MEDICAL AND BIOLOGICAL PHYSICS FOR MEDICAL STUDENTS

Minsk BSMU 2015
MEDICAL AND BIOLOGICAL PHYSICS FOR MEDICAL STUDENTS

Minsk BSMU 2015
МИНИСТЕРСТВО ЗДРАВООХРАНЕНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ МЕДИЦИНСКИЙ УНИВЕРСИТЕТ
КАФЕДРА МЕДИЦИНСКОЙ И БИОЛОГИЧЕСКОЙ ФИЗИКИ

МЕДИЦИНСКАЯ
И БИОЛОГИЧЕСКАЯ ФИЗИКА
ДЛЯ СТУДЕНТОВ-МЕДИКОВ
MEDICAL AND BIOLOGICAL PHYSICS FOR MEDICAL STUDENTS

Учебно-методическое пособие

Минск БГМУ 2015
Рекомендовано Научно-методическим советом университета в качестве учебно-методического пособия 16.09.2015 г., протокол № 1

А в т о р ы: Л. В. Кухаренко, О. В. Недзьведь, М. В. Гольцев, В. Г. Лещенко

Р е ц е н з е н т д-р физ.-мат. наук, проф. каф. физики твердого тела Белорусского государственного университета В. Г. Шепелевич


Издание содержит все разделы курса медицинской и биологической физики. В нем рассматриваются физические явления, лежащие в основе методов медицинской диагностики и лечения.

Предназначено для студентов 1-го курса, изучающих медицинскую и биологическую физику на английском языке.
Chapter 1. MATHEMATICS FUNDAMENTALS

1.1. THE FUNCTION DERIVATIVE

In addition to knowing the value of a function \( f(x) \) at a particular \( x \), one often want to know how fast the function is changing with \( x \). The derivative of a function represents an infinitesimal change in the function \( f(x) \) with respect to its variable \( x \).

Consider a function \( y = f(x) \) at two points: \( (x_0, f(x_0)) \) and \( (x_0 + \Delta x, f(x_0 + \Delta x)) \).

Here \( \Delta x \) means some small change of an argument and is called an increment of argument. Correspondingly a difference between the two values of a function \( f(x_0 + \Delta x) - f(x_0) \) is called an increment of function.

The derivative of function \( y = f(x) \) at a point \( x_0 \) is the limit:

\[
y' = \lim_{\Delta x \to 0} \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}.
\]

Derivative of a function \( f(x) \), denoted by \( f'(x) \), read «f prime of x», is defined as:

\[
y' = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x}, \quad \text{or} \quad \frac{dy}{dx} = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x}.
\]

(1.1)

The process of calculating the derivative of a function is called differentiation.

The definition of the derivative can be approached in two different ways. One is geometrical (as a slope of a curve) and the other one is physical (as a rate of change of a function).
A tangent is a straight line that just touches a curve. A secant is a straight line that cuts a curve. Hence, consider the secant line that cuts the curve \( f(x) \) at points \( A \) and \( B \) (fig. 1.1). Then the inclination of that secant with respect to the x-axis (the slope) can be given by:

\[
t\alpha = \frac{BD}{AD}.
\]

If to fix the point \( A \) and to move the point \( B \) towards \( A \), then \( \Delta x \) will unboundedly decrease and approach 0, and the secant \( AB \) will approach the tangent \( AC \). The slope of the tangent line at \( A \) is the limit of the change in the function \( \Delta y \) divided by the change in the independent variable \( \Delta x \) as that change approaches 0 (\( \Delta x \to 0 \)):

\[
t\alpha_A = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x}.
\]

(1.2)

**The geometric interpretation of the derivative** of a function is the following: the derivative of a function \( f'(x_0) \) at a specific point \( x_0 \) is equal to the slope of a tangent line to the graph of \( f(x) \) at that point \( x_0 \).

To give **physical interpretation of the derivative** of a function, let us consider a movement of a material point along a coordinate line. During the time interval from \( t_0 \) to \( t_0 + \Delta t \) the point displacement \( \Delta s \) is equal to:

\[
s(t_0 + \Delta t) - s(t_0) = \Delta s,
\]

and its **average velocity** is:

\[
v_{\text{aver}} = \frac{\Delta S}{\Delta t}.
\]

As \( \Delta t \to 0 \), then an average velocity value approaches the certain value, which is called an **instantaneous velocity** \( v(t_0) \) of a material point at a precise moment in time \( t_0 \). Thus,

\[
v_{\text{inst}} = \lim_{\Delta t \to 0} \frac{\Delta S}{\Delta t} = \frac{dS}{dt} = S'.
\]

(1.3)

Hence, \( v_{\text{inst}} = S'(t_0) \), i. e. a derivative of a coordinate with respect to time is a instantaneous velocity.

Similarly to this, acceleration \( a \) is a derivative of a velocity with respect to time:

\[
a = v'(t).
\]

(1.4)

If some value \( y \) depends on the spatial coordinates \( x \), the derivative \( \frac{dy}{dx} \) describes the rate of the spatial variation of \( y \). The derivative of function with respect to the spatial coordinate is called the **gradient of function**. Gradient of function is a vector, which is directed in the direction of increasing of value \( y \).
For example, some substance is nonuniformly distributed along the axis \( x \), thus its concentration \( C \) is a function of \( x \) (fig. 1.2). The derivative of concentration \( C \) with respect to the spatial coordinate \( x \) is called the *gradient of concentration* and determines the rate of change of concentration \( C \) along the spatial coordinate \( x \). Similarly, there are gradients of pressure, temperature and other variables.

![Fig. 1.2. The function gradient](image)

**Calculation of the derivative**

In practice, the derivatives of a few simple functions are known and shown in table 1.1, the derivatives of other functions are more easily computed using rules for obtaining derivatives.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1) ( C = \text{const}, (C)' = 0 )</td>
<td>6) ( \cos x)' = -\sin x )</td>
</tr>
<tr>
<td>2) ( (x^n)' = nx^{n-1} )</td>
<td>7) ( \cos x\ldots = \frac{1}{\cos^2 x} )</td>
</tr>
<tr>
<td>3) ( (a^x)' = a^x \cdot \ln a )</td>
<td>8) ( \sin x\ldots = -\frac{1}{\sin^2 x} )</td>
</tr>
<tr>
<td>3a) ( (e^x)' = e^x )</td>
<td>9) ( \arcsin x)' = \frac{1}{\sqrt{1-x^2}} )</td>
</tr>
<tr>
<td>4) ( \log_a x\ldots = \frac{1}{x \cdot \ln a} )</td>
<td>10) ( \arccos x)' = \frac{1}{\sqrt{1-x^2}} )</td>
</tr>
<tr>
<td>4a) ( \ln x\ldots = \frac{1}{x} )</td>
<td>11) ( \arctg x)' = \frac{1}{1+x^2} )</td>
</tr>
<tr>
<td>5) ( \sin x)' = \cos x )</td>
<td></td>
</tr>
</tbody>
</table>

**Differentiation rules**

The derivative of constant \( C \) times a function \( u \) is equal to the constant \( C \) times the derivative of the function \( u' \):

\[(Cu)' = C(u)' \]

**Sum rule:** the derivative of a sum of functions \( (u \text{ and } v) \) is equal to the sum of the function derivatives:

\[(u + v)' = u' + v' \]
**Product rule:** the derivative of a product of two functions \( (u \text{ and } v) \) is equal to the first function \( u \) multiplies the derivative of the second one \( v' \) plus the second function \( v \) multiplies the derivative of the first one \( u' \):
\[
(u \cdot v)' = u \cdot v' + u' \cdot v.
\]

**Quotient rule:** the derivative of the quotient of two functions \( (u \text{ and } v) \) is equal to the denominator multiplies the derivative of the numerator minus the numerator multiplies the derivative of the denominator all divided by the square of the denominator:
\[
\left(\frac{u}{v}\right)' = \frac{u' \cdot v - u \cdot v'}{v^2}.
\]

**Chain rule:** if \( f \) is a function of \( g \), and \( g \) is a function of \( x \), then the derivative of \( f \) with respect to \( x \) is equal to the derivative of \( f(g) \) with respect to \( g \) multiplies the derivative of \( g(x) \) with respect to \( x \):
\[
f'(x) = h'[g(x)] \cdot g'(x).
\]

One can use chain rule for a composite function mean argument of which is also a function: \( h(x) = g(f(x)) \).

1.2. **Maxima and minima of functions**

Suppose \( f(x) \) is a continuous function on a closed interval \([c, d]\). Then the function \( f(x) \) has a **local maximum** at point \( A \) if and only if \( f(x) \leq f(A) \) for all \( x \) in some open interval containing \( A \). The function \( f(x) \) has a **local minimum** at \( B \) if and only if \( f(B) \geq f(x) \) for all \( x \) in some open interval containing \( B \). At each of these points (\( A \) and \( B \)) the tangent to the curve \( f(x) \) is parallel to the \( x \)-axis so the derivative of the function is equal to zero: \( f'(A) = 0 \) and \( f'(B) = 0 \) (fig. 1.3). At points immediately to the left of a maximum point \( A \) the slope of the tangent is positive: \( f'(x) > 0 \). While at points immediately to the right the slope is negative: \( f'(x) < 0 \). In other words, at a point of maximum \( f'(x) \) changes from positive to negative. At a point of minimum \( f'(x) \) changes sign from negative to positive, respectively.

![Fig. 1.3. Local maximum and minimum](image)
If the function \( f(x) \) is differentiable at point \( x \) and \( f'(x) = 0 \), then we call \( x \) a critical point or stationary point of function \( f(x) \).

To find the maximum and minimum values of a function \( f(x) \) it is necessary:
1. To solve the algebraic equation:

\[
f'(x) = 0.
\]

The roots \( x_1, x_2, x_3 \ldots \) of this equation are the critical points.

2. To calculate the second derivative of function \( f''(x) \) and definite its sign at the critical points.

If the second derivative is positive at a stationary point: \( f''(x_1) > 0 \), the point \( x_1 \) is a local minimum; if it is negative: \( f''(x_2) < 0 \), the point \( x_2 \) is a local maximum; if it is equal to zero: \( f''(x_3) = 0 \), it may or may not be a local extremum. In this case it is necessary to find a sign of the first derivative \( f'(x) \) on the left side (\( x < x_3 \)) and on the right one (\( x > x_3 \)) from the \( x_3 \). If \( f'(x) \) changes from positive to negative at \( x_3 \), then \( f(x) \) has a local maximum at \( x_3 \). If \( f'(x) \) changes from negative to positive at \( x_3 \), then \( f(x) \) has a local minimum at \( x_3 \). If \( f'(x) \) does not changes the sign at \( x_3 \), then \( f(x) \) has no maximum or minimum at \( x_3 \).

3. To determine a value of function in points of maximum and minimum.

### 1.3. Differential of a function

Let \( y \) is function of \( x \) (\( y = f(x) \)). Imagine we change argument \( x \) to \( x + \Delta x \) with \( \Delta x \) very small. What is the corresponding change in function \( y \). The answer is that the change is \( dy \) given by:

\[
dy = y' \cdot dx. \tag{1.5}
\]

This formula (1.5) requires \( \Delta x \) to be very small. The differential of an argument \( dx \) is equal an infinitesimal increment of argument \( \Delta x: dx \approx \Delta x \). The differential of a function \( dy \) represents the principal part of the change of a function \( y = f(x) \) with respect to changes of the independent variable \( x \). The differential of function \( dy \) is equal to the product of the derivative of function \( y' \) by an increment (a differential) of argument \( dx \).

Differential \( dy \) of function is not equal to its increment \( \Delta y \) but it is regarded as a linear approximation to the increment of a function:

\[
\Delta y \approx dy = y' \cdot dx.
\]

In fig. 1.4. differential of the function \( dy \) is equal to \( CD \).

Properties of the differential of function. A number of properties of the differential follow from the corresponding properties of the derivative.

**Linearity:** for constants \( a \) and \( b \) and differentiable functions \( f \) and \( g \):

\[
d(af + bg) = a \cdot df + b \cdot dg.
\]

**Product rule:** for two differentiable functions \( f \) and \( g \):

\[
d(fg) = f \cdot dg + g \cdot df.
\]
**Chain rule:** if $y = f(u)$ is a differentiable function of the variable $u$ and $u = g(x)$ is a differentiable function of $x$, then:

$$dy = f'(u)du = f'(g(x)) \cdot g'(x)dx.$$ 

**Fig. 1.4.** Differential of function

**1.4. PARTIAL DERIVATIVES**

Let us consider a function $U = U(x, y, z)$ of several variables. Such a function can be studied by holding all variables except one constant and observing its variation with respect to one single selected variable. If we consider all the variables except $x$ to be constant, then

$$U'_x = \frac{\partial U}{\partial x} = \lim_{\Delta x \to 0} \frac{U(x + \Delta x, y, z) - U(x, y, z)}{\Delta x} \quad (1.6)$$

represents the partial derivative of $U(x, y, z)$ with respect to $x$. The variables held fixed are viewed as parameters. One might also define partial derivatives of function $U(x, y, z)$ with respect to $y$ and with respect to $z$ as follows:

$$U'_y = \frac{\partial U}{\partial y} = \lim_{\Delta y \to 0} \frac{U(x, y + \Delta y, z) - U(x, y, z)}{\Delta y} \quad (1.7)$$

$$U'_z = \frac{\partial U}{\partial z} = \lim_{\Delta z \to 0} \frac{U(x, y, z + \Delta z) - U(x, y, z)}{\Delta z} \quad (1.8)$$

The **partial derivative of a function of two or more variables** with respect to one of its variables is the ordinary derivative of the function with respect to that variable, considering the other variables as constants.

Calculating partial derivatives is usually just like calculating an ordinary derivative of one-variable calculus. To calculate the partial derivative $U'_z$ of
$U(x, y, z)$ with respect to $x$ one can simply view $y$ and $z$ as being a fixed numbers and calculate the ordinary derivative with respect to $x$. All the rules and formulas being true to the derivative of a function of one variable are true to a partial derivative of a function of several variables.

1.5. PARTIAL DIFFERENTIALS, TOTAL DIFFERENTIAL OF FUNCTION

For functions of more than one independent variable, the partial differential of $U(x, y, z)$ with respect to any one of the variables $x$ is the principal part of the change in $U(x, y, z)$ resulting from a change $dx$ in that one variable. The partial differential is therefore:

$$dU_x = \frac{\partial U}{\partial x} dx.$$  

The total differential of function is the sum of all partial differentials. For function $U(x, y, z)$ the total differential $dU$ is written as:

$$dU = U'_x dx + U'_y dy + U'_z dz,$$

or

$$dU = \frac{\partial U}{\partial x} dx + \frac{\partial U}{\partial y} dy + \frac{\partial U}{\partial z} dz. \tag{1.9}$$

Total differential of function $dU$ is the principal part of the change in $U(x, y, z)$ resulting from the changes in the independent variables.

1.6. ANTIDERIVATIVE FUNCTION, INDEFINITE INTEGRAL

The function $F(x)$ is called an antiderivative of function $f(x)$ on an interval if the derivative of this function $F(x)$ is equal to the initial function $f(x)$ for all $x$ in that interval:

$$F'(x) = f(x).$$

We consider the problem of finding all the antiderivatives of $f(x) = 2x$. Certainly $F(x) = x^2$ is an antiderivative of $2x$ since $(x^2)' = 2x$, by the power rule. But $G(x) = x^2 + 1$, and $H(x) = x^2 - 4$ are also antiderivatives of $2x$. Indeed, for any constant $C$, $F(x) = x^2 + C$ is an antiderivative of $f(x) = 2x$.

If $F(x)$ is an antiderivative of $f(x)$ on an interval, then the most general antiderivative of $f(x)$ on that interval is $F(x) + C$, where $C$ is an arbitrary constant. Graphs of the antiderivatives of a given function are vertical translations of each other and each graph’s location depending upon the value of $C$.

Indefinite integral of a function $f(x)$ is a set of all antiderivatives $F(x)$ of a function $f(x)$. The indefinite integral of the function $f(x)$ is written as:

$$\int f(x)dx = F(x) + C.$$
One can read it as «the indefinite integral of \( f(x) \) with respect to \( x \)». \( \int f(x)dx \) is a collection of functions, it is not a single function, nor a number. The operation of indefinite integration is an inverse of differentiation.

**Features of the indefinite integral.** The process of differentiation and integration are inverses of each other:

\[
(\int f(x)dx)' = f(x).
\]

The integral of a sum or difference of functions is the sum or difference of the individual integrals. This rule can be extended to as many functions as we need.

\[
\int (f(x) \pm g(x))dx = \int f(x)dx \pm \int g(x)dx.
\]

One can factor multiplicative constants out of indefinite integrals.

\[
[k \int f(x)dx] = k \int f(x)dx,
\]

where \( k \) is any constant.

Some indefinite integrals can be finding with use of elementary functions (table 1.2).

<table>
<thead>
<tr>
<th>Table 1.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. ( \int x^n dx = \frac{x^{n+1}}{n+1} + C ), ( n \neq -1 )</td>
</tr>
<tr>
<td>2. ( \int \frac{dx}{x} = \ln</td>
</tr>
<tr>
<td>3. ( \int a^x dx = \frac{a^x}{\ln a} + C )</td>
</tr>
<tr>
<td>4. ( \int e^x dx = e^x + C )</td>
</tr>
<tr>
<td>5. ( \int \sin x dx = -\cos x + C )</td>
</tr>
<tr>
<td>6. ( \int \cos x dx = \sin x + C )</td>
</tr>
<tr>
<td>7. ( \int \frac{dx}{\cos^2 x} = \tan x + C )</td>
</tr>
<tr>
<td>8. ( \int \frac{dx}{\sin^2 x} = -\cot x + C )</td>
</tr>
<tr>
<td>9. ( \int \frac{dx}{\sqrt{1-x^2}} = \arcsin x + C ), (</td>
</tr>
<tr>
<td>10. ( \int \frac{dx}{1+x^2} = \arctan x + C )</td>
</tr>
</tbody>
</table>

Sometimes it may be difficult or impossible to find an antiderivative which is an elementary function. There are different methods of integration in this case. The simplest methods are linear integration and integration by substitution.

**Linear integration** allows one to break complicated integrals into simpler ones.

*Example:*

\[
\int (5x + \sin x)dx = \int 5x dx + \int \sin x dx = \frac{5x^2}{2} - \cos x + C.
\]
Integration by substitution

In calculus, the substitution rule is an important tool for finding antiderivatives and integrals. It allows to find the antiderivative for composite function (like the chain rule for differentiation).

Example:

\[
\int \frac{\ln^2 x}{x} \, dx = \left[ t = \ln x \right] dt = \frac{1}{x} \, dx = x \, dt
\]

\[
= \int t^6 \, dt = \frac{t^6}{6} + C = \frac{\ln^6 x}{6} + C.
\]

1.7. DEFINITE INTEGRAL

Integration was introduced as the reverse of differentiation. A more rigorous treatment would show that integration is a process of adding or «summation».

Consider the graph of the positive function \( y(x) \) shown in figure 1.5. Suppose we are interested in finding the area of the region bounded above by the graph of \( y(x) \), bounded below by the \( x \)-axis, bounded to the left by the vertical line \( x_1 = a \), and bounded on the right by the vertical line \( x_n = b \).

![Fig. 1.5. Calculation of the area under a curve \( f(x) \)](image)

One way in which this area can be approximated is to divide it into a number of rectangles, find the area of each rectangle, and then add up all these individual rectangular areas. The sum of the areas of all \( n \) rectangles is then

\[
S_{ABDC} \approx \sum_{i=1}^{n} y_i \cdot \Delta x_i
\]

This quantity gives us an estimate of the area under the curve but it is not exact. To improve the estimate we must take a large number of very thin
rectangles. So, what we want to find is the value of this sum when \( n \) tends to infinity and \( \Delta x \) tends to zero. We write this value as

\[
S_{ABCD} = \lim_{\Delta x_i \to 0} \sum_{i=1}^{n} y_i \Delta x_i = \int_{a}^{b} y \, dx.
\]

(1.10)

Limit of the sum is called **the definite integral** of \( y \) from \( x = a \) to \( x = b \) and it is written \( \int_{a}^{b} y \, dx \).

Fundamental theorem of calculus (**the Newton–Leibniz formula**): Let \( f(x) \) be integrable over the interval \([a; b]\), and suppose there is an antiderivative \( F(x) \) of \( f(x) \) over the interval \([a; b]\). Then, the definite integral with integrand \( f(x) \) and limits \( a \) and \( b \) is equal to the value of the antiderivative \( F(b) \) minus the value of antiderivative \( F(a) \):

\[
\int_{a}^{b} f(x) \, dx = F(x) \bigg|_{a}^{b} = F(b) - F(a).
\]

(1.11)

The notation \( F(x) \bigg|_{a}^{b} \) means the following: at first substitute the upper limit \( b \) into the function \( F(x) \) to obtain \( F(b) \) and then from \( F(b) \) we subtract \( F(a) \), the value obtained by substituting the lower limit \( a \) into \( F(x) \). This **Newton–Leibniz formula** (1.11) allows us to easily solve definite integral, if we can find the antiderivative function of the integrand.

Unlike the **indefinite integral**, which is **the set of functions**, the **definite integral** is a **numerical value**, that represents the area under the curve \( f(x) \). **Features of the definite integral:**

1. \( \int_{a}^{b} f(x) \, dx = -\int_{b}^{a} f(x) \, dx \).
2. \( \int_{a}^{b} (f(x) + g(x)) \, dx = \int_{a}^{b} f(x) \, dx + \int_{a}^{b} g(x) \, dx \).
3. \( \int_{a}^{a} f(x) \, dx = 0 \).
4. \( \int_{a}^{b} k f(x) \, dx = k \int_{a}^{b} f(x) \, dx \).
5. \( \int_{a}^{b} f(x) \, dx = \int_{a}^{c} f(x) \, dx + \int_{c}^{b} f(x) \, dx \).
Example 1:

Consider the integral \( \int_{0}^{3} x \, dx \). The area under the line is the triangle (fig. 1.6). The area of any triangle is half its base times the height. It is: \( S = \frac{1}{2} \cdot 3 \cdot 3 = \frac{9}{2} \).

As expected, the integral yields the same result:

\[
\int_{0}^{3} x \, dx = \frac{x^2}{2} \bigg|_{0}^{3} = \frac{3^2}{2} - \frac{0^2}{2} = \frac{9}{2} - 0 = \frac{9}{2}.
\]

Fig. 1.6. The area \( S \) under the line \( y = x \)

Example 2:

Calculate area \( S \) limited the curve \( y = x^2 \), axis \( x \) and lines \( x_1 = -1 \) and \( x_2 = 2 \). On fig. 1.7 this area is cross-hatched.

\[
S = \int_{-1}^{2} x^2 \, dx = \frac{x^3}{3} \bigg|_{-1}^{2} = \frac{2^3}{3} - \frac{(-1)^3}{3} = \frac{8}{3} - \frac{1}{3} = \frac{9}{3} = 3.
\]

Fig. 1.7. The area \( S \) under the curve \( y = x^2 \)

1.8. Differential Equations

A differential equation is an equation involving derivatives of an unknown function and possibly the function itself as well as the independent variable.

Example:

If an object of mass \( m \) is moving with acceleration \( a \) and being acted on with force \( F \) then Newton’s second Law is:

\[ F = ma. \]
Remained that acceleration \( a \) is a derivative of velocity \( v \) with respect to time or second derivative of a coordinate \( x \) with respect to time:

\[
a = \frac{dv}{dt} = \frac{d^2x}{dt^2}.\]

So, with all these things in mind Newton’s second Law can be written now as a differential equation:

\[
F = m \frac{dv}{dt} = m \frac{d^2x}{dt^2}. \tag{1.11}
\]

The **order** of a differential equation is determined by the highest order of the derivatives of the unknown function appearing in the equation.

**Example:**
\( y' = \sin(x) \) — is the 1\(^{\text{st}}\) order differential equations.
\( y'' + y^3 + x = 0 \) — is the 2\(^{\text{nd}}\) order differential equation.

A **solution** to a differential equation is any function which satisfies upon substitution of this function and its derivatives into the differential equation. It is important to notice, that a solution to a differential equation is a **function**, unlike the solution to an algebraic equation which is (usually) a number. A solution of a differential equation with its constants undetermined is called a **general solution**. The solution of differential equation complete with the values of the constants is called a **particular solution**. For constants determination additional conditions are used. These conditions constrain values of the function at some particular value of the independent variable. For example, if the equation involves the velocity, the additional condition might be the initial velocity, the velocity at time \( t = 0 \).

**Example:**
\[
\frac{dy}{dx} = x^2.
\]
Separate the variables: \( dy = x^2 \, dx \).
Integrate both sides: \( \int dy = \int x^2 \, dx \)
A general solution: \( y = \frac{x^3}{3} + C \).
Apply additional conditions: if \( x = 0 \), than \( y = 1 \).
Thus, the particular solution we are looking for is: \( y = \frac{x^3}{3} + 1 \).

**Questions:**
1. Give a definition of the derivative; explain its physical and geometrical meaning. What is a function gradient? What is the direction of a function gradient?
2. What is the physical meaning of the second derivative of way with respect to time?
3. What is an extremum of function? Formulate the stages of the extremum function investigation.
5. Give a definition of the partial derivatives. What is their physical meaning?
6. What are partial differentials and total differential of function?
7. Give a definition of the antiderivative function. What is an indefinite integral?
8. Give a definition of the definite integral. What is the geometrical meaning of the definite integral?
10. What is a differential equation? What determines the order of the differential equation? What is the solution of the differential equation?
11. What is a difference between a general solution and a particular solution? How to obtain a particular solution from general one?
12. How to check whether function is a solution of differential equation?

Chapter 2. PROBABILITY THEORY

In the real world events cannot be predicted with total certainty. The best one can do is to say how likely they are to happen, using the idea of probability.

Probability theory is the branch of mathematics deals with analysis of random events. Random event is event which at realization of a complex of conditions may occur or may not occur. Events can be named with capital letters: \( A, B, C \)… Examples of random events: the birth of girl in the family; the birth of a child with a predicted weight; the emergence of epidemic disease in the region in a certain period of time.

2.1. CLASSICAL (THEORETICAL) AND STATISTICAL (EMPirical) PROBABILITY DEFINITION

For example, let’s consider tossing a fair die. There are six possible numbers that could come up («outcomes»), and, since the die is fair, each one is equally likely to occur. So one can say each of these outcomes has probability \( 1/6 \). Since the event «an odd number comes up» consists of exactly three of these basic outcomes, one can say the probability of «odd» is \( 3/6 \), i.e. \( 1/2 \).

More generally, if there is a situation in which there are \( n \) equally likely outcomes, and the event \( A \) consists of exactly \( m < n \) of these outcomes, one can say that the classical probability \( P(A) \) of the event \( A \) is:

\[
P(A) = \frac{m}{n}.
\]  

(2.1)

This definition can be applied in a situation in which all possible outcomes and the outcomes in the events can be counted.

Example:
A box contains 3 red marbles, 1 blue marble, and 4 yellow marbles. One marble is drawn at random. There are now 8 equally likely marbles that can be drawn:
\[
P(\text{draw one of the eight marbles and it is red}) = \frac{3}{8}.
\]
\[
P(\text{draw one of the eight marbles and it is blue}) = \frac{1}{8}.
\]
\[
P(\text{draw one of the eight marbles and it is yellow}) = \frac{4}{8}.
\]

So the classical probability definition is based on the physics of the experiment, but does not require the experiment to be performed. For
example, we know that the probability of a balanced coin turning up heads is equal to 0.5 without ever performing trials of the experiment.

The probability of event accepts value between zero and unit: $0 \leq P(A) \leq 1$. If $P(A) = 1$, event $A$ is certain event. A certain event is certain to occur. If $P(A) = 0$, it is impossible event. An impossible event has no chance of occurring. In other cases $A$ is a random event and its probability $0 < P(A) < 1$.

**Examples:**
The Christmas will be celebrated on the 25th of December this year. This is a certain event.

When a number cube is rolled 7 is an impossible event.
The sunny day in London is a random event.

**Statistical probability.** Since random experiments can be repeated as many times as we wish under identical conditions (in theory) we can measure the relative frequency of the occurrence of an event. Statistical (empirical) probability is based on long-run relative frequencies.

The relative frequency $f$ of event $A$ in a given set of $N$ trials is the ratio of the number $M$ of those trials in which $A$ occurs to the total number of trials $N$:

$$f = M/N.$$

**Statistical probability** of event is a limit to which relative frequency of event tends at unlimited increase of the general number of tests:

$$P(A) = \lim_{N \to \infty} \frac{M}{N};$$

where the value $\frac{M}{N}$ is a relative frequency of event $A$.

For example, one tosses a coin 100 times and observes heads on 52 of the tosses. Thus, the relative frequency is equal to 0.52. In case if the number of tosses aspires to $\infty$ (for example, 1000) and the number of head appearance is equal to 498, the statistical probability is equal to:

$$P(A) = \lim_{N \to \infty} \frac{498}{1000} = 0.5.$$

**2.2. Types of random events**

There are 3 main types of random events: disjoint events, independent events and dependent events.

1. Two events are **disjoint** if it is impossible for them to occur together.

   **Example:** anyone cannot be both male and female, nor can they be aged 20 and 30.

   Events $M_1$, $M_2$, ..., $M_k$ form the **full group of events** if at any tests there can be only one of them, and can’t be any other events.
Example 1:
Getting a student on one test mark «1», or «2», or «3», or «4», or «5», or «6», or «7», or «8», or «9» or «10» — the events are disjoint, since one of these marks exclude the other on the same exam. These events form a full group of events.

Example 2:
Let \( P(A) \) is the probability of death for some diseases; it is known and is equal to 2 \%. Then the probability of a successful outcome in this disease is 98 \%. These events form full group of events.

2. Two or more events are independent if the occurrence of one of the events does not change the probability of the other events. That is, the events have no influence on each other. Two events \( A \) and \( B \) are independent if when one of them happens, it doesn't affect the other one happening or not.

Examples: choosing a marble from a jar and landing on heads after tossing a coin; choosing a 3 from a deck of cards, replacing it, and then choosing an ace as the second card.

If two events are independent then they cannot be disjoint.

3. Two or more events are dependent if the result of one event is affected by the result of other events. For dependent events \( A \) and \( B \) two types of probabilities are known: conditional probability and unconditional one.

The unconditional probability \( P(B) \) of an event \( B \) is the probability that event \( B \) will occur before an event \( A \). The conditional probability \( P(B|A) \) of an event \( B \), in relation to event \( A \), is the probability that event \( B \) will occur given the knowledge that an event \( A \) has already occurred.

Example: taking out a marble from a bag containing some marbles and not replacing it, and then taking out a second marble are dependent events.

2.3. Probabilities Addition and Multiplication Rules

Probability addition rule:
1. When two events, \( A \) and \( B \), are disjoint, the probability that event \( A \) or event \( B \) will occur is the sum of the probabilities of each event:

\[
P(A \text{ or } B) = P(A) + P(B).
\]

(2.3)

2. For some disjoint events \( M_1, M_2, \ldots, M_k \):

\[
P(M_1 \text{ or } M_2 \text{ or } \ldots \text{ or } M_k) = P(M_1) + P(M_2) + \ldots + P(M_k).
\]

3. For full group of events:

\[
\sum_{i=1}^{n} P(M_i) = 1.
\]

(2.4)

Example:
A glass jar contains 1 red, 3 green, 2 blue, and 4 yellow marbles. If a single marble is chosen at random from the jar, what is the probability that it is yellow or green?

The probability of extracting of yellow marble is:

\[
P(\text{yellow}) = \frac{4}{10}.
\]
The probability of extracting of green marble is:

\[ P(\text{green}) = \frac{3}{10}. \]

The probability of extracting of yellow or green marble is:

\[ P(\text{yellow or green}) = P(\text{yellow}) + P(\text{green}) = \frac{4}{10} + \frac{3}{10} = \frac{7}{10}. \]

**Probability multiplication rule for independent events:**

1. If \( A \) and \( B \) are independent events, the probability of both events occurring is the product of the probabilities of the individual events:

\[ P(\text{A and B}) = P(A) \cdot P(B). \] (2.5)

**Example:**

A drawer contains 3 red paperclips, 4 green paperclips, and 5 blue paperclips. One paperclip is taken from the drawer and then replaced. Another paperclip is taken from the drawer. What is the probability that the first paperclip is red and the second paperclip is blue?

Because the first paper clip is replaced, the sample space of 12 paperclips does not change from the first event to the second event. The events are independent.

\[ P(\text{red and blue}) = P(\text{red}) \cdot P(\text{blue}) = \frac{3}{12} \cdot \frac{5}{12} = \frac{15}{144} = \frac{5}{48}. \]

2. For some disjoint events \( M_1, M_2, \ldots, M_k \):

\[ P(M_1 \text{ and } M_2 \text{ and } \ldots \text{ and } M_k) = P(M_1) \cdot P(M_2) \cdot \ldots \cdot P(M_k). \]

**Probability multiplication rule for dependent events:**

If \( A \) and \( B \) are dependent events, the probability joint appearance of two dependent events \( A \) and \( B \) is equal to product of the unconditional probability of first event by the conditional probability of another one:

\[ P(\text{A and B}) = P(A) \cdot P(B \mid A) \] (2.6)

or

\[ P(\text{B and A}) = P(B) \cdot P(A \mid B). \] (2.6a)

In second case the first occurs event \( B \) and its probability is equal \( P(B) \) and for event \( A \) the conditional probability \( P(A \mid B) \) is realized.

**Example:**

A drawer contains 3 red paperclips, 4 green paperclips, and 5 blue paperclips. One paperclip is taken from the drawer and is not replaced. Another paperclip is taken from the drawer. What is the probability that the first paperclip is red and the second paperclip is blue?

Because the first paper clip is not replaced, the sample space of the second event is changed. The sample space of the first event is 12 paperclips, but the sample space of the second event is now 11 paperclips. The events are dependent.

\[ P(\text{red and blue}) = P(\text{red}) \cdot P(\text{blue/red}) = \frac{3}{12} \cdot \frac{5}{11} = \frac{15}{132} = \frac{5}{44}. \]
2.4. Bayes Formula

Comparing two formulas (2.6) and (2.6a), one can find conditional probability of event $A$:

$$P(A/B) = \frac{P(A) \cdot P(B/A)}{P(B)}.$$  

Let’s consider an event $B$, that can happen with any events $A_k$ with a known conditional probability $P(B/A_k)$. Events $A_1, A_2, A_3, \ldots, A_n$ form the full group of events:

$$\sum_{i=1}^{n} P(A_i) = 1.$$

Then the probability of event $B$ appearance $P(B)$ with any events $A_k$ is given by the formula of a total probability of event $B$:

$$P(B) = \sum_{k=1}^{n} P(A_k) \cdot P(B/A_k).$$

Then the conditional probability of event $A_k$ appearance $P(A_k/B)$, given that the event $B$ happened is given by Bayes formula:

$$P(A_k/B) = \frac{P(A_k) \cdot P(B/A_k)}{\sum_{k=1}^{n} P(A_k) \cdot P(B/A_k)}.$$  \hspace{1cm} (2.7)

Bayes formula (2.7) is an important method for calculation conditional probabilities. It is often used to calculate posterior probabilities (as opposed to prior probabilities) given observations.

For example, a patient is observed to have a certain symptom, and Bayes’ formula can be used to compute the probability that a diagnosis is correct, given that observation. We illustrate this idea with details in the following example.

**Example:** Mammogram posterior probabilities.

Approximately 1 % of women aged 40–50 have breast cancer. A woman with breast cancer has a 90 % chance of a positive test from a mammogram, while a woman without has a 10 % chance of a false positive result. What is the probability $P(C/A)$ a woman has breast cancer given that she just had a positive test?

The probability that the woman has breast cancer is equal: $P(C) = 0.01$;

The probability that the woman has not breast cancer is equal: $P(N) = 0.99$;

The probability of a positive test with breast cancer is equal: $P(A/C) = 0.9$;

The probability of a positive test without breast cancer is equal: $P(A/N) = 0.1$.

$$P(C/A) = \frac{P(A/C) \cdot P(C)}{P(A/C) \cdot P(C) + P(A/N) \cdot P(N)} = \frac{0.9 \cdot 0.01}{0.9 \cdot 0.01 + 0.1 \cdot 0.99} = 0.083 \approx 8.3 \%.$$


**Questions:**
1. What is a random event? Give a classical and statistical definition of a random event probability.
2. What types of random events do you know?
3. Formulate a probability addition rule. What type of random events can be used for this rule?
4. Which events form the full group of events? What is the sum of probabilities of full group of events?
5. Formulate a probability multiplication rule for independent events.
6. What are independent events? What is a conditional probability?
7. Formulate a probability multiplication rule for dependent events.
8. Present the Bayes formula; interpret a meaning of values in this formula.
9. How can Bayes formula be used for disease diagnostics problems?

