$$

This shift can be registered if gamma-photons are absorbed resonantly, so that the fractional width of the absorption line is less than $\Delta\nu/\nu$. Conversely, it is important that there be no resonant absorption, if the frequency of the gamma-photon incident on the nucleus differs by $\Delta\nu = 10^{-15}\nu$ from that of the photon which the nucleus can absorb. In other words one needs gamma sources and detectors for which the relative line width would be less than or equal to $10^{-15}$.

This requirement has been satisfied through the use of the Moessbauer effect. In the basic Moessbauer experiment, use is made of a gamma-ray source and a gamma-ray detector. As long as the detector is held level with the source, the emitted gamma-rays are absorbed resonantly. No resonance absorption occurs when the detector is lifted 20 m above the source, because the frequency of the incident photon is now reduced by the gravitational red shift. It has been found that the resonance absorption can be restored by introducing an appropriate velocity between the source and detector; the resulting Doppler shift (see Sec. 59.8) compensates for the change in the source frequency. The velocity at which the detector should approach the source to provide the necessary Doppler shift can readily be determined by calculation. For the first time this experiment was carried out in 1960 (see Sec. 24.6) and came as a very precise confirmation of the theory of relativity under terrestrial conditions.

At present, the Moessbauer effect is widely used in nuclear spectroscopy, for precise measurements of energy levels in nuclei, and in the study of many fine effects in solid-state physics. They lie outside the scope of this book.

## 81.12. THEORY OF BETA-DECAY

1. At first sight, the proton-neutron structure of the nucleus rules out the emission of electrons by the nucleus for it has none. This is why a theoretical interpretation of beta decay has been a difficult problem for nuclear physics. The very term "decay" should be used with qualifications*.

Nuclear beta-decay is the process whereby an unstable nucleus transforms itself into a more stable nucleus of the same mass number by converting one of its protons into a neutron or vice versa. The emission of beta-particles is not unlike the emission of photons by atoms. Much as an excited atom contains no "ready-made" photons

* Beta-decay with emission of a negative electron is symbolized $\beta_-$, and that with emission of a positive electron, viz. a positron, is symbolized $\beta_+$ (see Sec. 82.2).

and they are produced as the atom undergoes a transition from state to state, so the nucleus contains neither electrons nor positrons. They are born as a given nucleon changes from one quantum state to another, say, from the neutron to the proton state by the emission of an electron. Incidentally this is one of the arguments in favour of the present-day concept that the proton and the neutron are simply two quantum states of the same nucleon (see Sec. 80.3).

2. A proper understanding of the mechanism by which beta-decay occurs was handicapped by the energies emitted by beta sources. The emission of a beta-particle is the outcome of a transition of a nucleus from one discrete energy state to another, because nuclei have quantized energy levels, as is confirmed by the quantized energies of alpha- and gamma-rays. It would appear that beta-particles should likewise have a discrete energy spectrum, that is, one representing a set of energy values allowed for electrons. Nothing of the kind is however observed in experiments. *The energy spectra of the electrons emitted in beta-decay are always continuous.*

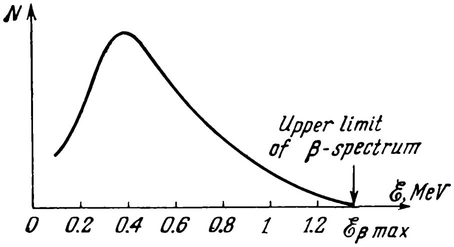


Fig. 81.8

The energy spectrum of electrons emitted by naturally radioactive potassium, ${}_{19}\mathrm{K}^{40}$, is shown in Fig. 81.8. The energy of beta-particles is laid as abscissa, and the number of beta-particles having the same energy as ordinate. As is seen, the beta electrons have a continuous spectrum. Identical nuclei emit electrons having all possible values of energy from zero up to a certain upper limit. The existence of this upper limit is very important. Instead of tending to zero asymptotically, the curve crosses the $x$-axis abruptly at $\mathscr{E}_{\beta\,\mathrm{max}}$. Each beta source has an upper limit of its own. For a given source, the electrons cannot have energies exceeding $\mathscr{E}_{\beta\,\mathrm{max}}$. Studies into the maximum energies of beta-particles from different sources showed that in all cases the *maximum energy of beta-particles was equal to the difference between the energy levels of the beta-emitting nuclei.*

This discovery only added to the difficulties in explaining why the electrons emitted in beta-decay displayed a continuous energy spectrum. For, if nuclei could only occupy certain energy states (as follows from the foregoing), why was it that beta-active nuclei which had certain definite energies before and after beta-decay released electrons having any values of energy? It could be conceived that all electrons as they were released from the nucleus had the same energy equal to the difference between the energy states in the emitting nucleus, but they lost varying amounts of energy as they went through the specimen because they collided with its atoms. If this assumption were true, the beta-active specimen would have warmed up itself. However, careful calorimetric tests failed to detect any warm-up. It appeared as if energy were lost in a mysterious way. Physicists faced the risk of admitting that the law of conservation of energy could be violated after all.

In fact, some investigators, including Bohr, assumed that the law of conservation of energy might apply to nuclear processes only in a statistical sense, that is, to a huge number of elementary processes. However, no prior advances of quantum theory in atomic and nuclear physics had disproved any of the conservation laws. On the contrary, the law of conservation of energy was consistently confirmed for elementary processes with an amazing accuracy. In those formative years, the theory of beta-decay came as a painful test for quantum mechanics and nuclear physics.

3. This difficulty was further aggravated by the spin of the nucleus. As has been shown (see Sec. 80.2), whether a nucleus has integral or half-integral spin (in units of $\hbar$) depends on its mass number. Nuclei with an even mass number have integral spin, while nuclei with an odd mass number have half-integral spin. Beta-decay leaves the mass number unchanged. This is also true of spin. A parent nucleus with integral spin decays into a daughter nucleus likewise having integral spin, while a parent nucleus having half-integral spin produces a daughter nucleus also having half-integral spin. In all cases, however, the emitted electron has half-integral spin. Because of this, one is forced to admit that beta-decay must change the spin of the nucleus, too. That is, a parent nucleus having integral spin should have produced a daughter nucleus having half-integral spin, and vice versa. Observations failed to prove that.

4. To account for the above discrepancy, Pauli suggested in 1931 that in addition to an electron another particle having no charge, a negligible rest mass and a spin of $\hbar/2$ (that is, as that of the electron) should be emitted in each beta-decay.

Fermi named this particle the *neutrino* and constructed a theory explaining the beta-decay process. According to the present-day concept, electron beta-decay proceeds by emitting not the beta-neutrino (symbolized as ${}_0\nu_e^0$), but a beta *antineutrino* (svmbolized as ${}_0\tilde{\nu}_e^0$) which, like the beta-neutrino, has spin $\hbar/2$, zero charge, zero or nearly zero rest mass, and a magnetic moment not exceeding $10^{-9}$ Bohr magneton (see Sec. 42.2).

The neutrino hypothesis resolved all the difficulties in understanding the beta-decay process. The missing energy, that is, the difference between the maximum beta-particle energy, $\mathscr{E}_{\beta\,\max}$, and that actually measured, was assumed to be carried away by the beta antineutrinos. The total energy given up by the nucleus as it ejects an electron was put equal to $\mathscr{E}_{\beta\,\max}$, or the upper limit of the beta-spectrum, but it may be shared between the electron and the antineutrino differently, according to the curve of Fig. 81.8. Notably, zero electron energy indicates that all of the energy associated with beta-decay is carried away by the antineutrino. At the cut-off point on the curve of Fig. 81.8, where the electron energy is $\mathscr{E}_{\beta\,\max}$, all energy associated with the beta-decay is carried by the electron, and the antineutrino has zero energy. At all intermediate points on the curve of Fig. 81.8 the energy of beta-decay is shared between the electron and the antineutrino so that the sum is always equal to $\mathscr{E}_{\beta\,\max}$.

The fact that both the electron and the neutrino have a spin of $\hbar/2$ has also resolved the difficulty associated with the spin of the nucleus. Since each beta-decay produces an electron and an antineutrino and since both have the same spin $\hbar/2$, the total spin of the two particles may be zero if the two spins are anti-parallel.

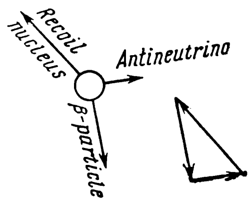


Fig. 81.9

5. The most difficult point in the neutrino hypothesis has been its experimental detection. Since it is assumed to have no charge or appreciable mass, the neutrino (or the antineutrino) could not possibly interact with matter.* This is why the experimental detection of the neutrino is based on the law of conservation of momentum. The idea of one of such experiments is as follows.** If the beta-decay proceeded by the emission of only an electron, the emitting nucleus would recoil in the direction opposite to that in which the electron is ejected, and its momentum would be numerically equal to that of the electron. If, on the other hand, both an electron and an antineutrino are emitted in each beta-disintegration, then, according to the law of conservation of momentum, the vector sum of the three momenta (electron, antineutrino and recoil nucleus) should remain equal to zero as it is before the decay (Fig. 81.9). Thus, if the antineutrino does exist and is emitted in beta-decay, the emitting nucleus will recoil at an angle to the direction in which the electron is ejected. Experiments have proved this.

6. To settle the matter of electron production in beta-decay, it should be recalled that according to the selection rules this process does not change the number of nucleons in the nucleus, while it increases its charge by unity. These conditions may be satisfied

* The ionizing ability of the neutrino is so small that one act of air ionization occurs over a distance of 500 km. This fact has been utilized in neutrino astronomy. Neither the Sun nor the Earth constitute any obstacle to neutrinos. Using the neutrinos emitted by nuclei inside the Sun, astrophysicists hope to get an insight into the Sun's otherwise inaccessible region.

** See also Sec. 83.7.

simultaneously only if a neutron ${}_0n^1$ in the nucleus decays to give a proton ${}_1p^1$, an electron ${}_{-1}e^0$ and an antineutrino ${}_0\tilde{\nu}_e^0$, which is illustrated by the following decay scheme:

$$ {}_0n^1 \rightarrow {}_1p^1 + {}_{-1}e^0 + {}_0\tilde{\nu}_e^0 \qquad (81.18)$$

This transformation conserves both charge and, as follows from the description of the experiment, momentum. The balance of mass numbers is preserved, too. It remains to be seen whether this transformation is feasible energetically. It should be accompanied by the release of an energy sufficient to produce an electron and an antineutrino in the case of natural beta-decay. The rest mass of a neutron is greater than the rest masses of a proton and an electron, or the mass of a hydrogen atom, by $0.837 \times 10^{-3}$ amu. According to Einstein's law, this mass, $\Delta m$, corresponds to an energy $\Delta \mathscr{E} = \Delta mc^2 = 782$ keV. This amount of energy is available for distribution between the emitted electron and antineutrino. Thus, there is a supply of energy for the transformation illustrated by Eq. (81.18) to proceed.

7. The foregoing suggests that the transformation described by (81.18) can well proceed with free neutrons and not only in the nucleus where they are bound. In fact, this type of transformation involving free neutrons was observed in 1959. Observations have shown that *a free neutron is a beta-particle* and that the decay half-life of free neutrons is $0.93 \times 10^3$ s. The electrons emitted by free neutrons have been found to have a continuous energy spectrum similar to that shown in Fig. 81.8, and the maximum electron energy has been found to be 782 keV, which is in agreement with the calculation presented above.

The decay half-life of beta-active nuclei differs from that of free neutrons because the neutrons in a nucleus are in states different from those occupied by free neutrons. This also applies to protons, for protons in the nucleus can change to a neutron state. As will be shown shortly, this constitutes $\beta_+$, or positive, beta-decay (or positron emission) occurring in some artificially radioactive nuclei (see Sec. 82.2).

# Chapter 82

# INDUCED TRANSMUTATIONS OF ATOMIC NUCLEI

## 82.1. TRANSMUTATION OF NITROGEN INTO OXYGEN. DISCOVERY OF THE NEUTRON

1. Studies into natural radioactivity have made huge contributions to our knowledge of the structure and properties of the nucleus. Using alpha-particles, investigators set about trying to probe the

insides of the nucleus. That was the beginning of studies into *nuclear reactions*, or artificial transmutations of nuclei induced by their interaction with charged particles or with one another.

The first conclusive evidence for artificially induced transmutation was obtained by Rutherford in 1919, in connection with some experiments on the scattering of alpha-particles emitted by polonium, $_{84}Po^{214}$, and having an energy of about 7.5 MeV. The apparatus used in these studies is shown in Fig. 82.1. In a gas-tight chamber, $K$, was mounted a radioactive polonium source, $C$. The scintillation produced by any particles striking a zinc sulphide scintillation screen, $S$, were observed with a microscope, $M$. With the gas pressure built up in the chamber, the ranges of alpha-particles were such that a thin aluminium foil, $F$, placed in front of the screen $S$ effectively absorbed all alpha-particles emitted by the polonium source. Thus, no scintillations caused by the alpha-particle would be observed in the microscope. Quite unexpectedly, scintillations were observed when the chamber was filled with nitrogen, while neither oxygen nor carbon dioxide produced such an effect.

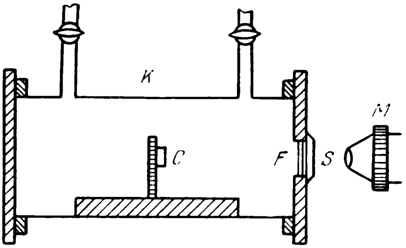


Fig. 82.1

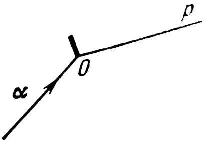


Fig. 82.2

2. Detailed investigations of the effect with an automated Wilson cloud chamber followed, in which photographs of the tracks left by particles were made and the ranges of the particles could be measured. The photographs showed that the alpha-track terminated in a fork, as shown in Fig. 82.2. At point $O$ the alpha-track disappears to give way to two branches, the shorter heavy track representing the oxygen ($_8O^{17}$) nucleus that had experienced a recoil, and the longer thin track which could not be obviously that of an alpha-particle. After the specific charge, $q/m$, of the particles leaving the thin track was found (with the aid of a magnetic field) to be equal to that of the hydrogen nucleus, they were identified as protons.

From this, Rutherford concluded that at point $O$ the alpha-particle had been absorbed into the nucleus, releasing a proton and forming a residual nucleus. This residual nucleus can be identified if we assume that in this transformation both mass number and charge are conserved as they are in natural radioactive decay. Then the mass number of the residual nucleus is $4 \times 14 - 1 = 17$ and its charge is $2 + 7 - 1 = 8$. Thus, the residual nucleus is the nucleus of an oxygen isotope, ${}_8O^{17}$. In the terminology of nuclear reactions, the events observed by Rutherford may be written as

$$ {}_2He^4 + {}_7N^{14} = {}_1p^1 + {}_8O^{17} \qquad (82.1)$$

This reaction was later confirmed from the application of the conservation of energy and momentum to the decay products. Among other things, it was found that the masses of the products were in the ratio 17 : 1, which corresponds to the nuclear reaction as described by Eq. (82.1).

For the reaction illustrated by Eq. (82.1) to occur, an alpha-particle should make a "direct hit" on the nucleus. This occurs rarely, because the nucleus is small (see Sec. 80.6) and because the concentration of target nuclei in gaseous nitrogen is likewise low. This is why photographs of forks on alpha-tracks can be observed seldom.

3. For some time after Rutherford's discovery, it was the only type of artificial transmutation known. Then the scale of research expanded, and alpha-bombardment was found capable of producing transmutations in boron, aluminium, fluorine, and potassium. In the general form, these nuclear reactions may be written as

$$ {}_ZX^A + {}_2He^4 \rightarrow {}_{Z+1}Y^{A+3} + {}_1p^1 \qquad (82.2)$$

where ${}_ZX^A$ is the source nucleus bombarded with an alpha-particle, ${}_{Z+1}Y^{A+3}$ is the end nucleus produced by the reaction. As an example, we shall write an equation describing the transmutation of the aluminium (${}_{13}Al^{27}$) nucleus into the silicon (${}_{14}Si^{30}$) nucleus:

$$ {}_{13}Al^{27} + {}_2He^4 \rightarrow {}_{14}Si^{30} + {}_1p^1 $$

In this reaction, the protons have a range of about 0.9 m, because the energy released in the reaction is about 2.26 MeV.

4. Artificial transmutations induced by alpha-particles culminated in the discovery of the neutron which, as has been shown, is part of the nucleus.

To begin with, it was noted in 1930 that alpha-bombardment caused some elements, among them beryllium, ${}_4Be^9$, to emit a radiation which was found to be more penetrating than any known gamma-ray. It was surmised that the beryllium nucleus, on capturing an alpha-particle, turned into a carbon isotope, ${}_6C^{13}$, which dropped to the ground state with the emission of hard gamma-rays. The energy estimates of this penetrating radiation made on the assumption that they were gamma-rays, yielded a figure of about 7 MeV.

In 1931, Mme. Joliot-Curie and F. Joliot discovered that this "beryllium radiation", on passing through hydrogenous substances, such as paraffin, produced a very marked increase in the number of protons with a range of up to 26 cm. From the range-energy relation for protons, they concluded that the gamma-quanta should have an energy of 55 MeV, and not 7 MeV. Further studies showed that this radiation, initially thought to be of electromagnetic origin, caused the production of recoil nuclei in the gases nitrogen, argon and even krypton. Moreover, from the range of recoil nuclei observed in argon it appeared that the energy of a gamma-quantum should be by about an order of magnitude greater than that actually found for the particles of the penetrating radiation discovered in beryllium.

In 1932, Chadwick proved that the ranges and velocities of recoil nuclei observed in various gases could only be accounted for if, instead of gamma-rays, one assumed that the radiation consisted of uncharged particles of about the same mass as the proton. These particles were named *neutrons*, now symbolized as ${}_0n^1$.

Chadwick's statement can be proved as follows. From the ordinary laws of elastic collisions, the maximum velocity $v$, that a neutron of rest mass $m_n$ and velocities of $v_0$ and $v_1$ before and after a central elastic collision can impart to a stationary nucleus of mass $M$, is:

$$v = v_0 2m_n/(m_n + M)$$

For neutrons passing through nitrogen ($M = 14$) and hydrogen ($M = 1$), the ratio of the velocities of recoil nuclei is

$$v_N/v_H = (m_n + 1)/(m_n + 14)$$

From experiments it has been found that the ratio $v_N/v_H$ is about 0.13. Hence, $m_n$ is about unity.

Being uncharged, neutrons are highly penetrating. While they have no electrical interaction with the atomic electrons or experience no Coulomb interactions with the nuclei, neutrons interact with the latter due to nuclear forces and the magnetic moments that both the nuclei and the neutrons have. This interaction is responsible for the production of recoil nuclei observed in experiments.

Thus, the nuclear reaction in which neutrons were produced for the first time may be written as:

$${}_4Be^9 + {}_2He^4 \rightarrow {}_6C^{12} + {}_0n^1 \qquad (82.3)$$

Later, neutrons played an outstanding part as particles inducing nuclear reactions.

## 82.2. INDUCED RADIOACTIVITY

1. As already noted (see Sec. 80.7), a stable nucleus is characterized by a definite number of protons and neutrons. An equation, (80.16), which defines the number of protons and, as a consequence, that of neutrons in a stable nucleus has been derived on the basis of the liquid-drop model. If this balance of protons and neutrons in a stable nucleus be upset by some artificial means, such as bombardment with some charged particles, the nucleus will display *induced radioactivity*. If a light nucleus has an excess electron, its neutron will turn into a proton according to Eq. (81.18) with emission of an electron.