**Chapter 3. RANDOM VARIABLES. DISTRIBUTION OF RANDOM VARIABLES**

When the numerical value of a variable is determined by a random event, that variable is called a **random variable**. As opposed to other mathematical variables, a random variable conceptually does not have a single, fixed value (even if unknown); rather, it can take on a set of possible different values, each with an associated probability. The value of the random variable will vary from trial to trial as the experiment is repeated.

There are two types of random variables: discrete and continuous.

A random variable is called **discrete random variable** if it may assume any of a specified list of exact values. For example, when two dice are rolled the total on the two dice will be 2, 3, ..., 12. The total on the two dice is a discrete random variable. This number cannot be 1.1 or 13. Within a range of numbers, discrete random variables can take on only certain values.

**Examples:** the number of children in a family, the number of patients in a doctor’s surgery.

A random variable is called **continuous random variable** if it can assume an infinite number of possible values in the possible interval. Suppose the temperature in a certain city in the month of June in the past many years has always been between 25º to 35º centigrade. The temperature can take any value between the ranges 25º to 35º. The temperature on some day may be 25.13 ºC or 25.14 ºC or it may take any value between 25.13 ºC and 25.14 ºC. When we say that the temperature is about 30 ºC, it means that the temperature lies between 29.5 ºC and 30.5 ºC. Any observation which is taken falls in the interval. In continuous random variable the value of the variable is located in some interval, the interval may be both small and big. The probability for continuous random variable is directly proportional to the value of this interval.
Examples include height, weight, the amount of glucose in an orange, the time required to run a mile.

Any representation of every possible value of random variable and the associated probabilities is called a probability distribution.

Probability distribution can be presented by table, graph or formula, that describes values a random variable can take on, and its corresponding probability (discrete random variable) or density (continuous random variable).

### 3.1. Discrete Probability Distribution

1. A probability distribution can be presented in a tabular form showing the values of the random variable \( X \) and the corresponding probabilities donated by \( p(x_i) \) (table 3.1). Suppose a discrete random variable \( X \) can assume the values \( x_1, x_2, \ldots, x_n \) with corresponding probabilities \( p(x_1), p(x_2), \ldots, p(x_n) \). The set of ordered pairs \([x_1, p(x_1)], [x_2, p(x_2)], \ldots, [x_n, p(x_n)]\) is called the probability distribution of discrete random variable.

<table>
<thead>
<tr>
<th>Values of random variable ( x_i )</th>
<th>( x_1 )</th>
<th>( x_2 )</th>
<th>( \ldots )</th>
<th>( x_i )</th>
<th>( \ldots )</th>
<th>( x_n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability ( p(x_i) )</td>
<td>( p(x_1) )</td>
<td>( p(x_2) )</td>
<td>( \ldots )</td>
<td>( p(x_i) )</td>
<td>( \ldots )</td>
<td>( p(x_n) )</td>
</tr>
</tbody>
</table>

The probabilities \( p(x_i) \) must satisfy the normalization condition:

\[
\sum_{i=1}^{n} p_i = p_1 + p_2 + \ldots + p_n = 1.
\] (3.1)

2. The discrete probability distribution may also be described by the probability polygon (fig. 3.1).

The probability polygon concludes all possible values of a random variable \( x_i \) on a horizontal axis and probabilities \( p(x_i) \) corresponding them on a vertical axis.

![Fig. 3.1. The probability polygon](image)
3. The discrete probability distribution can also be presented by some formula. One of these formulas as example is shown as:

\[ P_k(n) = (1 - p)^{n-k} \cdot p^k. \]

In practice probability distribution more often is given by table.

### 3.2. CONTINUOUS PROBABILITY DISTRIBUTION.

**Probability density function**

A continuous probability distribution differs from a discrete probability distribution. The probability that a continuous random variable will assume a particular value is always zero. As a result, a continuous probability distribution cannot be expressed in tabular form.

The probability distribution of a continuous random variable is represented by a certain function \( f(x) \), called the probability density function. The **probability density function** of a continuous random variable can be integrated to obtain the probability that the random variable takes a value in a given interval.

The probability density function \( f(x) \) must obey condition: the total probability for all possible values of the continuous random variable \( X \) is equal to 1:

\[
\int_{-\infty}^{+\infty} f(x) \, dx = 1. \tag{3.2}
\]

It is the **normalization condition** for continuous distribution.

The probability that value \( x \) takes values in the interval \([x_1, x_2]\) can be found as:

\[
P (x_1 < X < x_2) = \int_{x_1}^{x_2} f(x) \, dx. \tag{3.3}
\]

In other words, the probability density function \( f(x) \) allows to determine the probability that \( X \) takes values in any interval \([x_1, x_2]\). Graphically probability \( P(x_1 < X < x_2) \) is the area under the probability density function curve from \( x_1 \) to \( x_2 \) as shown in fig. 3.2. The total area \( S \) under the probability density function curve \( f(x) \), which corresponds to the total probability, is equal to 1:

\[
S = \int_{-\infty}^{+\infty} f(x) \, dx = 1.
\]

Function of probability distribution \( f(x) \) completely defines the distribution law of continuous random variables.
Example:

\[ f(x) = \frac{2}{x^2} \text{ for } 1 \leq x \leq 2. \]

It necessary to find probability of \( X \) being between 1.5 and 2.

Let’s check the fulfillment of normalization condition:

\[ \int_{1}^{2} \frac{2}{x^2} \, dx = 1. \]

To find the probability of \( X \) being between it is necessary to evaluate the following integral:

\[ P (1.5 \leq x \leq 2) = \int_{1.5}^{2} \frac{2}{x^2} \, dx = \left[ -\frac{2}{x} \right]_{1.5}^{2} = \frac{1}{3}. \]

So the probability of \( X \) being between 1.5 and 2 is equal to 1/3.

### 3.3. Random Distribution Characteristics

There are following characteristics of random distribution: the central tendency and the dispersion.

#### Estimation of central tendency

The central tendency of a random distribution is an estimate of the «centre» of values distribution. There are three major types of estimates of central tendency: 1) expectation; 2) mode; 3) median.

In probability theory, expectation (or population mean) \( \mu \) of a random variable is the weighted average of all possible values that this random variable can take on.

In case of a discrete random variable expectation \( \mu \) can be found as:

\[ \mu = x_1 p_1 + x_2 p_2 + \ldots + x_n p_n = \sum_{i=1}^{n} x_i p_i. \quad (3.4) \]

In case of a continuous random variable expectation \( \mu \) can be expressed as:

\[ \mu = \int_{a}^{b} x \cdot f(x) \, dx. \quad (3.5) \]
Mode \( \text{Mo}(X) \) is a value of random variable \( X \) that appears most often in a set of values. The mode of a discrete random variable is the value that is most likely to be sampled. The mode of a continuous random variable is the value \( X \) at which its probability density function attains its maximum value.

Median \( \text{Me}(X) \) of a random variable is the unique value \( X \) in the range of \( X \) such that divides all distribution to two equiprobable parts:

\[
P(X < \text{Me}(X)) = P(X > \text{Me}(X)) = 0.5.
\]

Therefore the median of a continuous random variable can be calculated from the formula:

\[
\int_{a}^{Me} f(x)dx = \frac{1}{2}.
\]

Graphically the median \( \text{Me}(X) \) is a value of a random variable \( X \) which ordinate divides the total area \( S \) under a probability density function into two equal areas \( (S_1 = S_2) \).

Estimation of dispersion

The determination of the center of the data set is one aspect observations. Another feature of the observations is to know how the observations are spread about the center. The observation may be closed to the center or it may be spread away from the center. If the observation is closed to the center (usually the arithmetic mean), one can say that dispersion is small. The dispersion is large if the observations are spread away from the center.

The statistical measure of dispersion is the variance \( \text{variance} \) of random variable \( x \) which is obtained by formula:

\[
\sigma^2 = \mu [x - \mu(x)]^2
\]

or

\[
\sigma^2 = \mu(x^2) - [\mu(x)]^2.
\]
For continuous random variables definite in the interval \((a, b)\) the variance is given as:

\[
\sigma^2 = \int_{a}^{b} [x - \mu(x)]^2 f(x) \, dx
\]

(3.7)

or

\[
\sigma^2 = \int_{a}^{b} x^2 f(x) \, dx - [\mu(x)]^2. \tag{3.7a}
\]

Variance is measured in square units. To overcome the problem of dealing with squared units, statisticians take the square root of the variance to get the standard deviation \(\sigma\): \((\sigma = \sqrt{\sigma^2})\).

### 3.4. Normal distribution

The normal (or Gaussian) distribution is a continuous probability distribution that has a bell-shaped probability density function, known as the Gaussian function or informally the bell curve. For normal distribution the probability density function \(f(x)\) is given as

\[
f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x - \mu(x))^2}{2\sigma^2}}.
\]

(3.8)

The parameter \(\mu(x)\) is the expectation (mean) and \(\sigma\) is the standard deviation, both these parameters \((\mu(x) \text{ and } \sigma)\) completely determine probability density function \(f(x)\). Area \(S\) under the probability density function curve \(f(x)\), which corresponds to the total probability, is equal to 1:

\[
S = \int_{-\infty}^{+\infty} f(x) \, dx = 1.
\]

Normal distribution is a symmetrical distribution: \((\mu(x) = Mo(x) = Me(x))\). The graph of the normal distribution depends on two factors — the expectation \(\mu(x)\) and the standard deviation \(\sigma\). The expectation \(\mu(x)\) of the distribution determines the location of the centre of the graph, and the standard deviation \(\sigma\) determines the height and width of the graph. When the standard deviation is large, the curve is short and wide; when the standard deviation is small, the curve is tall and narrow but area under the curve is always equal to one (fig. 3.4). The maximal density of probability is equal to \(\frac{1}{\sigma \sqrt{2\pi}} \approx 0.4\) and corresponds to an expectation \(\mu(x)\).
If the data distribution is approximately normal then about 68% of the data values are within one standard deviation of the expectation, about 95% are within two standard deviations, and about 99.7% lie within three standard deviations:

\[
P(\mu(x) - \sigma < x < \mu(x) + \sigma) = 0.6827 = 68.27\%.
\]

\[
P(\mu(x) - 2\sigma < x < \mu(x) + 2\sigma) = 0.9545 = 95.45\%.
\]

\[
P(\mu(x) - 3\sigma < x < \mu(x) + 3\sigma) = 0.9973 = 99.73\%.
\]

The formula (3.11) is known as three-sigma rule (fig. 3.5).

The expectation, median, and mode of the normal distribution are the same:

\[
\mu(x) = Mo(x) = Me(x).
\]
Example:
It is known the person pH blood is normally distributed quantity with mean 7.4 and a standard deviation 0.2. Find a range of pH values.
Let’s use the three-sigma rule.
\[ P(7.4 - 3 \cdot 0.2 < x < 7.4 + 3 \cdot 0.2) = 0.9973 = 99.73\% . \]
With a probability of equal to 99.73 % it is possible to approve, that the range of the person’s pH values makes 6.8÷8.

Questions:
1. What are the random variables? What are the differences between discrete and continuous random variables? Make examples.
2. How to specify probability distribution for discrete random variable? What is a normalization condition for the distribution for discrete random variable?
3. How to specify probability distribution for continuous random variable? What is a normalization condition in this case?
4. Give a definition of random distribution characteristics: expectation, mode, median, variance, standard deviation. Explain their meaning.
5. How to determine numeric parameters for discrete random variable distribution?
6. How to determine numeric parameters for continuous random variable distribution?
7. What features of continuous random variable normal distribution are known?

Chapter 4. MATHEMATICAL STATISTICS FUNDAMENTALS

To know the taste of melons, not necessarily eat all of it.
(old Eastern wisdom)

4.1. GENERAL POPULATION AND SAMPLE

Mathematical statistics deals with gaining information from data. In practice, data often contain some randomness or uncertainty. Statistics handles such data using methods of probability theory.

**General population** is the set of all objects (units), on which scientists are going to draw conclusions in the study of specific problems. General population consists of all objects that should be considered. **General population** consists of all objects that should be considered. The composition of the population depends on the objectives of the study. Sometimes the general population is the entire population of a given region (for example, when we study the attitudes of potential voters to the candidate), usually given several criteria that determine the object of research.

In practice it is not possible to investigate all the objects of interest to us. Then apply the **sampling method** — that is, limited to examining only some parts of objects.

In statistics, a **sample** is a subset of a population. The sample represents a subset of manageable size. Samples are collected and statistics are calculated
from the samples so that one can make inferences or extrapolations from the sample to on the entire population. However, this is necessary to select objects in the sample, subject to certain procedures. Without going into details, one can note that the **basic requirements for the sample** can be considered:

- representativeness (the ability to be a reflection of the general population);
- coincidence unit (each object the general population should have an equal probability of being selected);
- sufficiency level to obtain statistically significant results.

There are two types characteristics of the sample. A qualitative characteristic of the sample — which exactly one can choose and what methods of constructing the sample one can use for this. Quantification of the sample — how many cases we choose, in other words, the sample size.

**Sample size** is the number of cases included in the sample. From statistical considerations it is recommended that the number of cases is not less than 30–35. To calculate the sample size necessary to determine the permissible scope of sampling error, the level of confidence probability and the expected variance.

### 4.2. Statistical series types

**Simple statistical series** is the set of values of variable $X(x_i)$, where $i \in [1, n]$ which is presented as table 4.1.

<table>
<thead>
<tr>
<th>Number</th>
<th>1</th>
<th>2</th>
<th>...</th>
<th>$n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variant $x_i$</td>
<td>$x_1$</td>
<td>$x_2$</td>
<td>...</td>
<td>$x_n$</td>
</tr>
</tbody>
</table>

Resulting statistical material $x_1, x_2 \ldots x_i$ observation is the primary data on the size, subject of statistical analysis. Elements $x_1, x_2 \ldots x_i$ are called **variants**. Typically, such statistics are issued in the form of tables, charts, histograms, etc. If the sample volume $n$ contains $k$ various elements $x_1, x_2 \ldots x_k$, and $x_i$ found $m_i$ times, the number of $m_i$ is called the **frequency** of $x_i$ element. And the ratio $m_i/n = f_i$ is called the **relative frequency** of $x_i$ element.

**Variational series** is a table whose first row contains the $x_i$ elements in ascending order, and the second one contains their frequency $m_i$ (or relative frequency $f_i$). Variation series is presented in table 4.2.

<table>
<thead>
<tr>
<th>Variant, $x_i$ $(x_1 &lt; x_2 &lt; x_3 \ldots &lt; x_k)$</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>...</th>
<th>$x_k$</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency, $m_i$</td>
<td>$m_1$</td>
<td>$m_2$</td>
<td>$m_3$</td>
<td>...</td>
<td>$m_k$</td>
<td>( \sum_{i=1}^{k} m_i = n )</td>
</tr>
<tr>
<td>Relative frequency, $f_i = \frac{m_i}{n}$</td>
<td>$\frac{m_1}{n}$</td>
<td>$\frac{m_2}{n}$</td>
<td>$\frac{m_3}{n}$</td>
<td>...</td>
<td>$\frac{m_k}{n}$</td>
<td>( \sum_{i=1}^{k} \frac{m_i}{n} = 1 )</td>
</tr>
</tbody>
</table>
Frequencies (relative frequencies) polygon of sample is called a broken line with vertices $(x_i, m_i)$ or $(x_i, f_i)$ as illustrated in (fig. 4.1).

![Fig. 4.1](image)

Fig. 4.1. a — polygon of frequencies; b — polygon of relative frequencies

Ranked statistical series is the series containing the variants $x_i$ in ascending or descending order. Rankings is the process of converting a simple statistical series based on the ordering (clustering), the numerical values of the elements of a number of descending ($x_1 > x_2 > \ldots > x_i$) or ascending ($x_1 < x_2 < \ldots < x_i$), where $i$ is the rank of the element ranked a number of values of variable $i \in [1, n]$. As a result of the transformation a ranked series is obtained.

Depending on the type of trait distinguish discrete and interval variational series. Depending on the amount of input data and the field of values of one-dimensional quantitative trait, the frequency distribution is also divided into discrete and interval. If a different version of the very many (10–15), these options are grouped by selecting a certain number of intervals, grouping and thus obtaining an interval frequency distribution. Algorithm for grouping data set in interval variational series consists of the following steps:

1) to find the range of the variants $(x_{\text{max}} - x_{\text{min}})$;
2) to find the number of intervals $k$. The number of intervals $k$ varies usually from 5 to 25. There are formulas for determining the «optimal» values of $k$ and thus constructing the optimal allocation of frequencies:

$$k = \sqrt{n} \quad \text{or} \quad k \approx 1 + 3.32 \lg n. \quad (4.1)$$

For large sample size $n$, this formula gives a lower bound for $k$.

3) to find the interval width $h$:

$$h = \frac{x_{\text{max}} - x_{\text{min}}}{k}. \quad (4.2)$$

4) to find the boundary points of each interval:

$$x_0 = x_{\text{min}}; \quad x_1 = x_0 + h; \quad x_2 = x_1 + h; \ldots; \quad x_k = x_{\text{max}}.$$
5) to calculate the number of variant $m_i$, caught in the interval, with the variants that come on the border ranges, refer only to one of the intervals. The results are entered into the table (table 4.3).

<table>
<thead>
<tr>
<th>Interval</th>
<th>Frequency, $m_i$</th>
<th>Relative frequency, $m_i/n = f_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_0 - x_1$</td>
<td>$m_1$</td>
<td>$m_1/n$</td>
</tr>
<tr>
<td>$x_1 - x_2$</td>
<td>$m_2$</td>
<td>$m_2/n$</td>
</tr>
<tr>
<td>$x_2 - x_3$</td>
<td>$m_3$</td>
<td>$m_3/n$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$x_{k-1} - x_k$</td>
<td>$m_k$</td>
<td>$m_k/n$</td>
</tr>
</tbody>
</table>

For interval variational series **histogram** is a diagram consisting of rectangles whose bases are intervals of $\Delta x$ (x-axis). Frequency $m_i$ (relative frequency $m_i/n = f_i$ or probability density $m_i/n\Delta x$) is plotted on the vertical axis (y-axis).

\[ f_i = \frac{m_i}{nh} \]

To plot a histogram (fig. 4.2) it is necessary to:
1) divide the range of the data into intervals of equal width $h$;
2) count frequency $m_i$ (or relative frequency $m_i/n = f_i$ or probability density $m_i/n\Delta h$) in each interval;
3) draw a plot with the intervals on the horizontal axis and the frequencies $m_i$ (or relative frequency $m_i/n = f_i$ or probability density $m_i/n\Delta h$) on the vertical axis.

Let $x$ is a continuous random variable with unknown probability density $f(x)$. To assess $f(x)$ a sample $x_1, x_2, \ldots, x_n$ one can divide the range of values $x$ at intervals of width $h$. Denote middle value of intervals $x_i$, and through $m_i$ number of elements in the sample caught in a specified interval. Then $f_i = \frac{m_i}{nh}$ is an assessment of the probability density at $x_i$. In the rectangular coordinate
system rectangles are constructed with bases $h$ and heights $\frac{m_i}{n} h$ (fig. 4.2). Area of a rectangle equals to the relative frequency.

4.3. GENERAL POPULATION PARAMETER ESTIMATION

Conducting a research project always involves some errors. A common mistake is to study the difference between the true value (in the general population) of the observed variable and its observed value (in the sample).

To reduce the influence of random errors someone can measure values of some quantity $x$ several times. As a result of the measurements values of the quantity is obtained: $x_1, x_2, x_3, \ldots, x_n$.

With this sample, the result of measurements can be evaluated. Value that would be such an evaluation is denoted as $\bar{x}$. But as it is the importance of assessing the results of measurements would not represent the true value of the measured value, it is necessary to estimate its error. True value of $x$ lies in the interval $\bar{x} \pm \delta$, which is called a confidence interval (the permissible deviation of the observed values from the truth). In this case the detected result of the measurements can be written in the form:

$$\mu = \bar{x} \pm \delta.$$  

How frequently the observed interval contains the parameter of interest is determined by confidence level $\gamma$. Confidence level indicates the degree of confidence that the value of the observed element falls into the specified range of the confidence interval. Typically confidence level $\gamma = 95\%$ is used. To obtain more accurate data confidence level increases to $\gamma = 99\%$, but this entails a significant increase in sample size.

For example, measuring the length $l$ of a segment, the final result can be recorded in the form:

$$l = (8.34 \pm 0.02) \text{ mm}, \ (\gamma = 0.95).$$

This means that out of 100 chances −95 for what the true value of the length $l$ of the segment lies in the range from 8.32 to 8.36 mm.

The half confidence interval can be calculated by the following formula:

$$\delta = t_{\gamma, n} \frac{S}{\sqrt{n}}. \quad (4.4)$$

In this case the lower and the upper confidence limits are:

$$\bar{x} \pm \delta = \bar{x} \pm t_{\gamma, n} \frac{S}{\sqrt{n}}, \quad (4.4a)$$

where $t_{\gamma, n}$ is the Student's coefficient — special coefficient which depends on the confidence level $\gamma$ and the number of measurements $n$; $S$ is the sample
**standard deviation**, which may be calculated from the sample by using the following formula:

\[ S^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1} \]  \hspace{1cm} (4.5)

The coefficient Student’s \( t_{\gamma, n} \) is found from the special tables. Proceeding on the same lines the 95 and 99 percent confidence limits may be stated as:

\[ \bar{x} \pm \delta = \bar{x} \pm 1.96 \frac{S}{\sqrt{n}}, \quad (\gamma = 0.95) \]  \hspace{1cm} (4.6)

\[ \bar{x} \pm \delta = \bar{x} \pm 2.57 \frac{S}{\sqrt{n}}, \quad (\gamma = 0.99) \]  \hspace{1cm} (4.6a)

**Example:**
A random sample of 64 students made an average score of 60, with a standard deviation of 15. Construct 99 % confidence interval estimation for the mean score of entire class.

It’s known the following data: \( \bar{x} = 60, S = 15, n = 64, t_{\gamma, n} = 2.57 \) (from the table).

Using the formula for 99 % confidence interval may be written as:

\[ \mu = \bar{x} \pm 2.57 \frac{S}{\sqrt{n}}. \]

The lower confidence limit is:

\[ \bar{x} - 2.57 \frac{S}{\sqrt{n}} = 60 - 2.57 \frac{15}{8} = 60 - 4.82 = 55.18 \approx 55. \]

The upper confidence limit is:

\[ \bar{x} + 2.57 \frac{S}{\sqrt{n}} = 60 + 2.57 \frac{15}{8} = 60 + 4.82 = 64.82 \approx 65. \]

Hence, the 99 % confidence interval estimate for the mean will be, 55 < \( \mu \) < 65, i.e. the mean score is between 55 and 65.

**4.4. Correlation Analysis**

**Correlation** determines the degree of association between two or more variables, for example, the correlation between the height of parents and their children. Correlations are useful because they can indicate a predictive relationship that can be exploited in practice. There are two approaches for determination of the correlation: qualitative approach and quantitative one.

Qualitative determination of the correlation between random variables \( X \) and \( Y \) is based on the determination of the **scatterplot** form. The scatterplot is used to graphically display the relationship between two quantitative variables measured on the same cases. One variable is placed on the horizontal \((x)\) and another one is placed on the vertical \((y)\) axis. Each case is represented by a point \((x_i, y_i)\) on the graph, placed according to its values on each of the variables.

If the scatterplot is closed to the circle shape, there is not any relationship between two variables \( X \) and \( Y \). If the scatterplot is ellipse shaped, there is
relationship between two variables \(X\) and \(Y\). The plotted points \((x_i, y_i)\) can also cluster around a straight line. This line is called the regression line obtained by inspection. Scatter plots that show linear relationships between variables can differ in several ways including the slope of the line about which they cluster and how tightly the points cluster about the line (fig. 4.3). As long as the scattered points show closeness to a straight line of some direction, one can draw a straight line to represent the sample data (fig. 4.3, \(a, b\)). But when the points do not lie around a straight line, and from circle there is not any relationship between the two variables (fig. 4.3, \(d\)). If the plotted points cluster around a curve there is non-linear relationship between \(Y\) and \(X\) (fig. 4.3, \(c\))

![Fig. 4.3. Scatterplots](image)

**Quantitative determination** of the correlation between random variables \(X\) and \(Y\) is based on the definition of the correlation coefficient \(r\). For linear relationships between variables \(X\) and \(Y\) correlation coefficient \(r\) is defined as:

\[
r = \frac{\sum_{i=1}^{n} (x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{X})^2 \cdot \sum_{i=1}^{n} (y_i - \bar{Y})^2}}.
\]  

(4.7)

The terms \((x_i - \bar{X})\) and \((y_i - \bar{Y})\) are the deviation of an each observation from the mean.
In general, correlation coefficients can range from \(-1 \) to \( +1 \) \((-1 \leq r \leq 1\)), with the magnitude and the sign of \( r \) representing the **strength** and **direction** respectively of the relationship between the two variables. Depends on the correlation coefficient sign distinguish positive (with value \( X \) increase, the mean value of variable \( Y \) increases too) and negative (with value \( X \) increase, the mean value of \( Y \) decreases) correlation. If \(|r| < 0.3\) the linear relationship between \( X \) and \( Y \) is absent or extremely weak. Moderate correlation is observed when \(0.3 < |r| < 0.7\). There is a strong correlation if \(0.7 < r < 1\). There is a functional dependence between two variables if \(|r| = 1\).

**Questions:**
1. Give definition of a general population and a sample. What does representative sample mean? What are the basic requirements for the sample?
2. What are variant, simple statistical series, variational series?
3. What are the stages for variational series (discrete distribution) formation? How can variational series be represented graphically?
4. What are the stages for ranked statistical series (continuous distribution) formation? How ranked statistical series can be represented graphically?
5. Describe the central tendency and the dispersion characteristics for random variables.
6. Explain difference between sample parameters point estimate and sample parameters interval estimate.
7. What is confidence interval finding algorithm?
8. How to determine a required sample volume if the accuracy of interval estimate is known?
9. What is a correlation analysis major task?
10. What is a scatter diagram? What information about parameters correlation does this diagram contain?
11. How to calculate a correlation coefficient between parameters? What information does this coefficient contain?
12. What is the condition of reliable correlation coefficient?

**Chapter 5. BASICS OF BIOMECHANICS**

**5.1. DEFORMATION CHARACTERISTICS**

Forces acting on an extended body are assumed to produce deformation of body. The study of elasticity and plasticity is concerned with how bodies deform under the action of applied forces. Change in the size or shape of the body under the action of applied forces is called **deformation**. Elastic deformation and plastic one are distinguished.

**Deformation is elastic** if body returns to its original size and (or) shape when the load is removed. Elastic deformation is reversible. For example, rubber does large elastic deformation. **Plastic deformation** is defined as permanent, non-recoverable (irreversible) deformation: body does not return to its original
dimensions. For example, copper, silver, and gold have rather large plastic deformation. From an atomic perspective, in the case of elastic deformation when external force is removed atoms of the body return to their equilibrium (original) position. Plastic deformation corresponds to the breaking of bonds with original atom neighbors and then reforming bonds with new neighbors. After removal of the force, the large numbers of atoms that have relocated, do not return to original position.

Accordingly, the **elasticity** is the ability of a body to return to original size and (or) shape after the force that caused its deformation is no longer applied. **Plasticity** is the ability of body to irreversible changes of size and shape in response to applied forces.

There are different ways in which a body may change its dimensions when an external forces $F$ act on it. Depending on the direction of applied force $F$ the different types of deformation are distinguished. The main types of deformation are: compression, tension, bending, shear and torsion. These are shown in fig. 5.1. In fig. 5.1, $a$ to obtain a deformation of compression a body is compressed by four equal forces applied normal to its cross-sectional areas. If the body is stretched under the action of applied forces, the deformation body is known as deformation of tension (fig. 5.1, $c$). In order to obtain shear deformation two equal and opposite deforming forces should be applied parallel to the cross-sectional area of the body, as shown in fig. 5.1, $f$. The deformed body is illustrated by dashed line. In general, various types of deformation of the body can be reduced to two basic: **tensile deformation** and **shear deformation**.

![Fig. 5.1. Main types of deformation:](image)
If a force $F$ is exerted on a body, such as the vertically suspended metal rod shown in fig. 5.1, $c$, the length of the rod changes. One of the basic characteristics of a deformed body is absolute deformation $\Delta l$. Absolute deformation $\Delta l$ is the change in length of deformed body:

$$\Delta l = l - l_0,$$  \hspace{1cm} (5.1)

where $l_0$ is the original length. Absolute deformation has units of length [m].

Another basic characteristic of a deformed body is strain. Strain $\varepsilon$ (relative deformation) is a ratio of the change in length $\Delta l$ to the original length $l_0$ of the body. It is a dimensionless value, because it is only a ratio between two similar quantities:

$$\varepsilon = \frac{\Delta l}{l_0}, \text{ in percent } \varepsilon = \frac{\Delta l}{l_0} \cdot 100 \%. \hspace{1cm} (5.2)$$

When a deforming external force is applied to a body, it is deformed. The displaced particles within the deformed body try to attain their equilibrium position. This tendency gives rise to an internal force within the body. The force developed within the deformed body is called restoring force. This restoring force is equal in magnitude but opposite in direction to the applied force. The restoring force per unit area is known as stress $\sigma$. If $F$ is the force applied and $S$ is the area of cross section of the body, magnitude of the stress $\sigma$ is:

$$\sigma = \frac{F}{S}. \hspace{1cm} (5.3)$$

Stress $\sigma$ is measured in N/m$^2$. Thus, stress $\sigma$ is the internal resistance offered by the body to the external load applied to it per unit cross sectional area.

If deformation is elastic the restoring force $F_e$ is directly proportional to absolute deformation $\Delta l$:

$$F_e = k \cdot \Delta l. \hspace{1cm} (5.4)$$

Equation (5.4) is called Hooke’s Law. Restoring force (or other words force of elasticity) is always opposite to the direction of displacement $x$, therefore Hooke’s Law in vector form:

$$\vec{F} = -k \cdot \vec{x}, \hspace{1cm} (5.5)$$

where $x = \Delta l$, the coefficient $k$ is called the stiffness (in SI units: N/m). Stiffness $k$ depends on elastic properties of a material and its geometrical size.

If the deformation is elastic, Hooke’s Law may also be expressed in terms of stress $\sigma$ and strain $\varepsilon$:

$$\sigma = E \cdot \varepsilon. \hspace{1cm} (5.6)$$

where $E$ is Young’s modulus. $E$, sometimes referred to as the modulus of elasticity, is numerical constant, that describes the elastic properties of a solid undergoing tension or compression in only one direction. Because strain $\varepsilon$ is
dimensionless, $E$ has the same units as stress $\sigma$. The SI unit of modulus of elasticity is the Pascal ($1 \text{ Pa} = 1 \text{ N/m}^2$) The higher the $E$ value, the higher is the load required to stretch the body to the same extent, and thus the stiffer is the material. Compare, for example, the stiffness of metal wire with that of a rubber band or plastic sheet when they are loaded.

In general, elastic modulus $E$ is not the same as stiffness $k$. Elastic modulus is a property of the constituent material; stiffness is a property of a structure and depends on the geometric parameters. The formula for stiffness $k$ is:

$$k = \frac{ES}{l_0}.$$ (5.7)

So, $k$ is directly proportional to the elastic modulus $E$, cross-sectional area of the sample $S$ and inversely proportional to its length $l_0$.

5.2. Stress-strain diagram

The relation between the stress and the strain for a given material under tensile stress can be found experimentally. In a standard test of tensile properties, a test cylinder or a wire is stretched by an applied force. The fractional change in length (the strain) and the applied force needed to cause the strain are recorded. The applied force is gradually increased in steps and the change in length is noted. A graph is plotted between the stress (which is equal in magnitude to the applied force per unit area) and the strain produced. A stress-strain diagram is a graph derived from measuring load (stress $\sigma$) versus extension (strain $\varepsilon$) for a sample of a material (fig. 5.2). Analogous graphs for compression and shear stress may also be obtained. The stress-strain curves vary from material to material. These curves help to understand how a given material deforms with increasing loads. From the stress-strain diagram one can see the different mark points on the curve. It is because, when a material is subjected to tensile test, then it passes various stages before fracture. These stages are:

1) proportional limit;
3) elastic limit;
4) yield strength;
5) proof stress;
6) ultimate strength;
7) fracture of specimen.

From the graph, one can see that in the region between $O$ to $A$, the curve is linear: the strain is proportional to strain or elongation is proportional to the load. In this region, Hooke’s law is obeyed. This law of proportionality is valid up to a point $A$. The body regains its original dimensions when the applied force is removed. In this region, the solid behaves as an elastic body. This point $A$ is known as the proportionality limit ($\sigma_{pl}$). The linear portion of the curve is the elastic region and the slope is the modulus of elasticity or Young’s modulus:
\[ E = \text{tg} \alpha. \]

In the region from \( A \) to \( B \), stress and strain are not proportional. Nevertheless, material may still be elastic in the sense that the deformations are completely recovered when the load is removed. The point \( B \) in the curve is known as elastic limit. Elastic limit \( (\sigma_{\text{EL}}) \) is defined as the lowest stress at which permanent deformation can be measured.

If the load is increased further, the stress developed exceeds the elastic limit and strain increases rapidly even for a small change in the stress. The portion of the curve between \( C \) and \( K \) shows this. Beyond the elastic limit plastic deformation occurs and is not totally recoverable. There will be thus permanent deformation or permanent set when load is removed. The point \( C \) on the graph is yield limit. Yield strength \( (\sigma_{\text{YL}}) \) is defined as the stress at which a material begins to plastically deform. Interval \( CK \) is called proof stress: stress at which there are large increases in strain with little or no increase in stress.

Beyond point \( K \) stress increases rapidly with increase of strain. The highest point \( D \) of the diagram corresponds to the ultimate strength of a material. Stress which the specimen can withstand without failure is known as ultimate strength \( \sigma_U \).

Beyond point \( D \), additional strain is produced even by a reduced applied force and fracture occurs at point \( E \). If the ultimate strength and fracture points \( D \) and \( E \) are close, the material is said to be brittle. If they are far apart, the material is said to be ductile. As stated earlier, the stress-strain behaviour varies from material to material. For example, rubber can be pulled to several times its original length and still returns to its original shape.

\[ \text{Fig. 5.2. Stress-strain diagram} \]

**Ductile materials** are materials that are capable of undergoing large strains (at normal temperature) before failure. Ductile materials include mild steel, aluminum and copper and many others. **Brittle materials** exhibit very little inelastic deformation. Brittle materials include concrete, stone, glass. The stress-strain diagram for ductile and brittle material is shown in fig. 5.3.
Another important characteristic of a deformable body is the Poisson ratio $\mu$. When a specimen of material is stretched in one direction, it tends to contract in the other two directions perpendicular to the direction of stretch. Conversely, when a sample of material is compressed in one direction, it tends to expand in the other two directions.

The Poisson ratio is the ratio of the fraction of expansion divided by the fraction of compression, for small values of these changes. It connects the longitudinal relative deformation $\varepsilon$ and transverse relative deformation $\varepsilon_1$ of the specimen:

$$\mu = -\frac{\varepsilon_1}{\varepsilon}$$  \hspace{1cm} (5.8)

These deformations always have different signs. For example, during the tension (fig. 5.4)

$$\varepsilon = \frac{\Delta l}{l_0} > 0, \quad \varepsilon_1 = \frac{\Delta d}{d_0} > 0,$$

during the compression: $\varepsilon < 0, \varepsilon_1 > 0$, and Poisson ratio $\mu$ is always positive.

---

Fig. 5.3. The stress-strain diagram for ductile and brittle material

Fig. 5.4. The relation between longitudinal relative deformation and transverse relative deformation
\[ \mu = -\frac{\varepsilon_1}{\varepsilon} > 0. \]

Poisson’s ratio depends only on the properties of the material and determines the relative change in its volume \( V \):

\[ \frac{\Delta V}{V} = \varepsilon \ll 2\mu \varepsilon. \tag{5.8} \]

Most materials have Poisson ratio values ranging between 0 and 0.5. A perfectly incompressible material deformed elastically (\( \Delta V = 0 \)) at small strains would have a Poisson ratio of exactly 0.5.

On the molecular level, Poisson’s effect is caused by slight movements between molecules and the stretching of molecular bonds within the material lattice to accommodate the stress. When the bonds elongate in the direction of load, they shorten in the other directions. This behavior multiplied many times throughout the material lattice is what drives the phenomenon.