2. Light nuclei in which an excess number of neutrons above their normal complement is produced artificially, that is, the nuclei in which the condition of stability, Eq. (80.16), is upset, are usually negative beta emitters. A typical example is the transmutation of the stable sodium (${}_{11}Na^{23}$) isotope into the radioactive ${}_{11}Na^{24}$ isotope induced by neutron bombardment. The ${}_{11}Na^{24}$ isotope is a negative beta emitter and decays to a stable magnesium (${}_{12}Mg^{24}$) isotope by emission of an electron, a beta antineutrino ${}_0\tilde{\nu}_e^0$ (see Sec. 81.12) and a gamma-quantum:

$$ {}_{11}Na^{24} \rightarrow {}_{12}Mg^{24} + {}_{-1}e^0 + {}_0\tilde{\nu}_e^0 + \gamma $$

The radioactive carbon (${}_6C^{14}$) isotope produced from the stable nitrogen (${}_7N^{14}$) nucleus by neutron bombardment (with emission of protons) again decays to the stable nitrogen isotope according to the following decay scheme:

$$ {}_6C^{14} \rightarrow {}_7N^{14} + {}_{-1}e^0 + {}_0\tilde{\nu}_e^0 $$

The radioactive cobalt (${}_{27}Co^{60}$) isotope decays to the stable nickel (${}_{28}Ni^{60}$) isotope by electron emission.

3. The stability of nuclei, Eq. (80.16), may also be upset by adding excess protons to the nucleus. As a result, the energy of the nucleus will rise above its minimum value set by Eq. (80.16), and the nucleus will display radioactivity. Such nuclei may undergo radioactive transmutations in which the excess proton changes to a neutron according to the equation:

$$ {}_1p^1 \rightarrow {}_0n^1 + {}_1e^0 + {}_0\nu_e^0 \qquad (82.4) $$

As is seen, the decay products include one more particle symbolized ${}_{+1}e^0$, which indicates that it has a positive charge of one unit, a mass equal to that of an electron, and spin $\hbar/2$. It is called the *positron*. The transmutation proceeds by the ejection of a neutrino, ${}_0\nu_e^0$, an uncharged particle with zero rest mass, which follows from the same considerations as were set forth in Sec. 81.12 in connection with negative beta decay (negatron emission).

4. Induced radioactivity was discovered by Mme. Joliot-Curie and F. Joliot in 1934. Using alpha-bombardment of aluminium, boron and other light elements and investigating the decay products with a cloud chamber placed in a magnetic field, they found positrons. The emission of positrons did not stop even after alpha-bombardment was discontinued. Instead, it decreased exponentially with time according to the fundamental law of radioactive decay, $N = N_0 \exp(-\lambda t)$. The reaction leading to induced radioactivity they suggested was

$$_{13}\mathrm{Al}^{27} + {}_2\mathrm{He}^4 \rightarrow {}_{15}\mathrm{P}^{30} + {}_0\mathrm{n}^1$$

The isotope $_{15}\mathrm{P}^{30}$, they assumed, decays spontaneously by emission of positrons to a stable isotope of silicon by the reaction:

$$_{15}\mathrm{P}^{30} \rightarrow {}_{14}\mathrm{Si}^{30} + {}_{+1}\mathrm{e}^0 + {}_0\nu_e^0$$

Its decay half-life was observed to be $T = 3.25$ min.

In the same experiment, Curie and Joliot found that boron bombarded with alpha-particles also produced positron activity with a half-life of about 14 min. They attributed this activity to another isotope, $_7\mathrm{N}^{13}$ formed in the reaction:

$$_5\mathrm{B}^{10} + {}_2\mathrm{He}^4 \rightarrow {}_7\mathrm{N}^{13} + {}_0\mathrm{n}^1$$

and decaying to $_6\mathrm{C}^{13}$

$$_7\mathrm{N}^{13} \rightarrow {}_6\mathrm{C}^{13} + {}_{+1}\mathrm{e}^0 + {}_0\nu_e^0$$

As is seen, positive beta ($\beta_+$) decay proceeds by emission of a positron (hence the name 'positron emission') and a beta neutrino.

The above nuclear reactions were later confirmed by chemical analysis.

Induced radioactivity produces labelled or tagged atoms or even molecules. By observing the ejection of a beta-particles with a suitable counter (see Sec. 81.8), the labelled compound or its fragments may be followed through physical, chemical or biological processes. Owing to their decay life-time convenient for observation, radioactive isotopes have found wide use as tracers in both science and technology.

### 82.3. ELECTRON-POSITRON PAIR PRODUCTION AND ANNIHILATION

1. In 1932, two years before Mme. Joliot-Curie and F. Joliot discovered induced radioactivity, C. D. Anderson detected positrons in cosmic rays (see Sec. 83.2). Before long the appearance of positrons was interpreted as a consequence of pair production, that is, the conversion of a gamma-quantum into an electron and a posi-

tron when the quantum traverses a strong electric field, such as that surrounding a nucleus or an electron.

Experimentally, pair production can be observed in a cloud chamber set up in a magnetic field. Owing to Lorentz forces acting on a charge moving in a magnetic field, the electron and the positron of a pair deflect in opposite directions. A photograph of an electron-positron pair formed in a cloud chamber under the action of hard gamma-quanta is shown in Fig. 82.3.

2. The production of an electron-positron pair should obey the laws of conservation of energy and momentum. From the law of conservation of energy it follows that in order to create an electron-positron pair, a gamma-photon should have an energy of not less than $2m_0c^2$, that is, $h\nu \geqslant 2m_0c^2 = 1.022$ MeV, where $m_0c^2 = 0.511$ MeV is the rest energy of each of the two particles formed. The law of conservation of momentum imposes a further constraint on pair production. The point is that in the creation of two particles with a rest mass $m_0 \neq 0$ from a photon, the total momentum due to both particles turns out to be smaller than that of the photon, $h\nu/c$*. In other words, the law of conservation of momentum appears to be violated by the system "photon—electron/positron".

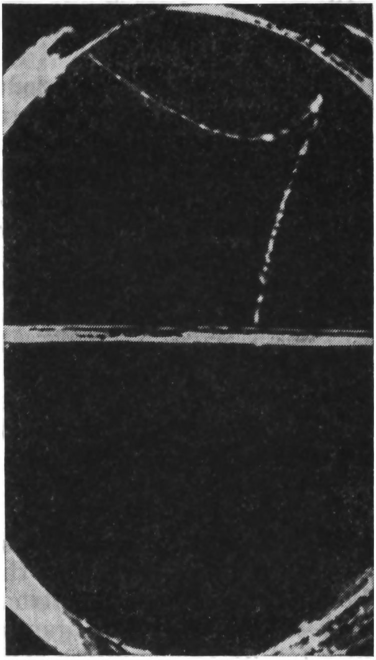


Fig. 82.3

To conserve momentum, one more particle is necessary in the process of pair production. Usually, this particle is the residual nucleus or an atomic electron in the medium in which hard gamma-rays are retarded. In the latter case, the recoil electron is imparted a momentum which may be detected from the track that the electron leaves in a cloud chamber. The momentum of the recoil nucleus left over from pair production can likewise be observed experimentally.

Apart from the conservation of energy and momentum, the production of an electron-positron pair requires that the photon have

* We leave it as an exercise for the reader to prove the point.

integral spin (in units of $\hbar$), that is, a spin of zero or unity. A number of forceful arguments (which lie outside the scope of this book) indicate that the photon cannot be a spinless particle. Therefore, its spin is $\hbar$.

3. There also exists a process which is the reverse of pair production. In this process, called *annihilation*, an electron and a positron meet and convert spontaneously into one or, mostly, two gamma-quanta or photons, by the reaction:

$$_{-1}e^0 + {}_{+1}e^0 \rightarrow 2\gamma \tag{82.5}$$

That two quanta are produced by pair annihilation results from the conservation of momentum.The point is that prior to their union the total momentum of the electron, $_{-1}e^0$, and the positron, $_{+1}e^0$, in the centre-of-mass coordinates of the system "electron-positron" is zero. In order to leave the momentum unchanged after pair annihilation, two quanta should be produced,with their momenta oriented in opposite directions. Then each of the gamma-quanta will carry with it an energy equal to

$$h\nu = mc^2 = 0.511 \text{ MeV}.$$

Experimentally, the existence of pair annihilation has been confirmed by L. A. Artsimovich (1928), A. I. Alikhanov (1904) and A. I. Alikhanian (1908) of the Soviet Union. They have used the fact that a positron can remain in a free state for a rather short time. On passing through a substance, it meets an electron, and the two convert into two gamma-quanta. In their experimental arrangement, a positron source was enclosed in a lead container placed between two counters. The thickness of the lead enclosure had been chosen such that all positrons inside it would combine with electrons and the resultant gamma-quanta would pass outside and be collected by the counters. These experiments have shown that each act of pair annihilation produces two gamma-quanta moving in opposite directions. From the intensity of the gamma-rays, it has been found that each quantum has an energy, $h\nu$, very close to 0.5 MeV.

4. The processes of pair production and pair annihilation are interesting also in that they illustrate the interrelations existing between different forms of matter. In these processes we see the conversion of matter existing as a substance into matter in the form of an electromagnetic field, and vice versa. Of course, these transformations obey all the laws of conservation. For example, the mass of the gamma-quantum from which a positron and electron pair is formed is precisely equal to the mass of the resultant particles.Therefore, any talk about mass or matter turning to nothing or being created from nothing in these processes has no ground whatsoever.

## 82.4. THE COMPOUND NUCLEUS. GENERAL OUTLINE OF NUCLEAR REACTIONS

1. In any nuclear reaction, except spontaneous nuclear fission (Sec. 82.8), nuclei collide with particles (neutrons, alpha-particles or protons) or with one another. These collisions, however, differ from those occurring between bodies and particles handled in classical mechanics (see Sec. 17.1), and also from the collisions of particles with the electronic shells of atoms and molecules. A non-nuclear collision involves the transfer of momentum and energy from the incident particle to any one particle of the target. For example, when an atom is excited or ionized by collision with an ion, the ion imparts its energy to any one electron in the atom. As a result, the electron is either excited to a higher energy state or removed from the atom.

2. A basically different situation arises in nuclear collisions. A nucleus is a closely packed formation, such that when a particle strikes it, it does not interact with any one nucleon. Once it enters the nucleus, any incoming particle almost immediately loses its identity, and quickly shares its energy with several of the nucleons, so that no single particle has sufficient energy to escape. Thus, it forms an intermediate or so-called *compound* nucleus. This is the first stage in an induced nuclear reaction.

3. The theory of the compound nucleus has been developed by N. Bohr, and by L. D. Landau and Ya. I. Frenkel who have shown that the energy brought in with the incoming particle is equally shared among the constituent nucleons in a fairly short time. According to their theory, a compound nucleus, like any other excited body (with a high mass number) may be treated as a statistical system of particles in a random motion, similar to that of particles in a drop of a liquid. The compound nucleus must be analogous to a liquid drop (see Sec. 80.7) because energy can be shared among the constituent nucleons quickly only if they collide at a frequent rate. which is characteristic of particles in a liquid. If the energy imparted to a compound nucleus of mass number $A$ is $\mathscr{E}$, then the average excitation energy per particle, $\mathscr{E}/A$, corresponds to a definite nuclear temperature, $T$, which can be defined from the relation

$$\mathscr{E}/A = 3kT/2$$

where $k$ is Boltzmann's constant (see Sec. 26.9). A measure of nuclear temperature is given by the average kinetic energy per particle of the compound nucleus. For example, at $A = 100$ and $\mathscr{E} = 10$ MeV, the nuclear temperature is of the order of $10^9$ K. The high value of $T$ is an indication of the artificiality of this concept.

4. The statistical approach to the compound nucleus on the basis of the liquid-drop model has proved immensely fruitful for a description of the regularities common to all nuclear reactions. It

has also been helpful in understanding the second stage of a nuclear reaction, that is, the decay or disintegration of the compound nucleus, which occurs some time later, when the excitation energy is again accidentally concentrated on some one particle or when it is lost by radiation. As a simple analogue of the second stage, we may consider a liquid drop from which some molecule can evaporate through an accidental concentration of energy or heat.

The lifetime of a compound nucleus, that is, the time interval between its formation and decay, is as a rule considerably longer than the so-called *nuclear traversal time*, or the time required for a nuclear particle with an energy of about 1 MeV and a velocity of $10^7$ or $10^8$ m/s to traverse the nucleus (a distance of about $10^{-15}$ m), that is,

$$\tau_N = 10^{-15}\ \text{m} \div 10^7\ \text{m/s} = 10^{-22}\ \text{s}$$

So far, this has been the shortest time interval encountered in nature. This time interval (sometimes called the *nuclear year*) characterizes the transfer of the strongest nuclear interactions (Sec.83.5).

Thus, the lifetime of a compound nucleus is ($10^6$ to $10^7$) $\tau_N$. This implies that the decay of a compound nucleus, the second stage of the nuclear reaction, will usually occur *independently* of the capture of an incoming particle by the target nucleus, that is, of the first stage of the reaction.

The two stages of a nuclear reaction may be written as follows:

$$_{Z_1}X^{A_1} + a \rightarrow {_{Z_2}Y^{A_2}} \rightarrow {_{Z_3}C^{A_3}} + b \tag{82.6}$$

where $_{Z_1}X^{A_1}$ is the target nucleus, $a$ is the incoming particle, $_{Z_2}Y^{A_2}$ is the compound nucleus, and $b$ is the particle ejected by the nucleus $_{Z_3}C^{A_3}$ as a result of the nuclear reaction.

If the emitted particle is identical with the incoming one, Eq. (82.6) describes the scattering of a particle (which may be inelastic or elastic, depending on whether the energy of the particle is the same or not before and after the scattering). If the emitted particle is not identical with the incoming one, Eq. (82.6) describes a true nuclear reaction.

5. Nuclear reactions may be classed in several ways, namely by the energies of the particles that induce them, by the kind of particles taking part in them and, finally, by the character of the nuclear transformations that take place. Accordingly, there are low-energy, medium-energy and high-energy nuclear reactions. Low-energy reactions (of the order of one electron-volt) are mainly neutron-induced. Medium-energy reactions (up to a few MeV) may additionally be induced by charged particles (protons and alpha-particles) and gamma-quanta. High-energy reactions (running into hundreds and thousands of MeV) produce fundamental particles usually non-existent in a free state, and are of special importance in

the study of the properties and structure of fundamental particles (Ch. 83).

6. The particles involved in nuclear reactions may be neutrons, charged particles (which, in addition to those listed above, may be deuterons and the multi-charge ions of heavy chemical elements). Apart from naturally radioactive elements, charged particles may be produced by particle accelerators (Secs. 41.4 and 41.6).

In shorthand notation, nuclear reactions are written as $(a, b)$, where $a$ stands for the particle inducing a reaction, and $b$ stands for the reaction product. For example, $(\alpha, p)$ and $(\alpha, n)$ describe reactions induced by alpha-particles and producing protons, ${}_1\mathrm{p}^1$, and neutrons, ${}_0\mathrm{n}^1$, respectively.

## 82.5. NEUTRON REACTIONS

1. Passing through a substance, neutrons have practically no interaction with the atomic or molecular electrons, because neutrons have no electrical charge. They can only interact with nuclei (which explains why neutrons are far more penetrating than charged particles).

Neutrons interact with the nucleus differently, according to whether they are *fast* or *slow*. Fast neutrons are those whose velocity $v$ is so high that the respective rationalized de Broglie wavelength, $\lambdabar = \hbar/mv$, is far shorter than the nuclear radius $R$, that is:

$$\hbar/mv \ll R \quad \text{or} \quad v \gg \hbar/mR$$

Fast neutrons have energies in the range from 0.1 MeV to 50 MeV. Slow neutrons are those for which $\lambdabar \gg R$, that is, $v \leqslant \hbar/mR$. Their energies do not exceed 100 keV.

2. For fast neutrons, the nucleus is a target which has a cross section equal to the geometrical cross section of the nucleus, that is, to the cross-sectional area of a spherical nucleus over the great circle.

For slow neutrons (often, though loosely, called *thermal* neutrons) the cross section of a nucleus is a measure of the probability of an interaction between the neutrons and the nucleus. Numerically, it is $10^2$ or $10^3$ times the geometrical cross section of the nucleus. It is relevant to draw an analogy with optics. If the linear dimensions of a light scatterer exceed the wavelength of the light, $\lambda$, the light is scattered according to the laws of geometrical optics; this corresponds to fast neutrons. If the dimensions of a light scatterer are such that $d \approx \lambda$, the light will be scattered according to the laws of wave optics; this corresponds to slow neutrons.

3. When they interact with nuclei, the neutrons can be either scattered or captured by the nuclei. In substances called *moderators*, neutrons are mainly scattered, and neutron capture is of minor importance.

Neutron moderators are chiefly graphite, heavy water ($D_2O$, HDO), and beryllium compounds. On passing through a moderator, fast neutrons undergo scattering by the nuclei and are slowed down until their energy $\mathscr{E}$ becomes equal to that of the thermal motion of atoms in the moderator, that is, $\mathscr{E} \approx kT$, where $T$ is the absolute temperature. The energy of thermal neutrons is mainly lost as the recoil energy of the nuclei. At room temperature, the energy of thermal neutrons is 0.025 eV. Further collisions of thermal neutrons with the nuclei of the moderator cannot reduce their energy any more, because now the neutrons are in thermal equilibrium with the surroundings, that is, they have an equal probability of either gaining or losing an amount of energy equal to $kT$. These collisions may only cause the neutrons to diffuse throughout the substance without losing any energy until they leave the moderator.

4. In 1934, in the course of this investigations of induced radioactivity, E. Fermi observed that the yield of radioactivity could be enhanced enormously by letting the neutrons pass through a hydrogen-containing substance, such as water or paraffin, which would slow them down. This increase in the yield of radioactivity was found to be due to the *capture* of some slow neutrons.

In other investigations it was established that the capture of slow neutrons would be especially intensive if the difference between the energy levels of a compound nucleus was the same as the neutron energy. Such neutron energies occur in very narrow bands called resonance-absorption bands for the reason that this form of capture is known as the *resonance absorption of neutrons*.

## 82.6. TRANSURANIC ELEMENTS

1. Neutron-induced reactions in uranium have played a special role in nuclear physics. Through them, scientists have run into nuclear fission (discussed in the next section) and devised ways and means of producing chemical elements with charge numbers in excess of 92, called *transuranic elements*. We shall dwell in brief on the production of some of them.

The resonance capture of slow neutrons by uranium-238, the most abundant of all uranium isotopes, produces a radioactive isotope, uranium-239, which decays by negatron ($\beta_-$) emission with a half-life of 23 min to an isotope of neptunium, a transuranic element, ${}_{93}Np^{239}$ by the reaction:

$${}_{92}U^{238} + {}_{0}n^{1} \rightarrow {}_{92}U^{239} \xrightarrow[23\text{ min}]{\beta-} {}_{93}Np^{239}$$

2. In turn, neptunium-239 decays by negatron emission with a half-life of 2.3 days to plutonium, ${}_{94}Pu^{239}$, by the reaction:

$${}_{93}Np^{239} \xrightarrow[2.3\text{ days}]{\beta-} {}_{94}Pu^{239}$$

Plutonium, $_{94}Pu^{239}$, is the most important of all transuranic elements, as a source of nuclear energy which it releases under the action of thermal neutrons (Sec. 82.7). Plutonium-239 decays by emission of alpha-particles with a very long half-life (24 000 years) to a stable isotope of uranium, $_{92}U^{235}$:

$$_{94}Pu^{239} \xrightarrow[2.4\times 10^{4}\text{ years}]{\alpha} {}_{92}U^{235}$$

3. The nuclear reaction of the (n, 2n) type turns uranium-238 to uranium-237, an induced radioisotope, which decays by $\beta_-$ emission to neptunium-237, in turn decaying by emission of alpha-particles with a huge half-life of 2.2 × $10^6$ years by the reaction:

$$_{92}U^{238} + {}_0n^1 \rightarrow {}_{92}U^{237} + 2{}_0n^1$$

$$_{92}U^{237} \xrightarrow[6.8\text{ days}]{\beta-} {}_{93}Np^{237}$$

Neptunium-237 starts one of the radioactive series (see Sec. 81.2).

Apart from the transuranic elements listed above, other elements can be produced by bombardment of uranium-238 and plutonium-239 with beams of accelerated neon nuclei (Sec. 82.8).

## 82.7. DISCOVERY OF FISSION

1. The experiments that resulted in the discovery of fission began in 1934 when Fermi was busy investigating neutron-induced radioactivity. It was then found that uranium bombarded with neutrons yielded radioactive products with several decay half-lives. At first, it was surmised that those were transuranic elements formed through the capture of a neutron by the uranium nucleus to form a heavier uranium isotopes decaying by emission of $\beta_-$ to a chemical element of atomic number 93. By the same token, element 93 was thought as decaying by $\beta_-$ emission to an element of atomic number 94. In fact, it is this striking ability of neutrons to produce $\beta_-$-unstable isotopes from naturally occurring elements that suggested to Fermi that it might be possible to produce the transuranic element we have examined in the previous section.

In the years 1936-1937, quite a number of reactions induced in uranium by neutrons widely differing in energy were investigated. To the investigators' surprise, however, they found that the materials thus produced appeared to have chemical properties similar to those of the elements in the middle of the Periodic Table. In 1938, using precise radiochemical identification, Hahn and Strassmann established beyond doubt that the neutron bombardment of uranium had produced barium, $_{56}Ba$, which is a chemical analogue of radium, $_{88}Ra$. More similar identifications followed, and the investigators were led to suggest that what they had thought to be "transuranic" elements might actually be isotopes of other, considerably lighter elements, possibly produced by an "explosion" of the uranium nucleus. Similar results were obtained when they bombarded thorium-232 with neutrons.

2. For the first time, a physical explanation of these striking results was proposed by Frisch and Meitner. According to them, the elements formed by the neutron bombardment of uranium might be quite *unstable*. Excited by the capture of a neutron, the heavy compound nucleus (say, the uranium nucleus) might break up into two approximately equal *fission fragments*. In the process, the nucleons of the compound nucleus must divide between the fission fragments so as to conserve charge and mass number:

$$Z_U = Z_1 + Z_2; \quad A_U + 1 \approx A_U = A_1 + A_2$$

where the subscripts "1" and "2" refer to the fission fragments, $Z_U$ and $A_U$ are the charge and mass number of the target nucleus of uranium.

Confirmation of their hypothesis came almost immediately from the experiments in which the neutron bombardment of uranium yielded, in addition to barium, radioactive isotopes of strontium, $_{38}$Sr, and yttrium, $_{39}$Y, and also a radioisotope of a chemically inert gas (krypton or xenon).

3. Simple considerations show that the *break-up of a uranium nucleus into two fragments must be accompanied by the release of a huge amount of energy*. The point is that, as has been shown in Sec. 80.4, the binding energy per nucleon in the elements occupying the middle part of the Periodic Table is about 8.7 MeV, while for the heavy nuclei it is 7.6 MeV. This point assumes a special importance now that a loosely packed uranium nucleus breaks up into two stable, closely packed fragments, because this fission should release an energy of 1.1 MeV per nucleon. The total release of energy from the fission of a uranium-238 nucleus containing 238 nucleons must be of the order of 200 MeV. The fission of the nuclei contained in one gram of uranium-235 should release a total of $8 \times 10^{10}$ J or 22 000 kWh of energy.

4. The greater proportion of the total energy, $\mathscr{E}$, should be released as the kinetic energy of the fission fragments. For, when a fission occurs and the fragments move a distance $r$ from one another, nuclear forces will be no longer operative, and the fission fragments, that is, the charged nuclei, will experience the electrostatic (Coulomb) force of repulsion. The potential energy of interaction between charges $Z_1e$ and $Z_2e$ is

$$U = Z_1Z_2e^2/4\pi\varepsilon_0r$$

Obviously, at the completion of a fission, the separation $r$ between the fragments will be

$$r = R_1 + R_2$$

where $R_1$ and $R_2$ are the nuclear radii of the fission fragments which can be found from Eq. (80.6):

$$R = 1.4 \times 10^{-15} A^{1/3}$$

Letting that $Z_1 = Z_2 = 92 \div 2 = 46$, $R_1 = R_2$ and $A_1 = A_2 = 238 \div 2 = 119$, we find that the expression for $U$ gives about 200 MeV. Obviously, the potential energy $U$ of repulsion between the fragment nuclei should be converted to their kinetic energy $K$, and the fragments should fly apart at high velocity. Observations have confirmed these considerations and estimates.

5. By mid-1939, it had been well established that fission occurs in uranium under both fast-neutron and slow-neutron bombardment, the latter being even much more effective. Mass spectrometry showed that thermal neutrons cause fission in uranium-235, while the activation energy, that is the minimum energy required to initiate a fission reaction in uranium-238 and also in the naturally occurring isotopes of thorium and protactinium is about 1 MeV.

The fragments resulting from uranium fission were observed in a cloud chamber into which uranium oxide applied to a thin film had been placed. Stereoscopic photographs clearly showed the tracks of heavy fragments flying in opposite directions upon fission. Later, similar results were obtained with nuclear emulsion blocks (see Fig. 15.1 in Vol. 1).

6. The fragments resulting from the fission of heavy nuclei must be negative beta emitters and may also emit neutrons. This stems directly from the composition of the original heavy nucleus and of the fission fragments. The nuclei of the chemical elements occurring in the middle of the Periodic Table have about as many neutrons, $N$, as there are protons, so that $N/Z \approx 1$. For heavy nuclei which have a neutron excess, the ratio $N/Z$ rises to about 1.6. Hence, the fission fragments should have a great neutron excess at the instant of formation and show $\beta_-$ activity. Because of their great neutron excess, the products of $\beta_-$ decay must likewise show $\beta_-$ activity. Furthermore, part of excess neutrons,representing the difference in their number between the original nucleus and the fragment nuclei will be emitted by the latter as so-called *prompt neutrons*.

Different numbers of neutrons may be produced in fission. Therefore, one uses the average number of neutrons per fission, $\bar{\nu}$. For plutonium-239 and uranium-235 in which fission is induced by thermal neutrons the number $\bar{\nu}$ is 3.0 and 2.5, respectively. Thus, fission is accompanied by *neutron multiplication*.

7. The theory of the fission process was worked out in detail by Ya. I. Frenkel of the Soviet Union and also by N. Bohr and J. A. Wheeler, who based their arguments on the liquid-drop model of the nucleus (Sec. 80.7 ) and on the hypothesis that heavy nuclei must

be unstable. We shall only give a brief outline of their theory, limiting ourselves to the energy balance of the fission process and some related matters. We shall be interested in the first stage of the process, namely in the formation of fission fragments and shall leave out their nuclear transformations.

As has been shown in Sec. 80.7, the liquid-drop model of the nucleus leads to a semiempirical expression for the total binding energy of the nucleus. This expression has several terms; two of them are of special interest in the examination of the first stage of the fission process. Above all, this is the surface energy, $\mathscr{E}_3$, of the "liquid-drop" nucleus, given by

$$\mathscr{E}_3 = 4\pi R^2 \sigma$$

where $\sigma$ represents the "surface tension". Secondly, this is the Coulomb energy, $\mathscr{E}_4$, due to the electrostatic repulsion of protons, given by

$$\mathscr{E}_4 = 3Z^2e^2/5 \times 4\pi\varepsilon_0 R$$

where $Ze$ is the nuclear charge and $R$ is the nuclear radius. The remaining terms describing various properties of nuclear forcaes depend on the total number $A$ of nucleons in the nucleus. Since the total number of particles remains unchanged, these terms will be immaterial to the energy balance of the fission process.

8. Let us find the energy released by the fission of the original nucleus of charge $Ze$ and mass number $A$ into two identical daughter nuclei or fragments, such that $A_1 = A_2 = A/2$ and $Z_1 = Z_2 = Z/2$.

Retaining only the most important terms, the expression for the total binding energy prior to fission may be re-written as follows

$$\mathscr{E} = 4\pi R^2 \sigma + \frac{3}{5} \frac{Z^2 e^2}{4\pi\varepsilon_0 R} \tag{82.7}$$

In the fission of the nucleus, the following conditions for the constancy of the liquid-drop volume should be satisfied:

$$\frac{4}{3}\pi R^3 = 2\,\frac{4}{3}\pi R_1^3$$

where $R_1$ is the radius of the fragment nucleus. Hence, $R_1 = R/\sqrt[3]{2}$.

The total binding energy for the two fragments can be written according to Eq. (82.7) as follows:

$$\mathscr{E}' = 2\left(4\pi R_1^2 \sigma + \frac{3}{5} \frac{Z_1^2 e^2}{4\pi\varepsilon_0 R_1}\right)$$

Substituting its expression for $R_1$ and also noting that $Z_1 = Z/2$, we get

$$\mathscr{E}' = 2\left(4\pi\sigma \frac{R^2}{\sqrt[3]{4}} + \frac{3}{5}\sqrt[3]{2}\,\frac{Z^2 e^2}{16\pi\varepsilon_0 R}\right) \tag{82.8}$$

The difference, $\Delta\mathscr{E}$, between Eqs. (82.7) and (82.8) represents the energy that is released by a symmetrical fission of a heavy nucleus as the kinetic energy of its fragments:

$$\Delta\mathscr{E} = \mathscr{E} - \mathscr{E}' = 4\pi R^2\sigma\left(1-\frac{2}{\sqrt[3]{4}}\right)+\frac{3}{5}\,\frac{Z^2e^2}{4\pi\varepsilon_0 R}\left(1-\frac{\sqrt[3]{2}}{2}\right) \qquad (82.9)$$

The first term in Eq. (82.9) has a minus sign because $1 - (2/\sqrt[3]{4}) = 1 - \sqrt[3]{2} \approx -0.26$. This represents the fact that, with the volume held constant, the fission of the nucleus results in an increase in the total surface area of the fission fragments*. Physically, this means that the balance of surface energy alone makes the fission process energetically unfeasible; in fact, it is energetically advantageous for the fragments to fuse together so as to reduce the free surface. The second term in Eq. (82.9) is positive, because the Coulomb energy due to the electrostatic repulsion of protons decreases in fission. The total energy balance of the fission process is determined by the sign of $\Delta\mathscr{E}$. At $\Delta\mathscr{E} > 0$, energy will be released; at $\Delta\mathscr{E} < 0$, it will be absorbed. Everything depends on the relation between the surface energy that must be expended in breaking up the "liquid-drop" nucleus, and the Coulomb energy that is released in fission.

9. The "liquid-drop" nucleus have certain critical values of $Z$, $R$ and $\sigma$ at which the fission process will occur *iso-energetically*, that is, without any change in the energy of the system ($\Delta\mathscr{E} = 0$), or is energetically advantageous ($\Delta\mathscr{E} > 0$). Let us write this condition as $\Delta\mathscr{E} \geqslant 0$. Then by Eq. (82.9\ we get

$$\frac{3}{5}\,\frac{Z^2e^2}{4\pi\varepsilon_0 R} \div 4\pi R^2\sigma \geqslant 2\,\frac{\sqrt[3]{2}-1}{2-\sqrt[3]{2}} \geqslant 0.70 \qquad (82.10)$$

The left-hand side of inequality (82.10) is the ratio of the Coulomb energy due to the repulsion of protons, tending to break up the nucleus, to the surface energy tending to oppose the spreading of the drop. Putting, according to Eq. (80.11), that $R = r_0A^{1/3}$, where $r_0 = 1.4 \times 10^{-15}$ m, we may re-write (82.10) into an equality as follows:

$$\frac{3}{5}\,\frac{e^2}{(4\pi)^2\,\varepsilon_0 r_0^3\,\sigma}\,(Z^2/A) = 0.70 \qquad (82.10')$$

The only variable in Eq. (82.10′) is the so-called *fission parameter*, $Z^2/A$. To find the values of this parameter at which the fission process is energetically advantageous, that is, the condition $\Delta\mathscr{E} > 0$ is satisfied, we substitute the numerical values of all constants in the preceding expression and obtain the following inequality:

$$Z^2/A > 17 \qquad (82.11)$$

* It is an easy matter to find that $2 \times 4\pi R_1^2 - 4\pi R^2 = 0.26\pi R^2$.

This condition is satisfied by all nuclei, starting with silver, ${}_{47}Ag^{108}$ for which $Z^2/A \approx 20$. The fission process becomes progressively more feasible as the fission parameter increases.

## 82.8. FISSION THRESHOLD. SPONTANEOUS FISSION

1. From the foregoing one might conclude that the nuclei of the chemical elements in the latter half of the Periodic Table ($Z \geqslant 47$) should be unstable to fission. However, observations show that for the most part the isotopes of mass numbers $A$ above 108 are stable, except the heaviest nuclei which are unstable to so-called spontaneous fission (which will be discussed shortly). This seeming contradiction arises from the fact that it is not enough to satisfy the condition $\Delta\mathscr{E} \geqslant 0$ in order that a fission may occur. For this to happen, one should expend an additional amount of energy, called the *fission activation energy* (or *fission threshold*), $\mathscr{E}_f$. The "liquid-drop" nucleus will be most stable when the sum of its surface and electrostatic energies is a minimum, and this will occur in a spherical nucleus. Let the nucleus in such a case have a radius $R$. Any elongation of the nucleus will increase its surface and bring about a proportionate increase in the surface energy, $\mathscr{E}_3$. At the same time, the electrostatic energy, $\mathscr{E}_4$, will decrease because the separation between the protons is a minimum and their energy of repulsion is a maximum only in a spherical nucleus.

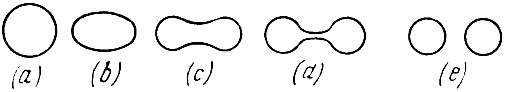


Fig. 82.4

2. Thus the stability of the nucleus will be upset already by small distortions. Yet, even a small distortion requires the expenditure of some excess energy. Thus, when there is no distortion, the nucleus is stable. Small distortions cause the nucleus to oscillate so that it alternately elongates and contracts. As the distortion approaches a certain critical value, the nucleus elongates until it breaks into two roughly equal fragments. The various stages in the fission process are illustrated in Fig. 82.4. As is seen the "neck" forms and extends during the intermediate stages (Fig. 82.4*c*). If the excitation energy is less than the fission threshold, the distortion will be insufficient to overcome the surface tension, and the nucleus will not break up. Instead, it will emit a gamma-quantum and regain its normal state.

3. Frenkel, and also Bohr and Wheeler deduced the relation between the activation energy (fission threshold) and the fission parameter, $Z^2/A$. They found that at some critical value of this ratio, $(Z^2/A)_{lim}$,

the nucleus becomes unstable to spontaneous fission. This critical value is

$$(Z^2/A)_{lim} \approx 49 \qquad (82.12)$$

At $Z^2/A < (Z^2/A)_{lim}$, one has to expend an amount of energy, the activation energy $\mathscr{E}_f$, so as to bring about the critical distortion of the nucleus which leads to fission. For heavy nuclei, $\mathscr{E}_f$ is 5 to 7 MeV, that is, it is of the same order of magnitude as the binding energy of a neutron in a heavy nucleus or, which is the same, as the binding energy per particle.

At $Z^2/A > (Z^2/A)_{lim}$, no nucleus can exist. Such a superheavy nucleus would break up spontaneously in a matter of $10^{-22}$ s. The condition (82.12) sets the limit for the stable existence of superheavy nuclei. The heaviest of all known nuclei are those of the transuranic elements produced artificially. They all have high values of the fission parameter. While for uranium-238, $Z^2/A = 35.6$, for americium, $_{95}\mathrm{Am}^{242}$, it is 37.3.

In 1964, the team working under Academician G. N. Flerov at the Joint Institute of Nuclear Research (Dubna, USSR) produced an isotope of kurchatovium, a new transuranic element with $Z = 104$ and mass number 260, by bombarding plutonium-242 with the nuclei of neon-22. As a result of this bombardment, the capture of a neon nucleus produced a compound nucleus of kurchatovium-264. In one case out of 10 000 million, this compound nucleus decayed with the emission of four neutrons to produce the nucleus of kurchatovium-260 by the reaction: *

$$_{94}\mathrm{Pu}^{242} + {}_{10}\mathrm{Ne}^{22} \rightarrow {}_{104}\mathrm{Ku}^{264} \rightarrow {}_{104}\mathrm{Ku}^{260} + 4{}_0n^1$$

Even in the strongest beams of accelerated neon nuclei, kurchatovium is produced at the rate of one nucleus in a few hours.

Also at Dubna, a fusion reaction involving uranium and neon nuclei produced atoms of the element with $Z = 102$. The reaction proceeded with the formation of an excited compound nucleus, $_{102}\mathrm{X}^{*260}$ which decayed by emission of four neutrons by the reaction:

$$_{92}\mathrm{U}^{238} + {}_{10}\mathrm{Ne}^{22} \rightarrow {}_{102}\mathrm{X}^{*260} \rightarrow {}_{102}\mathrm{X}^{256} + 4{}_0n^1$$

Every 100 million compound nuclei produce one nucleus of the isotope $_{102}\mathrm{X}^{256}$. The remaining nuclei break up into fragments and emit neutrons. Similar reactions have also produced the isotopes of the transuranic elements curium $_{96}\mathrm{Cm}$, berkelium $_{97}\mathrm{Bk}$, californium $_{98}\mathrm{Cf}$, mendelevium $_{101}\mathrm{Md}$, kurchatovium $_{104}\mathrm{Ku}$, and also the isotopes of the elements with charge numbers 102, 103, and 105.

4. At $Z^2/A < (Z^2/A)_{lim}$, nuclei may undergo spontaneous fission similar to alpha-decay (see Sec. 81.9), by way of the tunnel effect.

* The remaining nuclei broke up.

In 1939-40, G. N. Flerov and K. A. Petrzhak observed the spontaneous fission of uranium-238 for the first time. Using a very sensitive technique and an ionization chamber, they were able to register pulses produced by the fission fragments of uranium not bombarded by fission-inducing neutrons. According to Flerov and Petrzhak, the half-life of spontaneous fission should be $10^{16}$ or $10^{17}$ years. As will be recalled, for the natural alpha-decay of uranium-238 the half-life is $10^9$ years, or by seven orders shorter.

5. Observations have brought out a marked difference between the fission of uranium-238 and uranium-235. As already noted in Sec. 82.7, the fission of uranium-238 is induced by neutrons with a kinetic energy of at least 1 MeV; fission in the nuclei of uranium-235 is induced by the capture of thermal (that is, slowest) neutrons. Why this happens so can be explained as follows. The compound nucleus of uranium-239 formed by the capture of a neutron by the nucleus of uranium-238 has the ratio $Z^2/A = 35.46$ and a fission threshold $\mathscr{E}_f = 7.0$ MeV. The binding energy of the neutron captured by the nucleus of uranium-238 is about 6 MeV. Thus, fission in the nuclei of uranium-238 can be induced by neutrons with a kinetic energy of at least 1 MeV. For the compound nucleus of uranium-236 formed by the capture of a neutron by the nucleus of uranium-235, the fission parameter and the fission threshold are 35.9 and 6.6 MeV, respectively. From these values we may conclude that the conditions for neutrons to induce fission in uranium-235 are more favourable than in uranium-238. Besides, the excitation energy imparted to the nucleus of uranium-235 by the capture of a neutron is about 6.8 MeV.

Thermal neutrons induce fission in uranium-233 and plutonium-239, a transuranic element.

## 82.9. CHAIN REACTIONS

1. For the fission of heavy nuclei to have any practical value, it is important that each fission should release a large amount of energy and produce several (two or three) neutrons. Then, if each of these neutrons happens to induce fission in the neighbouring nuclei of the fissionable material, the number of fissions will grow cumulatively, thereby bringing about a self-sustaining, "chain" reaction, thus called by analogy with chemical chain reactions in which one of the agents necessary to the reaction is itself produced by the reaction so as to cause like reactions.

In the Soviet Union, this possibility was recognized for the first time by Ya. B. Zeldovich and Yu. B. Khariton, in 1939.