Questions:
1. Give a definition of a solid body deformation. What is the difference between elastic deformation and plastic one? Specify main types of deformation.
2. What is a quantitative measure of deformation? What is the unit of a strain. Give a definition of a stress and indicate its units.
3. Write and analyze the Hooke’s Law for tension (compression) deformation. What is the relation between stiffness and Young’s modulus?
5. Write and analyze the Hooke’s Law for shear deformation. What is the Poisson ratio?

Chapter 6. MECHANICAL OSCILLATIONS AND WAVES

6.1. HARMONIC OSCILLATIONS

Harmonic oscillation is the periodic process in which the parameter of interest is varied as sine or cosine. If there is no time-dependent force applied to the oscillator, then it is called a free oscillator.

The simple harmonic oscillator is a point mass connected to some elastic object (spring) of negligible mass that is fixed at the other end and constrained so that it may only move in one dimension.

Fig. 6.1 shows a point mass \( m \) connected to an elastic massless spring that is fixed at the other end and constrained so that it may only move in one dimension on smooth surface. If mass \( m \) will be displaced from the equilibrium position of a system by an amount \( x \) (displacement \( x \)), then the force will be developed in the spring that tries to restore the system to the equilibrium condition. The friction forces and air resistance are ignored. Let’s find the law, according to which the coordinate \( x \) changes with time \( x = f(t) \).
This restoring force acting in this system is elasticity force $F$, which is directly proportional to the displacement of mass from the equilibrium position of system $x$ and it will point in the opposite direction (fig. 6.1) (a condition known as Hooke’s Law):

$$F = -kx.$$ 

In this case the Newton’s Law, combined with Hooke’s Law for the behavior of a spring, states the following: $m\ddot{a} = \ddot{F}$, or in projection on horizontal axis $x$: $ma_x = -kx$, where $k$ is the spring constant, $m$ is the mass, $x$ is the position of the mass, and $\ddot{a}$ is its acceleration.

The velocity $\upsilon$, or the rate of change of the position with time, is defined as:

$$\upsilon = \dot{x}' = \frac{dx}{dt}.$$ 

The acceleration $a$, or the rate of change of the velocity, is defined as:

$$a = \dot{\upsilon}' = \dot{x}''; \quad a = \frac{d\upsilon}{dt} \Rightarrow a = \frac{d^2x}{dt^2}.$$ 

So, we’ve obtained the second order linear differential equation (DE). And its solution gives us the displacement $x$ of the mass as a function of time $t$. This DE describes the free harmonic oscillation:

$$m\frac{d^2x}{dt^2} = -kx \quad (6.1)$$

$$\downarrow$$

$$m\frac{d^2x}{dt^2} + kx = 0.$$
Introducing a constant $\omega_0^2 = \frac{k}{m}$, one can write
\[
\frac{d^2x}{dt^2} = -\omega_0^2x \quad \text{or} \quad \frac{d^2x}{dt^2} + \omega_0^2x = 0. \tag{6.2}
\]

The solution of this equation is harmonic function:
\[
x = A \sin(\omega_0 t + \phi_0). \tag{6.3}
\]

So, it is the general solution of the free harmonic oscillations, where $A$ is the amplitude; $\omega_0 = \sqrt{k/m}$ is the angular frequency and $\phi_0$ is the initial phase of oscillation; $(\omega_0 t + \phi_0)$ is the phase of oscillation; $t$ is the running time. The period of oscillation $T$ is equal to $T = 2\pi\sqrt{m/k}$; $\nu = 1/T$ is the frequency of oscillation.

The energy associated with a harmonic oscillator is the sum of the kinetic $E_k$ and potential $E_p$ energies. For a mass on a spring the total energy $E$ can be written as:
\[
E = E_k + E_p = \frac{mv^2}{2} + \frac{kx^2}{2}. \tag{6.4}
\]

The potential energy stored in a spring is equal to $E_p = \frac{kx^2}{2}$ since the force is given by Hooke’s Law. The displacement of a harmonic oscillator can be written as $x = A_0 \sin(\omega_0 t + \phi_0)$ and the velocity expression then becomes
\[
\nu = \frac{dx}{dt} = A_0 \omega_0 \cos(\omega_0 t + \phi_0).
\]

Substituting the displacement expression (6.3), the velocity expression and $\omega_0^2 = k/m$ to equation (6.4), the total energy can be written as:
\[
E = \frac{1}{2} m A_0^2 \omega_0^2 \left( \cos^2(\omega_0 t + \phi_0) + \sin^2(\omega_0 t + \phi_0) \right) = \frac{1}{2} m A_0^2 \omega_0^2. \tag{6.5}
\]

The total energy $E$ of a harmonic oscillator is directly proportional to the body mass $m$, to the amplitude square $A_0^2$, to the angular frequency square $\omega_0^2$ and doesn’t depend on time $t$.

6.2. DAMPED HARMONIC OSCILLATIONS

Let’s find the law, according to which the coordinate $x$ changes as a function of time $t$, when a body makes the oscillations overcoming the force of friction $F_{fr}$ (fig. 6.2).

For this reason, it is necessary to obtain and solve the differential equation, taking into account the force of friction. In real systems, together with the elastic force $F$ frictional force $F_{fr}$ is acting, which is proportional to the velocity $\nu$ and
its direction is opposite to the velocity \( \mathbf{v} \). This frictional force \( \mathbf{F}_{fr} \) hinders oscillatory process. Therefore: \( F_{fr} \sim -\mathbf{v} \Rightarrow F_{fr} = -r\mathbf{v} \), where \( r \) is the coefficient of proportionality.

![Fig. 6.2. Oscillations of negligible mass \( m \) overcoming the friction force \( F_{fr} \)](image)

In this case one can write the Newton’s second Law of motion as:

\[
m\ddot{x} = \ddot{F} + \ddot{F}_{fr}.
\]

In the projections on horizontal axis OX along which an oscillatory process occurs, the equation can be written as:

\[
ma = -kx - r\mathbf{v}.
\]

Or, in the differential form:

\[
m\frac{d^2x}{dt^2} = -r \frac{dx}{dt} - kx \tag{6.6}
\]

\[
\downarrow
\]

\[
m \frac{d^2x}{dt^2} + r \frac{dx}{dt} + kx = 0
\]

\[
\downarrow
\]

\[
\frac{d^2x}{dt^2} + \frac{r}{m} \frac{dx}{dt} + \frac{k}{m} x = 0.
\]

Taking into account that \( \frac{r}{m} = 2\beta; \frac{k}{m} = \omega_0^2 \), one can obtain:

\[
\frac{d^2x}{dt^2} + 2\beta \frac{dx}{dt} + \omega_0^2 x = 0 \tag{6.7}
\]

The general solution of this DE is the function:

\[
x = A_0 e^{-\beta \cdot t} \sin(\omega t + \varphi_0). \tag{6.8}
\]
The displacement $x$ of the body is going harmonically with frequency $\omega$, but the amplitude $A$ of the oscillations decreases exponentially with time:

$$A = A_0 e^{-\beta t}.$$

The rate of the amplitude decreasing is characterized by a damping factor $\beta$: $\beta = \frac{r}{2m}$, where $r$ is the coefficient of drag. The more damping factor $\beta$ the faster oscillation damping.

$$\omega = \sqrt{\omega_0^2 - \beta^2}$$ is the damping frequency. The oscillation process takes place if $(\omega_0^2 - \beta^2) > 0$. Fig. 6.3 represents the damped harmonic oscillation.

![Fig. 6.3. Displacement $x$ dependence on time $t$ during damped harmonic oscillation](image)

In practice an oscillation damping is characterized by the damping ratio $\delta$. **Damping ratio $\delta$** is the ratio of the amplitudes of two successive oscillations separated by oscillation period $T$:

$$\delta = \frac{A(t)}{A(t + T)}. \quad (6.9)$$

The damping ratio $\delta$ shows how oscillations decay in a system after a disturbance. The damping ratio is also related to the **logarithmic decrement $\lambda$**. The logarithmic decrement $\lambda$ is defined as the natural logarithm of the ratio of any two subsequent amplitudes:

$$\lambda = \ln \delta = \ln \frac{A_0 e^{-\beta t}}{A_0 e^{-\beta (t + T)}} = \ln e^{\beta T} = \beta T. \quad (6.10)$$
6.3. THE FORCED HARMONIC OSCILLATIONS

Let’s consider a damped harmonic oscillator driven by time-dependent external applied force $F_a = F_0 \sin \Omega t$, where $F_0$ is the driving force amplitude. In this case the Newton’s Law can be written as:

$$m \frac{d^2x}{dt^2} = -kx - r \frac{dx}{dt} + F_0 \sin \Omega t.$$  \hspace{1cm} (6.11)

Noting $\omega_0^2 = \kappa/m$, $2\beta = r/m$ and $f_0 = F_0/m$, the second order linear differential equation can be written as follows:

$$\frac{d^2x}{dt^2} + 2\beta \frac{dx}{dt} + \omega_0^2 x = f_0 \sin \Omega t.$$  \hspace{1cm} (6.12)

Fig. 6.4 shows a function, which is a solution of the differential equation (6.12).

![Fig. 6.4. Displacement x dependence on time t during forced harmonic oscillation](image)

The equation solution consists of two parts $X_1$ and $X_2$. The first one corresponds to unsteady oscillations, the second one — to steady oscillations:

$$X = X_1 + X_2.$$  

For the steady oscillations displacement is varied as a sine function with driving force frequency $\Omega$:

$$X_2 = A \sin(\Omega t + \phi).$$  \hspace{1cm} (6.13)

Steady amplitude $A$ depends on frequency of a natural oscillations $\omega_0$, a damping factor $\beta$, driving force characteristics ($f_0$, $\Omega$):

$$A = f_0 / \sqrt{(\omega_0^2 - \Omega^2)^2 + 4\beta^2 \Omega^2}.$$  \hspace{1cm} (6.14)

The amplitude has maximum value when a driving force frequency is determined by equation:

$$\Omega = \Omega_{res} = \sqrt{\omega_0^2 - 2\beta^2}.$$  \hspace{1cm} (6.15)
Mechanical resonance is the tendency of a mechanical system to oscillate at larger amplitude (to absorb more energy) when a driving force frequency $\Omega$ is determined by equation (6.15).

6.4. SUPERPOSITION OF HARMONIC OSCILLATIONS

The superposition of two harmonic (sinusoidal) oscillations $X_1 = A_1 \sin(\omega t + \varphi_1)$ and $X_2 = A_2 \sin(\omega t + \varphi_2)$ with the same frequency $\omega$ gives another harmonic (sinusoidal) oscillation $X = X_1 + X_2$ with the same frequency $\omega$ (but with a different amplitude and phase displacement):

$$X = X_1 + X_2 = A_1 \sin(\omega t + \varphi_1) + A_2 \sin(\omega t + \varphi_2) = A \sin(\omega t + \varphi).$$

(6.16)

The frequency of oscillation being obtained by superposition of harmonic oscillations with the same frequency is equal the frequency of added oscillations. Resulting amplitude $A$ depends on $A_1$ and $A_2$ amplitudes and initial phase displacement $\varphi_1$ and $\varphi_2$:

$$A = \sqrt{A_1^2 + A_2^2 + 2A_1A_2 \cos(\varphi_2 - \varphi_1)}.$$  

(6.17)

The initial phase $\varphi$ of resulting oscillation is determined by the following equation:

$$\tan \varphi = \frac{A_1 \sin \varphi_1 + A_2 \sin \varphi_2}{A_1 \cos \varphi_1 + A_2 \cos \varphi_2}.$$  

(6.18)

By superposing harmonic (sinusoidal) oscillations $X_1 = A_1 \sin(\omega_1 t + \varphi_1)$ and $X_2 = A_2 \sin(\omega_2 t + \varphi_2)$ with different frequencies $\omega_1$ and $\omega_2$ one can obtain a periodic function (fig. 6.5). Periodic function of the time $x(t) = x(t + T)$ is a function which repeat their course after a definite interval of time.

Fig. 6.5. The superposition of two harmonic vibrations $X_1$ and $X_2$ with different frequencies $\omega_1$ and $\omega_2$ (periods $T_1$ and $T_2$)
6.5. THE FOURIER THEOREM

According to the Fourier theorem: any complex periodic motion \( x(t) = x(t + T) \) (fig. 6.6) with period \( T \) can be represented as the sum of harmonic oscillations (harmonics) whose frequencies are multiples of the fundamental frequency \( \omega \) of the considered periodic process: \( \omega_k = k \cdot \omega \) \((k = 1, 2, 3 \ldots)\).

Each of sine and cosine term has specific amplitude and phase coefficients known as Fourier coefficients. In the brief description the Fourier theorem can be written as:

\[
x(t) = A_0 + \sum_{k=1}^{\infty} A_k \sin(k \omega t + \varphi_k),
\]

where \( A_0 \) is a constant component periodical function (in many cases it can be equal to zero); \( A_k \sin(k \omega t + \varphi_k) \) are harmonic components with amplitude \( A_k \), angular frequency \( k \omega \) and initial phase \( \varphi_k \). The first term (at \( k = 1 \)) describes harmonic of the fundamental frequency \( \omega \). Others components (at \( k = 2, 3, \ldots \)) are called overtones: \( 2\omega, 3\omega, 4\omega, \ldots \).

![Fig. 6.6. The complex oscillation with a constant component \( A_0 \)](image)

Frequency range from \( \omega \) to \( \omega_k = k \omega \) is called the frequency spectrum of complex oscillation. The harmonic spectrum is a set of frequencies and amplitudes of the harmonics corresponding to the complex oscillation (fig. 6.7).

![Fig. 6.7. The harmonic spectrum of the complex oscillation](image)
A standard result of Fourier analysis is that a function has a harmonic spectrum if and only if it is periodic.

### 6.6. Mechanical Waves. The Wave Equation

**Mechanical wave** is a process of propagation of mechanical oscillations in an elastic medium. The wave is characterized by the transfer of energy without the transfer of matter. Transverse waves are those with vibrations perpendicular to the direction of the propagation of the wave; examples include waves on a string, and electromagnetic waves. Longitudinal waves are those with vibrations parallel to the direction of the propagation of the wave; examples include most sound waves. Mathematically, the most basic wave is the harmonic wave (or sinusoid), described by the equation:

\[
S = A \sin \omega \left( t - \frac{x}{\upsilon} \right),
\]

where \( A \) is the amplitude of the wave (the maximum distance from the highest point of the disturbance in the medium (the crest) to the equilibrium point during one wave cycle); \( x \) is the space coordinate; \( t \) is the time; \( \omega \) is the angular frequency, and \( \upsilon \) is the wave velocity.

The basic characteristics of a wave are:
- the wavelength \( \lambda \) is the distance between two sequential crests (or troughs), and generally is measured in meters:
  \[
  \lambda = \upsilon T;
  \]
- the amplitude \( A \) is the maximum distance from the highest point of the disturbance in the medium (the crest) to the equilibrium point during one wave cycle;
- the period \( T \) is the time for one complete cycle of an oscillation of a wave and is measured in seconds;
- the frequency \( \nu \) is the number of periods per unit time (per second) and is measured in Hertz. Period \( T \) and frequency \( \nu \) are related by \( \nu = \frac{1}{T} \);
- the angular frequency \( \omega \) represents the frequency in radians per second.

It is related to the frequency by \( \omega = 2\pi \nu = \frac{2\pi}{T} \).

Thus the wave equation can be written as:

\[
S = A \sin 2\pi \left( \frac{t}{T} - \frac{x}{\lambda} \right).
\]

Wave propagates in a medium with a velocity \( \upsilon \) which is determined by the elastic modulus \( E \) and medium density \( \rho \):

\[
\upsilon = \sqrt{\frac{E}{\rho}}.
\]
The wave process is associated with transfer of the energy $E$ in space. The process of energy transfer is characterized by energy flux $\Phi$. In the case of uniform transfer of energy $E$ the energy flux $\Phi$ can be written as:

$$\Phi = \frac{E}{t}.$$  \hfill (6.23)

In general, the energy flux $\Phi$ is a derivative of energy with respect to time:

$$\Phi = \frac{dE}{dt}.$$  

$\Phi$ is the total rate of energy transfer and is measured in Watt ($W = J \cdot s^{-1}$).

Wave intensity $I$ (density of energy flux) is the rate of energy transfer $\Phi$ per unit area $S$ perpendicular to the wave propagation. Or other words, wave intensity $I$ is the energy transmitted per unit time through a unit area perpendicular to the wave propagation direction:

$$I = \frac{\Phi}{S} = \frac{E}{S t}.$$  \hfill (6.24)

The unit of intensity $I$ is $J \cdot m^{-2} \cdot s^{-1} = W \cdot m^{-2}$.

### 6.7. THE DOPPLER EFFECT

The Doppler effect is observed whenever the source of waves is moving with respect to an observer. A change in the observed frequency of a wave, as of sound or light, occurring when the source and observer are in motion relative to each other, with the frequency increasing when the source and observer approach each other and decreasing when they move apart. The relative changes in frequency can be explained as follows. When the source of the waves is moving toward the observer, each successive wave crest is emitted from a position closer to the observer than the previous wave. Therefore each wave takes slightly less time to reach the observer than the previous wave. Therefore the time between the arrival of successive wave crests at the observer is reduced, causing an increase in the frequency. While they are travelling, the distance between successive wave fronts is reduced; so the waves «bunch together». Conversely, if the source of waves is moving away from the observer, each wave is emitted from a position farther from the observer than the previous wave, so the arrival time between successive waves is increased, reducing the frequency. The distance between successive wave fronts is increased, so the waves «spread out».

In general, when the source and detector of wave are in motion with the velocity $v_s$ and $v_d$ correspondingly, the perceived wave frequency $v_d$ by detector can be written as:

$$v_d = \frac{v_w \pm v_d}{v_w \mp v_s} v_s,$$  \hfill (6.25)

where $v_w$ is the wave velocity in a medium.
Doppler effect is used to measure changes in sound waves therefore it is used to diagnose conditions related to circulation and blood flow. The Doppler ultrasound can actually measure how fast or slow blood is moving, which can indicate a circulatory problem. Blood clots can be found using Doppler ultrasound because the ultrasound will be able to detect slower blood flow or a lack of blood flow where the clot is located. Doppler ultrasound can also be used to identify narrowed arteries, plaque buildup in the blood vessels, or blocked arteries.

The fig. 6.8 shows a Doppler transducer placed on the skin and aimed at an angle, θ, towards a blood vessel, which contains blood flowing with a velocity of $v_b$ m/s, at any instant. Consideration of the Doppler effect for red cells is complicated by the fact that the red cell first acts as a moving receiver, and then as a moving source. The transducer emits ultrasound waves of frequency, $v_s$, and echoes generated by moving reflectors in the blood, e. g. red blood cells, have a frequency, $v_r$. The difference between these two frequencies, $\Delta v$ (so called the Doppler frequency), is related to the velocity of the flowing reflectors through the following equation:

$$v_r - v_s = \Delta v = \frac{2v_s v_b \cos \Theta}{v},$$

where $v$ is speed of sound in blood.

![Fig. 6.8. Illustration of blood flow detection using the Doppler effect with ultrasound waves](image)

The Doppler frequency $\Delta v$ is directly proportional to the velocity of the red blood cells multiplied by the cosine of the angle between the ultrasound beam and the blood flow. The Doppler frequency can be either positive or negative depending upon the direction of the blood flow. The equation for the Doppler frequency $\Delta v$ shows that one can directly relate the Doppler frequency to blood velocity. However, the equation also makes clear the major disadvantage of the technique as a method of calculating blood flow. One need to know the angle between the blood flow and the beam of ultrasound. An advantage of Doppler ultrasound is that it can be used to measure blood flow within the vessels without invasive procedures.
Questions:
1. Write harmonic oscillations equation and formula for displacement.
2. Write formula for displacement in the case of damped harmonic oscillations. What is the logarithmic decrement? What is the relation between a logarithmic decrement and a damping factor?
3. What does frequency of forced harmonic oscillations depend on?
4. Explain a phenomenon of resonance. What are the conditions of the resonance appearance?
5. Formulate the Fourier theorem. What is the harmonic spectrum of the complex oscillation?
6. Write the mechanical wave equation. Explain meaning of values in this equation.
7. What does the wave intensity depend on?
8. Explain the Doppler effect. How is it used for blood flow detection?

Chapter 7. ACOUSTIC

Acoustics is the field of physics, where the elastic vibrations and waves from the lowest frequencies to the highest ones ($10^{12}$–$10^{13}$ Hz) are investigated.

7.1. PHYSICAL AND PHYSIOLOGICAL CHARACTERISTICS OF SOUND

Sound is a longitudinal mechanical wave with frequency from 16 Hz to 20 000 Hz, which propagates in elastic media (solid, liquid and gas). As the sound waves travel through a media, the particles of the media start to vibrate producing the changes of density and pressure along the direction of propagation of the waves. The pulses of increased and decreased pressure reach ear, vibrating the eardrum. The ear sends nerve signals to the brain about the vibrations, and the brain interprets the signals as sounds.

Acoustic waves are divided into three categories that cover different frequency ranges. Frequency range from 16 Hz to 20 000 Hz corresponds to the sound. **Infrasound** has a frequency less than 16 Hz. **Ultrasound** has a frequency greater than 20 000 Hz.

The velocity of the sound wave depends on the medium properties (modulus of elasticity $E$, density $\rho$ and temperature $T$) through which the sound is traveling:

$$v = \sqrt{\frac{E}{\rho}}. \quad (7.1)$$

Sound velocity in the air is 340 m/s, in water and soft tissues — 1500 m/s, in bones 3000–6000 m/s.

Sounds are divided into tones, noise and sonic booms. **Simple or pure tone** is a sound, whose source makes harmonic oscillations (for example, a tuning fork). Simple tone has the only frequency.
Complex tone is the sound wave, whose source makes periodic non-harmonic oscillations (for example, musical sounds, vowel sounds of speech apparatus). Any complex tone can be decomposed into simple tones by Fourier’s theorem. The tone of lowest frequency is called the fundamental tone. The other tones are called harmonics (overtones) and the frequency of each is a whole-number multiple of the fundamental frequency.

![Complex tone spectrum](a)
![Noisy spectrum](b)

Fig. 7.1. Complex tone has a line spectrum (a). The noise spectrum is continuous (b)

The noise is a combination of disorderly varying complex tones of all frequencies (machine vibration, squeaking, rustling sound, consonant sounds of speech). The noise spectrum is continuous (fig. 7.1, b).

Sonic boom is a short-term sound effect of high intensity (explosions, thunder).

There is correlation between physics of acoustical stimuli and hearing sensations. Sound has physical characteristics as mechanical wave and physiological characteristics as human perception. Physical and physiological characteristics of sound are given in table 7.1.

<table>
<thead>
<tr>
<th>Physical characteristics</th>
<th>Physiological characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intensity, $I$, W/m², $I = \frac{E}{S \cdot t}$</td>
<td>Loudness, $E$, phones</td>
</tr>
<tr>
<td>Frequency, $\nu$, Hz</td>
<td>Pitch</td>
</tr>
<tr>
<td>Harmonic spectrum</td>
<td>Timbre</td>
</tr>
</tbody>
</table>
The \textit{intensity} $I$ of the sound wave has been defined as the energy $E$ which transmitted through a unit area $S$ perpendicular to the direction of sound propagation per unit time $t$:

$$I = \frac{E}{S \cdot t}.$$  \hspace{1cm} (7.2)

Unit of the intensity is W/m$^2$.

\textbf{Loudness $E$} is the psychological aspect of sound related to perceived intensity amplitude. Loudness is measured in phones. Relationship between the loudness and the intensity is logarithmic.

It is the frequency of the wave that determines its \textit{pitch}. Pitch refers to whether a sound is low or high.

\textit{Timbre} is determined by the harmonic spectrum of the sound. Timber characterizes tone quality and color.

\textbf{7.2. Audition diagram}

In the frequency range from 16 Hz to 20 000 Hz action of the acoustic waves leads to the formation of auditory sensations. Sensitivity of the ear varies with the frequency of a sound.

The minimum intensity that causes the sensation of sound is called \textit{the threshold of hearing} for a given frequency. The ear has a maximum sensitivity in the frequency range 1000–3000 Hz. At frequency $\nu = 1000$ Hz the threshold value of sound intensity is minimal ($I_0 = I_{\text{min}}$) and is equal to $I_0 = 10^{-12}$ W/m$^2$. At other frequencies, the intensity at the threshold of audibility is higher.

The maximum value of sound intensity above which pain occurs is called \textit{the threshold of pain}. The threshold of pain at frequency of 1000 Hz ($\nu = 1000$ Hz) is the lowest and is equal to $I_{\text{max}} = 10$ W/m$^2$. When the sound intensity greater than the threshold of pain the ear damage occurs. At other frequencies the pain threshold is higher.

Thus, one can hear sounds in the range of sound intensities: $I = 10^{-12} - 10^1$ W/m$^2$. Because one can hear sound intensities over an enormous range, it is convenient to use a logarithmic scale, where the \textit{intensity level} $L$ is defined by the equation:

$$L = n \lg \frac{I}{I_0}.$$  \hspace{1cm} (7.3)

The intensity level $L$ is measured in \textit{decibels (dB)} if $n = 10$ and in \textit{bels} if $n = 1$.

The relation between the \textit{intensity level} $L$ and the \textit{hearing threshold} $I_0 = 10^{-12}$ W/m$^2$ is the following:

$$L = 10 \lg \frac{10^{-12}}{10^{-12}} = 0 \text{ dB}.$$
The relation between the intensity level $L$ and the threshold of pain $I_{\text{max}} = 10 \text{ W/m}^2$ is the following:

$$L = 10 \log \frac{10}{10^{-12}} = 10 \log 10^{13} = 130 \text{ dB}.$$ 

At audible frequencies (16–20 000 Hz), the area between the threshold of hearing and the threshold of pain is called the auditory field. Figure 7.2 shows an audition diagram. Audition diagram is dependence of the intensity $I$ or level of intensity $L$ on the sound frequency $\nu$.

Fig. 7.2. Audition diagram

7.3. Weber–Fechner Law

*Weber–Fechner’s Law* represents the relation between the magnitude of physical stimulus and the magnitude of psychological perception. *Weber–Fechner Law* states that: if a sound stimulus (intensity $I$) varies as a geometric progression, the corresponding perception (loudness $E$) is altered in an arithmetic progression, i.e. physiological response (loudness $E$) does not increase directly proportional to intensity $I$ of the stimulus, but much weaker — is directly proportional to the decimal logarithm of $I$.

If the sound intensity is low, then a slight increase of intensity (the value of $\Delta I$) leads to a significant rise of the loudness (the value $\Delta E_1$ in fig. 7.3). As soon as the sound intensity is slightly higher than the threshold, a sound sensation appears. If the sound intensity is high, its increase by the same value $\Delta I$ results in a small rise of the loudness (the value $\Delta E_2$ in fig. 7.3).

**According to the Weber–Fechner Law loudness level (often called loudness) $E$** is related with a level of intensity $L$ by formula:

$$E = kL.$$  \hspace{1cm} (7.4)

$k$ is the coefficient of proportionality that depends on the frequency and sound intensity.
Since the hearing threshold depends on the frequency, the loudness level also varies with frequency. Therefore, there is a special unit called Phon to measure the loudness level. Loudness level $E$ in Phones and intensity level $L$ in dBs are equivalent only at $\nu = 1000$ Hz ($k = 1$), reflecting the fact that intensity and loudness do not map isomorphically, as loudness also depends on frequency. For other frequencies ratio between the intensity level and loudness level can be determined using the equal loudness curves (fig. 7.4).

![Fig. 7.3. Relationship between loudness and intensity of sound](image)

**7.4. REFLECTION AND ABSORPTION OF THE ACOUSTIC WAVES**

Going through the media interface, the acoustic waves are reflected and refracted by the laws similar to the reflection and refraction of light (fig. 7.5).

The reflection coefficient ($R$) is determined by a ratio of the intensity of reflected wave to intensity of incident one: $R = I_{ref} / I_{inc}$. Its value depends on
density of the first medium $\rho_1$ and density of the second one $\rho_2$ as well as on acoustic wave velocity in the first medium $v_1$ and in the second one $v_2$.

\[ Z = \rho v, \quad (7.5) \]

where $\rho$ is the density of the medium, and $v$ is the sound velocity in that medium. Taking into account that $v = \sqrt{E/\rho}$, acoustic impedance $Z$ can be defined as:

\[ Z = \sqrt{E \cdot \rho}. \quad (7.6) \]

For «normal» incidence the reflection coefficient $R$ is equal to:

\[ R = \left( \frac{Z_2 - Z_1}{Z_2 + Z_1} \right)^2 \quad \text{or} \quad R = \left( \frac{\rho_2 v_2 - \rho_1 v_1}{\rho_2 v_2 + \rho_1 v_1} \right)^2. \quad (7.7) \]

The amount of reflection from a boundary is determined by the acoustic impedance mismatch. The greater the acoustic impedance mismatch between media, the greater the amount of reflection.

When the sound wave propagates through the medium, its intensity $I$ decreases due to absorption and diffusion. The law of the wave intensity decrease is given by formula:

\[ I = I_0 e^{-kx}, \quad (7.8) \]

where $I$ is the wave intensity after passing of distance $x$ in a medium; $I_0$ is the incident wave intensity; $k$ is the absorption coefficient.

The absorption coefficient $k$ increases in all media with acoustic wave frequency increase.
7.5. ULTRASOUND IN DIAGNOSTIC AND THERAPEUTIC APPLICATIONS

Diagnostic ultrasonography is an ultrasound-based diagnostic imaging technique used to visualize subcutaneous body structures including tendons, muscles, joints, vessels and internal organs for possible pathology or lesions. It is possible to perform both diagnosis and therapeutic procedures, using ultrasound to guide interventional procedures (for instance biopsies or drainage of fluid collections).

Ultrasonography (sonography) uses a probe containing one or more acoustic transducers to send pulses of sound into a material. Whenever a sound wave encounters a material with a different density (acoustical impedance), part of the sound wave is reflected back to the probe and is detected as an echo. The time it takes for the echo to travel back to the probe is measured and used to calculate the depth of the tissue interface causing the echo. The greater the difference between acoustic impedances, the larger the echo is. If the pulse hits gases or solids, the density difference is so great that most of the acoustic energy is reflected and it becomes impossible to see deeper. This is a reason why it is impossible to recognize tissues under hollow organs. The frequencies used for medical imaging are generally in the range of 0.5 to 15 MHz. Higher frequencies have a correspondingly smaller wavelength, and can be used to make sonograms with smaller details with high spatial resolution. However, the attenuation of the sound wave is increased at higher frequencies, so in order to have better penetration of deeper tissues, a lower frequency is used. The choice of frequency is a trade-off between spatial resolution of the image and imaging depth: lower frequencies produce less resolution but image deeper into the body. Sonography is effective for imaging soft tissues of the body. Superficial structures such as muscles, tendons, testes, breast and the neonatal brain are imaged at a higher frequency, which provides better axial and lateral resolution. Deeper structures such as liver and kidney are imaged at a lower frequency with lower axial and lateral resolution but greater penetration.
The speed of sound is different in different materials. However, the sonographic instrument assumes that the acoustic velocity is constant at 1540 m/s. An effect of this assumption is that in a real body with non-uniform tissues, the beam becomes somewhat de-focused and image resolution is reduced.

Four different modes of ultrasound are used in medical imaging. These are:

- **A-mode**: A-mode is the simplest type of ultrasound. A single transducer scans a line through the body with the echoes plotted on screen as a function of depth. Therapeutic ultrasound aimed at a specific tumor or calculus is also A-mode, to allow for pinpoint accurate focus of the destructive wave energy;

- **B-mode**: in B-mode ultrasound, a linear array of transducers simultaneously scans a plane through the body that can be viewed as a two-dimensional image on screen;

- **M-mode**: M stands for motion. In m-mode a rapid sequence of A-mode scans whose images follow each other in sequence on screen in horizontal direction enables doctors to see and measure range of motion, as the organ boundaries that produce reflections move relative to the probe;

- **Doppler mode**: this mode makes use of the Doppler effect in measuring and visualizing blood flow. By calculating the frequency shift of a particular sample volume, for example a jet of blood flow over a heart valve, its speed and direction can be determined and visualised. This is particularly useful in cardiovascular studies (sonography of the vascular system and heart) and essential in many areas such as determining reverse blood flow in the liver vasculature in portal hypertension. Most modern sonographic machines use pulsed Doppler to measure velocity.

Most ultrasound procedures are done using a transducer on the surface of the body, but improved diagnostic confidence is often possible if a transducer can be placed inside the body. For this purpose, specialty transducers, including endovaginal, endorectal, and transesophageal transducers are commonly employed. Very small transducers can be mounted on small diameter catheters and placed into blood vessels to image the walls and disease of those vessels.

**Therapeutic applications** use ultrasound to bring heat or agitation into the body. Therefore much higher energies are used than in diagnostic ultrasound. For therapeutic applications maximum acceptable intensity is

\[ I_{\text{max}} \approx \frac{1}{s \cdot m^2} \]  

In many cases the range of frequencies used are also very different (0.8–3 MHz).

Ultrasound may be used to clean teeth in dental hygiene. Ultrasound sources may be used to generate regional heating and mechanical changes in biological tissue, e.g. in occupational therapy, physical therapy and cancer treatment. However the use of ultrasound in the treatment of musculoskeletal conditions has fallen out of favor. Focused ultrasound may be used to generate highly localized heating to treat cysts and tumors (benign or malignant). This is
known as Focused Ultrasound Surgery (FUS) or High Intensity Focused Ultrasound (HIFU). These procedures generally use lower frequencies than medical diagnostic ultrasound (from 250 kHz to 2000 kHz), but significantly higher energies. HIFU treatment is often guided by magnetic resonance imaging. Focused ultrasound may be used to break up kidney stones by lithotripsy. Ultrasound may be used for cataract treatment by phacoemulsification. Additional physiological effects of low-intensity ultrasound have recently been discovered, e.g. its ability to stimulate bone-growth and its potential to disrupt the blood-brain barrier for drug delivery.

As with all imaging modalities, ultrasonography has in list of positive and negative attributes. Strengths of this method are as follows. It images muscle, soft tissue, and bone surfaces very well and is particularly useful for delineating the interfaces between solid and fluid-filled spaces. It renders «live» images, where the operator can dynamically select the most useful section for diagnosing and documenting changes, often enabling rapid diagnoses. Live images also allow for ultrasound-guided biopsies or injections, which can be cumbersome with other imaging modalities. It shows the structure of organs. It has no known long-term side effects and rarely causes any discomfort to the patient. Equipment is widely available and comparatively flexible. Small, easily carried scanners are available; examinations can be performed at the bedside.

On the other hand, the method has some weaknesses. Sonographic devices have trouble penetrating bone. For example, sonography of the adult brain is very limited though improvements are being made in transcranial ultrasonography. Sonography performs very poorly when there is a gas between the transducer and the organ of interest, due to the extreme large reflection. For example, overlying gas in the gastrointestinal tract often makes ultrasound scanning of the pancreas difficult, and lung imaging is not possible (apart from demarcating pleural effusions). Even in the absence of bone or air, the depth penetration of ultrasound may be limited depending on the ultrasound frequency. Consequently, there might be difficulties imaging structures deep in the body, especially in obese patients. The method is operator-dependent. A high level of skill and experience is needed to acquire good-quality images and make accurate diagnoses. There is no scout image as there is with computer tomography and magnetic resonance imaging. Once an image has been acquired there is no exact way to tell which part of the body was imaged.

Questions:
1. What are the sound waves? What determines the sound velocity?
2. What physical characteristics determine physiological characteristics of sound?
3. What is a threshold of hearing? What does it depend on?
4. Formulate the Weber-Fechner Law. What is the relation between the intensity level and the loudness? What are the units for intensity level and loudness?
5. What is acoustic impedance? What does the reflection coefficient depend on?
6. Write the absorption law for acoustic waves in medium.
7. Describe ultrasound production and registration methods. Explain physical basics for the methods.
8. What is the essence of ultrasound diagnostics methods?
9. Why different frequencies are used for ultrasound diagnostics of different organs?

Chapter 8. PROPERTIES OF LIQUIDS. SURFACE FENOMENA

8.1. SURFACE TENSION

At the free surface that forms between a gas and a liquid, a fluid property called surface tension becomes important. Surface tension is a property of liquids that results from the tendency of liquids to minimize their surface area. By observation, it is known that the surface of a liquid tends to contract to the smallest possible area, behaving as though its surface were a stretched elastic membrane. For example, the surface tension of the water allows the insect to walk on the water without sinking (fig. 8.1).