2. Let us examine the possibility of neutron-fission chain reactions in greater detail. The assumption that each of the neutrons is captured by neighbouring nuclei is not realized. Some of the secondary neutrons are trapped by the nuclei of non-fissionable materials ine-

vitably present in the region where the fission reaction takes place. These are neutron moderators, coolants, and the like. Some neutrons may simply leave the reacting region, that is, the space where the fission reaction takes place.

It is obvious that a chain reaction will occur only if the supply of neutrons is continually restored, that is, if neutrons are multiplied. The necessary neutron multiplication can be determined with the aid of the *neutron multiplication factor*, $k$, sometimes called the effective multiplication constant and defined as the ratio of the number of neutrons present in a given stage of the reaction to that present in one stage earlier. The condition for a self-sustaining chain reaction is that $k$ be equal to or greater than unity, $k \geqslant 1$. The value of $k$ is determined, firstly, by the average number of neutrons produced per fission (see Sec. 82.7) and, secondly, by the probabilities of the various modes of interaction between neutrons and fissionable and non-fissionable nuclei. The geometry of the system also plays an important role.

The geometry of the system is important because the proportion of neutrons leaving the reacting region, or core, increases and the probability of a self-sustaining chain reaction decreases with a decrease in the core. The loss of neutrons is proportional to the surface area, and the generation of neutrons is proportional to the mass and, as a consequence, to the volume of the fissionable material. In the case of a spherical core (with $V \backsim R^3$, $S \backsim R^2$, and $S/V \backsim 1/R$), a decrease in $R$, that is, with a decrease in volume and mass of the fissionable material, the number of neutrons escaping from it will increase. The minimum size of the core that can just maintain the chain reaction is called the *critical size*. Accordingly, the minimum mass of the fissionable materials present in a system of critical size is called the *critical mass*.

To minimize escape of neutrons and the critical size of the fissionable material, it is surrounded by a *reflector* which is a layer of a non-fissionable material having a small cross-section for neutron capture and a large cross-section for their scattering. Thus, a reflector returns a greater proportion of the escaping neutrons back to the reaction. Most often, the materials used as reflectors are usually the same as for moderators (see Sec. 82.5), namely graphite, heavy water, and beryllium compounds.

3. A very important aspect of a neutron-fission chain reaction is its rate which, apart from the multiplication factor, depends on the mean time, $\tau$, between two consecutive fissions. Obviously, $\tau$ defines the mean life of a generation of neutrons, that is, the mean time between a fission and the capture of a neutron by a fissionable nucleus. More accurately, the time $\tau$ is a sum of the time required for a nucleus to undergo a fission, the time for which escape of a neutron is delayed after a fission, and the time between two captures.

4. Fast fission of an explosive type requires that the system have a satisfactory multiplication factor with prompt neutrons. In contrast, for a neutron-fission chain reaction of the controlled type the time $\tau$ must be made as long as practicable. That is one should seek to delay escape of neutrons after the fission and to extend the time between captures. The generation of delayed neutrons depends largely on the mechanism by which secondary neutrons are produced and is less amenable to control. The extension of time between captures depends on the interaction of the emitted neutrons with surrounding nuclei, that is, on the slow-down of their motion through the material. Control of the chain reaction mainly reduces to these processes.

### 82.10. NUCLEAR REACTORS

1. Controlled nuclear chain reactions are carried out in *nuclear reactors* (formerly called nuclear or atomic *piles*).

The source and fissionable materials usually employed in nuclear reactors are uranium-235, plutonium-239, uranium-238 and thorium-232. The naturally occurring mixture of uranium isotopes contains 140 times more uranium-238 than uranium-235. For an insight into operation of a nuclear reactor using this natural mixture of uranium isotopes, it is important to remember the difference in the conditions under which fission can be induced in them. From the energy spectrum of the neutrons emitted in fission it has been found that their average energy is about 0.7 MeV. Such neutrons can induce fission only in uranium-235. The few neutrons whose energy exceeds the fission threshold of uranium-238 have a greater probability of inelastic scattering and their energy drops below that which is necessary to bring about fission in uranium-238. Through a series of collisions with uranium nuclei, the neutrons lose their energy in small portions, are slowed down and are either captured by the nuclei of uranium-238 or absorbed by the nuclei of uranium-235. The absorption of neutrons by uranium-235 promotes the chain reaction, while absorption by uranium-238 switches them out of the reaction, and the chain reaction is interrupted. Calculations show that in the natural mixture of uranium isotopes the probability of the chain reaction being interrupted is greater than the probability of sustaining it. Thus, neither fast nor slow neutrons can sustain the chain reaction.

2. In nuclear reactors using slow neutrons, or *thermal reactors* as they are called, the conditions necessary for a chain reaction to proceed are provided by a moderator which reduces the capture of neutrons by the nuclei of uranium-238. At each collision with the moderator nuclei, the neutron loses its energy in large decrements, owing to which they "skip" the energy range within which the nuclei of uranium-238 are likely to capture neutrons. The moderator materials

are carbon (as graphite), deuterium (as heavy water), beryllium and beryllium oxide whose nuclei have the least cross-section for the capture of thermal neutrons.

3. Thermal reactors may be *homogeneous* or *heterogeneous*. In the former type, the fissionable material and moderator are combined in a mixture such that an effectively homogeneous medium is presented to the neutrons. In slowing down, the neutrons are always near the uranium nuclei distributed throughout the moderator volume. Because of this, the neutrons have a great probability of being absorbed by the uranium and not by the moderator nuclei; for the same reason, however, they have a greater probability of being captured by the nuclei of uranium-238. In a heterogeneous reactor, the fissionable material is distributed through the moderator in the form of blocks. Therefore, the thermal neutrons have a lower probability of being absorbed by uranium nuclei, but they have a higher probability of avoiding the capture by the nuclei of uranium-238 because they stay outside the blocks of fissionable material a greater part of the time when they have energies which may lead to capture.

In each case, it is important to reduce the loss of neutrons by leakage from the reactor. This can be done by providing a reflector and by increasing the critical size of the reactor.

4. The neutron-fission chain reaction proceeds at a high rate and releases a large amount of energy, so that overheating has to be regarded as a highly probable occurrence. To avoid it, a reactor is allowed to reach its design energy output,after which its multiplication factor is adjusted to and maintained at unity. This is done by inserting the so-called regulating rods of some material with a high neutron-absorption coefficient, such as boron or cadmium, into the reactor core.

5. The fission of uranium nuclei in reactors is accompanied by the formation of a great number of radioactive fragments. It has been calculated that the generation of 22 000 kWh of energy is accompanied by the production of about one gram of fission fragments, and the emission of beta- and gamma-rays. Besides, reactors incorporating moderators produce strong fluxes of thermal neutrons. It may be noted that these thermal neutrons are now utilized to produce a variety of artificial radioisotopes widely used in science and technology.

The neutron and gamma radiations emitted by reactors are highly intensive, have a high penetrating ability, and adversely affect human health.To reducc these radiations to permissible levels,shielding is provided in reactor design and reactor control is made fully automatic and remote.

6. An example of a thermal heterogeneous reactor is the world's first nuclear power plant built in the Soviet Union and put in operation on June 27, 1954. Its power level is 5000 kW. The neutron moderator is made of graphite. The reactor core is a graphite cylinder

1.7 m high and 1.5 m in diameter, surrounded by a graphite reflector. The core has 128 vertical channels into which fuel elements are inserted. Fuel is uranium alloy enriched with uranium-235 and fabricated into slugs put around steel tubes with water flowing inside to cool the uranium. The core also has 22 channels for control rods of boron carbide which has a high neutron-absorption coefficient. With these rods, the power level of the reactor is maintained at the desired point. Since the water that cools the reactor becomes radioactive, it is admitted into a steam-generator where it gives up its heat to the water circulating in an outer closed circuit. This generates steam under a pressure of 12.5 atm and a temperature of 260 °C, which drives a turbine. All units of the nuclear power plant are controlled automatically and remotely.

7. The first nuclear power plant served as a prototype for the USSR's biggest nuclear power station built at Byeloyarsk and named after I. V. Kurchatov. The first 100-MW power-generating unit of this station was put in operation in 1964. The station has a very high efficiency owing to the use of steam under supercritical conditions (a pressure of 250 atm and a temperature of 535 to 565 °C).

Thermal reactors burning uranium can supply power on a limited scale because of the limited supply of uranium-235. Even with all the available supplies of uranium-235 burned in nuclear reactors they will generate an amount of energy roughly equivalent to the supplies of conventional fuel on the Earth.

8. As a way of augmenting the supply of nuclear fuel, we may utilize the fact that when the nuclei of uranium-238 and thorium-232 capture neutrons, they produce plutonium-239 and uranium-233 which are both effectively fissionable materials. The reaction that produces plutonium (see Sec. 82.6) may be written as follows:

$$_{92}\mathrm{U}^{238}\,(n,\ \gamma) \rightarrow {}_{92}\mathrm{U}^{239} \xrightarrow[23\ \mathrm{min}]{\beta-} {}_{93}\mathrm{Np}^{239} \xrightarrow[2.3\ \mathrm{days}]{\beta-} {}_{94}\mathrm{Pu}^{239}$$

The reaction with thorium proceeds as follows:

$$_{90}\mathrm{Th}^{232}\,(n,\ \gamma) \rightarrow {}_{90}\mathrm{Th}^{233} \xrightarrow[23\ \mathrm{min}]{\beta-} {}_{91}\mathrm{Pa}^{233} \xrightarrow[27.4\ \mathrm{days}]{\beta-} {}_{92}\mathrm{U}^{233}$$

The separation of plutonium-239 from the parent uranium-238 can be carried out by chemical means.

Each fission of uranium-235 in a reactor produces an average of 2.5 neutrons, one of which is needed to sustain the chain reaction. If the remaining 1.5 neutrons are absorbed by uranium-238 or thorium-232, then 1.5 plutonium-239 or uranium-233 nuclei may be produced to replace the fissioned uranium-235 nucleus. If at least one new fissionable atom is produced for each atom destroyed, the process is called *breeding*, and reactors in which the number of fissionable atoms produced exceeds that of fissionable atoms destroyed (the ratio of the two numbers is called the *breeding ratio*) are called *breeder reactors*.

Thermal reactors operating on uranium cannot breed nuclear fuel, because out of 100 absorptions of thermal neutrons by uranium-235, fission occurs only in 84.5 cases, and the maximum breeding ratio that is theoretically possible is $2.5 \times 0.845 - 1 = 1.11$ instead of 1.5. In fact, the breeding ratio is reduced still more because the moderator absorbs some neutrons and a further number is lost by leakage from the core. In reactors with a moderator, the breeding ratio is usually less than unity. For example, it is as little as 0.32 in the first Soviet nuclear power plant.

Breeder reactors use fast neutrons and, therefore, no moderators. The core is a uranium alloy enriched with uranium-235, and a heavy metal (such as bismuth or lead) that has a low neutron-absorption coefficient. Reactor control is effected by moving the reflector or by varying the mass of the fissionable material.

9. Among the nuclear reactors built in the Soviet Union are fast reactors that generate high-intensity neutron fluxes used for irradiation purposes, isotope production, and material testing.

As a pioneer in nuclear power generation, the Soviet Union is doing much to apply nuclear energy and nuclear reactors to other peaceful uses. Under the agreement signed by the USSR and the USA in 1964, large quantities of fissionable materials are allocated for peaceful uses, including desalination of sea water. It has been found that a fast reactor with a power level of 2200 MW can support a 510-MW electric power plant and a water desalination unit with an output of 180 000 $m^3$ per day at a cost of 2 or 3 kopeks per cubic metre. With 10 000 to 20 000 MW reactors, the cost of desalinated water will be so low that it may well be used to irrigate arid lands.

Concurrently with large-scale nuclear projects, work is under way in the Soviet Union on small-scale units. A drastic reduction in reactor size is important in portable and mobile installations, notably in those used for propulsion. Nuclear propulsion plants have already been installed in submarines and ocean-going ice-breakers.

## 82.11. THE ATOMIC BOMB

1. The atomic bomb is in effect a fast reactor of an explosive or uncontrolled type. The explosive is a pure fissionable material, such as uranium-235, plutonium-239 or uranium-233. Its critical size and mass are chosen such that a fast-rate (explosive) chain reaction can take place. Some reduction in the critical size and mass is possible owing to the omission of a reflector. For uranium-235, plutonium 239 and uranium-233, the critical mass appears to be 10 to 20 kg. With a density of $\rho = 18.7$ g/cm$^3$, the critical mass occupies the volume of a sphere with a radius of 4 to 6 cm. A further reduction in the

critical mass may be secured by surrounding the nuclear explosive with a *tamper*, a shell of a metal having a high density and a low neutron-absorption coefficient.

2. Before an atomic bomb is to be detonated, its charge should be kept in a state that will prevent a spontaneous chain reaction. Then, at the instant of detonation, the charge should instantaneously be manipulated into a position in which an uncontrolled chain reaction is set off. One way of doing this is to divide the critical mass into two halves each of which is held under conditions preventing the initiation of a chain reaction. At the instant of detonation, one half is fired into the other, the two combine into the critical mass, and an explosive chain reaction follows almost momentarily.

3. An explosive chain reaction releases a huge amount of energy, so that the temperature rises to $10^7$ °C. The destructive power of the A-bomb dropped on Hiroshima was equivalent to an explosion of 20 000 tons of TNT. Later A-bombs have a TNT equivalent of hundreds of kiloton and more. If we add to this the huge amount of radioactive fission fragments, including long-lived ones, the proportion of the catastrophe caused by an atomic explosion is not difficult to visualize. This is why the Soviet Union has been putting in much effort in order to ban nuclear weapons.

## 82.12. THERMONUCLEAR REACTIONS

1. Apart from the fission of heavy nuclei, there is one more way for releasing nuclear energy, namely by fusion of hydrogen isotopes into helium.

Hydrogen has three isotopes: light hydrogen or protium of atomic weight 1.008, heavy hydrogen or deuterium of atomic weight 2.015, and superheavy hydrogen or tritium of atomic weight 3.017. The nuclei of these isotopes are respectively called the proton, the deuteron and the triton, symbolized as ${}_1H^1$ (or ${}_1p^1$), ${}_1H^2$ or ${}_1D^2$, and ${}_1H^3$ or ${}_1T^3$. The binding energy per particle of the helium nucleus (see Sec. 80.4) is much greater than that of the hydrogen nuclei. Therefore, the combination of hydrogen nuclei into a helium nucleus should release an amount of energy. As regards energy release, the following reaction appears to be attractive:

$${}_1D^2 + {}_1T^3 \rightarrow {}_2He^4 + {}_0n^1 \tag{82.13}$$

This reaction would release 17.6 MeV of energy.

2. Energy release per nucleon in a fusion reaction is several times that obtained with the fission of heavy nuclei. As already noted, the fission of uranium releases about 200 MeV, which works out to about 0.85 MeV per nucleon. In contrast, the reaction described by Eq. (82.13) releases about 3.5 MeV per nucleon, or four times as much. A still greater amount of energy can be released when a helium nuc-

leus is formed by the fusion of four protons:

$$4{}_1p^1 \rightarrow {}_2He^4 + 2{}_{+1}e^0 + 2{}_0\nu_e^0 \qquad (82.13')$$

The energy release in this reaction is 26.8 MeV, so that the energy per nucleon is 6.7 MeV.

3. For the fusion reaction to take place, the reacting nucleons should overcome the potential barrier due to Coulomb repulsion. Let us find the height of the Coulomb barrier.

In order that deuterons can combine to form a helium nucleus, they should be brought close together, so that the separation between them is not more than twice the radius of the hydrogen nucleus, which is $r \approx 3 \times 10^{-15}$ m. This involves the expenditure of energy equal to the electrostatic potential energy of the nuclei separated by this distance, or $U = e^2/4\pi\varepsilon_0 r$. Substituting the numerical values, we find that the height of the Coulomb barrier is about 0.1 MeV. The deuterons can overcome this barrier if they have an appropriate kinetic energy at collision. The average thermal kinetic energy of deuterons ($3/2\ kT$) is 0.1 MeV, which is sufficient for overcoming the Coulomb barrier at $T = 2 \times 10^9$ K, which is much greater than the temperature existing inside the Sun (which is estimated at $10^7$ K).

Thus one has to attain an immensely high temperature before the rate of reaction is usefully great. This is the reason why fusion nuclear reactions are also called *thermonuclear* reactions.

4. Yet, it is possible to carry out fusion nuclear reactions at temperatures below $10^9$ K. The point is that the velocities of nuclei obey Maxwell statistics, and so at, say, $T \approx 10^7$ K there is a number of nuclei whose energy exceeds the Coulomb barrier and which can therefore initiate a fusion nuclear reaction.

The particles with energies high upon the "tail" of the Maxwell energy distribution at $T \approx 10^7$ K have energies of the order of tens of keV. However, this is still below the Coulomb barrier. In nuclear reactions involving charged particles at room temperature, the probability of tunnelling through the Coulomb barrier at collision is low, but it rapidly increases with an increase in the energy of the colliding particles. For example, for two deuterons with an average energy of 1.7 keV (which corresponds to a temperature of $2 \times 10^7$ K), the probability is $10^{47}$ times that of fusion of two deuterons with an average energy of 17 eV ($T = 2 \times 10^5$ K) by tunnelling. Thus, a temperature of $10^7$ K is sufficient for a fusion nuclear reaction to be initiated by the tunnelling of the nuclei whose energies lie high upon the "tail" of the Maxwell energy distribution. With the accompanying rise in temperature, these nuclei experience collisions at a growing rate, and this too promotes the interpenetration of the nuclei through the Coulomb barrier.

5. The conversion of hydrogen into helium is believed to be the principal source of energy in the Sun and the stars. This belief is

based on at least two considerations. For one thing, the temperature on the Sun (or, rather, in its central part) is about $10^7$ K. For another, from the emission spectra of the Sun it has been found that hydrogen appears to constitute about 80% (by mass) of the total material, with helium accounting for about 20%, and the remainder (not over 1%) being carbon, nitrogen and oxygen. This is a huge amount because the Sun's mass is $1.99 \times 10^{30}$ kg. Under the circumstances, fusion reactions seem to be self-suggestive, supplying all the energy necessary to make up for that lost through the emission of radiation. Every second, the Sun emits $3.8 \times 10^{26}$ J of energy, which represents a decrease of 4.3 million tons in its rest mass. It may be added, though, that the specific energy release from the Sun, that is, the energy emitted per unit mass per second is rather modest, being $1.9 \times \times 10^{-4}$ W/kg, or as little as 1% of the specific energy release by a living organism due to metabolism. This figure explains why the light intensity from our natural luminary is today about the same as it was several thousand million years ago.

6. In 1938, it was suggested that one of the possible mechanisms by which fusion reactions can proceed on the Sun might be the so-called *proton-proton chain*. In one of its modifications, the chain is believed to start by the direct combination of two protons with the formation of a deuteron and emission of a positron and neutrino:

$${}_1p^1 + {}_1p^1 \rightarrow {}_1D^2 + {}_{+1}e^0 + {}_0\nu_e^0$$

Then the deuteron interacts with a proton to form a nucleus of the light helium isotope, ${}_2He^3$, and the excess energy is emitted as gamma-radiation:

$${}_1D^2 + {}_1p^1 \rightarrow {}_2He^3 + \gamma$$

It may be noted that the positron produced in the first stage of the proton-proton chain combines with an electron of the plasma to produce gamma-radiation.

According to a hypothesis advanced in 1951, the most likely continuation of the proton-proton chain is the combination of helium nuclei, ${}_2He^3$, with the formation of a ${}_2He^4$ nucleus (an alpha-particle) and two protons by the reaction:

$${}_2He^3 + {}_2He^3 \rightarrow {}_2He^4 + 2{}_1p^1$$

The chain is terminated by the combination of two hydrogen nuclei into a helium nucleus, which is accompanied by the liberation of energy.