![Fig. 8.1. There are the indentations in the water surface made by the feet of an insect known as a water strider, because it can stride or walk on the surface as a person can walk on a trampoline](image)

Figure 8.2 illustrates the molecular basis for surface tension by considering the attractive forces that molecules in a liquid exert on one another. Part a shows a molecule within the bulk liquid, so that it is surrounded on all sides by other molecules. The surrounding molecules attract the central molecule equally in all directions, leading to a zero net force. In contrast, part b shows a molecule in the surface. Since there are no molecules of the liquid above the surface, this molecule experiences a net attractive force $F$ pointing toward the liquid interior. This net attractive force $F$ causes the liquid surface to contract toward the interior until repulsive collisional forces from the other molecules halt the contraction at the point when the surface area is a minimum. This resultant force $F$ on molecules near the surface is inward, tending to make the surface area as small as possible.
Fig. 8.2. \( a \) — a molecule within the bulk liquid is surrounded on all sides by other molecules, which attract it equally in all directions, leading to a zero net force; \( b \) — a molecule in the surface experiences a net attractive force pointing toward the liquid interior, because there are no molecules of the liquid above the surface.

Because the molecule at the interface has an unbalanced attractive force, a certain amount of work is required to move a molecule from within the body of the liquid to interface. Thus, in order to increase the surface area of liquid by an amount \( \Delta S \), a quantity of work, \( A = \sigma \Delta S \), is required. This work \( A \) is stored as excess potential energy \( W_s \) of the molecules at the surface. Every molecule near the surface posses more potential energy as compared to the molecules in the bulk of the liquid. So if the surface area \( S \) of liquid increases, the surface potential energy \( W_s \) of the liquid also increases:

\[
W_s = \sigma S,
\]

where coefficient of proportionality \( \sigma \) depends on liquid properties and temperature and is called surface tension. The surface tension \( \sigma \) of a liquid can be defined as the excess potential energy per the unit area of the surface. Surface tension \( \sigma \) is measured in SI as joules per square meter (\( J/m^2 \)). The surface potential energy \( W_s \) is minimum for the smallest surface area (since mechanical systems try to find a state of minimum potential energy), hence a droplet of water tends to take a spherical shape, which has the minimum surface area for a given volume.

Let’s consider the apparatus (fig. 8.3) consisting of a U-shaped wire frame, on which is mounted a wire that can slide with negligible friction. The frame and sliding wire contain a thin film of liquid. Because surface tension causes the liquid surface to contract, a force \( F \) is needed to move the slider to the right and extend the surface. The surface tension \( \sigma \) can also be defined as the tensile force \( F_s \) exerted parallel to the surface of a liquid divided by the unit length \( l \) of the line over which the force acts

\[
\sigma = \frac{F_s}{l}.
\]

The surface tension is therefore measured in \( N/m \).

For the specific case illustrated in figure 8.3, there is an upper surface and a lower surface, as the blow-up drawing indicates. Thus, the force \( F \) acts along a total length of \( L = 2l \), where \( l \) is the length of the slider.
For water at 20 °C, \( \sigma = 0.073 \text{ N/m} \), while at 100 °C, \( \sigma = 0.059 \text{ N/m} \), as increasing temperature decreases the surface tension (the attractive force between water molecules). Soapy water has a surface tension reduced by a factor of 4 or 5. The fluid in our lung has a surface tension, of 0.05 N/m, a bit too large to allow the lungs to expand freely. In fact the lungs secretes a substance that can reduce the surface tension of that liquid by a factor of 10 when the lungs are fully expanded. Table 8.1 shows the surface tension of various liquids.

<table>
<thead>
<tr>
<th>Liquid</th>
<th>Temperature, °C</th>
<th>Surface tension, ( \sigma ), N/m</th>
</tr>
</thead>
<tbody>
<tr>
<td>Water</td>
<td>0</td>
<td>0.0756</td>
</tr>
<tr>
<td>Water</td>
<td>20</td>
<td>0.0725</td>
</tr>
<tr>
<td>Water</td>
<td>100</td>
<td>0.0589</td>
</tr>
<tr>
<td>Ethanol</td>
<td>20</td>
<td>0.0223</td>
</tr>
<tr>
<td>Mercury</td>
<td>20</td>
<td>0.47</td>
</tr>
<tr>
<td>Milk</td>
<td>20</td>
<td>0.050</td>
</tr>
<tr>
<td>Urine</td>
<td>20</td>
<td>0.066</td>
</tr>
</tbody>
</table>

Measurement of surface tension is important for diagnosis. For example, the surface tension of urine in normal state is 66 mN/m and in disease state is 56 mN/m.

### 8.2. PHENOMENON OF THE WETTING AND NONWETTING OF SOLIDS BY LIQUIDS

It was shown that surface tension arises because of the intermolecular forces of attraction that molecules in a liquid exert on one another. These forces, which are between like molecules, are called cohesive forces. A liquid, however,
is often in contact with a solid surface, such as glass. Then additional forces of attraction come into play. They occur between molecules of the liquid and molecules of the solid surface and, being between unlike molecules, are called adhesive forces.

The free surface of a liquid will form a curved surface when it comes in contact with a solid. Figure 8.4 shows two glass tubes, one — containing mercury (b) and the other — water (a). The free surfaces are curved, convex for mercury, and concave for water. Such a concave or convex surface shape is known as \textit{a meniscus}. The angle between the solid surface and the tangent to the liquid surface at the point of contact is called the \textit{angle of contact}, \( \theta \) (fig. 8.4).

![Fig. 8.4. The contact angle \( \theta \) of liquid in a vertical tube: 
\( a \) — phenomena of the wetting; \( b \) — phenomena of the nonwetting](image)

For liquids where the angle \( 0 \leq \theta < 90^\circ \) (for example, water on glass), the liquid is said \textit{to wet the surface} (fig. 8.5, a). For liquids where the angle \( 90^\circ < \theta \leq 180^\circ \) (for example, mercury on glass), the liquid is said \textit{not to wet the surface} (fig. 8.5, b). In the case of a liquid that forms a uniform film (i. e., where \( \theta = 0^\circ \)), the solid is said to be \textit{completely wetted by the liquid}, or that the liquid wets the solid.

![Fig. 8.5. Phenomena of the wetting (a) and nonwetting (b) of solids by liquids](image)

Water wets glass because the attractive forces between the water molecules and the glass molecules (\( F_{ls} \)) exceed the forces between water molecules (\( F_{ll} \)). In this case the cohesive force is weaker than the adhesive force (\( F_{ls} > F_{ll} \)). The reverse holds true for mercury: the cohesive force is larger than the adhesive force (\( F_{ls} < F_{ll} \)).

Surfaces which are wetting by liquids are called hydrophilic surfaces and surfaces which are nonwetting by liquids are called hydrophobic ones.
8.3. LAPLACE PRESSURE

In a sufficiently narrow tube of circular cross-section with radius \( R \), the interface between liquid and air forms a meniscus that is a portion of the surface of a sphere with radius \( r \) (fig. 8.6) due to phenomena of the wetting and nonwetting.

\[ \Delta P = \sigma \left( \frac{1}{R_1} + \frac{1}{R_2} \right), \]

where \( R_1 \) and \( R_2 \) are the curvature radii of curved liquid surface in two mutually perpendicular planes. If the surface of liquid is spherical, then \( R_1 = R_2 = r \) and the pressure difference \( \Delta P \) is:

\[ \Delta P = \frac{2\sigma}{r}, \]

where \( \Delta P \) is Laplace pressure; \( \sigma \) is the surface tension; \( r \) is the radius of curvature. The pressure difference \( \Delta P \) is proportional to the surface tension \( \sigma \) and inversely proportional to the effective radius \( r \) of the interface, it also depends on the contact (wetting) angle \( \theta \) of the liquid on the surface of the capillary. Laplace pressure arises from the surface tension of a liquid at its interface with a gas in a solid container. The pressure excess \( \Delta P \) is directed to the center of curvature.

In medicine formula for Laplace pressure is used in the context of respiratory physiology, in particular alveoli in the lung, where a single alveolus is modeled as being a perfect sphere. In this context, the pressure differential is a force pushing inwards on the surface of the alveolus. The Laplace law states that there is an inverse relationship between surface tension \( \sigma \) and alveolar
radius $r$. It follows from this that a small alveolus will experience a greater inward force than a large alveolus, if their surface tensions are equal. In that case, if both alveoli are connected to the same airway, the small alveolus will be more likely to collapse, expelling its contents into the large alveolus. This explains why the presence of surfactant lining the alveoli is of vital importance. Surfactant reduces the surface tension on all alveoli, but its effect is greater on small alveoli than on large alveoli. Thus, surfactant compensates for the size differences between alveoli, and ensures that smaller alveoli do not collapse.

An air embolism is the result of air bubbles forming in the vasculature and causing ischemia to tissues, resulting in brain damage or death. The air emboli can be classified as either arterial or venous. Venous emboli occur when air enters the systemic circulation and is transported to the lung via the pulmonary artery. This can cause pulmonary hypertension, or the emboli can travel to the heart and lead to cardiac failure. Arterial gas emboli can result from the over expansion of the lungs due to decompression barotraumas or cardiac bypass surgery.

Once a bubble has entered the bloodstream, let’s assume on the venous side, if its size is small enough not to directly occlude a vessel it will be transported by the blood and follow the normal circulation into the right heart and then the lungs. The bubble will travel at the velocity of blood provided it is small compared to the vessel cross-sectional area, with velocities from 0.03 cm/s in the capillaries to 40 cm/s in the aorta and 15 cm/s in the vena cavae.

If the air bubble enters the small vessel whose diameter is smaller than the diameter of the bubble, the menisci of bubble are deformed under the effect of the dynamic pressure of flowing blood (fig. 8.7).

![Fig. 8.7. Bubble behavior in the bloodstream](image)

It leads to the fact that the radii of curvature of the menisci will be different ($r_1 < r_2$). Excess Laplace pressures across the menisci will be different ($\Delta P_1 > \Delta P_2$) because of different radii of curvature of the menisci ($r_1 < r_2$). Laplace pressures are directed towards each other and resultant force applied to the bubble is directed against the blood flow, which can result in blocking blood flow. These bubbles can cause blockage in cerebral or cardiac vessels and cause life threatening problems.
8.4. Capillarity

Another phenomenon due to surface tension is capillarity. When a clean glass tube of radius $R$ is inserted into a dish of water, the water will rise inside the tube a distance $h$ above the surface (fig. 8.8, $a$). This happens because the attraction between glass and water molecules is greater than that between water molecules themselves, producing an upward force. The liquid rises until the weight of the liquid column balances the upward force due to surface tension: $\Delta P = \rho gh$, where $\Delta P$ is the excess Laplace pressure, $\rho$ is the liquid density, $g$ is the gravitational acceleration, $h$ is the height of liquid raising. Thus the height of liquid raising $h$ can be calculated as:

$$h = \frac{\Delta P}{\rho g} = \frac{2\sigma}{\rho g R} = \frac{2\sigma \cos \theta}{\rho g r}.$$  (8.4)

Connection between the radius of curvature $r$ and the capillary radius $R$ has been taken into account in equitation (8.4): $R = r / \cos \theta$.

![Fig. 8.8](image)

Fig. 8.8. $a$ — water rises in a glass capillary due to the surface tension of the water and the fact that the water wets the glass surface; $b$ — mercury falls in a glass capillary due to the surface tension of the mercury and the fact that the mercury does not wet the glass surface.

The contact length (around the edge) between the top of the liquid column and the tube is proportional to the diameter of the tube, while the weight of the liquid column is proportional to the square of the tube's diameter, so a narrow tube will draw a liquid column higher than a wide tube (fig. 8.9).

Conversely, if a glass tube is inserted into mercury, the level of the liquid in the tube falls (fig. 8.8, $b$). The mercury does not wet the tube and in this case cohesive force dominant over adhesive force, so the column of liquid in equilibrium is below the surrounding level.

Notice that capillary action is what allows plants to grow, as they need to bring up water well above the ground level. It is also what makes absorbent materials like sponges and paper towels absorb so well.
8.5. METHODS OF SURFACE TENSION MEASUREMENT

Pendant drop test
Surface tension can be measured using the pendant drop method. A simple way to form a drop is to allow liquid to flow slowly from the lower end of a vertical tube of small diameter $d$ (fig. 8.10). The surface tension $\sigma$ of the liquid causes the liquid to hang from the tube, forming a pendant. When the drop exceeds a certain size it is no longer stable and detaches itself. The falling liquid is also a drop held together by surface tension. The basic premise of the drop method is that surface tension can be calculated from physical drop characteristics as it forms at the end of a capillary tip of known external radius.

Thus a drop of liquid is suspended from the end of a tube by surface tension. The force due to surface tension is proportional to the length of the boundary between the liquid and the tube:

$$F_t = \sigma l,$$

where $l = 2\pi r = \pi d$ — is the length of the boundary between the liquid and the tube, where $d$ is the tube diameter.
When drop hangs from the end of the tube the net force (the force due to gravity \( F_g = mg \)) plus the force due to surface tension \( F_t = \sigma l \), acting on drop is equal zero. One can write that:

\[
mg = \sigma 2\pi r. \tag{8.5}
\]

This relationship is the basis of a convenient method of measuring surface tension. Thus surface tension can be found by the following formula:

\[
\sigma = \frac{mg}{2\pi r}. \tag{8.6}
\]

**The ring method**

The ring method is one technique by which the surface tension of a liquid can be measured. The advantage of this method is that the surface tension can be determined directly from the force required to pull the ring from a liquid (fig. 8.11). Surface tension for the ring method is the mechanical force \( F = F_t + mg \) necessary to detach a ring of known inner radius \( R_1 \) and outer radius \( R_2 \) at the liquid's surface. Therefore \( F = \sigma l + mg \).

\[
F = F_t + mg.
\]

![Fig. 8.11. The surface tension can be determined directly from the force \( F \) required to pull the ring from a liquid by the ring method. Surface tension for the ring method is equal to the mechanical force necessary to detach a ring of known inner radius \( R_1 \) and outer radius \( R_2 \).](image)

The force \( F \) required to detach the ring at the liquid's surface is measured and related to the liquid's surface tension (\( \sigma \)). Thus surface tension can be found from the equation:

\[
\sigma = \frac{F - mg}{l} = \frac{F - mg}{2\pi R_1 + 2\pi R_2}. \tag{8.7}
\]

**Method of maximum bubble pressure**

The method of maximum bubble pressure is based on the maximum pressure in a capillary or a maximum pressure difference between two capillaries of different radii necessary to produce and detach a bubble from
the capillary tip immersed in test solution. Maximum bubble pressure measuring apparatus for surface tension is shown in fig. 8.12.

Under pressure $P_{\text{atm}} - P_1 = \Delta P$ a bubble from the capillary tip immersed in test solution is produced. The pressure drop $\Delta P$ is created in the system by opening a tap and is measured by the manometer. Thus, $\Delta P = \rho gh$. The pressure drop $\Delta P$ and Laplace pressure are related as:

$$\Delta p = \frac{2\sigma}{r}.$$

Thus, one can write:

$$\rho gh = \frac{2\sigma}{r}. \quad (8.8)$$

The same procedure can be carried out with the reference liquid, for example distilled water, and obtain

$$\rho gh_0 = \frac{2\sigma_0}{r}. \quad (8.8a)$$

![Fig. 8.12. Maximum bubble pressure measuring apparatus for surface tension: 1 — is the tube containing test liquid; 2 — is the capillary tip immersed in test liquid; 3 — is the corked jar containing water; 4 — is the tap; 5 — is the manometer](image)
The attitude of these two equations (8.7) and (8.7a) is equal to:

\[
\frac{h}{h_0} = \frac{\sigma}{\sigma_0}.
\]

Thus equation for surface tension of test liquid is obtained:

\[
\sigma = \sigma_0 \frac{h}{h_0}.
\]

**(Questions):**
1. What is the surface tension energy? What does the surface energy depend on?
2. What is the physical meaning of surface tension coefficient? What does it depend on? What are its units?
3. Give the definition of surface tension force. What are the units of surface tension force?
4. Write Laplace formula for excess pressure under curved surface.
5. Explain the wetting and nonwetting phenomena. What is the contact (wetting) angle?
6. How to calculate the height of liquid raising in capillary?
7. What is the embolism? What are the conditions of its appearance?
8. Describe the surface tension coefficient determination methods.

**Chapter 9. BIOPHYSICAL PRINCIPLES OF BIORHEOLOGY AND HEMODYNAMICS**

**9.1. CONTINUITY EQUATION**

The fluid flow is characterized by streamlines. *Streamline* is a such line in the flow that the velocity of each particle on the line is tangential to the line (fig. 9.1, a).

*Fig. 9.1. a — streamline; b — streamtube*

*Streamtube* is an imaginary tube whose boundaries consist of streamlines (fig. 9.1, b).
Two main types of fluid flow are distinguished: laminar fluid flow and turbulent fluid flow. **Laminar fluid flow** is observed when the streamlines are continuous, and stirring of the fluid layers does not occur. **Turbulent fluid flow** is flow in which turbulence arises, fluid particles velocity varies randomly and the streamlines are discontinuous.

**An ideal fluid** is incompressible and nonviscous fluid.

Consider an ideal fluid flowing through a streamtube of varying cross sectional area \( S \) (fig. 9.2).

\[
S_1 \cdot v_1 = S_2 \cdot v_2 \quad \text{or} \quad S \cdot v = \text{const.} \quad (9.1)
\]

It states that the product of the linear velocity \( v \) of the fluid and cross-sectional area \( S \) through which it flows is constant for a given streamtube.

Two different characteristics commonly are used as measure of fluid flow: linear flow rate \( v \) and volume flow rate \( Q \) (discharge).

**Linear flow rate** \( v \) is the distance (\( L \)), passed by fluid particles per unit time (\( t \)):

\[
v = \frac{L}{t}. \quad (9.2)
\]

Linear flow rate \( v \) is measured in m/s.

**Volume flow rate** \( Q \) is the volume of fluid (\( V \)) passing a given cross-section in unit time (\( t \)):

\[
Q = \frac{V}{t}. \quad (9.3)
\]

Volume flow rate \( Q \) is measured in cubic meters per second m\(^3\)/s. Other common units are: liter/s = 10\(^3\) cm\(^3\)/s = 10\(^{-3}\) m\(^3\)/s.
The relation between linear $\nu$ and volume fluid flow rate $Q$ is the following:

$$Q = \frac{V}{t} = \frac{S \cdot L}{t} = \nu S. \quad (9.4)$$

It is another statement for continuity equation: the volume flow rate $Q$ must be the same for all cross sections of the streamtube: $Q = \text{const}$. According to the principle of continuity for hemodynamics: total volume flow $Q$ of blood must be the same at any level of arborization. In the cardiovascular system: the greater the total cross-sectional area of vessels is, the lower the linear flow rate $\nu$ will be (fig. 9.3). Volume flow rate $Q$, however, remains constant (fig. 9.3).

![Diagram](image)

Fig. 9.3. Volume flow rate $Q$ at systemic veins is equal to volume flow rate through systemic capillaries, which is equal to flow through the aorta: $Q = S\nu = \text{const}$. Because the same volume of blood must flow through each segment of the circulation each minute, the velocity of blood flow is inversely proportional to vascular cross-sectional area.

Because cross-sectional area increases greatly from arteries to the arterioles and to the capillaries, the lowest blood flow velocity occurs through the capillary network (fig. 9.3). This slow velocity through this exchange segment of the vascular system has the beneficial effect of allowing more time for exchange of material between the cardiovascular system and the extracellular fluid.
9.2. Bernoulli’s Equation

Bernoulli’s equation is one of the most important and useful equations in hemodynamic. Bernoulli’s equation describes the flow of a nonviscous fluid. It is valid for incompressible fluids steady flowing along a streamline, no energy loss due to friction, no heat transfer. This equation gives the relationship between velocity, pressure and elevation in a streamline.

Let’s consider streamtube of an ideal fluid with two chosen cross-sectional areas $S_1$ and $S_2$ (fig. 9.4); $u_1, u_2$ are the linear velocities of the fluid particles in the chosen sections $S_1$ and $S_2$; $P_1, P_2$ are the pressures at the sections $S_1$ and $S_2$; $h_1, h_2$ are the heights of the centers of chosen sections $S_1$ and $S_2$.

![Fig. 9.4. Streamtube of an ideal fluid with two chosen cross-sectional areas $S_1$ and $S_2$](image)

In the steady flow of a nonviscous, incompressible fluid of density $\rho$, the pressure $P$, the fluid speed $u$, and the elevation $h$ at any two points are related by:

$$\frac{mu_1^2}{2} + PV + mgh_1 = \frac{mu_2^2}{2} + P_2V + mgh_2,$$

(9.5)

where $\frac{mu^2}{2}$ is the kinetic energy of fluid, $PV$ is the fluid potential energy, which is defined by the liquid pressure, $mgh$ is the potential energy, provided by the location of the liquid at height $h$.

Dividing equation (9.5) by the volume $V$, one can obtain the Bernoulli equation:

$$\frac{\rho u_1^2}{2} + P_1 + \rho gh_1 = \frac{\rho u_2^2}{2} + P_2 + \rho gh_2 \quad \text{or} \quad \frac{\rho u^2}{2} + P + \rho gh = \text{const},$$

(9.6)

where $\frac{\rho u^2}{2}$ is the dynamic pressure, $\rho gh$ is the hydrostatic pressure, $P$ is the static pressure in the fluid.

Bernoulli equation states that the sum of dynamic, static and hydrostatic pressures is constant along a streamline during steady flow.
In a person with advanced arteriosclerosis, the Bernoulli effect produces a symptom called \textit{vascular flutter}. In this situation, the artery is constricted as a result of accumulated plaque on its inner walls. To maintain a constant flow rate, the blood must travel faster than normal through the constriction. It reduces the pressure in the artery, relative to the stationary extracellular fluid surrounding the artery. If the blood speed is sufficiently high in the constricted region, the artery may collapse under external pressure, causing a momentary interruption in blood flow. At this moment, there is no Bernoulli effect, and the vessel reopens under arterial pressure. As the blood rushes through the constricted artery, the internal pressure drops and again the artery closes. Such variations can be heard with a stethoscope. If the plaque becomes dislodged and ends up in a smaller vessel that delivers blood to the heart, the person can suffer a heart attack.

An \textit{aneurysm} is caused by the weakening of the arterial wall where a bulge occurs and the cross-section of a vessel increases considerably (fig. 9.5). An analysis as before will show that the flow velocity $\nu$ will be reduced at the cross-section of an aneurysm and the pressure $P$ will increase. The higher pressure may cause further expansion of the cross-section, which can lead to the bursting of the vessel at that site.

\textbf{Fig. 9.5. Normal aorta and aorta with aneurysm}

\section*{9.3. Fluid viscosity}

Between the molecules of real fluids interaction forces exist, which manifest themselves as friction between the moving particles of the fluid. The presence of internal friction forces in the fluids leads to the fact that its various layers are moving with different velocities.

Consider the experiment shown in fig. 9.6. A fluid is located between two parallel plates. The top plate is moved with constant velocity $\nu$ by the action of shearing force $F$, and the bottom plate is kept in place (velocity is zero).
It leads to the fact that the different layers of the fluid are moved with different velocities. Quantitatively the change in fluid layers velocity with distance is characterized by a velocity gradient \( \frac{d\nu}{dx} \), also is called the shear rate \( \gamma \):
\[
\gamma = \text{grad}\nu = \frac{d\nu}{dx}.
\]

The rate of shear \( \gamma \) is the relative displacement of one fluid layer with respect to the next. The unit of shear rate (velocity gradient) is \( 1/s \).

The ratio of the tangential force \( F \) needed to maintain the moving plate at a constant velocity \( \nu \) to the plate area \( S \) is the shear stress \( \tau \):
\[
\tau = \frac{F}{S}.
\]

The shear stress \( \tau \) per unit area is proportional to the velocity gradient \( \gamma \):
\[
\frac{F}{S} = \eta \frac{d\nu}{dx} \quad \text{or} \quad \tau = \eta \gamma. \tag{9.8}
\]

Equation (9.8) is called Newton’s equation. The proportionality constant between shear stress \( \tau \) and the velocity gradient \( \gamma \) is called the viscosity \( \eta \) of the fluid. Fluid viscosity \( \eta \) depends on nature of the fluid and fluid temperature (viscosity \( \eta \) usually decreases with temperature increase). The units of viscosity are: \( \text{Pa} \cdot \text{s} \) (in the SI system) and \( \text{poise} \) (in the CGS system). Conversions: \( 1 \text{Pa} \cdot \text{s} = 10 \text{P} \); \( 1 \text{mPa} \cdot \text{s} = 1 \text{cP} \). The viscosity of water at 20 °C is \( 1 \text{mPa} \cdot \text{s} = 1 \text{cP} \).

Fluids with a straight relationship between shear stress and shear rate are called \textit{Newtonian fluids}, i. e., viscosity does not depend on shear rate (\( \eta = \text{const} \neq f(d\nu/dx) \) at given \( T \)). Water and plasma are \textit{Newtonian fluids}.

\textit{Non Newtonian fluids} don’t follow the linear relation between shear stress and shear rate and their viscosity depends on the shear rate (\( \eta = f(d\nu/dx) \neq \text{const} \)). Blood and paint are \textit{Non Newtonian fluids}. 

\textit{Fig. 9.6. Fluid flow between two parallel plates}
9.4. Poiseuille’s Equation

Consider a smooth pipe of length $L$ and radius $r$ with a viscous fluid laminar moving within it, where $P_1$ and $P_2$ are the pressures at the ends of the pipe (fig. 9.7).

![Fig. 9.7. A smooth pipe with a viscous fluid laminar moving within it](image)

The flow volume $V$ of the fluid is proportional to fluid flow time ($V \sim t$); $V$ is proportional to a difference in pressures ($V \sim (P_1 - P_2)$); $V$ is proportional to the radius ($V \sim r^4 \sim S^2$); $V$ is inversely proportional to viscosity ($V \sim 1/\eta$); $V$ is inversely proportional to the length ($V \sim 1/L$). The flow volume $V$ of a viscous fluid laminar flowing through the smooth tube of length $L$ and radius $r$ during time $t$ is given by the Poiseuille’s equation:

$$V = \frac{\pi r^4 (P_1 - P_2)}{8\eta L} t.$$

(9.9)

Dividing both parts of the equation (9.9) by $t$, one can obtain the volume flow rate $Q$:

$$Q = \frac{\pi r^4 (P_1 - P_2)}{8\eta L}.$$

(9.9a)

The equation (9.9a) for the volume flow rate $Q$ can be written in more simple form as the Hagen–Poiseuille Law:

$$Q = \frac{(P_1 - P_2)}{X}.$$

In this formula $X$ is the hydraulic resistance of the pipe:

$$X = \frac{\Delta P}{Q} = \frac{8\eta L}{\pi r^4}.$$

(9.10)

Poiseuille’s Law corresponds to Ohm’s Law for electrical circuits $I = \frac{U}{R}$.

The pressure drop $\Delta P$ is analogous to the potential difference $\Delta \phi$ and the volume flow rate $Q$ is analogous to the current $I$. This similarity is presented in table 9.1.
Table 9.1

<table>
<thead>
<tr>
<th>Hydrodynamics</th>
<th>Electrodynamics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flows: $V$ — volume</td>
<td>$Q$ — charge</td>
</tr>
<tr>
<td>$Q = \frac{V}{t}$ is the volume flow rate</td>
<td>$I = \frac{q}{t}$ is the current</td>
</tr>
<tr>
<td>Driving force</td>
<td>Driving force</td>
</tr>
<tr>
<td>$P_1 - P_2$ is the pressure difference</td>
<td>$\varphi_1 - \varphi_2$ is the potential difference</td>
</tr>
<tr>
<td>$X = \frac{\Delta P}{Q} = \frac{8\eta L}{\pi r^4}$ is the hydraulic resistance</td>
<td>$R = \frac{\Delta \varphi}{I} = \frac{\rho l}{S}$ is the electrical resistance</td>
</tr>
<tr>
<td>$Q = \frac{P_1 - P_2}{X}$ is Hagen–Poiseuille Law</td>
<td>$I = \frac{\Delta \varphi}{I} = \frac{U}{R}$ is Ohm’s Law</td>
</tr>
</tbody>
</table>

Poiseuille law gives two of the most fundamentally important relationships used to describe blood flow rate and pressure in the cardiovascular system, which are:

$$Q = \frac{\Delta P}{X};$$
$$\Delta P = X \cdot Q.$$  

These equations indicate that blood flow rate is proportional to the pressure difference between the entrance and exit points of the tube and inversely proportional to the resistance (i.e., as resistance increases, blood flow rate decreases).

These equations express the most important of all the relations that one needs to understand to comprehend the hemodynamics of the circulation. In this equation resistance $X$ is the impediment to blood flow in a vessel. It must be calculated from measurements of blood flow rate $Q$ and pressure difference $\Delta P$ between the two ends of the vessel ($\Delta P$ also sometimes called «pressure gradient» along the vessel, which is the force that pushes the blood through the vessel).

Blood pumped by the heart flows from the high pressure part of the systemic circulation (i.e., aorta) to the low pressure side (i.e., vena cava) through many meters of blood vessels arranged in series and in parallel. The arteries, arterioles, capillaries, venules, and veins are collectively arranged in series. When blood vessels are arranged in series, flow through each blood vessel is the same and the total resistance to blood flow ($R_{\text{total}}$) can be calculated (in analogy with the total electrical resistance calculation) as the sum of the resistances of each vessel:

$$X_{\text{total}} = X_1 + X_2 + X_3 + \ldots + X_n.$$ (9.11)

The total peripheral vascular resistance is therefore equal to the sum of resistances of the arteries, arterioles, capillaries, venules, and veins. In the example shown in figure 9.8, the total vascular resistance is equal to the sum of $X_1$, $X_2$ and $X_3$.  

79
Blood vessels branch extensively to form parallel circuits that supply blood to the many organs and tissues of the body. This parallel arrangement permits each tissue to regulate its own blood flow, to a great extent, independently of flow to other tissues. For blood vessels arranged in parallel (in analogy with the total electrical resistance calculation) (fig. 9.9), the total resistance to blood flow is expressed as:

$$\frac{1}{X_{\text{total}}} = \frac{1}{X_1} + \frac{1}{X_2} + \frac{1}{X_3} + \ldots + \frac{1}{X_n}.$$  \hspace{1cm} (9.12)

That is, the reciprocal of the total resistance in vessels arranged in parallel is the sum of the reciprocals of the individual resistances.

It should be noticed that losing similar-size vessels in parallel in the arterial system raises vascular resistance, whereas adding similar-size vessels in parallel reduces resistance.

The resistance $X$ of the vessel decreases in proportion to the fourth power of the vessel radius $r^4$, in accordance with the following formula:

$$X = \frac{\Delta P}{Q} = \frac{8\eta L}{\pi r^4}.$$  

Length $l$ of the vessel and viscosity $\eta$ of the blood don’t normally change, but radius $r$ can and does. Slight changes in the diameter of a vessel cause tremendous changes in the vessel’s ability to conduct blood when the blood flow is streamlined. In the cardiovascular system, vessels get smaller as they proceed from arteries down to arterioles and capillaries, which tends to increase resistance to flow. Arterioles are of the greatest resistance (the «resistance component» of the cardiovascular system): the influence of small radius of arteriole (the first reason of resistance increase) predominates over the influence of parallel arterioles in the cardiovascular systems. However, the number of vessels arranged in parallel also increases dramatically in this direction, which
tends to decrease resistance. The effect of these two phenomena on the relative resistances of individual sections of the vasculature depends on whether the number of vessels added in parallel can compensate for the resistance effects of adding vessels that individually have a high resistance resulting from small radii. This interplay between series and parallel elements is a primary factor in creating the form of the pressure profile, or \( \Delta P \), along the arterial side of the circulation, as will be explained later in this chapter.

As the resistance of the vessels also depends on the blood viscosity, which increases toward the walls of blood vessels, the value of blood resistance will depend on the general area of blood vessels. For example, in the arterioles the blood flow rate is high—just a little bit less than the blood flow rate in the aorta—whereas the general area of the internal walls of the arterioles results in the increased blood resistance.

9.5. METHODS OF VISCOSITY MEASUREMENT

A viscometer (also called viscosimeter) is an instrument used to measure the viscosity of a fluid.

9.5.1. Falling sphere viscometers

The falling sphere viscometer is one of the earliest and simplest methods to determine the viscosity of a Newtonian fluid. In this method, a sphere is allowed to fall freely a measured distance through a viscous liquid medium, and its velocity is determined. The viscous drag of the falling sphere results in the creation of a restraining force, \( F_d \), described by Stokes’ law:

\[
F_d = 6\pi \eta rv,
\]

where \( F_d \) is the drag force of the fluid on a sphere, \( \eta \) is the fluid viscosity, \( v \) is the velocity of the sphere relative to the fluid, and \( r \) is the radius of the sphere.

If a sphere of density \( \rho_s \) is falling through a fluid of density \( \rho_f \) in a cylindrical container of infinite extent three forces exert on a solid sphere: \( F_b \), \( F_d \), and \( mg \) (fig. 9.10).

![Fig. 9.10. Falling sphere viscometer](image-url)
The first two forces arise from the buoyancy effect of displacing the fluid in question and from the viscous drag of the fluid on the sphere, respectively. Both forces act upwards — buoyancy $F_b$ tending to «float» the sphere and the drag force $F_d$ resisting the acceleration of gravity. The only force acting downwards is the body force resulting from gravitational attraction $mg$. By summing forces in the vertical direction one can write the following equation (at uniform falling of sphere):

$$F_b + F_d = mg.$$  

The buoyancy force can be written as:

$$F_b = V_s \rho_f g = \frac{4}{3} \pi r^3 \rho_f g,$$

where $r$ is the radius of the sphere, $\rho_f$ is the density of the fluid, $g$ is the gravitational acceleration.

When reaching the uniform motion of sphere, force of gravity $mg$ is equal to the sum of the drag force $F_d$ and the buoyancy force $F_b$:

$$\frac{4}{3} \pi r^3 \rho_s g = \frac{4}{3} \pi r^3 \rho_f g + 6\pi \eta \nu,$$  \hspace{1cm} (9.14)

where $m = \frac{4}{3} \pi r^3 \rho_s$, $\rho_s$ is the density of the sphere.

Knowing the terminal velocity $\nu$ of the uniform motion, the radius $r$, density of the sphere $\rho_s$ and the density of the fluid $\rho_f$, it is possible to calculate the viscosity of the fluid $\eta$:

$$\eta = \frac{2}{9} \frac{\rho_s - \rho_f}{r^2 g \nu}.$$  \hspace{1cm} (9.15)

As the sphere passes through the marked region of length $L$ at its terminal velocity $\nu$, a measure of the time $t$ taken to traverse this distance allows the velocity of the sphere to be calculated: $\nu = L / t$.

**9.5.2. U-tube viscometers**

U-tube viscometers are also known as glass capillary viscometers or *Ostwald viscometers*, named after Wilhelm Ostwald. Their geometry resembles a U-tube with two reservoir bulbs connected to a capillary tube passage (fig. 9.11). There is a bulb of volume $V$ above the capillary and a reservoir which is lower down on the other arm. In use, fluid is drawn into the upper bulb by suction, and then allowed to flow down through the capillary. Two marks (one above and one below the upper bulb) indicate a known volume $V$. The efflux time, $t$, is measured for that volume to flow through the capillary under gravity. Ostwald viscometers measure the viscosity of a fluid with a known density.
The viscosity of test fluid $\eta$ can be determined by measuring time required for reference fluid $t_0$ and for test one $t$ to pass through a bulb of volume $V$ (between two marks). The flow volume of reference fluid (water) $V_0$ and the flow volume of test one $V$ are the same $V_0 = V$:

$$\frac{\pi r^4 \rho_0 gh}{8\eta_0 L} t_0 = \frac{\pi r^4 \rho gh}{8\eta L} t,$$

(9.16)

where $\eta$ and $\eta_0$ are the viscosities of test fluid and reference one respectively; $\rho$ and $\rho_0$ are the densities of test fluid and reference one respectively; $t$ and $t_0$ are time required for the test fluid and reference one to pass between two marks indicating a known volume $V$.

The viscosity of the fluid $\eta$ can be calculated as:

$$\eta = \eta_0 \frac{\rho t}{\rho_0 t_0}.$$  

(9.17)

Thus in order to determine the viscosity of test fluid $\eta$ it’s necessary to measure time required for reference fluid $t_0$ and for test one $t$ to pass between two marks indicating a bulb of volume $V$.

9.5.3. Rotational viscometers

The main feature of this method is that it gives possibility to find the viscosity dependence on gradient velocity: $\eta = f(\frac{d\nu}{dx})$. It is very important for non-Newtonian fluids particularly for blood.