7. In 1939, H. Bethe suggested another mechanism by which fusion nuclear reactions can operate. It is called the *carbon-nitrogen* or *carbon cycle*. The carbon-nitrogen cycle requires the previous existence of a small amount of carbon; the carbon is not depleted in the cycle, but acts as a kind of catalyst for the fusion reaction. The cycle

starts with the penetration of a fast proton into a carbon (${}_6C^{12}$) nucleus, the formation of a nucleus of unstable radioisotope of nitrogen, ${}_7N^{13}$, and the emission of a gamma-quantum:

$${}_6C^{12} + {}_1p^1 \rightarrow {}_7N^{13} + \gamma$$

The ${}_7N^{13}$ nucleus decays with a half-life of 14 min by the reaction

$${}_1p^1 \rightarrow {}_0n^1 + {}_{+1}e^0 + {}_0\nu_e^0$$

into a nucleus of ${}_6C^{13}$ by the reaction:

$${}_7N^{13} \rightarrow {}_6C^{13} + {}_{+1}e^0 + {}_0\nu_e^0$$

About every 2.7 million years, the ${}_6C^{13}$ nucleus captures a proton to form a nucleus of ${}_7N^{14}$, the stable nitrogen isotope, by the reaction:

$${}_6C^{13} + {}_1p^1 \rightarrow {}_7N^{14} + \gamma$$

In about 32 million years, the nucleus of ${}_7N^{14}$ captures a proton and changes into an oxygen nucleus, ${}_8O^{15}$:

$${}_7N^{14} + {}_1p^1 \rightarrow {}_8O^{15} + \gamma$$

The unstable nucleus of ${}_8O^{15}$ with a half-life of 3 minutes emits a positron and a neutrino and changes into a nucleus, ${}_7N^{15}$:

$${}_8O^{15} \rightarrow {}_7N^{15} + {}_1e^0 + {}_0\nu_e^0$$

The cycle terminates in the absorption of a proton by ${}_7N^{15}$ which decays to a nucleus of ${}_6C^{12}$ and an alpha-particle in about 100 000 years:

$${}_7N^{15} + {}_1p^1 \rightarrow {}_6C^{12} + {}_2He^4$$

Another carbon-nitrogen cycle starts again with the absorption of a proton by a ${}_6C^{12}$ nucleus which occurs once in an average of 13 million years. The individual stages in the cycle are separated by enormously long intervals of time. However, it should be remembered that the cycle is closed and continuous. Therefore, the various reactions of the cycle occur on the Sun at the same time, although they start at different instants.

One carbon-nitrogen cycle produces a helium nucleus from four protons, with the emission of two positrons and gamma-radiation, to which should be added the radiation emitted when the positrons combine with the electrons of the plasma. The energy release is 26.8 MeV per helium nucleus. For one gram-atom of helium, energy output is 700 MWh. This amount is sufficient to make up for the radiation emitted by the Sun. Although the fusion nuclear reactions on the Sun use up hydrogen, its supplies on the Sun are sufficient to support these reactions and emission of light for thousands of millions of years.

8. From what has been said about the energy release by fusion nuclear reactions it is clear how important their realization under terrestrial conditions may be. The amount of deuterium contained

in one litre of common water would give up as much energy in a fusion reaction as we can obtain by burning about 350 litres of petrol.

For the first time a fusion nuclear reaction similar to those occurring on the Sun was brought about in the USSR and then in the USA in a *thermonuclear bomb*. In such a device, a self-sustaining fusion reaction of explosive character is initiated, with the aid of the enormous temperature developed by a "conventional" atomic (or fission) bomb. The material used for this reaction is a mixture of deuterium, $_1D^2$ and tritium, $_1H^3$.

9. A theoretical basis for *controlled* fusion nuclear reactions is provided by the reactions that take place in a high-temperature deuterium plasma. A major problem is to maintain rather than to initially create conditions necessary for the reaction. For a self-sustaining thermonuclear reaction it is essential that the rate of energy release by the system be not less than the rate of energy withdrawal from the system.

It has been calculated that for a self-sustaining controlled thermonuclear reaction to take place, the temperature of the deuterium plasma should be raised to several hundred million degrees. At temperatures of about $10^8$ K, thermonuclear reactions proceed at a sufficient rate and release huge amounts of energy. At this temperature, the power liberated by the combination of deuterium nuclei is about 3 kW/m$^3$, while at about $10^6$ K, the figure is only $10^{-17}$ W/m$^3$.

10. The principal cause of energy losses from the high-temperature plasma is its tremendous thermal conductivity which rapidly rises (in proportion to $T^{5/2}$) at the temperatures of interest. The plasma may lose its energy because the hot particles diffuse from the reacting region to the walls of the apparatus that contains the plasma. If the plasma is not isolated from any material surrounding it, the energy liberated by the fusion reaction will be lost through the walls, and it cannot be heated even to a few hundred thousand degrees. In other words, the problem is to contain the plasma within the reacting region for a sufficient length of time.

One approach to this problem, proposed in the Soviet Union in 1950, is to use a magnetic field for plasma containment. If a strong electric current be passed through a column of plasma, the current will set up a magnetic field of a shape usual for straight conductors. This field will produce electrodynamic forces that will cause the plasma to constrict so that it forms a pinched plasma column (see Sec. 12.8) and draw away from the walls of the container. Obviously, the plasma will pinch until the pressure due to electrodynamic forces is balanced by the gas-kinetic pressure of the plasma particles. In Fig. 82.5, the pinched plasma column, *2*, is kept away from the walls, *1*, by a magnetic field, *H*. The electric current, *I*, passed through the gas performs several functions, namely: (1) it produces the plasma through intensive ionization of the gas; (2) it causes the plasma to pinch; and (3) it heats up the plasma to a very high temperature owing to the liberation of Joule heat and constriction.

In the early experiments conducted in the USSR by L. A. Artsimovich (1909) and his co-workers, a strong pulse discharge was produced by a bank of high-capacitance capacitors in deuterium held at a pressure of 0.01 to 0.1 mm Hg. At the instant of the discharge, the maximum current was $10^5$ to $10^6$ A, with a rise time of 5 to 10 μs (from zero to a maximum). The resultant plasma rapidly pinched along the axis of the tube. At the end of the pinch effect the temperature of the plasma was $10^6$ K or even higher.

However, the plasma escapes this form of containment, because the pinched plasma column oscillates radially, that is, alternately expands and contracts, at a very high rate. Because of this instability, strains are produced in the plasma, which change its configuration. As a result, thermal insulation is practically non-existent, the deuterium is contaminated by the material of the container walls, and the plasma cools rapidly. All these events happen in a matter of a few microseconds, comparable with the time of a discharge pulse. By the time the current rises to its maximum, the temperature of the plasma has already fallen off its value it had when the plasma column first pinched.

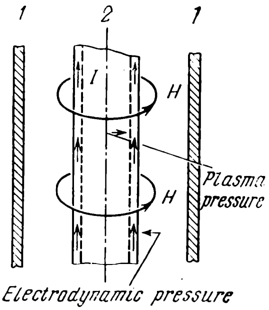


Fig. 82.5

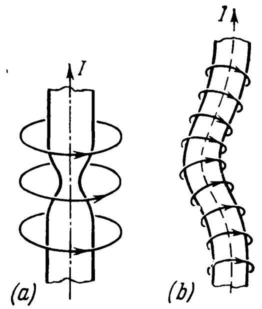


Fig. 82.6

For control of thermonuclear reactions, it is important to know the conditions under which a high-temperature plasma confined in a magnetic field of a suitable configuration will remain stable. Along with the search for ways and means by which the temperature of the plasma can be raised to a point required for a self-sustaining fusion reaction, this problem sets the course for further research in this direction.

The problem of plasma stability calls, above all, for a precise knowledge of the distortions that can occur in the pinched plasma column. Two simplest distortions, necking and twisting, are shown in Fig. 82.6. Without going into further detail, it may be poined out that at the necking (Fig.82.6*a*), the magnetic field intensity increases, as so do the electrodynamic forces that cause the plasma to pinch at that area. Meanwhile, the pressure of the plasma is the same at any point along its length and the plasma is free to flow along the column. This implies that the increased electrodynamic pressure at the necking will not be balanced by the plasma pressure, and the necking will continue until the column is ruptured. Similarly it may be shown that the twisting of the plasma column will continue until it assumes enormous proportions.

At present, many forms of plasma instability have been investigated, and several techniques have been devised to counteract them through the use of additional external magnetic fields not related to a current passing through the plasma column (see Sec. 48.8).

Among important advances in the field of controlled thermonuclear reactions was the production of a plasma with a controlled temperature of $10^8$ K by a team under G. I. Budker at the Siberian Branch of the USSR Academy of Sciences in 1964. The investigators used the pinch effect in the plasma and heating it by shock waves (see Sec. 30.8) produced by a rapid build-up of a magnetic field, over a time interval shorter than is required for plasma instabilities to develop. This build-up was produced by applying a power of about 200 GW from suitable spark-gaps in a matter of a few tenths of a microsecond. The resultant plasma had a density of $10^{13}$ to $10^{14}$ $m^{-3}$ and proved capable of sustaining a thermonuclear reaction. Later, similar results were obtained by Ye. K. Zavoisky and co-workers at the Kurchatov Institute of Atomic Energy.

At present, the most important problem is to increase the time during which a stable plasma can exist, and also its density.

Despite the fact that the unsolved problems are still many, the consistent efforts in the field of controlled thermonuclear reactions are bringing us nearer to the final goal—a practically inexhaustible source of energy.

# Chapter 83

# FUNDAMENTAL PARTICLES

## 83.1. TWO APPROACHES TO THE STRUCTURE OF FUNDAMENTAL PARTICLES

1. The term 'fundamental' or 'elementary' applies to particles which, at the present state of physics, cannot be treated as a system of other, "simpler" particles. In interactions with other particles or fields, a fundamental or elementary particle must behave as an entity.

Present-day physics deals with the nature, properties and mutual transformations of particles. The structure of the electron and the proton, particles well known already in the days of classical physics, is looked upon from two angles. Sometimes, these particles appear to be *structureless* material points possessing electric charge and mass. For example, if we are interested in the electric field of an electron away from it, the structure of the electron may well be ignored. The concept of a point-like fundamental particle is in agreement with relativity theory.

2. If a fundamental particle has finite dimensions, that is, has an extent, then, as a single whole, it should not be deformable because, by definition, deformation implies changes in the relative positions of the points of an assembly. As applied to a fundamental particle, this means that an external influence must momentarily be transmitted throughout the particle. However, that would contradict the basic tenet of relativity theory which asserts, that no interaction in nature can be transmitted at a velocity exceeding that of light in a vacuum. Thus, in the light of relativity theory, a fundamental particle must be a point.

Unfortunately, the concept of a point-like particle devoid of any extent leads to the entirely unsatisfactory conclusion that a substance occupying a definite volume, that is, a continuum, consists of particles having no dimensions.

3. It is known from electrodynamics that a stationary charged particle of charge $e$ produces an electrostatic field with a potential $\varphi$ equal to $e/4\pi\varepsilon_0 r$, where $r$ is the distance to the particle. For a structureless point-like particle, this means that its field at its location (that is, at $r \rightarrow 0$) has an infinite potential and, as a consequence, an infinite potential energy $U$. But then, the mass of the field due to the particle, will, according to the equation $m = U/c^2$, be likewise infinite. This is where "infinities" have made their way into the physics of structureless particles. In fact, the whole history of quantum mechanics in its application to fundamental particles has been one of a creation of various methods that would do away with "infinities" and define a structure for fundamental particles that would fit relativity theory.

4. For a proper insight into the structure of fundamental particles, one should deal with particles of extremely high energies. It is an easy matter to show that the interactions of particles having progressively increasing energies can furnish information on the processes that occur at progressively shorter distances between the particles. As an example, suppose that we are interested in what happens when two fundamental particles collide at a very short separation, $\delta$, between them. If the positions of the particles are defined with an uncertainty, then the uncertainty in their momentum will be $\Delta p \geqslant \hbar/\delta$, and that in their energy will be not less than $\Delta \mathscr{E} = \hbar c/\delta$ [see Eq. (16.25)]. Hence, it is seen that the shorter the distances related to the structure of elementary particles, the greater should

be their energy $\mathscr{E}$, because it cannot be smaller than $\Delta\mathscr{E}$ (see Sec. 16.7)

5. Like any division of science in the making, the physics of fundamental particles has not yet reached a state that it could be presented in a systematic way, especially in an elementary course. Therefore, in this closing chapter of the book, we shall limit ourselves to a brief outline of basic experimental observations, and their theoretical interpretation will only be touched upon in passing.

## 83.2. COSMIC RAYS

1. The Earth is continually showered with fast-moving highly energetic atomic nuclei and other particles that these nuclei produce in the terrestrial atmosphere. For the reason that they reach our home planet from outer space, they have come to be called *cosmic rays.*

2. The discovery of cosmic rays was made at the beginning of the present century as an outgrowth of studies into the ionization of the atmosphere. The method used was to mount an electroscope inside a closed thick-walled lead vessel and to observe the rate at which electric charge was lost by the electroscope. Despite all refinements and precautions against any accidental loss of charge, the conductivity of the air enclosed in the electroscope was found to be permanent. Finally it was found that the cause of this ionization was a highly-penetrating radiation of extra-terrestrial origin. That this radiation was of extra-terrestrial origin was confirmed by many observations. Among other things, it was noted that the ionization decreased only slightly to an altitude of 1000 m and then began to increase again, which could not be explained if one supposed that the source of ionization were inside the Earth. Further observations showed this radiation to be falling upon the Earth from all directions because the ionization of the air remained sensibly the same by night as by day.

The intensity of cosmic rays is determined by the number of particles that strike unit area per unit time (a second). The manner in which this intensity varies with altitude is shown in Fig. 83.1. The intensity, $I$, is expressed in relative units.

3. Cosmic rays have been found to be deflected by the Earth's magnetic field. Because of this, their observed intensity varies according to the latitude of the point of observation. The deflection is at its greatest in the equatorial region. This deflection affects the greatest number of particles which thus fail to reach the Earth's lower atmosphere. This is known as the *latitude effect.* At an altitude of 10 km, the latitude effect is 36%. It has also been noted that the Earth's magnetic field deflects the positive particles of cosmic rays in an easterly direction, and the negative particles in a westerly direction (the so-called *east-west effect*).

4. Cosmic rays outside the terrestrial atmosphere are called *primary cosmic rays*. Their composition is investigated with ionization chambers, counters, and nuclear emulsion blocks flown on balloons, in sounding rockets, and space probes. Observations have shown that primary cosmic rays contain atomic nuclei widely differing in mass number and that the energy per nucleon ranges from $10^9$ eV to $10^{20}$ eV. In primary cosmic radiation with energies under $10^{13}$ eV, protons account for 90% of the total material, helium nuclei for about 9%, the remainder (about 1%) being the nuclei of the heavier elements. The total energy brought in with cosmic rays every second (about $1.5 \times 10^6$ kW) is comparable with that due to the visible light of the stars. Some particles of primary cosmic rays may however have energies of the order of $10^{19}$ or $10^{20}$ eV.

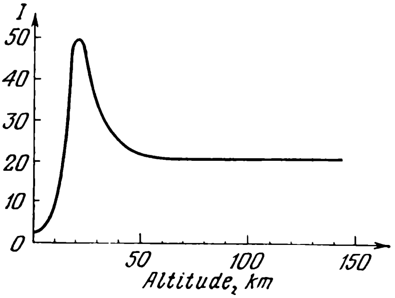


Fig. 83.1

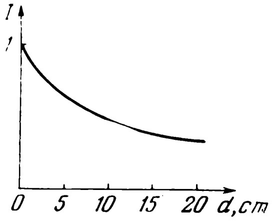


Fig. 83.2

5. It has been found that at altitudes in excess of 50 to 60 km, the intensity of cosmic rays remains practically constant (Fig. 83.1). Closer to the Earth's surface, the intensity suddenly changes. Today we know that this change is due to the so-called *secondary cosmic rays*. The point is that the particles of primary cosmic rays expend their enormous energy in inelastic collisions with the nuclei of nitrogen and oxygen in the Earth's upper atmosphere. As a result of these collisions and associated transformations, there appear secondary cosmic rays, and these reach the Earth's surface. From the ranges of the protons and heavy nuclei carried by primary cosmic rays it has been established that below an altitude of 20 km all of the cosmic radiation reaching the Earth's surface is secondary.

Secondary cosmic rays have a certain penetrating capacity. It can be measured by determining their intensity, $I$, after they have passed through a layer of lead with a thickness $d$. The results of such measurements are shown in the plot of Fig. 83.2 where the unit of intensity is that existing at $d = 0$. In the thickness range from 0 to 10-13 cm, secondary cosmic rays are rapidly attenuated, but any further increase in the thickness *d* leaves their intensity practically unchanged.

This observation suggests the existence of a *soft* and a *hard* component in secondary cosmic rays. The soft component is strongly absorbed by lead, while the hard component retains its penetrating capacity even in lead. The hard component consists of the heavier fast charged particles which lose their energy mainly through the ionization of the atoms they encounter on their way. The soft component consists of light charged particles such as electrons and positrons, and also photons.

6. Recent observations have discovered *radiation belts* near and around the Earth (the *Van-Allen belts*), inside which cosmic rays show an increased intensity (in comparison with that observed at low altitudes). These belts, it has been found, owe their origin to the capture of charged particles by the terrestrial magnetic field. The inner Van-Allen belt is within 600 to 6000 km of the Earth. At some places, such as opposite the magnetic anomalies in the southern Atlantic, it comes within 300 km of the Earth's surface. The outer Van-Allen belt is within $2 \times 10^4$ to $6 \times 10^4$ km of the Earth, being within 300 to 1500 km in some places (at latitudes of 55-70°). The inner Van-Allen belt contains mainly protons with energies from 10 or 20 MeV to 700-800 MeV. In the outer Van-Allen belt, the principal constituents are electrons with energies under 100 keV. It appears that all celestial bodies having a magnetic field should be surrounded by such radiation belts. Knowledge of these radiation belts is important to manned flights to space and, above all, to the Moon.

7. Several hypotheses have been put forward to explain the origin of primary cosmic rays. They are all based on data about the energies of the primary particles and information gleaned by radio astronomy. It is believed that the primaries gain their energy due to the acceleration they experience in the electromagnetic fields of the stars and the Sun. It is essential that this acceleration should be *gradual*. Otherwise, that is, if the heavy and superheavy nuclei carried by the primary radiation received their energy (which runs as high as $10^{13}$ eV and more) at once, as a result of some extremely fast processes, they would "evaporate" into the constituent nucleons. For the total binding energy would be insufficient to hold them together even at lower values of energy input. Rotation of the stars that have magnetic fields produces electric fields, and the protons and nuclei are trapped in closed paths within these electric fields by the stars' magnetic fields until they gain enormous accelerations.

In a theory proposed by Fermi, charged cosmic-ray particles are assumed to be partially trapped within the galaxy and accelerated through the operation of local variations caused in the general field by the motion of ion clouds in interstellar space. As a result, the primaries may be accelerated to the highest energies that can exist in such a field. The energy required to inject primaries into the field (the *injection threshold*) is supplied by shock waves (see Sec. 30.8) produced by collisions of gas masses as a result of explosions of *supernovae* (super-new-stars). The energy of such explosions is internuclear in origin.

## 83.3. THE MU-MESON

1. For a long time, interactions of cosmic rays with matter remained the only tool for the study of fundamental particles. Since the advent of particle accelerators (see Secs. 41.4 and 41.6), it has become possible to investigate the nature, mutual transformations and structure of fundamental particles under laboratory conditions. The powerful accelerator at Serpukhov near Moscow can impart particles energies up to 76 GeV (1 GeV = $10^9$ eV). Yet, some particles in primary cosmic rays may have still higher energies—up to $10^{20}$ eV. This is why cosmic rays have retained their importance in the study of fundamental particles and their transformations at very high energies.