There are a variety of rotational viscometers. Consider the principle of action of one of them. It consists of two cylinders with a common axis of
rotation (fig. 9.12). The inner cylinder is suspended by a thread, and the external can be rotated about its longitudinal axis with adjustable angular velocity $\omega$. The gap between the cylinders is filled with the test non-Newtonian fluid, such as blood. Due to the viscosity of the fluid the inner cylinder begins to rotate and reaches equilibrium at particular angle of rotation $\theta$. This angle can be easily measured. The higher the viscosity of the fluid $\eta$ and the angular velocity $\omega$ is, the greater the angle of rotation $\theta$ will be:

$$\theta = k \eta \omega,$$

where $k$ is the viscometer constant.

![Fig. 9.12. Schematic diagram of rotating cylinder viscometer](image)

Thus, the viscosity of the liquid $\eta$ can be calculated using the following formula:

$$\eta = \frac{\theta}{k \omega},$$

(9.18)

where $\theta$ is the rotation angle of inner cylinder; $k$ is the viscometer constant; $\omega$ is the angular velocity of outer cylinder rotation.

Actually, the shear rate profile across the gap between the cylinders depends on the relative rotational speed. Therefore the dependence of the fluid viscosity on the shear rate can be measured by rotating cylinder viscometer.

9.6. Factors affecting blood viscosity

Blood is not a homogeneous fluid consisting of cellular elements, such as the red blood cells, white blood cells, platelets in a fluid solution called plasma. The viscosity of whole blood in normal state is 4–5 centipoise (cP) and in
disease state is 1.7–22.9 cP. The viscosity of blood depends on the number of factors:

– temperature;
– hematocrit;
– shear rate (velocity gradient du /dx);
– arrangement of erythrocytes in the blood stream.

For Newtonian fluids, the viscosity $\eta$ decreases with increasing temperature $T$. Temperature change can lead to a change in the degree of aggregation of red blood cells and platelets and cause other changes in the blood structure. The influence of temperature on the viscosity of the blood is ambiguous.

Most of the volume of the cellular components of blood is composed of red blood cells, or erythrocytes. Therefore the red blood cells are the most important components of the cellular component with respect to the rheological properties of blood. Haematocrit ($Ht$) is ratio of the total volume of erythrocytes ($V_{rbc}$) to blood volume ($V_{bl}$) in which they are contained: $Ht = V_{rbc} / V_{bl}$ (or relative volume of blood occupied by erythrocytes). Normally haematocrit is about 46% for men and 38% for women. With increasing haematocrit the viscosity of the blood $\eta$ rises drastically. Venous blood viscosity is greater than arterial one. In the venous blood there is increased concentration of carbon dioxide and the red blood cells are larger in size (volume) and have spherical shape. The relation between hematocrit and viscosity is shown in fig. 9.13.

![Fig. 9.13. Dependence of blood viscosity on the hematocrit](image)

The viscosity of blood $\eta$ depends on shear rate (velocity gradient $du/dx$). More exactly formulated, when velocity gradient (shear rate) increases, viscosity $\eta$ reduces. In large and medium size arteries shear rates are higher than 100 s$^{-1}$,
so viscosity is relatively constant $\eta = \text{const} \approx 2 \text{ cP}$. At high shear rates blood is often modeled as a Newtonian fluid with a mean viscosity of 4–5 cP. When shear rate is reduced in small blood vessels, viscosity of blood $\eta$ gradually increases. At shear rates less than 1 s$^{-1}$ the viscosity growth is going rapidly. For extremely low shear rates formation of red blood cells aggregates may occur, thereby increasing viscosity to very high values (fig. 9.14).

![Fig. 9.14. Dependence of blood viscosity on shear rate](image-url)

Velocity flow profile for a Newtonian fluid is the ideal parabolic flow: the velocity is zero at the wall and reaches a maximum at the centreline of the vessel (fig. 9.15). Fluid nearer the wall moves more slowly. But for the circulating blood the velocity profile is relatively flat (blunt profile).

![Fig. 9.15. Velocity profile for Newtonian fluid and for blood](image-url)
In large arteries axial accumulation of red blood cells arises in the direction of movement (fig. 9.16). Thin plasma layer near the vessel wall is formed, which does not contain red blood cells and this plasma layer has low viscosity.

Fig. 9.16. Axial accumulation of red blood cells in large artery

### 9.7. Reynolds Number

A British scientist Osborne Reynolds established that the nature of the flow (laminar or turbulent) depends upon a dimensionless quantity, which is called the Reynolds number $Re$. It is given as follows:

$$\text{Re} = \frac{\rho ud}{\eta},$$

where $\rho$ is the density of the fluid; $u$ is the mean velocity of the flow; $d$ is the diameter of the tube; and $\eta$ is the fluid’s dynamic viscosity.

There is the critical Reynolds number ($Re_{cr}$) beyond which the flow can be considered turbulent. $Re_{cr}$ indicates whether flow is laminar or turbulent.

Laminar flow within smooth pipes, for example, will occur when the Reynolds number is below the critical Reynolds number of $Re_{cr \text{ pipe}} = 2300$ and turbulent flow will happen when it is above 2300. The Reynolds number depends on the pipe diameter and the mean fluid velocity $v$ within the pipe. The value of 2300 has been determined experimentally and a certain range around this value is considered the transition region between laminar and turbulent flow.

For blood the critical Reynolds number is varied from 900 to 1600. It should be emphasized that flow in the circulatory system is normally laminar, although flow in the aorta can destabilize briefly during the deceleration phase of late systole; however, this time period is generally too short for flow to become fully turbulent. Turbulent flow may occur in large blood vessels, but the distensible vessel wall and arterial narrowing diminish the disturbances in flow. Certain disease conditions can produce turbulent blood flow, particularly downstream of a vessel narrowing or distal to defective heart valves. Such a flow can damage the vessel wall and contribute to the further progression of a disease.
9.8. **Pulse Wave**

When the heart ejects blood into the aorta during systole, at first the proximal part of the aorta becomes distended, because aorta is elastic. The walls of large arteries are composed of smooth muscle cells, collagen fibers and elastine fibers. They give to arteries the ability to distend and recoil.

Because of these elastic properties, the arterial system dampens the dramatic pressure changes created when the ventricle ejects blood into the aorta (fig. 9.17). As the aorta close to the heart distends with each ventricular contraction, the reservoir is formed. After that during diastole the aorta near to the heart contracts and pushes blood along the aorta. Large arteries also function as pressure reservoirs; when they are expanded, they store pressure in their walls.

![Fig. 9.17. Systolic ejection of blood from left ventricle, aorta expands and than recoils — setting up pressure wave or pulse wave](image)

Then next part of the aorta distends and the wave of distention moves along the aorta. Ejection of blood from the left ventricle initiates the oscillations of pressure exerted by blood on walls of a vessel. These oscillations of pressure exerted by blood on walls of a vessel propagated along the vascular system are called the **pulse wave**.

The pulse wave velocity is expressed by the Moens–Korteweg equation:

\[
\nu = \sqrt{\frac{E \cdot h}{\rho \cdot d}},
\]

(9.20)

where \( E \) is the Young’s modulus; \( h \) is the vascular wall thickness; \( d \) is the diameter of vessel; \( \rho \) is the blood density. With increasing vessel stiffness \( E \), increasing the thickness \( h \) of vessel wall and decreasing vessel diameter \( d \), pulse wave velocity \( \nu \) increases.
The more rigid the wall of the artery, the faster the wave moves. The velocity of pulse wave in the aorta is 4–6 m/sec; in the artery 8–12 m/sec. In general, the greater the compliance of each vascular segment is, the slower the velocity will be.

Let’s compare the velocity of the pulse wave and blood velocity in the aorta. The velocity of transmission of the pressure pulse is much larger than the velocity of blood flow. It occurs because the pulse wave is a moving wave of pressure.

As the heart beats, the blood pressure rises and falls (fig. 9.18). **Systolic pressure (Ps)** is the maximum of the pressure during systole. Averages about 120 mm Hg. **Diastole pressure (Pd)** is the minimum pressure during diastole. Averages between 70–80 mm Hg. **Pulse pressure** is the difference between systolic and diastolic pressure (40 mm Hg). **Mean pressure** is the average pressure throughout each cardiac cycle (90–95 mm Hg), determined by formula:

\[
P_{\text{mean}} = \frac{1}{T} \int_{0}^{T} P(t) dt,
\]

where \( T \) is period pulse fluctuations; \( t \) is the current time.

![FIG. 9.18. Dependence of blood pressure on time](image)

At any point of circulatory system blood pressure \( P_{CS} \) depends on the atmospheric pressure \( P_0 \), hydrostatic pressure \( \rho gh \), and pressure \( P \) due to the pumping function of the heart:

\[
p_{CS} = p_0 + \rho gh + p,
\]

where \( p_0 \) is the atmospheric pressure; \( \rho gh \) is the hydrostatic pressure due to weight of the blood in the vessels (\( h \) is the height of blood column; \( \rho \) is the blood density); \( P \) is the pressure generated by the heart.
Transmural («transmural»: across the wall) pressure $P_{tm}$ is the pressure difference between the pressure inside $P_i$ and the pressure outside $P_e$ the vessel wall (interstitial pressure) (fig. 9.19):

$$P_{tm} = P_i - P_e.$$  \hspace{1cm} (9.23)

![Diagram of transmural pressure](image)

Fig. 9.19. Transmural («transmural»: across the wall) pressure

The interstitial pressure normally doesn’t play a significant role in determining transmural pressure in the systemic arteries (except in contracting muscle), but very important in the low pressure system (venous circulation).

Gravitational pressure also occurs in the vascular system of the human being because of weight of the blood in the vessels. When a person is standing, the pressure in the right atrium remains about 0 mm Hg because the heart pumps into the arteries any excess blood that attempts to accumulate at this point. However, in an adult who is standing absolutely still, the pressure in the veins of the feet is about +90 mm Hg simply because of the gravitational weight of the blood in the veins between the heart and the feet. The venous pressures at other levels of the body are proportionately between 0 and 90 mm Hg.

9.9. DISTRIBUTION OF BLOOD PRESSURE IN CARDIOVASCULAR SYSTEM

The heart is an intermittent pump; it generates high pressure within the ventricles when it contracts during systole, which then drops to near zero during diastole. However, because arteries are compliant, some of the ejected blood into the arteries distends these vessels, like the expansion of a water-filled balloon. During diastole, recoil of the arteries pushes blood forward against the downstream vascular resistance, generating a significant diastolic pressure. For this reason, diastolic pressure drops to only about 80 mm Hg in the aorta as compared with near zero in the ventricles.

Examination of the pressure profile across the cardiovascular system (fig. 9.20, b) shows that the largest drop of pressure (i. e. pressure difference $\Delta P$ between the two ends of the vessel) occurs across the arterioles, indicating that this is the site of greatest vascular resistance (fig. 9.20, a) in the cardiovascular system. Although there are many more arterioles than arteries in the cardiovascular system (resistances in parallel), this large pressure drop indicates that their reduction in individual size dominates over the addition of parallel vessels.
Similarly, although individual capillaries are small, so many of these lie in parallel that resistance across the capillaries is actually lower than that across the arterioles; hence, the pressure drop across the capillary segment of the circulation is less than that across the arterioles. The pressure drops still lower in the veins. The mean pressure falls below 0 mm Hg by the time it reaches veins what empty into the right atrium of the heart. Since the pressure drop in the main arteries is small, when the body is horizontal, the average arterial pressure is approximately constant throughout the body. If a person is standing erect, the blood pressure in the arteries is not uniform in the various parts of the body. The weigh of the blood affects on the pressure at various locations.

Fig. 9.20. a — vascular resistance of the vessels relative to the resistance of the aorta in the cardiovascular system; b — pressure and flow velocity profile in the systemic circulation. The arterial portion of the circulation is characterized by high, pulsatile pressure and high flow velocity. This profile changes to one of low pressure and velocity without pulsatile character in the veins. The largest drop in mean arterial pressure occurs across the arteriolar segment of the circulation, indicating that this is the sight of highest vascular resistance in the cardiovascular system.
Immediately following the arterioles are the capillaries. Though the radii of the capillaries are very small, the network of capillaries have the largest surface area in the vascular network. They are known to have the largest surface area in the human vascular network. The larger the total cross-sectional area, the lower the mean velocity as well as the pressure. The pressure in the capillaries is low enough for nutrients can diffuse easily through pores of the capillary walls to the tissue cells. In the capillaries the average blood pressure is only about 30 mm Hg.

The pressure drops still lower in the veins. The mean pressure falls below 0 mm Hg by the time it reaches veins what empty into the right atrium of the heart.

Since the pressure drop in the main arteries is small, when the body is horizontal, the average arterial pressure is approximately constant throughout the body. If a person is standing erect, the blood pressure in the arteries is not uniform in the various parts of the body. The weigh of the blood affects on the pressure at various locations.

### 9.10. Blood Pressure Measurement

The arterial blood pressure is an important indicator of the health of an individual. It almost always is measured in millimeters of mercury (mm Hg). There are invasive and non-invasive methods of blood pressure measurement. Let’s consider direct (invasive) method of blood pressure measurement. A vessel is punctured and a catheter (a flexible tube) is guided in. The most common sites are brachial and radial arteries but also other sites can be used e.g. femoral artery. This method is precise but it is also a complex procedure involving many risks.

In non-invasive method (palpatory method or Riva-Rocci method) an occlusive cuff is placed on arm and inflated to $P_{cuff} > $ systolic pressure, so that it stops arterial blood flow. When the cuff is deflated, there is a palpable pulse in the wrist. Only the systolic pressure can be measured by Riva-Rocci method.

Auscultatory method — is routine method for determining systolic and diastolic arterial pressures. A stethoscope is placed over the artery and a blood pressure cuff is inflated around the upper arm (fig. 9.20).

The cuff compresses the arm, pressure in the cuff elevates greater than systolic pressure, and blood can not pass. As long as this cuff pressure is higher than systolic pressure, the artery remains collapsed so that no blood jets into the distal artery during any part of the pressure cycle. No sounds are heard from the artery with the stethoscope. Then the cuff pressure decreases. Just as soon as the pressure in the cuff falls below systolic pressure, blood begins to slip through the artery beneath the cuff during the peak of systolic pressure, and one begins to hear tapping sounds from the artery in synchrony with the heartbeat. These sounds are called Korotkoff sounds. The cause of Korotkoff sounds is
turbulence in the vessel beyond the cuff. As soon as these sounds begin to be heard, the pressure level indicated by the manometer connected to the cuff is about equal to the systolic pressure. Then, finally, when the pressure in the cuff falls to equal diastolic pressure, the artery no longer closes during diastole, which means that the basic factor causing the sounds (the jetting of blood through a squeezed artery) is no longer present. Therefore, the sounds suddenly disappear. One notes the manometer pressure when the Korotkoff sounds disappear; this pressure is about equal to the diastolic pressure.

**Fig. 9.20. Auscultatory method**

### 9.11. HEART WORK AND HEART POWER

Blood flow is provided by the pumping action of the heart. Heart work $A$ during one cardiac contraction consists of work of right $A_r$ and left $A_l$ ventriculars:

$$A = A_l + A_r.$$  

Right ventricular work is normally about 0.2 of the left ventricle work:

$$A_r = 0.2 A_l$$  

$$A = 1.2 A_l$$

When heart ejects the stroke volume into the aorta, work done by left ventricle is spent into overcoming the pressure forces of blood in the vascular system and into kinetic energy of blood flow. Therefore heart work consists of static component and kinetic one. Static component can be calculated by formula:

$$A_{st} = P_{mean} \cdot V_s,$$
where $P_{\text{mean}}$ is the mean blood pressure in aorta: $P_{\text{mean}} = 100 \text{ mmHg} = 13.3 \text{ kPa}$; $V_s$ is the stroke volume at rest: $V_s \approx 60 \text{ ml} = 6 \cdot 10^{-5} \text{ m}^3$. Then the static component of heart work $A_{st}$ is equal to $\approx 0.8 \text{ J}$.

Kinetic component of heart work $A_k$ can be written as:

$$A_k = \frac{mV^2}{2} = \frac{\rho V_s \cdot \nu^2}{2},$$

where $\rho$ is the blood density ($\rho \approx 1.05 \cdot 10^3 \text{ kg/m}^3$); $\nu$ is the linear velocity of blood flow in aorta ($\nu \approx 0.5 \text{ m/s}$); $V_s$ is the stroke volume at rest: $V_s \approx 60 \text{ ml} = 6 \cdot 10^{-5} \text{ m}^3$. Then the kinetic component of heart work $A_k$ is equal to $\approx 0.008 \text{ J}$.

Total work of heart $A$ during one contraction can be calculated as:

$$A = 1.2 \left( PV_s + \frac{\rho V_s \nu^2}{2} \right) \approx 1 \text{ J}. \quad (9.24)$$

Total work of heart $A$ during one contraction divided over systole duration is the average heart power over the one contraction. At rest the average power over the one contraction can be calculated as:

$$P = 1 \text{ J} / 0.3 \text{ s} = 3.3 \text{ W}. \quad (9.25)$$

At rest, the proportion of the kinetic component of heart work is 1% of the total work $A$. With the rise of physical activity proportion of kinetic component in the total heart work increases due to increase of blood flow velocity $\nu$ and can reach 30%.

**Questions:**

1. Give the definition of the linear flow velocity and the volumetric flow rate? What is the relation between them?
2. What is the meaning of the continuity equation?
3. Write Bernoulli’s equation and characterize it.
4. Describe viscous fluid flow features. Write Newton’s equation. What is the fluid viscosity? What are the units of the fluid viscosity?
5. What is the difference between Newtonian fluids and non-Newtonian ones? Make examples of these fluids.
6. Write Poiseuille’s Equation. How to determine the vessel hydrodynamic resistance?
7. Compare advantages and disadvantages of viscosity determination methods.
8. What does Reynolds number describe? Write the formula for Reynolds number.
9. Specify blood viscosity values for norm and for pathological processes.
10. In which part of cardiovascular system does the most of the pressure drop occur? Why?
11. Write the pulse wave velocity formula. Compare the pulse wave velocity values for the aorta, arteries and veins.
12. In which parts of cardiovascular system is turbulent flow of blood observed?
13. Calculate heart work during one cardiac contraction. What is the heart power?
Chapter 10. PHYSICAL PROPERTIES AND FUNCTIONS OF THE BIOLOGICAL MEMBRANE

10.1. STRUCTURE AND PHYSICAL PROPERTIES OF THE BIOLOGICAL MEMBRANE

All biological cells are surrounded by a plasma membrane. In 1972, S. Singer and G. Nicolson proposed the Fluid Mosaic Model of membrane structure. The cytoplasmic membrane consists of phospholipids or glycolipids, cholesterol and protein molecules. Molecules of lipids consist of polar heads (the phosphate radical or glycerol that is soluble in water) and non-polar tails (the fatty acid radical that are insoluble in water) (fig. 10.1). Such molecules are called *amphipophilic*. The head of a phospholipid is attracted to water (it is *hydrophilic*), due to its polar nature. The nonpolar tail is the *hydrophobic*.

![Lipid molecule schematic representation and behavior of lipids on the water surface](image)

There are some self-organizing structures of lipids which depend on lipid concentration and lipid type. The self-organizing structures include *monolayers*, *micelles* and *vesicles*. Self-assembly occurs due to thermodynamics. If the phospholipids are in water (or other polar solution) the tails will want to be «away» from the solution. They could all go to the top (like oil on water), or they could have the tails point toward each other (fig. 10.2).

The phospholipid bilayer is arranged so that the polar parts of the molecules form the outermost and innermost surface of the membrane while the non-polar parts form the center of the membrane. The lipids of membrane are similar to liquid crystal in which the fluidity and plasticity of liquids referred to symmetry of crystal. The liquid-crystalline properties of membranes are explained by the fact that lipids are in molten state in case of normal blood-heat.

Except lipid molecules plasma membrane contains proteins and carbohydrates. Membrane proteins are divided into two categories, integral and peripheral, depending on their location in the membrane.
Proteins that go through the membrane are called **integral or transmembrane proteins**. They have hydrophobic (non-polar amino acids with alpha helix coiling) regions within the interior of the membrane and hydrophilic regions at either membrane surface. The interior and exterior «faces» of transmembrane proteins are comprised of different tertiary domains, as is the hydrophobic «core». Some integral proteins become «anchored» within the phospholipid bilayer by covalently bonding to fatty acids. **Peripheral proteins** are attached to the surface of the membrane, often to the exterior hydrophilic regions of the transmembrane proteins (fig. 10.3).
On the interior surface, peripheral proteins typically are held in position by the cytoskeleton. On the exterior, proteins may attach to the extracellular matrix. Peripheral proteins help give animal cell membranes strength. Usually carbohydrates are located on the extracellular surface of the plasma membrane. The membrane thickness is 8–9 nm.

### 10.2. Types of Lipids and Proteins Motion in the Cell Membrane

The cell membrane is not solid/static/fixed but rather elastic and adaptable to changing needs. Lipids and proteins are in constant thermal motion in the membrane. No strong bonds between neighboring phospholipids, so they fluidly move past one another. The lipids are mobile within their half of the lipid bilayer. Chaotic movements of lipids and proteins along the membrane surface are called the lateral diffusion. The rate of lipids lateral diffusion is about 5 μm/s. The rate of lateral diffusion of proteins is much less than lipids due to their large mass.

Lipids and proteins are participating in the rotational motion, called the rotational diffusion. The rotation angular velocity at normal temperatures for phospholipids is high (~ $10^9$ rad/s) and for proteins it is much less. For example, for rhodopsin the rotation angular velocity $\nu$ is equal to $10^6$ rad/s.

The transition of lipids from one membrane monolayer to another (this transition is called flip-flop) is very unlikely and happens very rarely, as in this case, the polar head must pass through the hydrophobic inner region of the membrane, where it is not soluble. The probability of such flip-flop transitions is $10^{10}$ times smaller than the probability of lateral diffusion. Flip-flop movement needs enzymes (flippases) to speed flip-flop.

Protein mobility can vary greatly. Some proteins are free to move. Others may be tethered to structures in the cytoplasm or extracellular spaces, thus restricting their movement. Some types of cell junctions (e.g., tight junctions) can restrict protein movements to a specific membrane domain.

### 10.3. Transport of Molecules and Ions Through the Membrane

The cytoplasmic membrane is a selectively permeable membrane that determines what goes in and out of the cell (fig. 10.4).

Water-soluble ions generally pass through small pores in the membrane. All other molecules require carrier molecules to transport them through the membrane. There are two major types of the membrane transport: passive and active.

Transport of substances through the membrane is called passive if it does not expend metabolic energy stored in the cell. Passive transport does not require the electrochemical energy of the hydrolysis ATP — adenosine triphosphate. The main types of the passive transport are the following:
1) simple diffusion through the membrane lipid bilayer;
2) simple diffusion through a protein channels-pores in membrane;
3) facilitated diffusion through the membrane with the help of special carrier molecule.

**Fig. 10.4. Movement of substances across cell membranes**

**Diffusion** is a spontaneous process of substances penetration from the region of higher concentration to the region of lower concentration due to the energy of thermal motion. The main driving force for passive transport is the gradient of concentration (more exactly — electrochemical potential gradient) across the membrane.

**Simple diffusion through the lipid bilayer**

One of the most important factors that determine how rapidly a substance diffuses through the lipid bilayer is the lipid solubility of the substance. Hydrophobic molecules and (at a slow rate) very small uncharged polar molecules can diffuse through the lipid bilayer. For instance, the lipid solubilities of oxygen, nitrogen, carbon dioxide, and alcohols are high, so that all these can dissolve directly in the lipid bilayer and diffuse through the cell membrane. For obvious reasons, the rate of diffusion of these substances through the membrane is directly proportional to their lipid solubility. Especially large amounts of oxygen can be transported in this way; therefore, oxygen is delivered to the interior of the cell almost as though the cell membrane did not exist. Membrane permeability for nonpolar organic compounds is high, since
the membrane lipids are well dissolved nonpolar substance. Large polar molecules and ions cannot pass through phospholipid bilayer. One can conclude:
- membrane permeability for the organic molecules decreases when the number of polar groups (hydroxyl, carboxyl and amine) increases;
- membrane permeability for the organic molecules increases when the number of non-polar groups (methyl, ethyl and phenyl) increases.

**Simple diffusion through a protein channel**

Inorganic polar molecules and ions are insoluble in lipids, so they can pass through the membrane only if there are special channels-pores that exist in the membrane. However, the number of such channels is relatively small, so the membrane permeability for ions and polar compounds is in a hundred times worse than for non-polar compounds.

An ion channel is an integral membrane protein or more typically an assembly of several proteins. The size, shape and charge of each channel acts a **selective filter** that allows only certain types of ions to pass. Some types of channel proteins are always open. They allow specific ions to continually pass through the pore’s selective filter using the kinetic energy of the ions. Other channel proteins are gated. Access to the ion is governed by «gates», which can be opened or closed by chemical or electrical signals, or mechanical force, depending on the dimensions of channel (fig. 10.5). If the conformational state of protein channels depends on difference in ionic charges on two sides of membrane, the channels are called **voltage-gated channels**. If the conformational state depends on binding of specific molecule (ligand) to outer or inner surface of channels, the channels are called **chemically-gated channels**.

![Fig. 10.5. The gating of ion channels](image-url)
Facilitated diffusion (by carrier molecule)

Facilitated diffusion is also called carrier-mediated diffusion because a substance transported in this manner diffuses through the membrane with a specific carrier protein helping it to do so. That is, the carrier facilitates the diffusion of the substance to the other side.

There are two kinds of facilitated diffusion:

1) **transfer of a substance with a movable carrier** — a carrier molecule is combined with the transported substance on one side of the membrane, and with it moves through the lipid bilayer to the other side of the membrane;

2) **relay transfer** — in this case, the carrier molecules do not make shuttle movements in the membrane and are embedded in the membrane of each other, forming a bridge to it. Capturing a substance transported, extreme carrier molecule transfers it neighboring molecule, and so on «in the relay».

Facilitated diffusion involves proteins known as carriers, which are specific for a certain type of ions and can transport substances in either direction across the membrane. However, unlike channels, they facilitate the movements of solutes across the membrane by physically binding to them on one side of the membrane and releasing them on the other side. The direction of the solute’s net movement simply depends on its concentration gradient across the membrane. If the concentration is greater in the cytoplasm, the solute is more likely to bind to the carrier on the cytoplasmic side of the membrane and be released on the extracellular side, and there will be a net movement from inside to outside.

A characteristic feature of carrier-mediated transport is that its rate is saturable. Facilitated diffusion differs from simple diffusion through an open channel in the following important way: although the rate of diffusion through an open channel increases proportionately with the concentration gradient of the diffusing substance, in facilitated diffusion if the concentration gradient of a substance is progressively increased, the rate of transport of the substance will increase up to a certain point and then level off. Further increases in the gradient will produce no additional increase in rate. The reason for this is that there is a limited number of carriers in the membrane. When the concentration of the transported substance is raised high enough, all of the carriers will be in use and the capacity of the transport system will be saturated. This difference between simple diffusion and facilitated diffusion is demonstrated in fig. 10.6, showing that as the concentration gradient of the substance increases, the rate of continues to increase proportionately, but there is limitation of facilitated diffusion to the $v_{max}$ level.

Carrier-mediated diffusion has **three essential characteristics**:

– it is specific, with only certain molecules or ions transported by a given carrier;

– the direction of net movement being determined by the relative concentrations of the transported substance inside and outside the cell;
it may become saturated if all of the protein carriers are in use.

Among the most important substances that cross cell membranes by facilitated diffusion are glucose and most of the amino acids.

![Facilitated diffusion vs Simple diffusion graph](image)

**Fig. 10.6.** The transfer rate $\nu$ dependence on the transported molecule concentration difference $\Delta C$ across the membrane in simple and facilitated diffusion

10.4. **Mathematical Description of the Passive Transport**

**Electrochemical potential** $\mu$ is the free energy of one mole of solution. Free energy is the thermodynamic potential, which determines the ability of a physical-chemical system to perform useful work. All useful work that can be done in one mole of a substance is due to decrease of its electrochemical potential. For solutions of substances electrochemical potential can be expressed as

$$\mu = \mu_0 + RT\ln C + ZF\theta,$$

(10.1)

$\mu_0$ is the part of chemical potential of one mole of solution which is determined by the energy of chemical bonds of the solute with the solvent; $R$ is the universal gas constant; $T$ is the absolute temperature of the solution; $C$ is the molar concentration of the solute; $Z$ is the electric charge of the dissolved ions, which is expressed in units of electron charge; $F$ is the Faraday number; $\theta$ is the electric potential of the solution.

Let’s imagine that the membrane separates two solutions of identical composition but different ion concentration (fig. 10.7). If the values of electrochemical potential on both sides of the membrane are different $\mu_e \neq \mu_i$, then the system is thermodynamic non-equilibrium. In this case electrochemical potential gradient appears across the membrane: $d\mu/dx = \Delta \mu/d$, where $d$ is the membrane thickness.

The thermodynamic equilibrium of system is characterized by the equality of thermodynamic potentials including electrochemical ones: $\mu_e = \mu_i$. The process of transition from the non-equilibrium to equilibrium state in the case of biological membranes is always accompanied by substance diffusion.
from the region of greater value of the electrochemical potential into the region with its lower value.

![Diagram](image)

*Fig. 10.7. Connection between the diffusion flux direction and the electrochemical potential distribution across the membrane*

Mathematically, the process of substance transfer is described by the Theorell’s equation:

$$\Phi = -CU \frac{d\mu}{dx},$$  \hspace{1cm} (10.2)

where $\Phi$ is the diffusion flux density (amount of substance transported through the unit membrane area per second); $C$ is molar concentration of the solution; $U$ is the mobility; $d\mu/dx$ is the electrochemical potential gradient.

Let’s find the electrochemical potential gradient $d\mu/dx$. Taking into account that on both sides of biomembranes solvent is always the same — water, so $\mu_0 = \mu_e = \text{const}$, one can obtain:

$$\frac{d\mu}{dx} = RT \frac{1}{C} \frac{dC}{dx} + ZF \frac{d\phi}{dx}. \hspace{1cm} (10.3)$$

Let’s substitute this expression (3) in (1) and write Nernst–Planck equation describing diffusion of ions across the membrane:

$$\Phi = -URT \frac{dC}{dx} - CUZF \frac{d\phi}{dx}. \hspace{1cm} (10.4)$$

First summand in this equation describes diffusion which is due to the concentration gradient $dC/dx$, second summand describes electrodiffusion which is due to electric potential gradient $d\phi/dx$ through membrane.
In case of the uncharged particles diffusion ($Z = 0$) the second term in the equation (10.4) vanishes and the passive transport of such substances is described by *Fick Law*:

$$\overline{\Phi} = -D \frac{dC}{dx},$$

where $D$ is the diffusion coefficient. The diffusion coefficient depends on the mobility of the substance $U$ and the absolute temperature of the medium $T$:

$$D = URT.$$  

(10.6)

It is possible to simplify Fick equation, if concentration gradient is expressed as

$$\frac{dC}{dx} \sim \frac{\Delta C}{\Delta x} = \frac{|C_i - C_e|}{d},$$

(10.7)

where $d$ is the membrane thickness; $C_i$ and $C_e$ are concentration absolute values on the interior and exterior membrane surfaces:

$$\Phi = p \frac{|C_i - C_e|}{d},$$

(10.8)

where coefficient $p = D/d$ is *permeability coefficient*.

**10.5. ACTIVE TRANSPORT OF IONS**

Active transport typically moves molecules or ions through a membrane from the region of low concentration to high one in the direction of the electrochemical potential increase. Passive transport of substances has always gone from the region of large values of the electrochemical potential to the region of its lower values, resulting in electrochemical potential gradient decrease. *Active transport* of substances is going in the opposite direction and leads to an increase of the electrochemical potential difference on both side of the membrane, so energy is required. Active transport is mediated by carrier proteins that undergo conformational changes in order to move substance across membranes. Many of the carrier proteins involved in active transport are referred to as pumps. The ATP-dependent pump uses the energy derived from adenosine triphosphate (ATP) hydrolysis to adenosine diphosphate (ADP) and inorganic phosphate ($P_i$):

$$ATP \rightarrow ADP + P_i + E (E = 45 \text{ KJ/mol}).$$

Active transport of substances can be divided into two types:

1) active transport of ions;

2) active transport of organic compounds, mainly amino acids and carbohydrates.

All ATP-dependent pumps (ATPases) share a common feature. They transport substances from the side where they are less concentrated to the side where they are more concentrated by utilizing the free energy associated with
ATP hydrolysis. There are several types of ATPases, and they function by distinct mechanisms. The best-studied ATPase is the Na, K-ATPase, also known as the sodium-potassium pump.

The sodium-potassium (Na-K) pump is a transport process that pumps sodium ions outward through the cell membrane of all cells and at the same time pumps potassium ions from the outside to the inside (fig. 10.8). This pump is responsible for maintaining the sodium and potassium concentration differences across the cell membrane as well as for establishing a negative electrical voltage inside the cells.

During one cycle of pumping, the sodium-potassium pump exports three ions of sodium and imports two ions of potassium through the cell membrane.

After binding sodium ions on the interior of the cell, ATP is hydrolyzed and the phosphate is transferred to the pump protein. The phosphorylated protein undergoes a conformational change, delivering the sodium ions to the exterior of the cell and exchanging them for potassium ions. The protein is then dephosphorylated and undergoes an additional conformational change, returning to its original state and delivering the potassium ions to the interior of the cell. The fact that the Na-K pump moves three Na\(^+\) ions to the exterior for every two K\(^+\) ions to the interior means that one positive charge is moved from the interior of the cell to the exterior for each cycle of the pump. This creates positivity outside the cell but leaves a deficit of positive ions inside the cell; that is, it causes negativity on the inside.

Fig. 10.8. Scheme of the sodium-potassium pump
Questions:
1. Characterize the physical properties of cell membrane lipids. What are the functions of the cell membrane lipids and proteins?
2. Describe types of the lipids and proteins motion in the cell membrane (lateral diffusion, rotational diffusion, flip-flop).
3. What types of passive transport across the cell membrane are known?
4. Which substances can move across cell membranes? Specify membrane channels properties.
5. What is a facilitated diffusion? Describe the types of facilitated diffusion.
6. What is the meaning of electrochemical potential? Write the Theorell equation, Nernst-Planck equation and Fick Law. What is the cell membrane permeability?
7. What is the active transport? Explain ions active transport mechanism by the sodium-potassium pump.

Chapter 11. MEMBRANE POTENTIALS OF THE CELL

The cell membrane acts as a barrier which prevents the intracellular fluid from mixing with the extracellular fluid. These two solutions have different concentrations of their ions. Furthermore, this difference in concentrations leads to a difference in charge of the solutions. This creates a situation whereby one solution is more positive than the other. The membrane potential ($\theta_m$) of an excitable cell is defined as the potential at the inner surface ($\theta_i$) relative to that at the outer ($\theta_e$) surface of the membrane, i.e. $\theta_m = (\theta_i) - (\theta_e)$. This definition is independent of the cause of the potential, and whether the membrane voltage is constant, periodic, or nonperiodic in behavior. If the potential outside is taken to be zero, then the interior resting membrane potential varies from $-60$ mV to $-100$ mV depending on the type of cell.

A nerve cell conducts an electrochemical impulse because of membrane potential changes. These changes allow movement of ions through the membrane, setting up currents that flow through the membrane and along the cell. Similar impulses travel along muscle cells before they contract.

11.1. THE NERNST EQUATION

Find the equilibrium membrane potential, which arises due to the diffusion of ions through the cell membrane. Suppose that in a rest the membrane is permeable to one type of ion ($K^+$). The concentration of potassium ions is higher inside cells than outside due to the active transport of potassium ions. Most cells have potassium-selective ion channel proteins that remain open all the time. There will be net movement of positively-charged potassium ions through these potassium channels with a resulting accumulation of excess negative charge inside of the cell. The outward movement of positively-charged potassium ions is due to its diffusion and continues until enough excess negative charge accumulates inside the cell to form a membrane potential which can balance
the difference in concentration of potassium between inside and outside the cell. «Balance» means that the electrical potential that results from the build-up of ionic charge, and which impedes outward diffusion, increases until it is equal in magnitude but opposite in direction to the tendency for outward diffusive movement of potassium. This balance point (the equilibrium state) is characterized by the equality of electrochemical potentials on both sides of the membrane \( \mu_e = \mu_i \) and the net transmembrane flux (or current) of \( K^+ \) is zero (\( \Phi_{K^+} = 0 \)).