2. The fact that a collision of a fast charged cosmic primary (such as a proton) with a nucleus of atmospheric nitrogen or oxygen can produce further particles stems from the mass-energy relations. Since a proton has an energy of $10^4$ GeV, which is about $10^4$ times its rest mass, its collision with a nucleus not only breaks it up into the nucleons, but also imparts the fragments a kinetic energy sufficient to produce further particles having rest mass and also only having field mass (photons).

3. In 1937, Anderson and Neddermeyer reported the discovery of a new particle. In their observations they had used lead filters to stop cosmic rays, and a cloud chamber placed in a magnetic field. From the curvature caused by the magnetic field in the track of the newly discovered particle they found that it was positively charged. Its ionization effect and energy loss were such that its rest mass appeared tobe about 200 $m_e$,where $m_e$isthe rest mass of an electron. Since this mass is intermediate between those of the electron and the proton, the new particle was called the *meson*. Later, however, several kinds of particles with masses in this range but with a negative charge were discovered, and also called mesons. As a distinction, the meson discovered by Anderson and Neddermeyer has come to be known as the *mu-meson* (or *muon*).

4. From changes in the intensity of the hard cosmic-ray component with altitude it has been found that the intensity of mu-mesons at sea level is markedly lower than it is at the top of a high mountain (that is, at an altitude). This decrease in intensity in the time interval during which the mu-mesons travel a distance $H$, equal to the height of the mountain appeared to be caused by the spontaneous

decay of a mu-meson into other particles. The decay time can be estimated as follows. Assuming that mu-mesons travel at a velocity close to that of light in a vacuum, their transit time is $t = H/v \approx$ $\approx H/c$. The fundamental law of radioactive decay has been found to be expressed by Eq. (81.3) which we will write again for convenience:

$$N = N_0 \exp(-\lambda t) \tag{83.1}$$

From the known number of mu-mesons at the top of the mountain, $N_0$, and at sea level, $N$, we can find the decay constant, $\lambda$, and, finally, the mean life, $\tau_\mu$, of a mu-meson. By definition, the mean life of a mu-meson is the average time during which the number of mu-mesons reduces to $1/e$ of their original number, that is, $N = N_0/e$. Then by Eq. (83.1) we have

$$N_0/e = N_0 \exp(-\lambda\tau_\mu)$$

whence

$$-1 = -\lambda\tau_\mu$$

or

$$\tau_\mu = 1/\lambda$$

That is, Eq. (83.1) may be re-cast as follows

$$N = N_0 \exp(-t/\tau_\mu) \tag{83.2}$$

On taking a logarithm, we obtain from Eq. (83.2) that

$$\tau_\mu = t/\ln(N_0/N) \tag{83.2$'$}$$

According to (83.2), $\tau_\mu \approx 10^{-5}$ s. More accurate observations with the use of lead filters to absorb the hard cosmic-ray component and better mu-meson detectors have given $\tau_\mu = 2.2 \times 10^{-6}$ s.

5. Mu-mesons may decay according to the reactions:

$$\mu^+ \rightarrow {}_{+1}e^0 + {}_0\nu_e^0 + {}_0\widetilde{\nu_\mu^0} \tag{83.3}$$

$$\mu^- \rightarrow {}_{-1}e^0 + {}_0\widetilde{\nu}_e^0 + {}_0\nu_\mu^0 \tag{83.3$'$}$$

where ${}_0\nu_e^0$ and ${}_0\nu_\mu^0$ stand for the neutrino and the neutretto, respectively (and those with the ~ sign, for their "anti"-counterparts).

According to present-day data, the neutrino ${}_0\nu_e^0$ and the antineutrino ${}_0\nu_e^0$ emitted with electrons differ from the neutretto ${}_0\nu_\mu^0$ and the antineutretto ${}_0\widetilde{\nu}_\mu^0$ emitted with mu-mesons. Accordingly, they are sometimes called the electron-type neutrino (or antineutrino) and 'muon-type neutrino' (or antineutrino).

The electrons, ${}_{-1}e^1$, and the positrons, ${}_{+}e^0$, emitted in the decay of mu-mesons were readily detected with nuclear emulsion blocks. The energy of an electron (or a positron) produced in the reactions described by Eqs. (83.3) and (83.3′) was found not to exceed 50 MeV and to be less than the energy of a positive or a negative mu-meson. This requires that in addition to an electron (or a positron) the decay of a mu-meson should produce two more particles. Application of the conservation laws to the decay of a mu-meson has led to the reactions described by (83.3) and (83.3′). From these equations it follows that mu-mesons, like electrons, should have a spin of $\hbar/2$, because a neutrino and an antineutrino have a spin of $+\hbar/2$ and $-\hbar/2$, respectively, and these spins therefore cancel out (provided each spin has no preferred orientation).

6. Observations have shown that mu-mesons have weak interactions with nuclei. In fact, they are nuclearly inactive particles. Among other things it has been found that the time of interaction between mu-mesons and lead nuclei is $10^{-8}$ s, which is $10^{14}$ times the nuclear year ($10^{-22}$ s) which is characteristic of internucleonic interactions (Sec. 82.4). In other words, the interaction of mu-mesons with nuclei is $1/10^{14}$ of what is required for the transmission of the short-range nuclear forces between nucleons. This observation has suggested a very important conclusion that mu-mesons cannot be the nuclear field quanta that accomplish the exchange interaction of nucleons in the nucleus (Sec. 80.5).

7. The weak interaction of mu-mesons with nuclei is similar to the similarly weak interaction of electrons, positrons, neutrino and antineutrino with nuclei. On this basis, they are classed in a single group called *leptons* (see Table 83.2). To-day, weak interactions are treated as a special type of interaction or coupling, observed in nature along with electromagnetic interactions or couplings (Sec. 83.5).

All particles in the lepton class are assigned a lepton number of +1. In contrast, all antileptons are said to have a lepton number of −1. The remaining particles have no lepton numbers at all. The processes involving leptons are extremely slow and proceed so that the sum of lepton numbers is conserved (the law of conservation of lepton number).

From the fact that $\mu$-mesons are nuclearly inactive it follows that they cannot be the cosmic-ray primaries that interact with the nuclei of atmospheric gases. By 1946, the physics of cosmic rays had accumulated a wealth of data suggesting that primary cosmic rays should contain nuclearly active particles strongly interacting with nuclei and having masses intermediate between those of the $\mu^{\pm}$-meson and the proton.

## 83.4. THE PI-MESON

1. In 1946, W. M. Powell and his collaborators found tracks in nuclear emulsion blocks, that could only be produced by particles heavier than $\mu$-mesons, that is, those with a rest mass close on 300 $m_e$.

In schematic form, one such track is shown in Fig. 83.3. Here may be seen that coming in from the upper right a particle (found to to have a rest mass of about 300 $m_e$) stops at point $A$, producing a particle (with a rest mass of about 200 $m_e$) which moves as far as point $B$ where it stops, too. The track $AB$ is that of a $\mu^+$-meson which decays at point $B$. The primary particle decaying at point $A$ into a $\mu$-meson was called the *pi-meson*, or *pion*. Schematically, the "$\pi$ -$\mu$-e" decay is shown in Fig. 83.4.

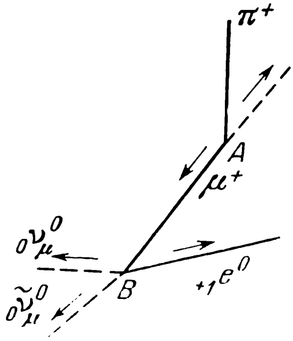


Fig. 83.3

After the energies of the two particles had been determined from the range in the emulsions it became apparent that, in order to conserve energy and momentum, the decay of a pi-meson should produce one more particle that would carry away with it a greater proportion of energy than the $\mu^+$-meson. According to the conservation laws, its mass should be considerably smaller than the rest mass of the meson and the electron. This particle could not be a gamma-photon because no electron-positron pairs had appeared on its track in the emulsion (see Sec. 82.3). Finally, it was established that this particle was the neutretto or muon-type meson neutrino, $_0\nu_\mu^0$, produced by the reaction:

$$\pi^+ \rightarrow \mu^+ + {}_0\nu_\mu^0 \qquad (83.4)$$

2. In the same emulsions the investigators noted several cases in which a meson track terminated in a "star" such as shown in Fig. 83.5.

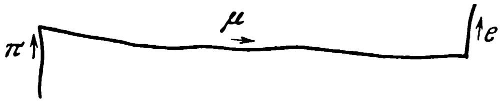


Fig. 83.4

This occurrence was interpreted as the disintegration of a nucleus which had captured the meson. From the conservation of energy and mass and also on the basis of the binding and kinetic energies of all the particles, the rest energy of a pi-meson was found to be close to 140 MeV which corresponds to a rest mass of about 270 $m_e$.

3. From further observations it was concluded that the "star"-producing mesons were negative pi-mesons. This explains why they are captured by a nucleus before they can decay. In contrast, positive pi-mesons need to have a much greater kinetic energy in order to overcome the Coulomb repulsion from the nucleus, enter it and

cause a "star" break-up. Most often, positive pi-mesons merely decay near a nucleus by the reaction described by Eq. (83.4).

In cosmic rays, pi-mesons are produced as fast cosmic particles (protons and alpha-particles) cause the disintegration of the nuclei of atmospheric gases.

4. Soon after pi-mesons had been discovered in cosmic rays, they were produced artificially in a laboratory. Calculations which we omit show that the fast particles used for pi-meson production must have very high energies. For example, the energy of protons must be of the order of 300 MeV.

In sketch form, pi-meson production is illustrated in Fig. 83.6. A beryllium (or carbon) target, $A$, placed inside the dee of a cyclotron, was bombarded with fast protons, ${}_1p^1$ (and also alpha-particles) which knocked pi-mesons out of the target at arbitrary angles. The magnetic field of the cyclotron caused the ejected pi-mesons to trace out circular paths with radii determined by their velocities. Any pi-mesons ejected from the target in the forward direction were separated into two streams so that the negative ones were coupled out of the dee, and the positive ones were deflected inside the dee. Pions ejected in the rearward directions were deflected the other way around (this is not shown in Fig. 83.6).

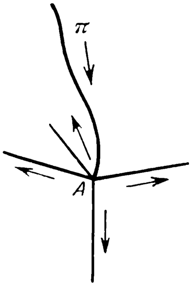


Fig. 83.5

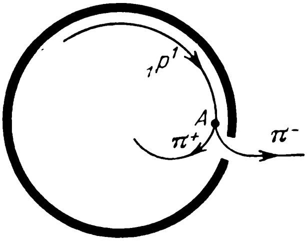


Fig. 83.6

The pi-mesons coupled out of the cyclotron were identified for energy, momentum and mass from the deflection of the pi-meson beam in an electric and a magnetic field. The lifetime of pi-mesons, that is the time interval between the instant when a pi-meson is produced and the instant when it disintegrates into a mu-meson and a neutrino or antineutrino, was measured with scintillation counters (see Sec. 81.8).

It has been found that positive and negative pi-mesons have the same lifetime which is by two orders of magnitude shorter than that of mu-mesons. This is in agreement with the observation that cosmic rays at sea-level contain considerably fewer pi-mesons than mu-mesons.

A negative pi-meson decays according to Eq. (83.4):

$$\pi^- \to \mu^- + {}_0\tilde{\nu}^0_\mu \qquad (83.5)$$

where $\mu^-$ is a negative mu-meson and ${}_0\tilde{\nu}^0_\mu$ is a muon-type antineutrino (antineutretto).

Experiments on artificially produced pi-mesons have yeilded a very accurate value for their rest mass.

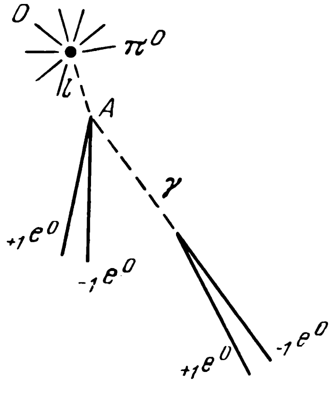


Fig. 83.7

5. In addition to positive and negative pi-mesons, those having no charge have also been discovered. These neutral pions, or $\pi^0$-mesons, have a very short lifetime and decay into two gamma-photons according to the reaction:

$$\pi^0 \to \gamma + \gamma \qquad (83.6)$$

Then these gamma-photons produce electron-positron pairs (see Sec. 82.3) as illustrated below:

$$\begin{tikzcd}
 & & {}_{+1}e^0 + {}_{-1}e^0 \\
 & \gamma \arrow[ur] & \\
\pi^0 \arrow[ur] \arrow[dr] & & \\
 & \gamma \arrow[dr] & \\
 & & {}_{+1}e^0 + {}_{-1}e^0
\end{tikzcd} \qquad (83.7)$$

This reaction provides a basis for determining the lifetime of neutral pi-mesons. Since neither a $\pi^0$-meson nor a gamma-photon leave any tracks in emulsions, the lifetime of a $\pi^0$-meson can conveniently be found by measuring the distance $l$ in the emulsion from the point $O$ where a $\pi^0$-meson is produced by interaction of fast charged particles with a nucleus, to the nearest point $A$ where an electron-position pair, ${}_{+1}e^0$-${}_{-1}e^0$, appears (Fig. 83.7). This technique yields a lifetime of about $5 \times 10^{-15}$ s for the neutral pi-meson. More accurate measurements give

$$\tau_{\pi^0} = 0.80 \times 10^{-16} \text{ s}$$

These figures illustrate the degree of accuracy with which the characteristics of fundamental particles can be measured by present-day nuclear physics.

6. As is with any neutral particle, the mass of the neutral pi-meson cannot be determined from the curvature of its path in a magnetic field. Instead, resort is made to the laws of conservation of energy and momentum and the interactions of the neutral particle with other particles. On interacting with a proton, a negative pi-

meson decays to a neutral pi-meson, and the proton to a neutron by the reaction:

$$\pi^- + {}_1p^1 \rightarrow \pi^0 + {}_0n^1$$

The neutral pi-meson decays into two gamma-quanta

$$\pi^0 \rightarrow \gamma + \gamma$$

Now one can readily determine the rest mass of the neutral pi-meson because the masses and energies of the proton, neutron and pi$^-$-meson participating in these transformations are known, the energy of the gamma-photons may be measured, and data on the momenta of the participating particles are available.

7. The spin of the neutral pi-meson may be identified from its decay, Eq. (83.6). As is seen, the spin of the pi$^0$-meson is not integral, for otherwise it would not decay to two gamma-photons each of which has a spin of $\hbar$. Now it is known that the neutral pi-meson has zero spin. This may be interpreted as an indication that in the decay described by Eq. (83.6) the spins of the photons cancel out. Available data point out that positive and negative pi-mesons have likewise zero spin.

## 83.5. CLASSIFICATION OF COUPLINGS IN NUCLEAR PHYSICS

1. At present, nuclear physics treats the difference in rest mass between charged ($\pi^\pm$) and neutral ($\pi^0$) pi-mesons, and also the difference in rest mass between the protonic and electronic states of a nucleon in the nucleus in the light of the *charge independence of nuclear forces* and all the related observations.

It has been pointed out in Sec. 80.5 that the nuclear forces operating between nucleons are charge-independent. As regards charge, the difference between the proton and the neutron manifests itself only in electromagnetic and not in nuclear interactions. The charge independence of nuclear forces imposes certain limitations on the interaction of pi-mesons with nucleons. The short range of nuclear forces can then be explained as an exchange of pi-mesons between nucleons (see Sec. 80.5).

2. Calculations show that for nuclear forces to be charge-independent, the interaction of nucleons with charged pi-mesons must be identically independent of the polarity of the pionic charge. If the nucleus were free from electromagnetic coupling and *pi-meson nuclear coupling* were the only form of interaction possible, the charge independence of nuclear forces would require the same value of mass for the proton and the neutron and the same mass for all pi-mesons. This is not so, however, and the explanation lies in the fact that in addition to nuclear interaction, there is also electromagnetic

coupling arising from the interaction of particle charge with the Maxwell field. Since the energies of interacting charged particles differ from those of neutral particles, their rest masses differ, too. Much as the effect of spin on electrons in an atom results in the splitting of the energy levels of electrons (Sec. 72.6), so the addition of electromagnetic interaction to the nuclear one is responsible for the fact that the dual (protonic-neutronic) state of a nucleon splits up into two states as regards rest mass, and the neutron, $_0n^1$, and the proton, $_1p^1$, appear to have different rest masses. For the same reason, instead of the same rest mass, charged pi-mesons have a rest mass slightly different from that of the neutral pi-meson.

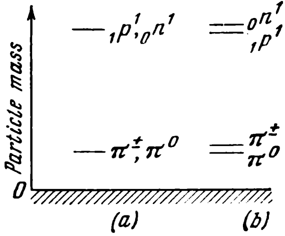


Fig. 83.8

In approximate form, these masses, with and without allowance for electromagnetic interaction (*a* and *b*, respectively) are shown in Fig. 83.8.

3. Because of the charge independence of nuclear forces and the presence of additional electromagnetic coupling in the nucleus, there is a difference in mass between neutral and charged particles. This difference is of electromagnetic origin. In a way, the mass of a particle is a sum of a basic contribution which has nuclear-mesonic origin, and an additional mass electromagnetic in origin. It is believed that the additional mass may be either positive, that is, related to an increase in energy, or negative if there is a decrease in energy caused by electromagnetic interaction. For example, this additional energy is negative in the case of neutrons because their mass, $m_n$. is by 2.53 $m_e$ greater than that of protons. It is positive with neutral pi-mesons whose mass is by 8.98 $m_e$ smaller than that of charged ($\pi^\pm$) pi-mesons. With charged ($\pi^\pm$) pi-mesons, it is the same, and so these pi-mesons have the same rest mass.

4. The charge independence of nuclear forces and all the related corollaries are characteristic of so-called *strong interactions*. Apart from nuclear forces between nucleons, examples of strong interactions are the production of mesons by nuclear interactions at high energies. All processes involving strong interactions proceed at a very high rate, typically inside the time interval corresponding to the nuclear year approximately equal to $10^{-22}$ s (see Table 83.1).

5. Further forms are *electromagnetic interactions* and *weak interactions*. One example of weak interactions is the interaction of mu-mesons with the nucleus. Another example is offered by processes leading to the beta-decay of nuclei (both positive and negative). Electromagnetic interactions arise from the fact that fundamental particles have charge. It is the cause of Coulomb repulsion of pro-

tons in nuclei, and also the production and annihilation of electron-positron pairs. Charge independence does not apply to electromagnetic and weak interactions, because the forces of interaction in these forms depend on whether the interacting particles have charge.

A comparison of the three forms of interactions between fundamental particles is given in Table 83.1.

*Table 83.1*

| Interaction | Relative strength of interactions | Typical time constant, s |
|---|---|---|
| Strong | 1 | $10^{-23}$-$10^{-22}$ |
| Electromagnetic | 1/137 | $10^{-20}$-$10^{-18}$ |
| Weak | $10^{-14}$ | $10^{-10}$-$10^{-8}$ |

83.6. K-MESONS AND HYPERONS

1. In 1949, the existence of heavy unstable particles, with a rest mass of about 1000 $m_e$, was established. They were called *K-mesons*, or *kaons*. A considerable number of events have been observed in photographic emulsions in which K-mesons produce "star" transformations. The products from these transformations are pi- and mu-mesons, and also electrons and positrons.

At present, it is a fact that there are charged K-mesons, namely positive ($K^+$) and negative ($K^-$), neutral K-mesons ($K^0$), neutral anti-K-mesons ($\tilde{K}^0$) (antiparticles will be discussed in greater detail in Sec. 83.7). Furthermore, it has been found that there are two types of neutral K-mesons, namely $K^0_S$ and $K^0_L$ differing in lifetime (see Table 83.2).

2. Their tracks in photographic emulsions have yielded enough evidence that, as regards rest mass, charge and spin, all K-mesons are close relatives. Like pi-mesons, K-mesons have zero spin.

3. Information about the various K-mesons has mainly been derived from the study of their decays and the mechanisms by which they are produced. As an example, we shall examine the decay of a positive $K^+$-meson. More often, a $K^+$-meson decays into a positive mu-meson and a neutretto or muon-type neutrino, ${}_0\nu^0_\mu$, by the reaction:

$$K^+_{\mu 2} \rightarrow \mu^+ + {}_0\nu^0_\mu$$

The subscripts of the $K^+$-meson indicate the product particles and their number. The same convention will be used in other reactions.

A fairly often occurrence is the decay of a $K^+$-meson into two

*Table 83.2*

| Particle and antiparticle | Symbol | Charge, $e$ | Mass, MeV | Lifetime, s | Spin, $\hbar$ |
|---|---|---|---|---|---|
| Photon | $\nu$ | 0 | 0 | stable | 1 |
| **LEPTONS:** | | | | | |
| *Neutrino*: | | | | | |
| electron-type, antineutrino | $\nu_e\ \tilde{\nu}_e$ | 0 0 | 0 | stable | 1/2 |
| meson-type (neutretto) antineutretto | $\nu_\mu\ \tilde{\nu}_\mu$ | 0 0 | 0 | stable | 1/2 |
| *Electrons*: | | | | | |
| electron, positron | $e^-\ e^+$ | $-1\ +1$ | 0.511 | stable | 1/2 |
| *Muons*: | | | | | |
| $\mu^-$-meson, $\mu^+$-meson | $\mu^-\ \mu^+$ | $-1\ +1$ | 106 | $2.2\times10^{-6}$ | 1/2 |
| **MESONS:** | | | | | |
| *Pions*: | | | | | |
| $\pi^+$-meson, $\pi^-$-meson | $\pi^+\ \pi^-$ | $+1\ -1$ | 140 | $2.6\times10^{-8}$ | 0 |
| $\pi^0$-meson | $\pi^0$ | 0 | 135 | $0.8\times10^{-16}$ | 0 |
| *Kaons*: | | | | | |
| $K^+$-meson, $K^-$-meson | $K^+\ K^-$ | $+1\ -1$ | 494 | $1.2\times10^{-8}$ | 0 |
| $K^0$-meson, anti-$K^0$-meson | $K^0\ \tilde{K}^0$ | 0 0 | 498 | $K^0_S : 0.86\times10^{-10}$<br>$K^0_L : 5.38\times10^{-8}$ | 0 |
| $\eta^0$-meson | $\eta^0$ | 0 | 549 | $2.4\times10^{-19}$ (?) | 0 |
| **BARYONS:** | | | | | |
| *Nucleons*: | | | | | |
| proton, antiproton | $p\ \tilde{p}$ | $+1\ -1$ | 938.2 | stable | 1/2 |
| neutron, antineutron | $n\ \tilde{n}$ | 0 0 | 939.6 | $0.93\times10^3$ | 1/2 |
| *Hyperons* | | | | | |
| $\Lambda^0$-hyperon, anti-$\Lambda^0$-hyperon | $\Lambda^0\ \tilde{\Lambda}^0$ | 0 0 | 1116 | $2.5\times10^{-10}$ | 1/2 |
| $\Sigma^+$-hyperon, anti-$\Sigma^+$-hyperon | $\Sigma^+\tilde{\Sigma}^+$ | +1 1— | 1189 | $0.8\times10^{-10}$ | 1/2 |

*Table 83.2 (Continued)*

| Particle and antiparticle | Symbol | Charge, e | Mass, MeV | Lifetime, s | Spin, ℏ |
|---|---|---|---|---|---|
| Photon | $\nu$ | 0 | 0 | stable | 1 |
| $\Sigma^-$-hyperon, anti- $\Sigma^-$-hyperon | $\Sigma^-\ \tilde{\Sigma}^-$ | —1 +1 | 1197 | $1.5\times10^{-10}$ | 1/2 |
| $\Sigma^0$-hyperon, anti- $\Sigma^0$-hyperon | $\Sigma^0\ \tilde{\Sigma}_0$ | 0 0 | 1192 | $<10^{-14}$ | $^1/_2$ |
| $\Xi^-$-hyperon, anti- $\Xi^-$-hyperon | $\Xi^-\ \tilde{\Xi}^-$ | —1 +1 | 1321 | $1.7\times10^{-10}$ | 1/2 |
| $\Xi^0$-hyperon, anti- $\Xi^0$-hyperon | $\Xi^0\ \tilde{\Xi}^0$ | 0 0 | 1315 | $3\times10^{-10}$ | 1/2 |
| $\Omega^-$-hyperon, anti- $\Omega^-$-hyperon | $\Omega^-\ \tilde{\Omega}^-$ | —1 +1 | 1672 | $1.3\times10^{-10}$ | 3/2 (?) |

pi-mesons:

$$K^+_{\pi2}\rightarrow\pi^++\pi^0$$

There are four more types of decay available to a K⁺-meson:

$$K^+_{\pi3}\rightarrow\pi^+\ +\pi^-+\pi^+$$
$$K^+_{\pi3}\rightarrow\pi^+\ +\pi^0+\pi^0$$
$$K^+_{\mu3}\rightarrow\mu^+\ +\pi^0+{}_0\nu^0_\mu$$
$$K^+_{e3}\rightarrow{}_{+1}e^0+\pi^0+{}_0\nu^0_e$$

These decays occur with both natural K-mesons found in cosmic rays and with artificial positive K-mesons produced in particle accelerators.

4. Photographic emulsions have also supplied evidence for the existence of a large group of particles called *hyperons* (or Y-particles), with a rest mass exceeding that of nucleons. Observations in bubble chambers placed in a magnetic field have revealed that these are positive, negative and neutral hyperons. All hyperons have a spin of $\hbar/2$, except the $\Omega^-$-hyperon which has a spin of $3\hbar/2$.

Protons, neutrons and hyperons are classed in the same group of heavy particles called *baryons* (see Table 83.2).

All fundamental particles may be identified on the basis of their baryonic (nucleonic) number. Assuming that this number is unity for baryons,—1 for antibaryons, and zero for particles not belonging to baryons, we may formulate the law of conservation of baryonic (nucleonic) number as follows: in all nuclear transformations in an isolated system, the baryonic number remains unchanged.

Like the law of conservation of electric charge, that of conservation of baryonic number applies to both strong (nuclear) and electromagnetic interactions. For example, if we had one baryon, say a neutron, with a baryonic number of unity before a transformation, the transformation should produce one baryon, which may be a neutron, a proton or one of hyperons, with a nuclear charge of unity, too.

5. Hyperons are extremely unstable and decay in a matter of $10^{-11}$ to $10^{-10}$ s. As an example, the decay scheme of a $\Lambda^0$-hyperon is shown in Fig. 83.9. It is seen that the $\Lambda^0$-hyperon decays to a proton and to a pi$^-$-meson which in turn produces a star:

$$\Lambda^0 \rightarrow {}_1p^1 + \pi^-$$

The decay schemes, such as shown in Fig. 83.9, provide all the information necessary to determine the energies of the proton and the pi-meson and also the mass of the $\Lambda^0$-hyperon. The lifetime of hyperons can accurately be determined from the instants of their production and decay observed in bubble chambers. All of these data about hyperons and other fundamental particles are presented in Table 83.2. It should be noted that the masses of the particles are stated in MeV, units of energy, and not in units of electronic mass, as was practised before. Since the mass of an electron is 0.5 MeV, the electronic mass of any particle can be found by multiplying its mass in MeV by two.

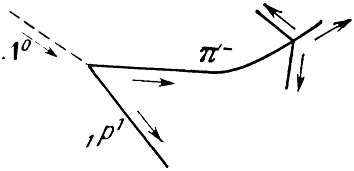


Fig. 83.9

6. Both K-mesons and hyperons have been found to have unusual properties widely differing from those of other fundamental particles. Because of this, they have been called *strange particles.*

It is beyond doubt that a strange particle is produced by a strong interaction because the time during which it is produced is comparable with the time constant of strong interaction ($10^{-23}$ to $10^{-22}$ s). At the same time, decaying to nuclearly active pi-mesons according to the schemes given above, that is, to particles characterized by strong interactions, K-mesons have lifetimes ($10^{-10}$ to $10^{-8}$ s) longer than is observed for nuclearly inactive mu-mesons, that is, for weak interactions. Then, it has been established that K-mesons and hyperons are always produced as doublets, and then in preferred combinations.

Last but not least, a marked distinction has been found in the conditions under which K-mesons are produced and in their interactions with other particles. For example, at energies of a few GeV, the number of $K^+$-mesons produced is $10^2$ times as great as that of $K^-$-mesons; while a $K^+$-meson may make a doublet with both a $K^-$-meson and a hyperon, a $K^-$-meson does so with a $K^+$-meson only.

7. A present, the nature of the strange particles is explained on the basis of charge independence. That is, the principle of charge inde-

pendence is treated as applying not only to the pi-meson exchange between nucleons in the nucleus, but also to the interactions of K-mesons and hyperons. A more detailed examination of this matter lies outside the scope of this book, as do many other aspects of the physics of fundamental particles.

## 83.7. ANTIPARTICLES

1. At present it is recognized that, with a few exceptions, for each fundamental particle there should exist an antiparticle. In fact, the exceptions include particles which are each its own antiparticle. Thus, the photon, the neutral pi-meson, and the neutral K-meson are each its own antiparticle. They may be called absolutely or truly neutral. This ought not to be confused with electrical neutrality, because an electrically uncharged particle may well have an antiparticle. Conversely, the lack of an electric charge is not enough for a particle to be absolutely or truly neutral.

The most common particle-antiparticle pairs are the electron and the positron, the positive and negative mu-meson, the positive and negative pi-meson, the electron-type and muon-type neutrinos ${}_0\nu_e^0$ and ${}_0\nu_\mu^0$, and their respective antineutrinos, the positive and negative K-meson.

A particle and its antiparticle have the same rest mass, the same spin, and the same lifetime. The electric and nuclear charges are also the same in magnitude, but opposite in sign, which is also true of their magnetic moments.

2. For the first time the concept of an antiparticle was proposed in 1927-1928 when Dirac had shown on the basis of quantum theory that the electron should have spin and that two continuums of states of total energy $\mathscr{E}$ should be available to a free electron, one extending from $m_ec^2$ to $+\infty$, and the other from $-m_ec^2$ to $-\infty$, where $m_e$ is the rest mass of an electron. Negative values for the total energy of a free electron implied the existence of negative mass, and this entailed a considerable controversy. Among other things, it was argued that in a state with a negative mass an electron exposed to an external force should be accelerated in a direction *opposite* to the applied force. However, such arguments could not shatter the validity of Dirac's theory. The point is that negative mass is postulated by relativity theory. More specifically (see Sec. 16.3), the energy $\mathscr{E}$, momentum $p$, and rest mass $m_e$ of an electron are connected by the relation

$$\mathscr{E}^2 = p^2c^2 + m_e^2c^4$$

whence, at $p = 0$, we have that $\mathscr{E} = \pm m_ec^2$. Fig. 83.10 shows two continuums of allowed energies separated by a forbidden energy gap of magnitude $2m_ec^2$. According to Dirac, in quantum mechanics the probability of a transition from a positive-energy state to a negative-energy state is *other than zero*. In classical physics, negative energies were ignored and, since a transition from the continuum of positive-energy states (point $A$ in Fig. 83.10) into that of negative-energy states (point $B$) requires a jump across a gap $2m_ec^2$ wide, it was ruled out.

For a better understanding of the interpretation that Dirac gave to his result, it should be noted that for an electron ($-e$) with a negative mass ($-m_e$), the specific charge, $-e/-m_e$, is equal to that of a positive electron ($+e$) with a positive mass ($+m_e$), that is:

$$-e/-m_e = +e/+m_e$$

3. Dirac hypothesized that the continuum of negative-energy states (see Fig. 83.10) should have quantized levels, all of which should be filled full with electrons that should occupy them according to the Pauli exclusion principle. The electrons in the negative levels, it was argued, produce a constant (infinite) charge which is undetectable if the levels are filled full; that is, the electrons in negative levels show no action. In contrast, the energy levels in the continuum of positive-energy states (above point $A$ in Fig. 83.10) are only partly filled with electrons. Therefore, if an electron in a negative level is imparted an energy $\mathscr{E} \geqslant 2m_ec^2$, it will move into the continuum of positive-energy states where it will behave like a "normal" electron. The "hole" or vacancy it has left in the continuum of negative-energy states will act like a *positive electron*, that is, a *positron*.

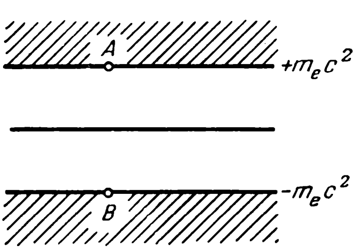


Fig. 83.10

Dirac's hypothesis was experimentally verified in 1932 when the positron was discovered in cosmic rays.

The picture presented by Dirac is in effect the production of an electron-positron pair (see Sec. 82.3). In Dirac's scheme, the annihilation of an electron-positron pair corresponds to a transition of an electron from a positive energy level to a vacancy in a negative energy state, which is accompanied by the conversion of the energy and mass of the combining particles into the energy and mass of the electromagnetic field (which is manifested by the emission of two gamma-photons).

4. The discovery of a particle-antiparticle (electron-positron) pair suggested that there should be a kind of *symmetry* as regards the sign of the charges carried by fundamental particles. This symmetry is known as *charge conjugation*. Basically, charge conjugation means that charged fundamental particles should exist in pairs. That is, for each charged particle there should be an antiparticle, or a particle

with an equal but opposite charge. Thus, the antiproton, $_{-1}\tilde{p}^1$, is the antiparticle of the proton.

With time, the charge conjugation principle has been extended to include other characteristics of fundamental particles, such as barionic and leptonic charge. Of special importance has been the generalization of this principle to the neutron and the neutrino, for this suggests the existence of an antineutron, $_0\tilde{n}^1$, an electron-type antineutrino $_0\tilde{\nu}_e^0$, and a muon-type antineutrino (or antineutretto) $_0\tilde{\nu}_\mu^0$.

5. When a particle combines with its antiparticle, the energy release is not less than twice the rest energy of each. The production of a particle-antiparticle pair calls for an input of energy more than twice the rest energy of each particle, because some momentum and kinetic energy have to be imparted to the pair being produced.

Calculations show that in a coordinate system such that one of the nucleons is at rest, the least energy required to produce a proton-antiproton pair is $6m_pc^2$, where $m_p$ is the rest mass of the proton, or 5.6 GeV (see Sec. 21.7). Practically, the figure is brought down to 4.3 GeV by some effects (such as internal motion of the nucleons in the target nuclei).

6. A distinction of antiparticles is their tendency to combine with their respective particles in a very short time interval. The point is that all substances around us are built up of particles (electrons, protons and neutrons) which are thus available in numbers greatly exceeding those of antiparticles (positrons, antiprotons and antineutrons). The latter inevitably run into their more numerous counterparts in matter and cease to exist, giving birth to other particles and fields in the process, according to conservation laws.

By a slight stretch of imagination, one can readily visualize the existence of "inverted" atoms, or antimatter, in which antiprotons, antineutrons and associated antiparticles make up negatively charged nuclei surrounded by positrons in outer shells. Under the circumstances, the usual electrons, protons and neutrons would likewise tend to annihilate with their "anti" counterparts. Thus, the stability of the "normal" particles and the instability of their antiparticles is a relative term. In a vacuum, the antiparticles are as stable as the respective particles.

7. Heavy antiparticles were first observed in 1950s.

The antiproton, $_{-1}\tilde{p}^1$, was observed in late 1955 by O. Chamberlain, E. Segrè, R. Trip, C. Wiegand and T. Ypsilantis by bombarding a copper target with protons accelerated by a bevatron (at Berkeley, USA) to energies of the order of 6 GeV. The arrangement used in the experiment is shown in Fig. 83.11. The accelerated beam of protons bombarded a copper target, *T*. The negative particles knocked out of the target were deflected by the magnetic field of the bevatron and travelled through an additional magnetic field due to two magnetic lenses, $M_1$, which could only pass particles with a momentum of 1.19 GeV s/cm. As a result, quite a number of negative pi-mesons could pass through the magnetic field along with the particles that were expected to be antiprotons.For example, with protons accelerated to 6.2 GeV, there were 62000 negative pi-mesons per antiproton.

The principal difficulty in antiproton identification was to separate them from the negative pi-mesons and to determine their mass. The mass was determined on the basis of momenta (as found from the curvature of their tracks in a magnetic field) and velocities. The latter were determined by two methods, namely from the time of transit and with Cerenkov counters. The beam of negative particles was passed through a focusing magnetic field, $Q_1$, and fell on a scintillation counter, $S_1$. Then, on passing through a magnetic lens, $Q_2$, a deflecting magnetic field, $M_2$, a second scintillation counter, $S_2$, and two Cerenkov counters, $C_1$ and $C_2$, the particles were counted by a third scintillation counter, $S_3$. The first Cerenkov counter, $C_1$, passed only particles for which $v/c > 0.79$. The second Cerenkov counter, $C_2$, filtered out negative pi-mesons and passed only particles for which the ratio $v/c$ was in the range $0.75 < v/c < 0.78$. This condition was satisfied by antiprotons, $_{-1}\tilde{p}^1$, for which with a momentum of 1.19 GeV s/cm, the ratio $v/c$ is 0.78, while for negative pi-mesons this ratio is 0.99 at the same momentum. Also, the transit time, $t$, of particles between $S_1$ and $S_2$ was measured. It was found to be $5.1 \times 10^{-8}$ s for antiprotons and $4 \times 10^{-8}$ s for negative pi-mesons.

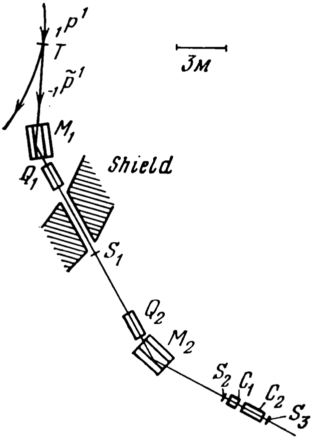


Fig. 83.11

A particle was identified as an antiproton only if the counters $S_1$, $S_2$, $C_2$ and $S_3$ operated all at the same time while $C_1$ did not, and the transit time measured between $S_1$ and $S_2$ fit. A few tens of antiprotons were detected in the experiment, and the relative yield of antiprotons and negative pi-mesons was plotted as a function of the energy of the primary protons, which ranged from 4.3 to 6.2 GeV.

As a cross-check on the identification of the antiproton and also to prove that its mass was the same as that of the proton, the experimenters used a very ingenious trick: they reversed the sense of all magnetic fields in the set-up and injected protons with a momentum of 1.19 GeV s/cm into the machine. The counters responded exactly as they had during the passage of what were thought to be antiprotons. Thus, the existence of the antiproton had been proved beyond any shade of doubt.

Measurements of the magnetic moment showed that it has a charge opposite in sign to that of the proton. In the early determinations which were lacking refinement and accuracy, the magnetic moment was found to be $-1.8\mu_N$ as against the theoretically expected value of $-2.79\ \mu_N$.

8. The antineutron, ${}_1\tilde{n}^1$, was for the first time observed by B. Corck, G. Lambertson, O. Piccioni and G. Wenzel in 1956 from the change of charge on an antiproton, as it and a proton were caused to interact with the production of an antineutron and a neutron:

$$ {}_{-1}\tilde{p}^1 + {}_1p^1 \rightarrow {}_0\tilde{n}^1 + {}_0n^1 \tag{83.8}$$

The production of an antineutron was accertained from its combination with the neutron which should release an energy of $\mathscr{E} = 2m_nc^2 = 1900$ MeV, where $m_n$ is the rest mass of a neutron (or antineutron). This energy mainly goes to produce negative pi- and both positive and negative K-mesons in the ratio 95 to 5%. From the "star" tracks produced in photographic emulsions, it was found that each time a neutron combined with an antineutron, an average of three charged pi-mesons were produced, each of which carried an energy of about 250 MeV. In addition to the charged pi-mesons, the other particles produced were neutral pi-mesons and neutral K-mesons, sharing between them the remainder of the energy released in the union. The absorption of the pi-mesons released by the union suggested that the antineutrino should combine with a neutron located near the surface of the nucleus.