The electrochemical potential inside cell can be written as:

\[
\mu_i = \mu_{0i} + RT \ln C_i + ZF \phi_i
\]

and the electrochemical potential outside is:

\[
\mu_e = \mu_{0e} + RT \ln C_e + ZF \phi_e.
\]

The chemical potential of the water is the same on both sides \( \mu_{0i} = \mu_{0e} \), and condition of the equilibrium state has the form:

\[
RT \ln C_i + ZF \phi_i = RT \ln C_e + ZF \phi_e.
\]

This equation can be rearranged to give:

\[
RT(\phi_i - \phi_e) = RT(\ln C_i - \ln C_e).
\]

The Nernst equation for equilibrium membrane potential is obtained from the last equation:

\[
\phi_i - \phi_e = -\frac{RT}{ZF} \ln\frac{C_i}{C_e},
\]

where \( \phi_i - \phi_e \) is the equilibrium potential for ion; \( R \) is the universal gas constant; \( T \) is the absolute temperature; \( Z \) is the number of elementary charges; \( F \) is the Faraday constant; \( C_e \) is the extracellular concentration of ion; \( C_i \) is the intracellular concentration of ion.

The equilibrium potential for a given ion depends only upon the concentrations on either side of the membrane and the temperature. The Nernst equation is widely used in physiology to relate the concentration of ions on either side of a membrane to the electrical potential difference across the membrane. Usually the outside solution is set as the zero voltage (\( \phi_e = 0 \)). Then the difference between the inside voltage and the zero voltage is determined. At physiological temperature, about 29.5 °C, and physiological concentrations (which vary for each ion), the calculated equilibrium potentials are approximately +67 mV for Na\(^+\), +90 mV for K\(^+\), –86 mV for Cl\(^-\) and +123 mV for Ca\(^{2+}\).
11.2. RESTING MEMBRANE POTENTIAL

In mammalian cells sodium Na\(^+\), potassium K\(^+\) and chloride Cl\(^-\) ions play large roles for the resting membrane potential. The resting membrane potential is determined by the equilibrium potentials for every ion to which the membrane is permeable, weighted by the permeability (P), via the Goldman–Hodgkin–Katz voltage equation:

\[
\varphi_m = \frac{-RT}{F} \ln \frac{P_K C_i(K^+) + P_{Na} C_i(Na^+) + P_{Cl} C_i(Cl^-)}{P_K C_e(K^+) + P_{Na} C_e(Na^+) + P_{Cl} C_e(Cl^-)},
\]

(11.2)

where \(R\), \(T\), and \(F\) are as above; \(P_K\), \(P_{Na}\), \(P_{Cl}\) are the membrane permeabilities for \(K^+\), \(Na^+\), \(Cl^-\) ions, respectively; \(C_e(K^+)\), \(C_e(Na^+)\), \(C_e(Cl^-)\) are the extracellular concentrations for \(K^+\), \(Na^+\), \(Cl^-\) ions, respectively; \(C_i(K^+)\), \(C_i(Na^+)\), \(C_i(Cl^-)\) are the intracellular concentrations for \(K^+\), \(Na^+\), \(Cl^-\) ions, respectively.

If the permeabilities of \(Na^+\) and \(Cl^-\) are zero, the membrane potential reduces to the Nernst potential for \(K^+\) (as \(P_K^+ = P_{tot}\)). Usually, under resting conditions \(P_{Na^+}\) and \(P_{Cl^-}\) are not zero, but they are much smaller than \(P_K^+\), which renders \(\varphi_m\) close to the equilibrium potential for potassium. Normally, permeability values are reported as relative permeabilities with \(P_K\) having the reference value of one (because in most cells at rest \(P_K\) is larger than \(P_{Na}\) and \(P_{Cl}\)). Hodgkin and Katz experimentally found that for the giant axon of squid the attitude of the membrane permeability for \(K^+\), \(Na^+\) and \(Cl^-\) ions in a rest is \(P_K : P_{Na} : P_{Cl} = 1 : 0.04 : 0.45\). Medical conditions such as hyperkalemia in which blood serum potassium (which governs \([K^+]_o\)) is changed are very dangerous since they offset the equilibrium potential for potassium, thus affecting resting membrane potential \(\varphi_m\). This may cause arrhythmias and cardiac arrest.

Because the electric field in the resting cell is zero, there is no net charge in the fluid. Positive ions are neutralized by negative ions everywhere except at the membrane. A layer of charge on each surface generates an electric field within the membrane and a potential difference across it. Measurements with a microelectrode show that the potential within the cell is about 60–100 mV less than outside. If the potential outside is taken to be zero, then the interior resting potential is \((-60) – (-100)\) mV. If the potential drops 80 mV and if the membrane thickness is 8 nm, then the electric field within the membrane is assumed to be constant:

\[
E = \frac{\varphi_0}{d} = \frac{80 \text{ mV}}{8 \text{ nm}} = \frac{80 \cdot 10^{-2} \text{ V}}{8 \cdot 10^{-9} \text{ m}} = 10^7 \text{ V/m}. \tag{11.3}
\]

11.3. ACTION POTENTIAL IN EXCITABLE CELLS

All cells exhibit a potential difference across the cell membrane. Nerve cells and muscle cells are excitable. They have the ability to generate and propagate electrical signals. The origin of the membrane potential is the same in
nerve cells as in muscle cells. In both cell types, the membrane generates an impulse as a consequence of excitation. This impulse propagates in both cell types in the same manner.

**An action potential** is the brief reversal in the potential difference across a plasma membrane (as of a nerve cell or muscle fiber) that occurs when a cell has been activated by a stimulus (fig. 11.1).

![Schematic representation of an action potential in an excitable cell](image)

Fig. 11.1. Schematic representation of an action potential in an excitable cell

The course of the action potential can be divided into five parts: the **rising phase**, the **peak phase**, the **falling phase**, the **undershoot phase**, and finally the **refractory period**. When the excitable cell membrane is stimulated so that the membrane potential rises and reaches the threshold, the sodium and potassium ionic permeabilities of the membrane change. The sodium ion permeability increases very rapidly at first, allowing sodium ions to flow from outside to inside, making the inside more positive. During the rising phase the membrane potential \( \phi_m \) depolarizes (becomes more positive). The sharp rise in membrane potential and sodium permeability correspond to the rising phase of the action potential. Hodgkin and Katz experimentally found that for the giant axon of squid the attitude of the membrane permeability for \( K^+ \), \( Na^+ \) and \( Cl^- \) ions **during the rising phase** is \( P_K : P_{Na} : P_{Cl} = 1 : 20 : 0.45 \).

The point at which depolarization stops is called the peak phase. At the peak of the action potential, the sodium permeability is maximized and the membrane potential is nearly equal to the sodium equilibrium voltage \( \phi_{Na} \). At this stage, the membrane potential reaches a maximum.

Subsequent to this, there is a falling phase. The same raised voltage that opened the sodium channels initially also slowly shuts them off, by closing their
pores; the sodium channels become inactivated. This lowers the membrane’s permeability to sodium, driving the membrane potential back down. At the same time, the raised voltage opens voltage-sensitive potassium channels; the increase in the membrane’s potassium permeability drives the membrane potential \( \phi_m \) towards the potassium equilibrium voltage \( \phi_K \). The efflux of potassium ions decreases the membrane potential thus returning the membrane potential to its resting value or hyperpolarizes the cell. Combined, these changes in sodium and potassium permeability cause the membrane potential \( \phi_m \) to drop quickly, repolarizing the membrane and producing the «falling phase» of the action potential.

The raised voltage opened many more potassium channels than usual, and these do not close right away when the membrane returns to its normal resting voltage. The potassium permeability of the membrane is transiently unusually high, driving the membrane potential \( \phi_m \) even closer to the potassium equilibrium voltage \( \phi_K \). Hence, there is an undershoot, a hyperpolarization, that persists until the membrane potassium permeability returns to its usual value. The undershoot phase is the point during which the membrane potential becomes temporarily more negatively charged than when at rest. While at rest, following activation, the Na-K pump restores the ion concentrations inside and outside the membrane to their original values.

Each action potential is followed by a refractory period, which can be divided into an absolute refractory period, during which it is impossible to evoke another action potential, and then a relative refractory period, during which a stronger-than-usual stimulus is required. These two refractory periods are caused by changes in the state of sodium and potassium channel molecules. When closing after an action potential, sodium channels enter an «inactivated state», in which they cannot be made to open regardless of the membrane potential — this gives rise to the absolute refractory period. Even after a sufficient number of sodium channels have transitioned back to their resting state, it frequently happens that a fraction of potassium channels remains open, making it difficult for the membrane potential to depolarize, and thereby giving rise to the relative refractory period. Because the density and subtypes of potassium channels may differ greatly between different types of neurons, the duration of the relative refractory period is highly variable.

Duration of the depolarization is small in any cases. For nerve cells and muscle cells this duration is 0.5–1 ms. Duration of the repolarization depends essentially on the type of cells: for the nerve cells and skeletal muscle cells duration of the repolarization is 0.5–10 ms, for the heart muscle cells — about 300 ms.

The action potential amplitude is equal to the sum of absolute values of the resting potential \( \phi_0 \) and the maximum achieved membrane potential \( \phi_{\max} \) and is \( \sim 90–120 \) mV:

\[ \phi_a = \phi_{\max} - \phi_0 = \phi_{\max} + |\phi_0|. \]  (11.4)
Currents produced by the opening of voltage-gated channels in the course of an action potential are typically significantly larger than the initial stimulating current. Thus the amplitude, duration, and shape of the action potential are largely determined by the properties of the excitable membrane and not the amplitude or duration of the stimulus. The action potentials are generated anew along excitable stretches of membrane and propagate without decay.

11.4. PROPAGATION OF ACTION POTENTIAL ALONG AN UNMYELINATED AXON

The nerve cell may be divided on the basis of its structure and function into three main parts:
- the cell body, also called the soma;
- numerous short processes of the soma, called the dendrites;
- the single long nerve fiber, the axon.

The long nerve fiber, the axon, transfers the signal from the cell body to another nerve or to a muscle cell. The long cylindrical axon has properties that are in some ways similar to those of an electric cable. Its diameter may range from less than one micrometer (1 μm) to as much as 1 mm for the giant axon of a squid; in humans the upper limit is about 20 μm. Pulses travel along it with speeds ranging from 0.6 to 100 m s⁻¹, depending, among other things, on the diameter of the axon. The axon core may be surrounded by either a membrane (for an unmyelinated fiber) or a much thicker sheath of fatty material (myelin) that is wound on like tape.

Let’s consider the action potential propagation along unmyelinated axon. At resting potential there is positive charge on the outside of axon membrane and negative charge on the inside, with high sodium ion concentration outside and high potassium ion concentration inside (fig. 11.2).

![Fig. 11.2. Unmyelinated axon in a rest. There is no net transport of the ion through the membrane](image)

If the membrane stimulus is insufficient to cause the membrane potential to reach the threshold, then the membrane will not activate. The response of
the membrane to this kind of stimulus is essentially passive. If the excitatory stimulus is strong enough, the membrane potential reaches the threshold, and the membrane produces a characteristic electric impulse, the nerve impulse. This potential response follows a characteristic form regardless of the strength of the transthreshold stimulus. When stimulated, voltage-dependent sodium ion channels open, and sodium ions flow into the axon, depolarizing the membrane. The potential difference \((\varphi_{\text{max}} - \varphi_0)\) between excited and unexcited regions of an axon would cause small currents, called local circuit currents, to flow between them in such a direction that they stimulate the unexcited region (fig. 11.3).

![Diagram of sodium ion channels and local electric current](image)

**Fig. 11.3.** The propagation of action potential along an unmyelinated axon

Meanwhile, in the earlier excited region potassium ions leave the axon, repolarizing the membrane. The currents flowing inwards at a point on the axon during an action potential spread out along the axon, and depolarize the adjacent sections of its membrane (fig. 11.4). The action potential generated at the axon propagates as a wave along the axon.
An important physical property of the axon membrane is the change in sodium conductance due to activation. The higher the maximum value achieved by the sodium conductance, the higher the maximum value of the sodium ion current and the higher the rate of change in the membrane voltage. The result is a higher gradient of voltage, increased local currents, faster excitation, and increased conduction velocity. The decrease in the threshold potential facilitates the triggering of the activation process. Conduction speed can be increased by reducing the internal resistance of the axon. Impulse transmission can be speeded up by increasing the diameter of the axon. However, there are limitations on the size of an axon. Transmission speed can reach 25 m per sec if the diameter of the unmyelinated axon is 1 mm.

**11.5. Propagation of Action Potential Along a Myelinated Axon**

The evolutionary need for the fast and efficient propagation of electrical signals in nervous system resulted in appearance of myelin sheaths around neuronal axons. The myelin sheath is not continuous but divided into sections with the size of 2–3 mm, separated at regular intervals by the nodes of Ranvier with the length of 1μm. A typical human nerve might contain twice as many unmyelinated fibers as myelinated. The myelin gives a faster impulse conduction speed for a given axon radius.
A myelinated axon, surrounded by the myelin sheath, can produce a nerve impulse only at the nodes of Ranvier. This myelin sheath makes the axon impermeable to ions so they are unable to diffuse between the tissue fluid and the neurone, so action potentials cannot be generated by the myelinated regions (it acts as an insulator). Action potentials can only be generated at the nodes of Ranvier, so the local currents involved in nerve impulse transmission flow over longer distances. An action potential at one node of Ranvier causes inwards currents that depolarize the membrane at the next node, provoking a new action potential there; the action potential appears to «jump» from node to node. Thus action potential seems to «jump» from node to node, as illustrated in fig. 11.5. Since the intervening parts of the axon membrane do not have to be successively depolarised it takes less time for the action potentials to pass from node to node. This results in nerve impulse transmission that is much faster, the consequence of which is that smaller myelinated nerves can transmit impulses much faster than larger unmyelinated ones (120 m/sec compared to 25 m/sec along unmyelinated axon). Another advantage of this is that energy is saved as sodium potassium pumps are only required at specific points along the axon. Such a propagation is called saltatory conduction. The process of excitation and conduction in myelinated nerve fibers is characterized by its discontinuous and saltatory features.

Fig. 11.5. Propagation of action potential along a myelinated axon

The cytoplasm of an axon is electrically conduction and because myelin inhibits charge leakage through the membrane, depolarization at one node of Ranvier is sufficient to elevate the voltage at a neighboring node to the threshold for action potential initiation. Even if one node is damaged, transmission can still effectively bypass that node. Nodes of Ranvier contain a significantly higher density of voltage gated sodium channels than is found in unmyelinated axons (4 orders of magnitude higher).
**Questions:**
1. How is resting membrane potential generated?
2. Obtain the Nernst Equation.
4. What is the condition of cell excitement?
5. What processes occur in cell membrane during action potential generation?
6. Characterize depolarization phase and repolarization one. Plot the graph for action potential.
7. What does determine the sodium channel permeability?
8. What are the refractory periods? Describe the types and duration of the refractory periods for different cells.
9. Describe the propagation of action potential along an unmyelinated axon.
10. Characterize the propagation of action potential along a myelinated axon.

**Chapter 12. ELECTRICAL FIELDS OF THE ORGANS AND TISSUES. METHODS OF THEIR REGISTRATION**

**12.1. ELECTRICAL FIELD AND ITS CHARACTERISTICS**

There are two types of observed electric charge, which are designated as positive and negative. The symbol for charge is \( q \). The SI unit of charge is **coulomb** (C). The charge of electron is the smallest charge found in nature, it is given the symbol \( e \) and is often referred to as the **elementary charge**. Electron \( e \) has the charge magnitude \( 1.6 \times 10^{-19} \) C. According to the law of conservation of electric charge, the total charge in an isolated system always remains constant. The total electric charge of a system is equal to the algebraic sum of electric charges located in the system.

When the charges are likely there is a repulsive force between them and, opposite, when the charges are unlikely, there is attractive force between them. The force between two charged bodies was studied by Coulomb. **Coulomb’s Law** states that the electrostatic force \( F \) of attraction or repulsion in vacuum between two point charges \( q_1 \) and \( q_2 \) is directly proportional to the product of the magnitude of the charges and inversely proportional to the square of the distance \( r \) between them:

\[
F = k \frac{q_1 q_2}{r^2},
\]

where \( k = 9 \times 10^9 \frac{N \cdot m^2}{C^2} \) is a proportionality constant.

The direction of forces is always along the line joining the two point charges, and it is attractive if the charges are opposite and repulsive if the charges are like (fig. 12.1).
Fig. 12.1. Direction of the interaction force between two charges depends on whether the charges have the same or opposite signs. The force is attractive if the charges are of different signs \((a)\), and repulsive if they have the same sign \((b)\).

Electric field is said to exist in the region of space around a charged object: the source is a charge. The presence of an electric field around a charge cannot be detected unless another charge is brought towards it. When a test charge \(q_0\) is placed near a charge \(q\), which is the source of electric field, an electrostatic force \(F\) will act on the test charge (fig. 12.2).

\[
F \sim q_0, \quad \text{or} \quad F = \vec{E} q_0 \quad (12.2)
\]

Thus, the ratio

\[
\vec{E} = \frac{
\vec{F} \quad q_0
}{q_0} \quad (12.3)
\]

does not depend on the point charge \(q_0\) and is called the **electric field strength** or **electric field intensity**.

The electric field strength \(\vec{E}\) at a point in space is a vector whose direction is the direction of the force acting on a positive test charge \(q_0\) placed at that point, and whose magnitude is the force per unit charge. Thus \(\vec{E}\) has SI units of newton per coulomb \((N/C)\).

Since the electric field strength \(\vec{E}\) is a vector, it is sometimes referred to as vector field. Lines of electric field strength are a convenient way of visualizing the electric field. The electric field strength lines indicate the direction of the force due to the given field on a positive test charge. For a positive point
charge, the electric field strength lines are directed radially outward from the charge (fig. 12.3, a).

For a negative point charge they point radially inward toward the charge because that is the direction the force would be on a positive test charge in each case (fig. 12.3, b). Since the electric field strength is the electric force per unit charge, the electric field strength lines are sometimes called lines of force.

The electric field strength $E$ at a distance $r$ from single point charge $q$ can be written as:

$$E = \frac{F}{q_0} = k \frac{q_0 q}{r^2} = k \frac{q}{r^2}. \quad (12.4)$$

If the electric field strength $\vec{E}$ is due to more than one charge ($q_1$, $q_2$, $q_3$, ..., $q_n$), the individual electric field strengths fields (call them $E_1$, $E_2$, ..., $E_n$) due to each charge are added vectorially to get the total electric field strength $\vec{E}$ at any point. It is the principle of superposition:

$$\vec{E} = \vec{E}_1 + \vec{E}_2 + \vec{E}_3 + \ldots + \vec{E}_n. \quad (12.5)$$

The work $A$ done by an electric force $F$ or «field» in moving a positive test charge $+q_0$ along the electric field line at a distance $d$ is:

$$A = F \cdot d = q_0 E \cdot d. \quad (12.6)$$

Therefore, any test charge in electric field is said to have an electric potential energy: $W_{pot} = A \sim q_0$, which is directly proportional to the magnitude of the charge $q_0$. The electric potential energy depends upon the charge placed in the electric field. To quantify the potential energy in terms of only the field itself it is more useful to define it per unit charge:

$$\varphi = \frac{W_{pot}}{q_0}. \quad (12.7)$$
**Electric potential** $\varphi$ is potential energy per unit charge. Electric potential $\varphi$ is a scalar characteristic of an electric field, independent of any other charges. Unit of electric potential $\varphi$ is Volt (V) (1 Volt = 1 Joule per Coulomb (J/C)).

The potential from a collection of $n$ charges is the algebraic sum of the potential due to each charge separately (this is much easier to calculate than the net electric field, which would be a vector sum). Potential due to a group of point charges:

$$\varphi = \varphi_1 + \varphi_2 + \varphi_3 + \ldots + \varphi_n. \quad (12.8)$$

If all the points of a surface are at the same electric potential, then the surface is called an **equipotential surface**.

In case of an isolated point charge, all points equidistant from the charge are at the same potential. Thus, equipotential surfaces in this case will be a series of concentric spheres with the point charge as their centre (fig. 12.4). The potential will however be different for different spheres.

![Fig. 12.4. The equipotential lines and electric field lines](image)

If the charge is to be moved between any two points on an equipotential surface through any path, the work done is zero. This is because the potential difference between two points $A$ and $B$ is defined as $\varphi_B - \varphi_A = W_{AB} q$. If $\varphi_B = \varphi_A$ then $W_{AB} = 0$. Hence the electric field lines must be normal to an equipotential surface.

Let $+q$ be an isolated point charge situated in air at point $O$; $B$ is a point at a distance $r$ from $+q$. The electric potential $\varphi$ at the point $B$ due to the charge $+q$ is the total work done in moving a unit positive charge from infinity to that point:

$$\varphi = k \frac{q}{\varepsilon r}. \quad (12.9)$$
Suppose charge $q_0$ is moved from point 1 to point 2 through a region of space described by electric field $E$ (fig. 12.5).

![Diagram of electric field with path from point 1 to point 2](image)

*Fig. 12.5. The charge $q_0$ is moved from point 1 to point 2 through a region of space described by electric field $E$.*

The work $A$ done by the electric field in bringing the charge $q_0$ from point 1 to point 2 can be written:

$$A = W_{pot1} - W_{pot2} = q_0 (\phi_1 - \phi_2) = q_0 U. \quad \text{(12.10)}$$

The potential difference $U$ between point 1 and point 2 is:

$$U = \phi_1 - \phi_2 = E \cdot d. \quad \text{(12.11)}$$

**The potential difference** between two points in an electric field is defined as the amount of work done in moving a unit positive charge from one point to the other against the electric force. The unit of potential difference $U$ is **Volt**. The potential difference between two points is 1 Volt if 1 joule of work is done in moving 1 Coulomb of charge from one point to another against the electric force. The electric potential in an electric field at a point is defined as the amount of work done in moving a unit positive charge from infinity to that point against the electric forces. The work $A$ does not depend upon the exact path chosen to move charge from point 1 to point 2 and is determined by the potential difference between point 1 and point 2.

The electric field is equal to the negative gradient of potential:

$$\vec{E} = - grad \phi. \quad \text{(12.12)}$$

The negative sign in the formula (12.12) indicates that the electric field is pointing to the direction of potential decrease. The unit of electric intensity can also be expressed as V·m$^{-1}$. 

118
12.2. ELECTRIC DIPOLE AND ITS FIELD

A system of two equal and opposite \((q^+ \text{ and } q^-)\) charges separated by a certain distance \(l\) is called an electric dipole (fig. 12.6). It has an electric dipole moment \(\vec{P}\). \(\vec{P}\) is a vector that points from the negative charge to the positive charge.

![Diagram of Electric Dipole](image)

**Fig. 12.6.** A system of two equal and opposite \(q^+\) and \(q^-\) charges separated by a certain distance \(l\)

**Dipole moment** \(\vec{P}\) is a measure of the strength of electric dipole. The magnitude of the dipole moment is given by the product of the magnitude of the one of the charges \(q\) and the distance \(l\) between them:

\[
\vec{P} = q\vec{l}.
\]

The unit of dipole moment \(\vec{P}\) is C m.

Let \(A\) be the point at a distance \(r\) from the midpoint of the dipole \(O\) and \(\delta\) be the angle between \(AO\) and the axis of the dipole. Let \(r_1\) and \(r_2\) be the distances of the point \(A\) from \(+q\) and \(-q\) charges respectively (fig. 12.7).

![Diagram of Dipole Field](image)

**Fig. 12.7.** Potential of dipole field. \(r_1\) and \(r_2\) are the distances of the point \(A\) from \(+q\) and \(-q\) charges respectively

Total potential \(\varphi_A\) at point \(A\) due to dipole is an algebraic sum of potential \(\varphi_+\) at point \(A\) due to charge \((+q)\) and potential \(\varphi_-\) at point \(A\) due to charge \((-q)\):

\[
\varphi_A = \varphi_+ + \varphi_- = k \frac{q}{\varepsilon r_1} - k \frac{q}{\varepsilon r_2} = k \frac{q}{\varepsilon} \left( \frac{1}{r_1} - \frac{1}{r_2} \right) \approx k \frac{q}{\varepsilon} \frac{(r_2 - r_1)}{r_2^2}.
\]
Rewrite this for special case $r >> l$: $(r_2 - r_1) \approx l \cos \delta$.

Than, electric potential at a point due to an electric dipole is determined by formula:

$$\varphi_A = k \frac{ql \cos \delta}{\varepsilon r^2} = k \frac{p \cos \delta}{\varepsilon r^2}. \quad (12.14)$$

Electric field lines and equipotential surfaces for the dipole are illustrated on fig. 12.8.

![Electric field lines and equipotential surfaces for the dipole](image)

*Fig. 12.8. Electric field lines and equipotential surfaces of the dipole*

The dashed lines represent equipotential (points of equal potential) lines and the solid lines are the electric field lines.

The potential difference $U_{AB}$ between point $A$ and point $B$ created by dipole (fig. 12.9) is directly proportional to the projection of the dipole moment $P \cos \alpha$ on the line connecting these points and inversely proportional to the square of the distance $r$ between the dipole and line:

$$U_{AB} = 2k \frac{p \sin \beta \cos \alpha}{\varepsilon r^2} \frac{p \cos \alpha}{r^2}. \quad (12.15)$$

![Connection between the dipole moment $P$ and potential difference $U_{AB}$](image)

*Fig. 12.9. Connection between the dipole moment $P$ and potential difference $U_{AB}$. Points $A$ and $B$ are located at a distances $r_A$ and $r_B$ from dipole with the dipole moment $P$*. 

120
12.3. ELECTROCARDIOGRAPHY

An electrocardiogram (ECG) is a recording of electrical activity (potentials) produced by the conduction system and the myocardium of the heart during its depolarization made from electrodes placed on the surface of the skin.

The beating heart generates an electric signal that can be used as a diagnostic tool for examining some of the functions of the heart. This electric activity of the heart can be approximately represented as a vector quantity (an equivalent dipole) called the electric heart vector (fig. 12.10). The heart consists of an electric dipole with dipole moment $P$ located in the partially conducting medium of the thorax. In electrocardiography this dipole moment is known as the cardiac vector. As we progress through a cardiac cycle, the magnitude and direction of $P$ vary because the dipole field varies.

![Electric heart vector](image)

*Fig. 12.10. The heart can be approximately represented as a vector quantity (an equivalent dipole) called the electric heart vector with dipole moment $P$ located in the partially conducting medium of the thorax.*

In 1901 a Dutch physiologist, Willem Einthoven, developed a galvanometer that could record the electrical activity of the heart. He found that a tracing can be produced as action potentials spread between negatively and positively charged electrodes. A third electrode serves to ground the current. He found that tracings varied according to the location of the positive and negative electrodes, and subsequently described three angles or leads in the form of a triangle with the heart in the middle. This is known today as Einthoven’s triangle, and the three electrode arrangements are known as the standard limb leads I, II, and III (fig. 12.11). If the two electrodes are located on different equal-potential lines of the electric field of the heart, a nonzero potential difference or voltage is measured.

Thus, Einthoven’s triangle is a hypothetical triangle created around the heart when electrodes are placed on both arms and the left leg. The heart is considered to be at the center of a equilateral triangle, each corner of which serves as the location for an electrode for two leads to the ECG recorder. The three standard limb leads are I, II, and III. Lead I records the potential
difference between right arm and left arm. **Lead II** records the potential difference between right arm and left leg. **Lead III** records the potential difference between left arm and left leg. Leads I, II and III are referred to as bipolar leads, because the measured signal is the difference in potential between two electrodes.

According to **Einthoven’s Law** these lead voltages have the following relationship:

\[ U_{II}(t) = U_{I}(t) + U_{III}(t). \]  

As research continued throughout the 20th century, additional arrangements were discovered that enable physicians to analyze electrical events as they spread in many directions through the heart. Today, the cardiologist uses a **12-lead ECG system consisting of the following 12 leads**, which are:
- I, II, III — three bipolar limb;
- \( \alpha VR, \alpha VL, \) and \( \alpha VF \) — augmented unipolar limb leads;
- \( V_1, V_2, V_3, V_4, V_5, V_6 \) — six chest leads.

**Three augmented unipolar limb leads** are used for frontal plane measurements (fig. 12.12). Unipolar measurements are made by recording the potential at one electrode with respect to the average of the other two potentials. These are referred to as \( aVR, aVL \) and \( aVF \) (augmented vector right, left, and foot).

For measuring the potentials close to the heart, Wilson introduced **the precordial leads (chest leads)** in 1944. These leads, \( V_1-V_6 \) are located over the left chest as described in the fig. 12.13. Each of the six precordial leads is unipolar (1 electrode constitutes a lead) and is designed to view the electrical activity of the heart in the horizontal or transverse plane.
Cardiac conduction system

To fully appreciate electrical impulses and the information provided by an ECG, let’s review fundamental concepts regarding electrical membrane potentials. All cardiac cell membranes are positively charged on their outer surfaces because of the relative distribution of cations. This resting membrane potential is maintained by an active transport mechanism called the sodium-potassium pump. When the cell is stimulated, ion channels open, allowing a sudden influx of sodium and/or calcium ions and thereby reversing the resting potential. This period of depolarization is very brief because sodium channels close abruptly, denying further influx of sodium. Simultaneously, potassium channels open and allow intracellular potassium to diffuse outward while sodium ions are actively pumped out. This reestablishes a positive charge to the outside of the membrane, a process called repolarization that returns the membrane to its resting membrane potential. The processes of depolarization and repolarization are referred to collectively as an action potential. This event self-propagates as an impulse along the entire surface of a cell and from one cell to another, provided that their membranes are connected.
Mechanical contraction of the heart is caused by the electrical excitation of the myocardial cells. The heart is largely autonomous having the ability to initiate its own beat with a regular period, so that it will continue to beat after being removed from the body. Cells capable of initiating electrical activity are called pacemaker cells, and exist in several places throughout the heart. Only those pacemaker cells with the fastest rate of pacemaker discharge control the electrical activity of the entire heart. The region of tissue with the shortest period of spontaneous electrical activity is the sinoatrial (SA) node which is located on the atrial wall near the junction of the superior vena cava and the right atrium (fig. 12.14). Action potentials are normally generated here at the rate of 60 to 100 per minute. From the SA node, the action potential is propagated from cell to cell through firstly the right atrium, followed closely by the left atrium at a conduction velocity of approximately 1 m s\(^{-1}\) until it reaches the atrioventricular (AV) node. The AV node consists of similar pacemaker-type cells as are found in the SA node but, because they beat spontaneously at a slower rate (approximately 40 to 55 beats per minute), they are paced by the excitation propagating from the SA node. In the event that the SA node is removed or destroyed, or that conduction is slowed through the atria, the cells in the AV node will take over as pacemaker for the ventricles. Conduction through the AV node is at a much slower rate (around 0.05 m s\(^{-1}\)) giving time for the atria to contract and pump blood into the ventricles before the action potential conducts through the ventricles and causes them to contract. Under normal conditions, the AV node is the only electrical connection between the atria and the ventricles with electrical propagation exciting the bundle of His’ which forms the upper portion of the ventricular conduction system and runs down the right side of the septum. This common bundle divides after a short distance into right and left bundle branches. The right branch continues down the right septal wall, and the left perforates the septum and splits into two (or three) further main branches on the left septal wall. All of these branches then continue to subdivide into a complex network of fibres called the Purkinje fibre network, spreading across the endocardial surface of both ventricles and into the subendocardial region of the ventricular myocardium. Due to this extensive arrangement of Purkinje fibres, the septum (except its basal region) is activated first and normally pushes in towards the left ventricular wall. The papillary muscles are also activated early thereby preventing the AV valves from inverting during systole. The fast conduction of approximately 1.5 to 4 m s\(^{-1}\) through the bundle branches and Purkinje fibres cause the entire endocardium to be excited almost simultaneously. The apical regions contract first and the basal regions are usually the last regions to be excited. Excitation spreads outwards through the ventricular wall at a rate of approximately 0.3 to 0.5 m s\(^{-1}\), and the first epicardial region to be excited is the thinnest portion of the right ventricular wall.
sinoatrial node (SA node) is the «pacemaker», located in right atrial wall causes depolarization and contraction of the atria, and depolarization of the AV node; atrioventricular node (AV node) — there is a slight delay at the AV node to allow the atria to contract completely before the ventricles contract, after the delay the bundle of His (AV bundle) depolarizes within the atroventricular septum; Bundle branches (right and left) — carry the depolarization to the purkinje fibers in the right and left ventricles; purkinje fibers — depolarize and contract the ventricles

The ECG tracing
A typical normal signal recorded between two electrodes is shown in fig. 12.15. The baseline of an ECG tracing is called the isoelectric line and denotes resting membrane potentials. The main features of this wave form are identified by the letters P, Q, R, S and T. The shape of these features varies with the location of the electrodes. A trained observer can diagnose abnormalities by recognizing deviations from normal patterns.

The P Wave and Atrial Depolarization. Atrial excitation results from a wave of depolarization that originates in the SA node and spreads over the atria.

The PQ Segment and Atrioventricular Conduction. After the P wave, the ECG returns to the baseline present before the P wave. The ECG is said to be isoelectric when there is no deflection from the baseline established before the P wave. During this time, the wave of depolarization moves slowly through the AV node, the AV bundle, the bundle branches, and the Purkinje system. The isoelectric period between the end of the P wave and the beginning of the QRS complex, which signals ventricular depolarization is called the PQ segment.
The ECG with interval and segment measurements

**The QRS Complex and Ventricular Depolarization.** The QRS complex is larger than P wave because of greater muscle mass of ventricles. The depolarization wave emerges from the AV node and travels along the AV bundle (bundle of His), bundle branches, and Purkinje system; these tracts extend down the interventricular septum. The small downward deflection produced on the ECG is the Q wave. The normal Q wave is often so small that it is not apparent. The wave of depolarization spreads via the Purkinje system across the inside surface of the free walls of the ventricles. The Q, R, and S waves together are known as the QRS complex and show the progression of ventricular muscle depolarization. The duration of the QRS complex is roughly equivalent to the duration of the P wave, despite the much greater mass of muscle of the ventricles. The relatively brief duration of the QRS complex is the result of the rapid, synchronous excitation of the ventricles.

**The ST Segment.** The ST segment is the period between the end of the S wave and the beginning of the T wave. The ST segment is normally isoelectric, or nearly so.

**The T Wave and Ventricular Repolarization.** Repolarization, like depolarization, generates a dipole because the voltage of the depolarized area is different from that of the repolarized areas. The T wave has a longer duration than the QRS complex because repolarization does not proceed as a synchronized, propagated wave. Instead, the timing of repolarization is a function of properties of individual cells, such as numbers of particular K⁺ channels.

**The QT Interval.** The QT interval is the time from the beginning of the QRS complex to the end of the T wave. If ventricular action potential and QT interval are compared, the QRS complex corresponds to depolarization, the ST segment to the plateau, and the T wave to repolarization (fig. 12.16). The relationship between a single ventricular action potential and the events of the QT interval are approximate because the events of the QT interval represent
the combined influence of all of the ventricular action potentials. The QT interval measures the total duration of ventricular activation.

![ECG and action potential of myocardial cell in ventricles](image)

**Fig. 12.16.** The ECG and action potential of myocardial cell in ventricles

**Questions:**
1. What types of electric charges are known? Specify their units. Write Coulomb’s Law.
2. What are the main characteristics of the electric field? Write formulas for electric field strength and potential.
3. What is the relation between the electric field strength and potential?
4. What is the electrical dipole? How to calculate electric potential at a point due to an electric dipole? Write the formula.
5. Write formula for the potential difference between two points created by dipole.
6. Why do certain organs and tissues create electrical fields? What is the electric heart vector? What is the electrogram?
7. What are Einthoven’s theory fundamentals? What are standard bipolar limb leads; augmented unipolar limb leads?
8. Give the approximated ECG. Give the relation between the physiological processes and main waves and intervals of ECG.
Chapter 13. ELECTROCONDUCTIVITY OF TISSUES AND LIQUIDS FOR DIRECT CURRENT

13.1. DIRECT CURRENT IN ELECTROLYTES

The current is defined as the rate of flow of charges across any cross sectional area of a conductor. The conditions for current flow are following: the presence of free charge carriers and the presence of an electric field. If a net charge $q$ passes through any cross section of a conductor in time $t$, then the current $I = q/t$, where $q$ is in coulomb and $t$ is in second. The current $I$ is expressed in ampere (A). If the rate of flow of charge is not uniform, the current varies with time and the instantaneous value of current $i$ is given by $i = dq/dt$. Current is a scalar quantity. The direction of conventional current is taken as the direction of flow of positive charges or opposite to the direction of flow of electrons. When the current in a circuit has a constant magnitude and direction the current is called direct current (DC).

Current density at a point is defined as the quantity of charge passing per unit time through unit area, taken perpendicular to the direction of flow of charge at that point. The current density $j$ for a current $I$ flowing across a conductor having an area of cross section $S$ is

$$j = \frac{I}{S}. \quad (13.1)$$

Current density $j$ is a vector quantity. It is expressed in A·m$^{-2}$.

Ohm’s Law established the relationship between potential difference $U$ and current $I$ states that, at a constant temperature, the steady current $I$ flowing through a conductor is directly proportional to the potential difference $U$ between the two ends of the conductor:

$$I = \frac{U}{R}. \quad (13.2)$$

The current $I$ is measured in amperes, the voltage $U$ in volts and the resistance $R$ in ohms ($\Omega$).

The resistance of a conductor $R$ is directly proportional to the length of the conductor $l$ and is inversely proportional to its area of cross section $S$:

$$R = \frac{\rho l}{S},$$

where $\rho$ is called specific resistance or electrical resistivity of the material. The unit of $\rho$ is Ohm·m ($\Omega$·m).

Ohm’s law from an electromagnetic field point of view

To derive Ohm’s Law at a point from Ohm’s Law for resistors, it is necessary to relate the circuit quantities (voltage $U$ and current $I$) to the field quantities (electric field strength $E$ and current density $j$).
After substitution of resistance $R = \frac{\rho l}{S}$ into the equation of Ohm’s (13.2) it is easy to obtain: $I = \frac{US}{\rho l}$.

Taking into account that $j = \frac{I}{S}$ and $E = \frac{U}{l}$, one can write equation for current density $j$: $j = \frac{E}{\rho}$.

The reciprocal of electrical resistivity, is called **electrical conductivity** $\sigma$:

$$\sigma = \frac{1}{\rho}.$$  

The unit of conductivity $\sigma$ is siemens [S] = [Ω⁻¹].

Thus, **Ohm’s Law at a point** can be obtained as:

$$j = \sigma E.$$  \hspace{1cm} (13.4)

**The electric current in electrolytes**

Let’s find the dependence of electrical conductivity on the electrolyte’s properties. In conductive liquids both positive and negative charges (ions of both signs) carry current.

Consider a cylindrical conductive liquid with a charge carrier density of $n = n_+ + n_-$ in which a current $I$ is flowing. This constitutes an average drift velocity $v_+, v_-$ of each charge carrier.

![Fig. 13.1. The selected volume of the electrolyte](image)

Each charge carrier (positive $q^+$ or negative $q^-$ ion) moves on average a distance $l$:

$$l_+ = v_+ t = \mu_+ Et \quad \text{and} \quad l_- = v_- t = \mu_- Et,$$

where mobility $\mu = \frac{v}{E}$ is velocity $v$ of the charge carrier per unit electrical field strength $E$.

The total charge $Q = Q_+ + Q_-$ transferred during time $t$ through the cross-sectional area $S$ is:

$$Q = Q_+ + Q_- = q_+ n_+ S l_+ + q_- n_- S l_- = (q_+ n_+ \mu_+ + q_- n_- \mu_-) S t E.$$  \hspace{1cm} (13.5)
The equation for current density \( j \) can be written as:
\[
 j = \frac{I}{S} = \frac{Q}{St} = \frac{(q_+ n_+ \mu_+ + q_- n_- \mu_-)StE}{St} = (q_+ n_+ \mu_+ + q_- n_- \mu_-)E. \quad (13.6)
\]

Therefore the conductivity \( \sigma \) of an electrolyte is equal:
\[
 \sigma = q_+ n_+ \mu_+ + q_- n_- \mu_. \quad (13.7)
\]

The charge and concentration of the positive and negative ions is equal in the case of dissociation: \( |q_+| = |q_-| = q \) and \( n_+ = n_- = \alpha n \), where \( \alpha \) is the coefficient of the dissociation.

The equation for electrolyte conductivity \( \sigma \) can be written as:
\[
 \sigma = qn\alpha(\mu_+ + \mu_-). \quad (13.8)
\]

Thus, the conductivity \( \sigma \) depends on the value of the ion charge \( q \), the concentration \( n \), the coefficient of the dissociation \( \alpha \) and mobility of the ions \( \mu \).

### 13.2. Features of Electrical Conductivity of Biological Tissues

Biological tissue, actually display some characteristics of both insulators and conductors because they contain dipoles as well as charges that can move, but in a restricted manner. For materials that are heterogeneous in structure, charges may become trapped at interfaces.

Mechanism of direct current passing through the living tissue is presented in fig. 13.2.

**Fig. 13.2.** The movement of ions in extracellular liquid and cytoplasm. \( I \) is the principal current; \( I' \) is the interstitial polarization current

The principal tissue current \( I \) is determined by motion of ions in the extracellular liquid under the applied potential difference. Inside cellular
structures positive and negative ions also start to move in opposite directions under the applied field. The ions are accumulated on the cell membranes since the membranes have low conductivity. Charge separation results in appearance of interstitial polarization current $I'$ inside tissue, which is in opposite direction to the principal current $I$. It creates additional resistance to the principal current.

Electrical properties of tissues and organs differ greatly. Epidermis, conjunctive tissues, bone without periosteum, chordas have a high electrical resistance. These tissues can be related to dielectrics. Body liquid media have low-resistance and good electric conductivity. The following tissues have small direct-current resistance: cerebrospinal fluid, blood, blood plasma, extracellular fluids.

### 13.3. SOME THERAPEUTIC METHODS BASED ON THE USE OF DIRECT CURRENT

**Galvanization** is a method of direct current medical use: low voltage current (less than 80 V) and small amperage (less than 50 mA). A maximum value of current density used is $0.1 \text{ mA/cm}^2$. Stainless steel, conductive rubber or fabric is used as electrodes. Gaskets are necessary to eliminate the possibility of chemical burn patient electrolysis products formed between the electrode and the skin during the course of DC.

The primary physical mechanisms of direct current action on tissue is caused by the motion of the ions, their separation and the change in ion concentrations in different tissue cells. When a direct current is applied to the tissue by means of two electrodes, the ions will move away from or towards the electrodes: the cations (+) will move towards the cathode (−) and the anions (−) will move towards the anode (+). Polarization is provided by accumulation of equal signs ions on plasmolemma, basement membranes and fascias different surfaces, interstitial polarization appears, which causes appearance of opposite direction current in relation to principal current. It creates additional resistance to active current, but at the same time these zones are places of the most active current action (after epidermis). Simultaneously with ions movement electric current modifies membrane permeability of tissues and increases passive transport of large protein molecules and other matters. Moreover, physiological diffusion and osmosis in human tissues are intensified due to DC action.

The thermal effect is negligible when galvanization is used as the current density is low (less than 0.1 mA/cm$^2$).

**Iontophoresis** is a process of delivery of ionic (charged) drugs into the body by the use of electric current. Iontophoresis is an alternative to oral or parenteral (e. g., needle injection) methods of drug delivery. This method is called electropharmacological because of combination of physical (electrical current) and chemical (ionic (charged) drugs) factors. This factors together increase effects of each other. Electric current acting on the receptors of tissue
excites them and effect of the medicine may be increased or weakened. The drug effect is more significant even in small concentration under current acting. Low amperage currents appear to be more effective as a driving force than currents with higher intensities.

The drug is administered through an electrode (active) which has the same charge as the drug. This is very important. If the polarity of the electrode is not the same as the ions, then penetration through the skin may not occur. During the procedure drug goes not so deeply and is concentrated in skin, partly in subcutaneous fat. It is possible to make a superficial pathological region of a high concentration drug and to have a local effect. Iontophoresis has wide applications in dermatology, ophthalmology, allergic conditions even in cardiac and neurological situations, but its greatest advantage is in the transport of protein or peptide drugs which are very difficult to transport trasdermally due to their hydrophilicity and large molecular size.

Questions:
1. Derive Ohm law in differential form.
2. What is the relation between electrical conductivity and electrical resistivity?
3. What is the interstitial polarization current? What is the reason of its appearance?
4. Whether the physiotherapy based on the direct current is attended by noticeable heat effect? Why?
5. What is the difference between galvanization and iontophoresis?

Chapter 14. THE ALTERNATING CURRENT. THE ELECTRICAL IMPEDANCE OF LIVING TISSUE

14.1. MAIN CHARACTERISTICS OF THE ALTERNATING CURRENT

The alternating currents (AC) varying according to harmonic law have the most important practical significance. The instantaneous value of voltage and current is given by:

\[ U = U_m \sin \omega t; \quad I = I_m \sin(\omega t + \varphi), \quad (14.1) \]

where \( U_m \) is the amplitude value of voltage, \( I_m \) is the amplitude value of current, \( \omega = 2\pi \nu = 2\pi / T \); \( \omega \) is angular frequency (radians/sec); \( \nu \) is the frequency (measured in Hertz, 1/sec); \( T \) is the period; \( \varphi \) is the phase difference between current and voltage (radians).

Averaged over the period magnitudes of alternating currents and voltages (\( I_{eff} \) and \( U_{eff} \)) determine their effect and are called effective values of an AC. \( I_{eff} \) and \( U_{eff} \) are related to the amplitude values of AC (\( U_m \) and \( I_m \)) by the following expressions:

\[ U_{eff} = \frac{U_m}{\sqrt{2}}; \quad I_{eff} = \frac{I_m}{\sqrt{2}}. \quad (14.2) \]
The average power $P$ of an AC circuit is also called the true power of the circuit and is given by:

$$P = I_{\text{eff}}U_{\text{eff}}\cos\phi; \quad P = \frac{1}{2}I_mU_m\cos\phi,$$

(14.3)

where $\cos\phi$ is the power factor. The average power $P$ depends strongly on the phase difference $\phi$ between current and voltage.

### 14.2. AC Circuit with Resistor

Let an alternating source of voltage be connected across a resistor of resistance $R$ (fig. 14.1).

![AC circuit with a resistor](image)

**Fig. 14.1.** AC circuit with a resistor

The instantaneous value of the applied voltage $U$ is:

$$U = U_m\sin\omega t.$$  

(14.4)

The current $I$ through the circuit at the instant $t$:

$$I = \frac{U}{R} = \frac{U_m}{R} \sin\omega t = I_m \sin\omega t.$$  

(14.5)

Equation (14.5) gives the instantaneous value of current in the circuit containing $R$. From the expressions of voltage and current given by equations (14.4) and (14.5) it is evident that in a resistive circuit, the applied voltage and current are in phase with each other (fig. 14.2). Average power $P$ ($\phi = 0$ and $\cos\phi = 1$) is maximal:

$$P_R = I_{\text{eff}}U_{\text{eff}} = \frac{1}{2}I_mU_m.$$  

![14.2. The phasor diagram of AC circuit with a resistor representing the phase relationship between the current and the voltage](image)

**Fig. 14.2.** The phasor diagram of AC circuit with a resistor representing the phase relationship between the current and the voltage

133
14.3. AC CIRCUIT WITH A CAPACITOR

An alternating source of voltage is connected across a capacitor of capacitance \( C \) (fig. 14.3). It is charged first in one direction and then in the other direction.

![AC circuit with a capacitor](image)

The charge \( q \) in the capacitor will vary according to the law:

\[
q = CU = CU_m \sin \omega t.
\]

But the current \( I \) is the time derivative of the charge:

\[
I = \frac{dq}{dt} = CU_m \omega \cdot \cos \omega t = I_m \sin \left( \omega t + \frac{\pi}{2} \right),
\]

where \( I_m = C \omega U_m \).

\[\begin{align*}
X_C &= \frac{U_m}{I_m} = \frac{1}{\omega C} \\
\end{align*}\]

is the resistance offered by the capacitor. It is called capacitive reactance.

From equation (14.4), it follows that in an AC circuit with a capacitor, the current leads the voltage by a phase angle of \( \pi/2 \) and average power \( P_c = 0 \) \((\varphi = \pi/2, \cos \varphi = 0)\). This is represented graphically in fig. 14.4.

![The phasor diagram of AC circuit with a capacitor](image)

**Fig. 14.4.** The phasor diagram of AC circuit with a capacitor representing the phase relationship between the current and the voltage
14.4. AC CIRCUIT WITH AN INDUCTOR

Let an alternating source of voltage be applied to a pure inductor of inductance $L$ (fig. 14.5). The inductor has a negligible resistance $R = 0$.

\[ \varepsilon_L = -L \frac{dI}{dt} = -U_m \sin \omega t. \]

The solution of this differential equation for current is:

\[ I = -\frac{U_m}{\omega L} \cdot \cos \omega t = I_m \sin(\omega t - \frac{\pi}{2}). \]  

(14.8) 

\[ X_L = \omega L \]  

(14.9) 

is the resistance offered by the inductor.

It is clear from equation (14.8) that in an AC circuit containing a pure inductor the current $I$ lags behind the voltage $U$ by the phase angle of $\pi/2$ and average power $P_L = 0$ ($\phi = -\pi/2$, $\cos \phi = 0$). This fact is presented graphically in fig. 14.6.

\[ \text{Fig. 14.6. The phasor diagram of AC circuit with a inductor representing the phase relationship between the current and the voltage} \]

14.5. RESISTOR, INDUCTOR AND CAPACITOR IN SERIES

Let an alternating source of voltage $U$ be connected to a series combination of a resistor of resistance $R$, inductor of inductance $L$ and a capacitor of capacitance $C$ (fig. 14.7).
The expression

$$Z = \frac{U_m}{I_m} = \sqrt{R^2 + (X_C - X_L)^2} = \sqrt{R^2 + \left(\frac{1}{\omega C} - \omega L\right)^2} \quad (14.10)$$

is the net effective opposition offered by the combination of resistor, inductor and capacitor known as the **impedance** of the circuit and is represented by \(Z\). Its unit is Ohm.

The amplitude value of current in a series \(RLC\) circuit is given by

$$I_m = \frac{U_m}{Z} = \frac{U_m}{\sqrt{R^2 + (\omega L - \frac{1}{\omega C})^2}}. \quad (14.11)$$

At a particular value of the angular frequency, the inductive reactance and the capacitive reactance will be equal to each other (i.e.) \(\omega L = \frac{1}{\omega C}\), so that the impedance becomes minimum and it is given by \(Z = R\) i.e. \(I\) is in phase with \(U\). The particular frequency \(\omega_{\text{res}} = \frac{1}{\sqrt{LC}}\) at which the impedance of the circuit becomes minimum and therefore the current becomes maximum is called resonant frequency of the circuit (fig. 14.8).
Maximum current flows through the circuit, since the impedance of the circuit is merely equal to the ohmic resistance of the circuit. i.e. $Z = R$.

14.6. ELECTRICAL IMPEDANCE OF BIOLOGICAL TISSUES
FOR ALTERNATING CURRENT

The cell is the basic unit of living tissues. Its basic structure (a phospholipid bilayer membrane that separates the intracellular medium from the extracellular medium) determines the tissue electrical impedance.

From the electrical point of view, the extracellular medium can be considered as a liquid electrolyte (ionic solution). By far, the most important ions are $\text{Na}^+ (\sim 140 \text{ mM})$ and $\text{Cl}^- (\sim 100 \text{ mM})$. Thus, the electrical properties depend on all physical or chemical parameters that determine their concentration or mobility.

The cell membrane has a passive role (to separate the extra and the intracellular media (the lipid bilayer)) and an active role (to control the exchange of different chemical species (ionic channels and pumps)). Its intrinsic electrical conductance is very low and it can be considered as a dielectric.

In the case of the intracellular medium, the important charge carriers are $\text{K}^+$, protein- and $\text{HPO}_4^{2-} + \text{SO}_4^{2-} + $ organic acids. Besides the ions and other charged molecules, inside the cell it is possible to find numerous membrane structures with a completely different electrical response. These membranes are formed by dielectric materials and their conductivity is very low. Thus, the impedance of the intracellular medium must be a mixture of conductive and capacitive properties. However, for simplification, it is generally accepted that the intracellular medium behaves as a pure ionic conductor.

The electrical behavior of biological tissues can be modeled with a series of nested RC circuits where $C$ is a pseudo-capacitance (fig. 14.9). It is known that the electrical impedance of biological tissue decreases with increase of current frequency and this dependence on frequency is due to the cell membrane, which behaves like a capacitor (fig. 14.10). The extracellular and intracellular constituents of tissue can be related to the electrical equivalent circuit, as shown in figure 14.9.

*Fig. 14.9. The equivalent circuit of living tissue*

*Fig. 14.10. A typical dependence of living tissue impedance on the current frequency*
Electrical model of biological tissue concludes resistance $R_1$ of the extracellular space, resistance $R_2$ of the intracellular space and membrane pseudo-capacitance $C$.

*At low frequencies* $(< 1 \text{ kHz})$ the current is blocked by the capacitance and the current is only capable to flow through $R_1$. *At high frequencies* $(> 1 \text{ MHz})$ the membrane capacitance is no impediment to the current and it flows indiscriminately through the extra and intracellular media.

The impedance $Z$ for this circuit is determined by:

$$Z = \frac{R_1\sqrt{R_2^2 + X_c^2}}{\sqrt{(R_1 + R_2)^2 + X_c^2}}. \quad (14.12)$$

At very high frequencies $X_C = \frac{1}{\omega C} \to 0$ the expression for the impedance is:

$$Z_2 = \frac{R_1R_2}{R_1 + R_2}. \quad (14.13)$$

At medium and high frequencies the impedance $Z$ can be written as:

$$Z = \sqrt{R_2^2 + \left(\frac{1}{\omega C}\right)^2}. \quad (14.14)$$

The electrical impedance of a living tissue can be continuously measured in order to determine its pathophysiological evolution. The ratio of the impedance value at low to impedance at high frequencies is called *polarization coefficient* $K$ (fig. 14.11):

$$K = \frac{Z_L (\nu = 10^3 \text{ Hz})}{Z_H (\nu = 10^6 \text{ Hz})} > 1. \quad (14.15)$$

*Fig. 14.11*. The determination of tissue viability
Measurement of polarization coefficient allows healthy tissues to be differentiated from pathological (malignant and benign) tissues with high reliability. The structure of biological tissue can be assessed impedancometrically during surgical intervention.

Questions:
1. What is an alternating current? What are the amplitude, instantaneous and averaged over the period voltage and current values?
2. What is the phasor diagram? What is it used for?
3. Describe a relationship between an alternating current, a voltage and a power in circuit with a resistor; with an inductor; with a capacitor?
4. Describe a relationship between an alternating current, a voltage and a power in a series combination of an inductor, a capacitor and a resistor.
5. What is an electric impedance? Write the formula.
6. Describe the dependence of living tissue impedance on the current frequency. Give the equivalent circuit of living tissue and characterize.
7. What is a polarization coefficient? What does it characterize?

Chapter 15. ELECTROSTIMULATION OF THE TISSUES AND ORGANS

The response of excitable cells to naturally occurring or artificial stimuli is a subject of great importance in understanding natural function of nerve and muscle, because most stimuli are produced by the natural system itself. Electrostimulation is the application of various types of low-frequency (v ≤ 200 Hz) electrical current to stimulate the body’s organs and systems for clinical diagnosis, therapy, and rehabilitation. A current, arising from an external stimulator or natural source, is introduced into a cell or its neighborhood. The current creates transmembrane voltage in nearby membrane. The membrane responds passively (i.e., with constant membrane resistance), so long as the voltage produced is below a threshold level. When the threshold level is reached, the membrane responds with an action potential, or some other active response. Very often electrostimulation is used in order to provide neurostimulating and voluntary and involuntary muscle stimulation activity, to help strengthen muscles and improves their tone, to stimulate secretory and motoric function of the gastrointestinal tract. The clinical effects of electrostimulation are anti-inflammatory, analgesic, sedative, tranquilising, spasmolytic, asodilating, metabolic.

For electrodiagnostics and electrostimulation realization the pulse electrical current of various form is used. The rectangular pulse currents have the most simple form and are used for nervous system stimulation. Pulse can be defined as an isolated electrical event separated by a finite time from the next event, or represents a finite period of charged particle movement.
15.1. Characteristics of a rectangular pulse

For a complete description of the rectangular form pulse current (fig. 15.1) it is necessary to indicate its amplitude and the two time parameters: pulse duration and pulse period (or interpulse interval).

![Graph of a rectangular pulse current with parameters](image)

*Fig. 15.1. The rectangular form pulse current and its parameters*

*Pulse amplitude* \((I_0, U_0)\): is the maximal magnitude of a pulse parameter, such as the voltage \([\text{mV}]\), current \([\text{mA}]\).

*Pulse duration* \((t_u)\): is the period of time during which a pulse is present; [ms].

*Interpulse interval* \((t_0)\): is the time between the end of one pulse and the beginning of the next pulse in a series, or other words \(t_0\) is the period of time between pulses during which there is no current flow; [ms].

*Period* \((T)\): the period is equal to the pulse duration plus the interpulse interval; [ms]:

\[
T = t_u + t_0. \tag{15.1}
\]

*Pulse repetition frequency or frequency* \(v\) is the number of pulses per time unit \([\text{Hz}]\):

\[
v = \frac{1}{T}. \tag{15.2}
\]

*The fill factor* \(k\) is defined as the ratio of the pulse duration \((t_u)\) to the period \((T)\) of a rectangular pulse. The fill factor is the proportion of time during which a current is operated. The fill factor can be expressed as a ratio or as a percentage:

\[
k = \frac{t_u}{T}. \tag{15.3}
\]

*The duty cycle* \(Q\) shows how many times pulse period is more than a pulse duration:

\[
Q = \frac{T}{t_u} = \frac{t_u + t_0}{t_u} = 1 + \frac{t_0}{t_u}. \tag{15.4}
\]
15.2. Characteristics of an arbitrary pulse

For an arbitrary pulse current description (fig. 15.2) it is necessary to enter some additional parameters characterizing the shape of the pulse. For this purpose the auxiliary lines are drawn at $0.1I_0$ and $0.9I_0$.

![Diagram](image)

Fig. 15.2. The arbitrary pulse current and its parameters

**Pulse rise time** ($t_{rt}$): is the period of time during which a pulse rises from ten percent of its amplitude value ($0.1I_0$) to 90 percent of its amplitude value ($0.9I_0$).

**Pulse fall time** ($t_{ft}$): is the period of time during which a pulse falls from ninety percent of its amplitude value ($0.9I_0$) to ten percent of its amplitude value ($0.1I_0$).

**Pulse peak time** ($t_{pt}$): is the period of time during which a pulse $I \geq 0.9I_0$.

**Steepness of the pulse** ($K$) determines the rate of current rise in the time from $0.1I_0$ to $0.9I_0$:

$$K = \frac{0.9I_0 - 0.1I_0}{t_{rt}} = \frac{0.8I_0}{t_{rt}} = \tan \alpha$$  \hspace{1cm} (15.5)

In this case the pulse duration $t_u$ can be obtained as a sum:

$$t_u = t_{rt} + t_{pt} + t_{ft}.$$  \hspace{1cm} (15.6)

15.3. Weiss–Lapicque Law

In electrical stimulation, current induced must be of sufficient amplitude and duration to bring excitable cells to the threshold of depolarization. The lowest current disturbance that causes tissue excitation is called the **threshold current** $I \geq I_{thr}$. If the stimulus is lower than the threshold, no activation will be initiated. But current magnitude does not exceed the let-go current:

$$I_{thr} < I < I_{let-go}.$$  \hspace{1cm} (15.7)
The threshold current $I_{thr}$ dependence on the rectangular pulse duration $t_u$ is given by Weiss–Lapicque Law:

$$I_{thr} = \frac{a}{t_u} + b,$$

where $a$ and $b$ are constants depending on the type of living tissue.

The minimum current $I_{chr}$ required above a certain threshold for tissues stimulation is inversely proportional to the duration of the electrical pulse $t_u$. A plot of the inverse relationship between the threshold stimulus current-pulse amplitude and its duration is known as the strength-duration curve (fig. 15.3).

Lapicque introduced two new terms to define the tissue stimulation threshold:

– the rheobase $R$ is the lowest current required to reach threshold as the stimulation duration grows long (conceptually, as $t_u \to \infty$).

– the chronaxie $t_{chr}$ is that pulse duration at which the threshold value is twice that of the rheobase.

![Fig. 15.3. The strength-duration curve shows the inverse relationship between the lowest stimulus current pulse required to produce a response versus its duration](image)

Values of the constants ($a$ and $b$) can be related to rheobase $R$ and chronaxie $t_{chr}$, which are determined experimentally.

1. If $t_u \to \infty$, the threshold current $I_{thr}$ is equal to rheobase $R$:

$$I_{thr} \to b,$$ it means $b = R$.  $[b] = mA$

2. If $t_u = t_{chr}$, than $I_{thr} = 2R$, and according to Weiss–Lapicque Law:

$$2R = \frac{a}{t_{chr}} + R.$$  \(15.9\)

Thus, $a = Rt_{chr}$.  $[a] = C.$
The pulse duration \( t_u \)

The pulse duration should be \( t_u \geq t \to \infty \), than in this case the threshold is minimum (is equal to rheobase). The pulse duration \( t_u \) depends on the type of living tissue.

The pulsed current frequency \( \nu \)

For the electrical excitation of tissues the pulse period should be more than the absolute refractory period \( T_{ref} \). The absolute refractory period \( T_{ref} \) is the time during which the cell can not be excited by any stimulus. That is why the maximal excitation cell frequency is \( \nu_{\text{max}} = 1/T_{ref} \).

For neural tissue \( \nu_{\text{max}} = 500 \div 1000 \text{ Hz} \) \( (T_{ref} = 1 \div 2 \text{ ms}) \);

for skeletal muscle \( \nu_{\text{max}} = 100 \div 200 \text{ Hz} \) \( (T_{ref} = 5 \div 10 \text{ ms}) \);

for heart muscle \( \nu_{\text{max}} = 3.3 \text{ Hz} \) \( (T_{ref} = 300 \div 350 \text{ ms}) \).

Steepness of the pulse \( K \)

The dependence of threshold current \( I_{thr} \) on the rate of pulse steepness increase is reflected in the Law of Du Bois–Reymond: a motor nerve responds, not to the absolute value, but to the alteration of value from moment to moment, of the electric current; rate of change of intensity of the current is a factor in determining its effectiveness.

The threshold current value \( I_{thr} \) decreases when pulse steepness \( K \) increases.

15.4. ELECTRICAL STIMULATION OF THE HEART MUSCLE

Defibrillation is used to treat cardiac arrest or fibrillation loss of coordinated contraction of heart muscle fibers. Death occurs in minutes if left untreated. Fibrillation is arrhythmia resulting from an abnormal spread of excitation, causing parts of the myocardium to contract while other regions of the cardiac muscle are relaxing. The functional fragmentation can be both localized in atria and in ventricles. In ECG the fluctuation associated to ventricular fibrillation are very irregular, changing rapidly in frequency, shape and amplitude. Cardiac arrest is the complete cessation of cardiac activity, either electrical, mechanical, or both.

\textbf{Defibrillation} involves the application of a powerful single current pulse of duration \( t_u = 2\text{–}5 \text{ ms} \) to the heart which leads to depolarization of the most of the heart cells simultaneously, which often reestablishes coordinated contractions and a normal sinus rhythm.

Parameters of the electric pulse used are:

– on the chest: voltage \( U = 5\text{–}7 \text{ kV} \), current \( I \approx 1 \text{ A} \);
– on the heart of the nude: voltage \( U = 1.5\text{–}2.5 \text{ kV} \), current \( I \approx 1 \text{ A} \).

\textbf{Cardioverter-defibrillator}

Cardioversion is used in persons who have heart rhythm problems (arrhythmias), which can cause the heart to beat too fast (tachycardia) or too
slowly (bradycardia). There are implantable cardioversion defibrillation and external defibrillator. An implantable cardioverter-defibrillator (often called an ICD) is a device that briefly passes an electric current through the heart. It is implanted in the chest to constantly monitor and correct abnormal heart rhythms (arrhythmias). Pacing circuit consists of a power source (pulse generator), one or more conducting (pacing) leads, and the myocardium (fig. 15.4). Electrical signal (stimulus) travels from the pacemaker, through the leads, to the wall of the myocardium. Myocardium is «captured» and stimulated to contract. External defibrillators are typically used in hospitals or ambulances, but are increasingly common outside the medical areas. As automated external defibrillators become safer and cheaper. There are synchronized cardioversion and non-synchronized one, automated defibrillator and semi-automated defibrillator. Parameters of the electric pulse used for heart stimulation are: pulse duration $t_u = 0.5\text{–}8 \text{ ms}$; frequency $\nu = 1\text{–}1.2 \text{ Hz}$; voltage $U \approx 6 \text{ V}$; current $I \approx 1\text{–}10 \text{ mA}$.

**Fig. 15.4.** Implantable cardioversion defibrillation

**Questions:**
1. What is the electrostimulation?
2. What is the electrical current used for electrostimulation?
3. Describe parameters of a rectangular pulse.
4. Specify main characteristics of an arbitrary pulse.
5. What do electrostimulation pulse duration, frequency and amplitude depend on?
6. Give the strength-duration curve and characterize it.
7. What are the rheobase and the chronaxie?
8. Write Weiss–Lapicque Law. What is a relationship between the law constants and the rheobase and the chronaxie?
10. What is the defibrillation method? Specify main parameters and features the method.
Chapter 16. HIGH FREQUENCY ELECTROMAGNETIC FIELDS USE IN MEDICINE

The concept of producing heat deep within the tissues, beyond the reach of infrared and other forms of superficial heat is appreciated by physical therapists. Therapeutic heating causes vasodilation, increases the rate of enzymatic biological reactions, increases nerve conduction velocity, and increases soft tissue extensibility. These physiologic effects underlie the benefits of therapeutic heating for promoting tissue healing, reducing pain and increasing range of motion.

The tissue can be heated due to the effect of electric current. When an electric current is passed through a tissue, the tissue gets heated up and here the electrical energy is converted into heat energy. The heat $Q$ produced in a tissue is directly proportional to the tissue resistance $R$, the duration of the electric current action $t$ and to the square of the applied current $I$:

$$Q = I^2Rt.$$  \hspace{1cm} (16.1)

Let’s $j$ is current flux density i. e. current flowing through a unit area:

$$j = \frac{I}{S}.$$  \hspace{1cm} (16.2)

The effectiveness of any thermal procedures is determined by the specific heat $q$. The specific heat $q$ is the heat produced in a tissue unit volume per unit time:

$$q = \frac{Q}{Vt}.$$  \hspace{1cm} (16.3)

16.1. DIATHERMY

**Diathermy** is the application of high-frequency electrical current that is primarily used to generate heat in body tissues. Two large electrodes placed at each side of the body are used in diathermy (fig. 16.1). Value of applied current $I$ is 1–2 A and frequency $\nu$ is 0.5–2.0 MHz.

![Fig. 16.1. Scheme of diathermy](image)
In order to determine which tissues are heated by diathermy let’s write the following equation:

\[ q = \frac{Q}{V_t} = \frac{I^2Rt}{V_t} = \frac{I^2\rho l}{SIS} = \frac{I^2\rho}{S^2} = j^2\rho, \]  

(16.4)

where \( \rho \) is electrical resistivity (also known as specific electrical resistanc).

Heating occurs in proportion to the square of the current density \( j \) and in direct proportion to the electrical resistivity \( \rho \) of the tissue. The higher the tissue resistivity is, the more heat will be released in the tissue when the current passes through it. From the equation (16.4) it follows that under diathermy for the same current density, the tissue with high resistivity (i.e. skin and subcutaneous adipose tissue) is heated more efficiently.

Surgical diathermy is based on the diathermy. Surgical diathermy is the application of a high-frequency electric current to biological tissue as a means to cut or coagulate tissue. There are two basic varieties of surgical diathermy: monopolar and bipolar.

In monopolar surgical diathermy the area of one electrode (so called active electrode or pointed probe) is much more smaller then another one (passive electrode) (fig. 16.2). Since area of electrode is inversely proportional to current density \( j = \frac{I}{S} \), the current density is very high at the point of contact between the active electrode and the tissue. High current density leads to the concentration of heat in the immediate vicinity of this electrode, thus producing the desired effects, i.e. coagulation or dissection, depending on the intensity of current. In bipolar surgical diathermy both electrodes are active ones, whereby both pointed probes have contact with the surgical field.

![Fig. 16.2. Monopolar technique of surgical diathermy](image)

In the case of tissue electrosection current density \( j \) is equal to \( \sim 40 \) mA/mm\(^2\) and in the case of tissue coagulation current density \( j \) is equal to \( \sim 6–10 \) mA/mm\(^2\). Electrocoagulation is ideal for clotting small blood vessels (less than 2 to 3 mm in diameter) in deep and superficial surgery. Usually, a 2- to 5-mm metallic sphere at the end of a treatment electrode is the optimal tip for
hemostasis of small vessels. In electrosection, the electrode is used to cut tissue. An electrode tip in the shape of a fine needle, wire loop, diamond, ellipse, or triangle is advanced slowly through the tissue, causing a steam envelope to advance around the tip and producing a smooth cutting effect with little sense of pressure against the tissue by the operator.

16.2. **Inductothermy**

**Inductothermy** is a form of physical therapy in which deep tissues heating is based on the use of an oscillating magnetic field $B = B_0 \sin 2\pi vt$ with frequency $v = 10$–20 MHz. This procedure uses an induction coil wrapped around the body part to produce an oscillating magnetic field within the body (fig. 16.3). The frequency employed is usually equal to 13.56 MHz ($v = 13.56$ MHz). The magnetic field penetrating the tissues, induces electrical currents in tissues named as induction currents, vortical currents or currents Foucault. The more is the electroconductivity of a tissue, the current of greater force is formed in it. The occurrence of vortical currents is accompanied by heating of tissues. Thus, to induce a current into the underlying tissue and organs strong and rapidly changing magnetic field must be generated by the coil.

Faraday law can be explained by the current generated in the closed loop circuit if an electric conductor, which forms a closed circuit, is linked by a time-varying magnetic flux $\Phi$. This current is due to the electromotive force ($\varepsilon$) induced by the time-varying flux. The magnitude of $\varepsilon$ depends upon the rate of the magnetic flux change $d\Phi/dt$. $\varepsilon$ has such direction that the time-varying magnetic field is always opposite to that of $d\Phi/dt$. Therefore,

$$\varepsilon = -\frac{d\Phi}{dt},$$

(16.5)

where $\varepsilon$ is an electromotive force [V], $\Phi$ is a magnetic flux [Wb], $t$ is a time [s]. Magnetic flux $\Phi$ can be written as $\Phi = BS$, where $B$ is a magnetic field, $S$ is an area.
Induced Foucault current (vortical current) \( I \) can be written as:

\[
I = \frac{\varepsilon}{R}.
\]  

(16.6)

To determine which tissues are heated by inductothermy let’s write the equation for the specific heat \( q \):

\[
q = \frac{Q}{V_t} = \frac{\varepsilon^2 R t}{R^2 V_t} = \frac{\varepsilon^2 t}{\rho l^2} = \frac{\varepsilon^2}{\rho l^2},
\]

(16.7)

where \( \rho \) is an electrical resistivity.

From the equation (16.7) it follows that under inductothermy the tissues with high resistivity \( \rho \) are less heated. Thus, inductothermy preferentially heats low impedance tissues (e. g., skeletal muscle, blood, synovial fluid).

### 16.3. ULTRA HIGH FREQUENCY THERAPY

**Ultra high frequency therapy** (UHF-therapy) is a form of physical therapy in which deep heating of tissues based on the use of an oscillating electric field \( E = E_0 \sin 2\pi v t \) (\( E \) is the electrical field strength) \((v = 30–60 \text{ MHz})\). The frequency employed is usually equal to 40.68 MHz.

To obtain the electromagnetic vibrations of various frequency the generator of high frequencies is used in physiotherapy devices. The basic element of the device is the LC circuit (or oscillating circuit) (fig. 16.4).

![Fig. 16.4. Scheme of oscillating circuit](image)

An oscillating circuit is a resonant circuit that consists of an inductor \( L \) and a capacitor \( C \). When connected together, an electric current can alternate between them at the circuit’s resonant frequency. An \( LC \) circuit can store electrical energy vibrating at its resonant frequency. A capacitor stores energy in the electric field between its plates, depending on the voltage across it, and an inductor stores energy in its magnetic field, depending on the current through it.