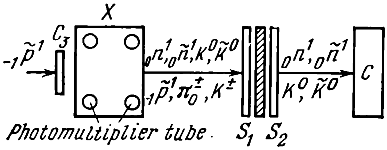


Fig. 83.12

The experimental set-up used to detect the antineutrino is shown in Fig. 83.12. Antiprotons were produced by bombarding a beryllium target with a beam of protons accelerated to 6.2 GeV in the bevatron. Using the method already described, a beam of antiprotons was then coupled out with an intensity of 350 to 600 antiprotons per hour. After its passage to the final counter in the set-up shown in Fig. 83.11*, the beam of antiprotons was admitted to a converter, $X$, where the antiprotons changed charge according to Eq. (83.8). The converter was a vessel filled with a scintillating organic fluid. The products of the process taking place in the converter were then picked up by four photomultiplier tubes.

* This counter also served to verify whether the antiprotons were travelling in the right direction.

The antiprotons produced in the converter along with other particles were then passed through two scintillation counters, $S_1$ and $S_2$, separated by a lead screen. These counters separated all the charged particles, gamma-quanta, and also neutral pi-mesons, and neutral $S$- and $L$-kaons. The final Cerenkov counter, $C$, made of a lead glass, registered all acts of union between antineutrons and neutrons from the strong Cerenkov radiation that was emitted by the products of the process, which mainly were pi-mesons. The occurrence of antineutrons was identified by noting simultaneous response from the counters $C_3$ and $C$ and from the converter $X$, while there should be no response from the scintillation counters, $S_1$ and $S_3$, which are not responsive to uncharged antineutrons. In the first experiment, a total of 60 antineutrons were observed, with a yield of about 0.0028 per antiproton.

Like a free neutron, a free antineutron decays with a decay half-life of $(1.01 \pm 0.03) \times 10^3$ s (see Sec. 81.12), according to the following scheme:

$$_0\tilde{n}^1 \rightarrow {}_{-1}p^1 + {}_{+1}e^0 + {}_0\nu_e^0$$

where $_0\nu_e^0$ stands for a neutrino.

9. A good deal of interest and a lot of difficulties were associated with the direct observation of the neutrino and also the search for the answer to whether the neutrino, $_0\nu_e^0$, and the antineutrino, $_0\tilde{\nu}_e^0$, were identical or different particles, that is, whether the neutrino was an absolutely neutral particle (one having no antiparticle except itself).

Advances in nuclear physics and reactor engineering made the observation of the antineutrino a practical proposition. For the fission fragments of heavy nuclei have a great neutron excess and undergo negative beta-decay with the emission of antineutrinos, $_c\tilde{\nu}_e^0$.

Using sufficiently strong beams of antineutrinos, experiments were conducted to investigate interactions between the antineutrino and the proton. The idea of the experiments was to detect the capture of an antineutrino by a proton. This reaction was assumed to proceed according to the following scheme:

$$_0\tilde{\nu}_e^0 + {}_1p^1 \rightarrow {}_0n^1 + {}_{+1}e^0 \tag{83.9}$$

Similarly, the capture of a neutrino by a neutron was expected to occur by the reaction:

$$_0\nu_e^0 + {}_0n^1 \rightarrow {}_1p^1 + {}_{-1}e^0 \tag{83.9'}$$

It may be shown that such reactions are feasible if a neutron changes to a proton and a proton changes to a neutron by the following schemes (see Secs. 81.12 and 82.2):

$$ {}_0n^1 \rightarrow {}_1p^1 + {}_{-1}e^0 + {}_0\tilde{\nu}_e^0 $$

$$ {}_1p^1 \rightarrow {}_0n^1 + {}_{+1}e^0 + {}_0\nu_e^0 $$

10. The experiments intended to detect an antineutrino causing a proton to change to a neutron and a positron according to Eq. (83.9) were made by F. Reines and C. Cowan in 1953-1954, using beams of antineutrinos from a plutonium-producing nuclear reactor. The target and detector used in their arrangement was a chamber with a volume of about 1 m$^3$, filled with a scintillation fluid which included hydrocarbons and dissolved $CdCl_2$. Watch on the reaction that proceeded according to Eq. (83.9) was kept by a great number of photomultiplier tubes. The positrons produced by the reaction annihilated with the atomic electrons of the fluid with emission of two gamma-quanta, which fact was indicated by a flash of the scintillating fluid. The neutrons produced in the reaction were slowed down by the hydrogen and captured by the cadmium (a radiative capture). The cascade of gamma-quanta that followed this radiative capture produced a second flash. These flashes positively confirmed that the reaction was that described by Eq. (83.9), and also the existence of the antineutrino.

The refined experiments conducted in 1956 have proved that the neutrino interacts with matter differently than the antineutrino, for which reason they should be treated as different particles. The reactions described by Eqs. (83.9) and (83.9′) have been uniquely identified, confirming that the neutrino and the antineutrino are different particles. The neutron differs from the antineutrino in lepton number and other characteristics which we will leave out.

11. The question whether the neutrino and the neutretto (and, also, the antineutrino and the antineutretto) were different particles arose in connection with studies into the decays of charged (±) pi-mesons according to the scheme described by Eq. (83.4) and (83.5). It has been found out that if the resultant neutrino and antineutrino are separated and then let to be captured, say the neutrino by a neutron, the reaction (83.9′) will not occur; instead, the capture will proceed as follows:

$$ {}_0\nu_\mu^0 + {}_0n^1 \rightarrow {}_1p^1 + \mu^- \qquad (83.9'') $$

which is a proof that the ${}_0\nu_e^0$ and the ${}_0\nu_\mu^0$ are different particles.

12. Antiparticles have also been detected among hyperons, and the antihyperons have shown that they meet the requirements common to all antiparticles.

A far greater amount of energy is required to produce a hyperon-antihyperon pair than a nucleon pair. For example, the energy required to produce a $\tilde{\Lambda}^0$-hyperon, the lightest of all hyperons, is 1 to

1.5 GeV greater than is necessary for the production of antinucleons. This increase is explained by the greater rest mass of hyperons.

13. Recent years have seen the discovery of a great number of new particles called *resonance particles*, *resonances*, or *resonons*, with a very short lifetime (about $10^{-22}$ s).

Resonons may be treated as particles on the ground that in both production and decay they behave like a particle having a definite spin, electric charge, baryon number, and other characteristics we have not taken up here. Also, resonons have definite momenta and energies.

For the first time, the term resonance was applied to particles back in the 1950s, when in experiments on the scattering of pi-mesons with an energy of about 200 MeV by protons a sudden increase was noted in the scattering due to resonance. The resonance effects similar to the pion-proton resonance, were found to have properties of particles, and the name stuck. In fact, each heavy strongly interacting particle has been noted to have a resonance of its own, having a great mass. Meson resonances have also been detected.

The number of fundamental particles and resonances discovered to date is so great that several attempts have been made to reach a unified model of particles. Unfortunately, this matter is still in a state that makes its presentation in a text-book difficult.

14. In fact, the very concept of fundamental particles is now subject to a serious doubt. Any of existing definitions is lacking precision or fails to fit observation in some cases. Yet, an exhaustive definition can hardly be thought up at present.

As physics probes deeper and deeper into the properties of "elementary" particles, more weight is added to Lenin's prediction that the electron is as inexhaustible as the atom.

## 83.8. STRUCTURE OF NUCLEONS

1. In Sec. 83.1 it is noted that relativity theory requires the elementary particles to be point-like, that is, devoid of any structure. Any extent of a particle in space, or any signs of structure in it would run counter to relativity theory. On the other hand, structureless fundamental particles, or material particles having no extent appear to be unsatisfactory from philosophical and physical points of view. As is noted in Sec. 72.5, the expression for the classical radius of an electron implies that the electron should have definite dimensions and, as a consequence, a certain structure. In modern physics, the classical models of elementary particles have given way to attempts to find a structure for these particles that would not contradict relativity theory. To-day, there is a considerable body of experimental evidence for the existence of some structure

in fundamental particles, and techniques have been proposed for its study.

2. One method is to investigate the elastic scattering of pi-mesons with an energy of about 7 GeV by protons. Another method is to use the elastic scattering of electrons by protons and neutrons. Both methods have been able to throw some light on the structure of the nucleon. It has been observed that in the scattering of pi-mesons by protons, the pi-mesons deviate but little from the original trajectory, while the recoil of the protons is likewise small, that is, the momentum, $\Delta p$, imparted by colliding pi-meson to the proton is negligible. From the Heisenberg uncertainty principle it follows that this process should take place in a region of space such that $a \gg \hbar/\Delta p$. Further studies appear to indicate that this process should be treated as the diffraction of pi-mesons by some pion-absorbing region which defines the size of the nucleon.

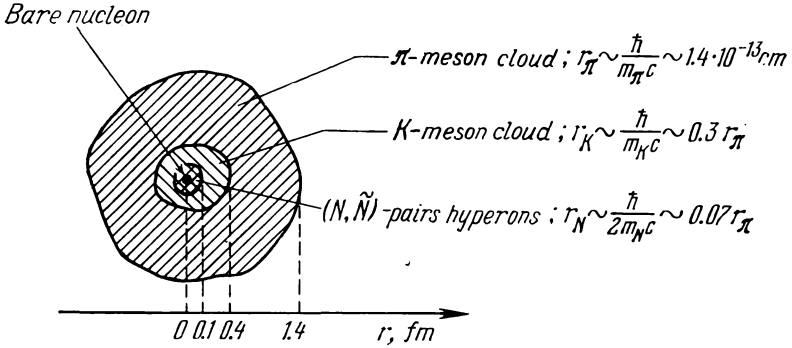

Bare nucleon
$\pi$-meson cloud; $r_\pi \sim \frac{\hbar}{m_\pi c} \sim 1.4 \cdot 10^{-13}$ cm
K-meson cloud; $r_K \sim \frac{\hbar}{m_K c} \sim 0.3\, r_\pi$
$(N, \tilde{N})$-pairs hyperons; $r_N \sim \frac{\hbar}{2 m_N c} \sim 0.07\, r_\pi$
0
0.1
0.4
1.4
r, fm

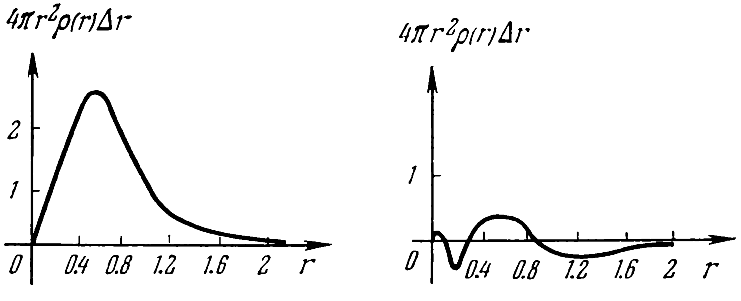
Fig. 83.13

3. In Sec. 80.5, we have used the concept of virtual pi-mesons to get an insight into the interactions of nucleons. It appears that this concept may be as fruitful in describing the structure of nucleons.

Observations indicate that a nucleon has a central core, $N$, (the bare nucleon) with an estimated size of $10^{-16}$ m, surrounded by clouds of pi-mesons. Inside the pi-meson clouds is a cloud of virtual K-mesons. The virtual emission or absorption of a K-meson is accompanied by the production of a Y-hyperon. Besides, virtual nucleon-antinucleon ($N\tilde{N}$) pairs are formed around the core. The arrangement of these clouds of $N\tilde{N}$ pairs, K-mesons and pi-mesons around the core is shown in Fig. 83.13. There are also virtual photons accounting for the electromagnetic interactions of the nucleons.

Virtual processes of particle emission and absorption are going on continually in the nucleon, so that it must be treated as a complex, incessantly changing composition of virtual particles. The composition existing at a particular instant cannot be looked upon as an independent state, as it instantaneously gives way to another one.

Experiments on the elastic scattering of pi-mesons and electrons by nucleons and a comparison of observations with theory have given an idea about the distribution density ρ (*r*) of electric charge inside a nucleon. This distribution obtained by a group of researchers at Standford University, USA, for the proton (*a*) and the neutron (*b*) is shown in Fig. 83.14. In both the charged (protonic) and the uncharged (neutronic) state, the nucleon contains definite clouds of charge. In the proton these clouds add together to give a charge $+e$; in the neutron they cancel out.



(*a*) (*b*)

Fig. 83.14

4. The structure of the nucleon, as it appears above, explains why the neutron has a negative magnetic moment of $p_{mn} = -1.9\ \mu_N$, although its charge is zero and its magnetic moment ought to be likewise zero, and also why the proton has an anomalous magnetic moment of 2.79 $\mu_N$, which is 1.79 $\mu_N$ greater than the value (1 $\mu_N$) it should have by analogy with the electron whose magnetic moment is one Bohr magneton, $\mu_B$. It is believed that the nucleon can undergo virtual dissociation by the schemes:

$${}_1p^1 \rightleftarrows {}_1n^0 + \pi^+$$

and

$${}_0n^1 \rightleftarrows {}_1p^1 + \pi^-$$

From theoretical deductions based on observed values of magnetic moments of the proton and neutron it follows that about 20% of the time each particle is in a dissociated state and about 80% of the time in the "bare" protonic or neutronic state. In the protonic state, the positive pi-meson cloud produces a magnetic moment additional to that of the core and of the same sign; the result is the anomalously great value of magnetic moment. In the neutronic state, the negative pi-meson cloud produces a negative magnetic moment.

Studies into the structure of the nucleons have led to the discovery of heavy mesons, notably the $\eta^0$-meson listed in Table 83.2.

The concept of a nucleon having a definite structure is very fruitful. Among other things, it attributes the difference in mass between the proton and the neutron to electrostatic and magnetic interactions between the nucleon core and the charge clouds.

5. There is also a body of theoretical and experimental evidence that the electron, too, has a structure which appears to be similar to that of the nucleon. The central core of the electron is surrounded by a system of shells formed by pairs of particles and antiparticles (photons, electron-positron, pi-mesons, nucleon-antinucleon, and the like). As follows from Table 83.1, electromagnetic coupling is much weaker than strong coupling. Therefore, the shell in the structure of the electron has a lower density than the shells in the nucleon.

At present, studies are under way into electron-electron collisions, using particle beams colliding head-on. These experiments involving huge energies will throw more light on the structure and size of the electron.

As a closing remark, it should be stressed that the physics of fundamental particles is developing in big strides both in theory and in experimental techniques.

# CONCLUSION

The book has covered all the most important divisions of classical and modern physics. In the presentation of the subject-matter, it has been sought to show that there is no unsurmountable gap between them and that physics is a continually developing science in which physical concepts, theories and results change one another in a natural way.

Starting with Newtonian mechanics and relativity theory, we have moved through the fundamentals of thermodynamics and molecular physics, electrodynamics, vibrations and wave motion, including electromagnetic waves and optics. Every effort has been made to show that the concepts of the special theory of relativity make up the backbone of all of modern physics and present in a new perspective many of the subdivisions of classical physics. From the outset, it has been our desire to stress that the atomic world imposes certain limitations on a purely classical description of microscopic entities. Quite a lot of space has been devoted to the present-day physics of atoms, to molecules and crystals, and to the principles underlying nuclear physics and the physics of fundamental particles.

Any course on modern physics must of necessity give a progressively deeper insight into all facts of nature, into the laws that govern the processes occurring around us. Of course, mechanics starts at the megascopic level. Then comes the molecular level at which we gain knowledge about atoms, molecules and their aggregations. In fact molecular physics with its specific combination of statistical and thermodynamical methods is the first step into the microscopic world. Already at the molecular level, we are sometimes brought to abandon some of the methods applied in macrophysics.

Methodologically, this is not unexpected, though, for, according to dialectical materialism, a change in a quantitative scale of necessity leads to qualitatively new regularities. In short, the microscopic world must obey other laws than the megascopic world does.

This is fully confirmed by electrodynamics which explains all phenomena by the behaviour of charged particles—electrons and ions—in vacuum and in the bulk of substances, that is, at the intermolecular level. During its long history, classical electrodynamics has built up a considerable body of laws and relations that describe various phenomena involving electricity, magnetism and optics. But it is in this field that classical theory proved inadequate to describe the interaction of electromagnetic fields with matter. On the other hand, failure of attempts to deal with thermal radiation and its interaction with matter in terms of the classical theory of radiation, classical statistical physics and electron theory had finally led to quantum theory. When used alone, classical physics also failed as regards the electrical properties of solids—metals and, especially, semiconductors. Examples of this kind might be multiplied.

When backed up by the special theory of relativity, classical theory can in some cases give a correct interpretation of observations and regularities in the field of electricity, magnetism and interactions of light with matter. This is a manifestation of an organic continuity between classical and modern physics and a proof that there is no Chinese wall between them.

Probing into the structure and properties of atoms, molecules and solids on the basis of quantum physics and advances in our knowledge of the nucleus and fundamental particles are the achievements of the 20th century's physics which owes a good deal of its progress to major breakthroughs in experimental techniques and apparatus.

The first quarter of the 20th century saw the advent of relativity theory and quantum mechanics. These important directions in physical science have established the laws that govern the microscopic world of motions at velocities close to that of light. The general theory of relativity has enabled physics to encompass the megascopic world—the stars, the Universe, and other objects.

To-day, quantum mechanics and relativity theory are not simply theories giving deep insight into all that happens around us. Relativity theory has since long been serving as a basis for practical ideas in accelerator, nuclear-reactor and nuclear-power engineering. Quantum mechanics has found many successful applications in the design of nuclear reactors, electronic and semiconductor devices, masers and lasers. In fact, it has become the basis of many other fields of present-day technology and industrial engineering.

At first sight, the basic ideas of relativity theory and quantum mechanics appear unusual and conflicting with all that man is accustomed to see in everyday life. The difficulty with which the student accepts these new ideas is to a great extent the result of the traditional presentation of physics at school. Of course, we cannot ignore the fact that new concepts are not reducible to the familiar ones or that sometimes we do not find analogies which would help the student to grasp these ideas—all this does handicap the teaching of modern physics at first. By far the greater proportion of this difficulty, however, arises in cases where insufficient emphasis is given to inherent logical ties between classical and modern physics and between the various aspects of physical phenomena. A major factor contributing to this difficulty is also that the basic principles of modern physics are introduced too late in its course. The only remedies here are time and patience—without them one cannot grasp new ideas, and not only new.

As has already been noted, the deep, inherent ties between classical and modern physics are adequately expressed by the correspondence principle—in the limit of high quantum numbers the predictions of quantum theory agree with those of classical physics.

The edifice of classical and modern physics, for all its complexity, rests firmly on the foundation of conservation laws. All the conservation laws established by classical physics remain applicable to the microscopic world—as we have seen they are as valid for the processes involving individual particles of matter. This universality of the conservation laws reflects the unity and diversity of natural phenomena. The physics of fundamental particles has brought with it new conservation laws which bear out all that Lenin has said on absolute and relative truth and the continuity in the process of cognition from the less deep to deeper entities.

Modern physics is one of the fast advancing sciences. Its dynamic character is especially felt in the physics of the nucleus and fundamental particles, and solid-state physics, and also in adjacent fields such as biophysics. With its growth, it gives birth to ever new disciplines. A few decades ago hardly any one could guess that there would be plasma mechanics, magnetohydrodynamics, or quantum radio engineering or other divisions of modern physics that we have today.

These closing remarks and, as the authors hope, the material presented in their book serve to stress the importance of physics in present-day education.

# NAME INDEX

# SUBJECT INDEX

*Printed in the Union of Soviet Socialist Republics*