Physiotherapy devices consist of technical and therapeutic \( LC \) circuits (fig. 16.5). In order to have effective heating of tissues technical and therapeutic circuits have to work in resonance: \( T_{tech} = T_{ther} \), i. e. oscillation periods of technical and therapeutic circuits should be equal. \( C_{ther} \) is a variable tuning capacitor which is used to adjust the resonant frequency to the desired value.
The specific heat $q$ produced in conductive tissue under the UHF-therapy can be characterized as:

$$q = \frac{Q}{Vt} = \frac{U^2Rt}{R^2Vt} = \frac{U^2}{\rho l^2} = \frac{U}{\rho} = \frac{E^2}{\rho}. \quad (16.8)$$

The specific heat $q$ produced in dielectric tissue under the UHF-therapy can be written as:

$$q = \varepsilon_0 E^2 \omega \tan \delta, \quad (16.9)$$

where $\varepsilon$ is a permittivity of medium, $\varepsilon_0$ is a permittivity of vacuum, $E$ is the electrical field strength, $\omega$ is the frequency, $\tan \delta$ is the dielectric loss tangent.

The most significant difference between these two expressions (16.8 and 16.9) is that the heating of electrolytes does not dependent on frequency and heating of dielectric increases with an increase of the electromagnetic field frequency $\omega$ (fig. 16.6).

*Fig. 16.6. Frequency dependence of specific heat for electrolyte and dielectric under the UHF-therapy*
At frequencies $\nu < \nu_1$ (fig 16.6) electrolyte is heated more efficiently than dielectric, and on the contrary at frequencies $\nu > \nu_1$ (fig 16.6) dielectric is heated more efficiently. At frequencies $\nu = 30–60$ MHz the electrical field energy is absorbed mainly in tissues having the large capacitor resistance i. e. in tissues badly conducting an electrical current — in dielectric tissues.

16.4. THE MICROWAVE THERAPY

The microwave therapy is a treatment of organism tissues by an electromagnetic field of super high frequency (microwave) $\nu = 300–2500$ MHz.

Centimeter wave therapy is a form of physical therapy in which tissues heating is based on the use of an electromagnetic waves with frequency $\nu = 2375$ MHz. In this case wavelength $\lambda$ is equal to ~ 12 cm ($\lambda = c/\nu$). It is necessary to dose centimeter wave therapy because of standing waves formation. They are formed at reflection of a wave from border of two environments and imposing reflected on the next falling wave. Such process occurs repeatedly in the same place. Under the laws of physics the «standing» wave is formed in case if distance between borders of two environments makes more than a quarter of length of a wave. This situation can arise at thickness of subcutaneous fatty layer more than 2 cm. At formation of «standing» waves there is a significant local increase of temperature of a tissue down to a burn. This overheating of a tissue is accompanied by sensation of bursting open, burning, rheumatic pains that requires immediate reduction of a doze of influence or termination of procedure. The uncontrollable overheating can arise at influence on hydropic tissue. That can lead to local burn inside of body. The energy of microwaves is absorbed mainly by molecules of water; their dielectric permeability in this connection is insignificant. Depth of penetration is ~ 3–5 cm.

Decimeter wave therapy is a form of physical therapy in which heating of tissues is based on the use of an electromagnetic waves with frequency $\nu = 460$ MHz. Wavelength $\lambda$ is equal to ~ 65 cm ($\lambda = c/\nu$). The microwaves of a decimeter range are approximately 2 times less intensively reflected by a surface of skin. They to a lesser degree, than wave of a centimetric range, are absorbed by water, as the phenomena of a resonance of dipoles of water at this frequency of an electromagnetic field are less expressed. The energy of these waves in the process of penetration into depth of tissues fades twice more slowly in comparison with centimetricwaves. This therapy is used for heating of tissues containing water. Depth of penetration is ~ 8–9 cm.

Extremely high frequency therapy is a form of physical therapy which is based on the use of the electromagnetic waves with frequency $\nu = 3\cdot10^{10}–3\cdot10^{11}$ Hz. The wavelength $\lambda$ of extremely high frequency radiation is equal to ~ 1–10 mm ($\lambda = c/\nu$). This therapy is used to obtain nonheating tissue effect (resonance energy absorption). It is well know that each molecule has a distinctive frequency fingerprint of emitting and absorbing electromagnetic
waves. Cell membranes participate in synchronized coherent high-frequency oscillations (10s of GHz up to 100s of GHz): therefore dynamic biological functionality can be influenced by weak electromagnetic radiation at certain narrow-band frequency. The mechanism of the effects of extremely high frequency therapy on the body is still unclear, but it is believed that it is related to the resonant absorption of radiation by water molecules, for which the strongest resonance is observed at a frequency of 50 GHz.

16.5. DARSONVALISATION

Darsonvalisation is a treatment of organism tissues by sine pulse electrical current of high frequency ($\nu = 110$ KHz), high voltage ($U = 10–30$ kV) and small current ($I = 10–15$ mA). The frequency of pulses is 50 Hz. The electrical current is modulated on amplitude.

A glass vacuum electrode, in which air is rarefied up to 0.1–0.5 mm mercury (fig. 16.7) is used. The name of this electrode is condenser. A condenser has resistance of $R = 10^6–10^7$ Om. The value of current can be calculated as:

$$I = \frac{U}{R} = \frac{20 \times 10^3 V}{10^7 \text{Om}} = 20 \times 10^{-4} \text{A}.$$

![Fig. 16.7. Scheme of darsonvalisation](image)

As very small current is used, the heating of tissues does not occur.

Air in an electrode is ionized under the action of high voltage and the electric current passes through the ionized gas. It is possible to assimilate conducting part of electrode and a body of the patient to facings of the condenser, and the glass is dielectric. The spark discharge arises in air between the electrode and the tissue under the procedure. The main darsonvalisation effect is connected with irritating action of the spark discharge on nervous receptors of a skin and mucous environments. It results in sedative and analgesic effects.
Questions:
1. Give the diathermy parameters. Which tissues are heated under diathermy.
2. What is the surgical diathermy? Describe monopolar technique and bipolar one.
3. What are inductothermy parameters? Which tissues are heated under this procedure?
4. What is the ultra high frequency therapy? Which tissues are heated more efficiently under UHF-therapy?
5. Explain technical and therapeutic circuits function in ultra high frequency apparatus.
6. What is the microwave therapy? Why centimeter wave therapy is dose dependent procedure?
7. What is the darsonvalisation? What is the mechanism of therapeutic effects under darsonvalisation?

Chapter 17. BIOPHYSICAL SIGNALS MONITORING

17.1. THE SENSORS

Some important biosignals do not have the character of the electrical potential or voltage (blood pressure, core body temperature, blood flow, cerebrospinal fluid pressure). The monitoring can be carried out only with the sensors transforming that signal as a physical quantity to some form of the electrical signal. Sensors are devices, which transform measured parameter into electrical signal.

There are two major types of sensors: active and passive. Active sensors generate electric current or voltage directly in response to measured parameter (physical stimulus, such as thermal energy, electromagnetic energy, acoustic energy, pressure, magnetism, or motion). Amplitude or frequency of generated signal is proportional to measured parameter. Examples of active sensors are thermocouples and piezoelectric accelerometers. Thermocouples produce a voltage related to a measured temperature of two metals and if the two junctions are at different temperatures, electricity is generated. Passive sensors produce a change in some their electrical characteristics, such as resistance, capacitance, or inductance, as a result of measured parameter action. These usually require additional electrical energy. The examples of passive sensors are resistance temperature detectors and thermistors.

Sensors have following characteristics:
– the sensitivity of the sensor is the minimum input of physical parameter that will create a detectable output change. This is defined as the ratio of the incremental output (Δy) to incremental input (Δx), i.e.

\[ S = \frac{\Delta y}{\Delta x}. \]

In some sensors, the sensitivity is defined as the input parameter change required to produce a standardized output change. In others, it is defined as an output voltage change for a given change in input parameter;
the measured range (MR) of the sensor is the maximum and minimum values of applied parameter that can be measured. The measured range is defined as the difference of the maximum input and the minimum input, \( x_{\text{max}} - x_{\text{min}} = \text{MR} \);

the accuracy of the sensor is the maximum difference that will exist between the actual value (which must be measured by a primary or good secondary standard) and the indicated value at the output of the sensor. The accuracy describes the closeness with which the measurement approaches the true value of the variable being measured. And it is given by

\[
\varepsilon_a = \frac{x_m - x_t}{x_t} \cdot 100 \%,
\]

where \( x_t \) is a true value, \( x_m \) is a measured value. The accuracy can be expressed either as a percentage of full scale or in absolute terms;

the resolution: it is defined as the smallest incremental change in the input that would produce a detectable change in the output. This is often expressed as percentage of the measured range, MR. For a detectable output \( \Delta y \), if the minimum change in \( x \) is \( \Delta x_{\text{min}} \), then the maximum resolution is

\[
R_{\text{max}} = \frac{\Delta x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \cdot 100 \%.
\]

the response time can be defined as the time required for a sensor output to change from its previous state to a final settled value within a tolerance band of the correct new value. Sensors do not change output state immediately when an input parameter change occurs. Rather, it will change to the new state over a period of time, called the response time.

17.2. Sensors of Temperature

There are many different sensors of temperature, but three find particularly wide application to biomedical problems — the resistance temperature detector, thermistors, and thermocouples.

The thermocouple is a temperature sensor which generates an electrical voltage or electromotive force directly dependent on the temperature without an additional power source because of its thermo-electric properties. Thermocouple systems are based on the thermoelectric effect, discovered by Thomas Seebeck. Seebeck found that when two different metals were joined and a temperature difference was present, a voltage was produced. Thermocouples contain two electrical conductors made of different materials which are connected at one end. The end of the conductors which will be exposed to the process temperature is called the measurement junction. The point at which the thermocouple conductors end has reference temperature is called the reference junction (fig. 17.1). When the measurement and reference junctions of a thermocouple are at different temperatures, a millivolt electromotive force is formed within the conductors. Knowing the type of thermocouple used, the magnitude of the
millivolt electromotive force within the thermocouple, and the temperature of
the reference junction allows the user to determine the temperature at the
measurement junction.

The electromotive force $\varepsilon$ generated by the thermocouple is largely
proportional to the difference between the temperature of the object under test
and the reference temperature:

$$\varepsilon = \alpha(T_2 - T_1),$$

where $\varepsilon$ is the electromotive force, $\alpha$ is the coefficient of proportionality, $T_2 - T_1$
is the temperature difference.

![Fig. 17.1. Schematic representation of temperature measurement with the thermocouples](image)

**The resistance temperature detector**, or the RTD, employs the property
that the electrical resistance of metals varies with temperature. The electric
resistance of a piece of metal or wire generally increases as the temperature of
that electric conductor increases (fig. 17.2). A linear approximation to this
relationship is given by

$$R = R_0[1 + \alpha(T - T_0)],$$

where $R_0$ is the resistance at temperature $T_0 = 0$; $\alpha$ is the temperature coefficient,
which depends only on the nature of the metal and $T$ is the temperature at which
the resistance is being measured.

The temperature coefficient of resistance is widely used to characterize
RTD: $TCR = \frac{1}{R_0} \frac{dR}{dT}$. The unit of TCR is $1/K°$.

RTD are positive temperature coefficient (PTC) sensors whose resistance
increases with temperature. The main metals in use are platinum and nickel.
The RTD exhibits behavior which is more accurate and more linear over wide
temperature ranges than a thermocouple. Unlike a thermocouple, however, an RTD is a passive sensor and requires current excitation to produce an output voltage.

Fig. 17.2. Resistance versus temperature for a typical conductor

Similar in function to the RTD, thermistors are constructed of solid semiconductor materials. The electric resistance of a typical intrinsic (non-doped) semiconductor decreases exponentially with the absolute temperature $T(°K)$ (fig. 17.3):

$$R(T) = Ae^{BT},$$  

(17.3)

where $A$ (Ohm) and $B$ (°K) are constants depending only on the semiconductor material being used in thermistor.

Fig. 17.3. Resistance versus temperature for a typical semiconductor

The most commonly used thermistors are those with a negative temperature coefficient. The thermistor high sensitivity allows it to detect minute variations in temperature which could not be observed with an RDT or thermocouple. The temperature coefficient of thermistors does not decrease linearly with increasing temperature as it does with RDTs; therefore, linearization is required
for all but the narrowest of temperature ranges. Thermistors have the most sensitivity but are the most non-linear.

Let’s compare a structure of semiconductors and conductors. A metal consists of a lattice of atoms, each with a shell of electrons. This can also be known as a positive ionic lattice. The outer electrons are free to dissociate from their parent atoms and travel through the lattice, creating a «sea» of electrons, making the metal a conductor. A conductor may be described as a substance in which the number of «free» electrons per unit volume is the same order of magnitude as the number of atoms per unit volume. When an electrical potential difference (a voltage) is applied across the metal, the electrons drift from one end of the conductor to the other under the influence of the electric field. Semiconductors are materials which lie between conductors and insulators. Let’s consider silicon as an example of semiconductors. Silicon forms crystal of the diamond type in which there is a co-valent bond between the atoms. Each atom shares its electrons with four other atoms, making a cubic lattice. When the thermal energy of the system is zero in a perfect crystal, all the electrons are in their proper positions and the substance is an insulator. With increasing temperature an increasing fraction of the electrons is displaced by thermal movement. In a pure metallic conductor the electrons are «free» even when the material contains no thermal energy. The proximity of the atoms, coupled with the small binding forces, leave the valency electrons «floating» in the material. As the temperature increases the kinetic energy of the atoms increases, and the electrons driven through the material by an electric field suffer larger energy losses at collisions and their mobility is therefore diminished. Thus the electric resistance of metal increases as the temperature increases. The opposite is the case with a semiconductor because the number of electrons increases with temperature. The electric resistance of intrinsic semiconductor decreases as the temperature increases.

17.3. BIOPOTENTIAL AMPLIFIER

Biopotential signals usually have amplitudes of the order of a few millivolts or less. Such signals must be amplified to levels compatible with recording and display devices. Amplifiers are an important part of modern instrumentation systems for measuring biopotentials. The essential function of a biopotential amplifier is to take a weak electric signal of biological origin and increase its amplitude so that it can be further processed, recorded, or displayed (fig. 17.4).

Fig. 17.4. Scheme of ECG signal amplifying
Thus, the amplifier is an electric device that increases the input voltage \( (U_{\text{in}}) \) (power \( P_{\text{in}}, \) current \( I_{\text{in}}) \) by a factor \( K \); that is, the output voltage \( U_{\text{out}} \) (power \( P_{\text{out}}, \) current \( I_{\text{out}}) \) is \( U_{\text{out}} = KU_{\text{in}}; P_{\text{out}} = KP_{\text{in}}; I_{\text{out}} = KI_{\text{in}}. \) The amplification factor (gain) \( K \) is determined by the ratio of the output and input voltages (power, current). Depending on the nature of the input and output signals, one can have different types of amplifier gain: current gain (current out/current in); voltage gain (voltage out/voltage in), power gain (power out/power in):

\[
K = \frac{I_{\text{out}}}{I_{\text{in}}}; K = \frac{U_{\text{out}}}{U_{\text{in}}}; K = \frac{P_{\text{out}}}{P_{\text{in}}}. \tag{17.4}
\]

Main requirement for the amplifier is to increase signal amplitude without distortion of the signal form.

The biosignals can be considered as periodic ones. The signal is periodic if signal waveshape is repeated periodically. Any periodic biosignal can be written as a Fourier series. According to the Fourier theorem any periodic function may be expressed as the sum of harmonic components at integer multiples of the fundamental frequency \( \nu_0 = 1/T \) (or angular frequency \( \omega_0 = 2\pi\nu_0 \) (a series of sine and cosine terms called the Fourier series). Each of sine and cosine term has specific amplitude and phase coefficients known as Fourier coefficients:

\[
\varepsilon(t) = a_0 + a_1\cos(\omega_0t + \varphi_1) + a_2\cos(2\omega_0t + \varphi_2) + \ldots + a_n\cos(n\omega_0t + \varphi_n) \tag{17.5}
\]

or in the brief description

\[
\varepsilon(t) = a_0 + \sum_{m=1}^{n} a_m\cos(m\omega_0t + \varphi_m), \tag{17.6}
\]

where \( a_0 \) is a constant signal component (in many cases it can be equal to zero); \( a_m\cos(m\omega_0t + \varphi_m) \) are harmonic signal components with amplitude \( a_m \) \((m = 1, 2, 3, \ldots, n),\) angular frequency \( m\omega_0 \) and initial phase \( \varphi_m. \)

The first term (at \( m = 1 \)) describes harmonic of the fundamental frequency \( \nu_0. \) This fundamental frequency \( \nu_0 \) is equal to the frequency of investigated biosignal and is called fundamental tone. Others components (at \( m = 2, 3, \ldots, n \)) are called overtones.

Frequency range from \( \nu_0 \) to \( \nu_n = n\nu_0 \) is called the frequency spectrum of signal. Signal frequency spectrum and the corresponding harmonic amplitudes determine the harmonic spectrum of the biosignal under consideration (fig. 17.5).

In order to satisfy the basic requirement for an amplifier (input signal amplitude increase without distortion of its form) gain \( K \) of the various harmonics of the amplified signal should be the same. In this case gain \( K \) is constant and does not depend on frequency: \( K_0 = K_1 = K_2 = \ldots = K_n = \text{const} = K, \) if the initial phase of the harmonics do not change with amplifying. The output signal will be a new periodic function:
\[ E(t) = K \left\{ a_0 + \sum_{m=1}^{n} a_m \cos(m\omega_0 + \phi_m) \right\} = K \varepsilon(t), \quad (17.7) \]

that is, the output will be \( K \) times as the input signal. The amplifier satisfying these conditions (\( K = \text{const} \)) can be considered as an ideal amplifier.

In fact for real amplifier the gain depends on frequency and amplitude of input signal: \( K = K(v, a_m) \neq \text{const} \) that leads to the frequency and amplitude distortion of the amplified signal.

Gain dependence on frequency is called frequency characteristic of the amplifier. For ideal amplifier frequency characteristic is a line parallel to axis of abscissas and the gain is independent on signal frequency: \( K = \text{const} \).

For a real amplifier gain is constant only in a certain range of frequencies (fig. 17.6).
The range of frequencies within which the gain $K$ is greater or equal to \(0.7K_{\text{max}}\) is called **frequency bandwidth**. If the frequency spectrum of the amplified signal completely falls into the bandwidth, frequency distortions of an output signal are negligible and the amplifier can be used for medical diagnostic.

Another important characteristic of an amplifier is its amplitude characteristic. **Amplifier amplitude characteristic** is the dependence of the output signal amplitude (i.e. voltage, current, power) on the input signal amplitude. For an ideal amplifier, this dependence is always linear because the gain is constant and independent on the input signal amplitude: $U_{\text{out}} = K U_{\text{in}}$. For a real amplifier, this dependence is linear only in a certain range of input voltages, namely, at $U_{\text{in1}} \leq U_{\text{in}} \leq U_{\text{in2}}$ (fig. 17.7).

![Fig. 17.7. Amplitude characteristic of amplifier](image)

Only within this range, the gain is constant and amplitude distortion are absent. Therefore, input voltage range between $U_{\text{in1}}$ and $U_{\text{in2}}$, within which the amplitude characteristic is linear, is called the **dynamic range** $D$ of amplifier and is expressed in decibels (dB). For voltage amplifier dynamic range $D$ is:

$$D = 20 \log_{10} \frac{U_{\text{in2}}}{U_{\text{in1}}}. \quad (17.8)$$

To amplify the signal without distortion by use of a real amplifier it is necessary to comply with two conditions:

- the frequency spectrum of the amplified signal should completely be within the bandwidth;
- amplitude range of amplified signal should completely be within the amplifier dynamic range $D$.

The table 17.1 shows the ranges of amplitudes and frequencies covered by several of the common biopotential signals.
**Table 17.1**

The ranges of amplitudes and frequencies covered by ECG, EEG and EMG biosignals

<table>
<thead>
<tr>
<th>Electrogram</th>
<th>Electrocardiogram ECG</th>
<th>Electroencephalogram EEG</th>
<th>Electromyogram EMG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequencies range, Hz</td>
<td>0.5–400</td>
<td>1–1000</td>
<td>1–10000</td>
</tr>
<tr>
<td>Amplitude range, mV</td>
<td>0.1–5</td>
<td>0.01–0.5</td>
<td>0.1–50</td>
</tr>
</tbody>
</table>

### 17.4. Differential Amplifier

A differential amplifier is a type of electronic amplifier that multiplies the difference between two inputs by some constant factor $K$. The differential amplifier feature is the presence of two inputs (a non-inverting input terminal and an inverting input terminal) and one common output. The gains for these inputs are equal in magnitude but opposite in sign:

$$U_{out1} = -KU_{in1}; \quad U_{out2} = +KU_{in2}. \quad (17.9)$$

The output voltage $U_{out}$ is proportional to the difference between the voltages $U_{in2}$ and $U_{in1}$ appearing at the two input terminals:

$$U_{out} = U_{out2} + U_{out1} = K(U_{in2} - U_{in1}). \quad (17.10)$$

The measurement of ECG signal provides an excellent example to demonstrate the need for use of a differential amplifier. The magnitude of the ECG signal on the body surface is very small: often less than 1 mV. On the other hand, due to the surrounding power supply lines, there is a strong 50 Hz noise signal on the body surface, and the magnitude of this noise is usually 1000 times larger than that of the ECG signal. To eliminate the noises mentioned above, differential amplifier can use as shown in fig. 17.8.

![Fig. 17.8. Driven-right-leg circuit for minimizing common-mode interference](image-url)
Fig. 17.8 shows that three electrodes, two of them (on the right and left arms) picking up the ECG signal and the third (on the right leg) providing the reference potential, connect the subject to the amplifier. The large 50 Hz noise \( U_{\text{noise}} \) in both \( U_{\text{in}2} = \varphi_2 - \varphi_0 + U_{\text{noise}} \) and \( U_{\text{in}1} = \varphi_1 - \varphi_0 + U_{\text{noise}} \) (which is called the common mode signal) will be cancelled out, and the ECG signal — the voltage drop between left leg and right arm (which is called the differential signal) is amplified:

\[
U_{\text{out}} = K(\varphi_1 - \varphi_0 + U_{\text{noise}} - \varphi_2 + \varphi_0 - U_{\text{noise}}) = K(\varphi_1 - \varphi_2). \tag{17.11}
\]

Strong rejection of the common mode signal is one of the most important characteristics of a good biopotential amplifier.

**Questions:**

1. Give the main sensor characteristics. What is the difference between passive sensors and active ones?
2. What is a thermocouple? Describe temperature measurement by thermocouple.
3. Explain dependence of electric resistance of conductor on temperature. Give the graph.
4. What is the difference between resistance temperature detectors and thermistors?
5. Describe purpose and constitution of biopotential amplifier.
6. What is a harmonic spectrum of the periodic signal? What is harmonic spectrum determined by? Specify amplitude and frequency parameters for ECG, EEG and electromiogram.
7. Give real amplifier frequency characteristic and amplitude one. How to determine a frequency bandwidth and amplifier dynamic range?
8. How is it possible to amplify the signal without distortion by using a real amplifier?
9. Describe the differential amplifier construction.

---

**Chapter 18. ELECTROMAGNETIC WAVES. LIGHT POLARIZATION**

**18.1. ELECTROMAGNETIC EQUATION. THE ELECTROMAGNETIC SPECTRUM OF RADIATION**

For centuries the nature of light was disputed. In the 17th century, Isaac Newton proposed the «corpuscular theory» stating that light is composed of particles. Other scientists, like Robert Hooke and Christian Huygens, believed light to be a wave. Today it is known that light behaves as both a wave and as a particle. Light undergoes interference and diffraction, as all waves do, but whenever light is emitted, it is always done so in discreet of packets called photons.

*Light is a transverse electromagnetic wave made up of mutually perpendicular, fluctuating electric and magnetic fields with the same amplitude and frequency.*

According to Maxwell’s theory a varying electric field sets up a magnetic one which is also varying. This varying magnetic field sets up an electric field, and so on. Thus, if one use oscillating charges to produce a varying (alternating)
electromagnetic field, then in the space surrounding the charges a sequence of mutual transformations of an electric and a magnetic field propagating from point to point will appear. This process will be periodic in both time and space and, consequently, will be an electromagnetic wave.

This sinusoidally varying electric and magnetic fields can be written as:

\[ E = E_0 \sin \omega(t - \frac{x}{v}), \quad B = B_0 \sin \omega(t - \frac{x}{v}), \]  

(18.1)

where \( E_0 \) and \( B_0 \) are the amplitude values of the electric field and the magnetic induction respectively; \( \omega = 2\pi v \) is the angular frequency; \( t \) is the time; \( v \) is the velocity; \( x \) is the coordinate.

If the wave is sinusoidal, then the period \( T \), frequency \( v \), and wavelength \( \lambda \), are related by \( v = 1/T, \lambda = c/v \). The velocity \( v \) of electromagnetic waves is determined by formula:

\[ v = \frac{1}{\sqrt{\varepsilon_0 \mu_0}}, \]  

(18.2)

where \( \varepsilon \) is a permittivity; \( \varepsilon_0 = 8.85 \cdot 10^{-12} \text{ F/m}; \mu \) is a permeability; \( \mu_0 = 1.43 \cdot 10^7 \text{ G}\text{n/m}. \)

In a vacuum (i.e. when \( \varepsilon = \mu = 1 \)), the velocity \( c \) of electromagnetic waves is maximum. Light travels in a vacuum with a velocity \( c \approx 3 \cdot 10^8 \text{ m s}^{-1} \). When light travels through matter, its velocity \( v \) is less than this and is given by \( v = c/n \), where \( n \) is the index of refraction of the substance. The value of the index of refraction \( n \) is determined as

\[ n = \frac{c}{v} = \sqrt{\frac{\varepsilon \mu}{\varepsilon_0 \mu_0}} \]  

(18.3)

and depends on both the composition of the substance and the wavelength of the light.

The orderly distribution of electromagnetic waves according to their wavelength or frequency is called the electromagnetic spectrum. Electromagnetic spectrum covers a wide range of wavelengths or frequencies. The whole electromagnetic spectrum has been classified into different parts in order of increasing wavelength and type of excitation. The electromagnetic spectrum includes radio waves, infrared, visible, and ultraviolet light, X-rays and gamma rays. All of these are fundamentally similar in that they move at 300 000 km s\(^{-1}\) the speed of light. The only difference between them is their wavelength (or frequency), which is directly related to the amount of energy the waves carry. The shorter the wavelength of the radiation, the higher the energy.

On one end of the spectrum are radio waves with wavelengths billions of times longer than those of visible light. On the other end of the spectrum are
gamma rays. These have wavelengths millions of times smaller than those of visible light. The following are the basic categories of the electromagnetic spectrum, from longest to shortest wavelength:

**Radio waves** are used to transmit radio and television signals. Radio waves have wavelengths \( \lambda > 10^{-3} \text{ m} \) that range from less than a centimeter to tens or even hundreds of meters.

**Infrared (IR)** is the region of the electromagnetic spectrum that extends from the visible region to about one millimeter (in wavelength \( 10^{-3} \text{ m} > \lambda > 0.76 \cdot 10^{-6} \text{ m} \)). Infrared waves include thermal radiation. Infrared radiation can be measured using electronic detectors and has applications in medicine.

**Visible light**: the rainbow of colors is known as visible light is the portion of the electromagnetic spectrum with wavelengths between 400 to 760 nanometers \((760 \text{ nm} > \lambda > 400 \text{ nm})\). It is the part of the electromagnetic spectrum that we see, and coincides with the wavelength of greatest intensity of sunlight.

**Ultraviolet (UV)** radiation has a range of wavelengths \( 400 \text{ nm} > \lambda > 80 \text{ nm} \). A small dose of ultraviolet radiation is beneficial to humans, but larger doses cause skin cancer and cataracts. UV is used to destroy the bacteria and for sterilizing surgical instruments.

**X-rays** are high energy waves which have great penetrating power and are used extensively in medical applications (as a diagnostic tool) and in inspecting welds. The wavelength range is \( 80 \text{ nm} > \lambda > 10^{-5} \text{ nm} \).

**Gamma rays** have wavelengths of less than \( \lambda < 10^{-5} \text{ nm} \). They are more penetrating than X-rays. Gamma rays are generated by radioactive atoms and in nuclear explosions, and are used in many medical applications (for example, for treatment of cancer).

Light was discovered to have both particle properties and electromagnetic wave properties at the same time. A traveling of light can be described by a wave with wavelength \( \lambda = c/\nu \). As light moves from one medium into another where it travels with a different speed, the frequency remains the same. The wavelength changes as the velocity changes. According to the quantum theory light may at times exhibit properties like those of particles in their interaction with matter. Each particle of light or photon has energy \( E \). The energy of each photon (a «particle» concept) is related to its frequency (a «wave» concept) by

\[
E = h\nu = h \frac{c}{\lambda}. \tag{18.4}
\]

The proportionality constant \( h \) is called the Planck’s constant. It has the numerical value \( h = 6.63 \cdot 10^{-34} \text{ J s} = 4.14 \cdot 10^{-15} \text{ eV s} \). The electron volt (eV) is a unit of energy. It is the energy acquired by an electron that moves through a potential difference of 1 V: 1 eV = 1.6 \cdot 10^{-19} \text{ C} \cdot 1 \text{ V} = 1.6 \cdot 10^{-19} \text{ J}.

It should be noticed the shorter the wavelength of the radiation is, the higher the energy and the more harmful for biological objects will be.
18.2. POLARIZATION OF LIGHT

Light can be represented as a transverse electromagnetic wave made up of mutually perpendicular, fluctuating electric and magnetic fields. Figure 18.1 shows the sinusoidally varying electric field \( E \). The magnetic field \( B \) varies sinusoidally in the direction perpendicular to that of the electric field \( E \) and to the direction of propagation of the wave, as shown in figure 18.1. The electric field, the magnetic field, and the wave direction are all mutually perpendicular.

The plane perpendicular to the direction of propagation of the wave is called the plane of observation (fig. 18.1).

Polarization is a property of electromagnetic waves that describes the orientation of oscillations of electric field vector \( E \). Conventionally, when considering polarization, the electric field vector \( E \) is described and the magnetic field \( B \) is ignored since it is perpendicular to the electric field and proportional to it. Vector \( E \) is called light vector.

Light from an ordinary light bulb is not polarized. In addition to unpolarized light, there is partially polarized light and totally polarized light. Coherent laser light is an example of polarized light. There are three different types of polarization states: linear, circular and elliptical. Each of these commonly encountered states is characterized by a differing motion of the electric field vector \( E \) with respect to the direction of propagation of the light wave. The direction of polarization of each individual wave is defined to be the direction in which the electric field is vibrating.

A wave is said to be linearly polarized (or plane-polarized) if the electric field vector \( E \) oscillates in the same direction (plane) at all times at a particular point. The plane formed by the vibrating electric field and the direction of propagation is called the plane of polarization of the wave. In figure 18.1 the special case of vertically polarized light is represented by a vertical arrow in the observation plane. The tip of the vector \( E \) traces out a single line in the plane of observation as illustrated in fig. 18.2.
The light is **circularly polarized** if electric field vector $\mathbf{E}$ appears to be rotating around the direction of propagation. The magnitude of electric field vector $\mathbf{E}$ remains constant. In this special case the electric field vector $\mathbf{E}$ traces out a circle in the observation plane (fig. 18.3). There are two directions of circularly polarized light. **Right-hand circularly polarized light** is defined such that the electric field is rotating clockwise as seen by an observer towards whom the wave is moving. **Left-hand circularly polarized light** is defined such that the electric field is rotating counterclockwise as seen by an observer towards whom the wave is moving.

The light is **elliptically polarized** if the magnitude of the electric field vector $\mathbf{E}$ varies as it rotates around the direction of propagation. In this case the tip of the electric field vector $\mathbf{E}$ describes an ellipse in an observation plane (fig. 18.4). In case of circular or elliptical polarization, the plane of polarization rotates, in contrast to linear polarization where the plane of polarization is fixed.

Most light sources contain waves in which the electric fields are oriented (oscillated) in all possible directions and this light is referred to as **unpolarized**. All directions of vibration from a wave source are possible. Thus **unpolarized light** is represented as a linear superposition of linearly polarized waves.
The resultant electromagnetic wave is a superposition of waves with electric fields vibrating in many different directions. In fig. 18.5 the arrows show a few possible directions of vibration of the electric fields in the beam.

![Fig. 18.5. Unpolarized light](image)

**Partly polarized light** can be considered as a mixture of unpolarized light and plane-polarized one (fig. 18.6). Ratio of intensity of polarized component \((I_{\text{min}}, I_{\text{max}})\) of a beam to the total intensity is the degree of polarization \(P\):

\[
P = \frac{I_{\text{max}} - I_{\text{min}}}{I_{\text{max}} + I_{\text{min}}}. \tag{18.5}
\]

For plane-polarized light \(I_{\text{min}} = 0\) and therefore the degree of polarization equals unit \((P = 1)\). For natural light \(I_{\text{min}} = I_{\text{max}}\) and \(P = 0\). For partly polarized light the degree of polarization \(P\) is less than unit \((P < 1)\).

![Fig. 18.6. Partly polarized light](image)

It is possible to transform unpolarized light into polarized light. The process of transforming unpolarized light into polarized light is known as **polarization**. It is possible to obtain a linearly polarized light from an unpolarized one by removing all waves from the light except those whose electric field vectors oscillate in a single plane. Processes for accomplishing this include:

– reflection;
– birefringence (double refraction) in crystalline materials;
– selective absorption (dichroism);
– scattering.

**18.3. Polarization by Reflection**

When unpolarized light strikes a boundary between two dielectric surfaces (for example on an air-glass boundary) at any angle other than perpendicular, the reflected and transmitted components are partially plane polarized. Light
with the perpendicular electric field $E$ oscillations to the plane of incidence (is said to be s-polarized) predominates in the reflected ray (in fig. 18.7 these oscillations are denoted by dots). The component with electric field $E$ oscillations parallel to the plane of incidence (is said to be p-polarized) is preferentially transmitted or absorbed (in the fig.18.7 these oscillations are depicted by double-headed arrows).

The amount of polarization in the reflected light depends on the angle, varying from no polarization at normal incidence to 100 % polarization at an angle known as the polarizing angle $\theta_B$. This angle is related to the index of refraction of the two materials on either side of the boundary by the equation:

$$\operatorname{tg} \theta_B = \frac{n_2}{n_1},$$  \hspace{1cm} (18.6)

where $n_1$ is the index of refraction of the medium in which the light is traveling, and $n_2$ is that of the medium beyond the reflecting boundary.

The polarizing angle $\theta_B$ is called Brewster’s angle, and equation (18.6) Brewster’s Law.

![Fig. 18.7](image)

Fig. 18.7. At $\theta_B$ the reflected light is plane-polarized parallel to the plane of incidence, and $\theta_B + \theta_r = 90^\circ$, where $\theta_r$ is the refraction angle. The large dots represent electric field oscillations parallel to the reflecting surface and perpendicular to the page, double-headed arrows represent electric field oscillations perpendicular to those represented by the dots (parallel to the page).

A simple polarizer can be made by tilting a stack of glass plates at Brewster’s angle to the beam (fig. 18.8).

For visible light in air and typical glass, Brewster’s angle is about 57°. Some of the s-polarized light is reflected from each surface of each plate. For a stack of plates, each reflection depletes the incident beam of s-polarized light, leaving a greater fraction of p-polarized light in the transmitted beam at each stage. After one interface the refracted beam will be partially polarized, having lost some of its s-polarized component. If the stack contains many plates, then
the refracted beam will have a high degree of polarization, since at each interface the same fraction of the remaining s-polarization is lost. This pile-of-plates polarization mechanism is used in many polarizing beam splitters, where many layers of dielectric thin film are laid onto the interior prism angle of the beam splitter.

![Diagram of polarization](image)

*Fig. 18.8. A stack of plates at Brewster’s angle to a beam reflects off a fraction of the s-polarized light at each surface, leaving a p-polarized beam. Full polarization at Brewster’s angle requires many more plates than shown.*

18.4. **Optical Anisotropy**

**Isotropic media** is a medium in which light behaves the same way no matter which direction it is traveling. **Anisotropic media**, that is, media in which light behaves differently depending on which direction the light is propagating. The behavior of a light ray that propagates through an anisotropic material is dependent on its polarization. An anisotropy crystal, such as calcite, will divide an entering ray of monochromatic light into two rays having orthogonal polarizations. The rays will usually propagate in different directions and have different propagation speeds. One of the rays is called an **ordinary** and designated by the symbol **o-ray**. Its propagation speed \( v_o = \frac{c}{n_o} \) is constant in different directions. For the other ray, called an **extraordinary ray** and designated by **e-ray**, propagation speed \( v_e = \frac{c}{n_e} \) is various in different directions.

In an anisotropic crystal light going in one or two special directions has the same speed (index of refraction) independent of its polarization. These special directions are called the **optic axes**. Optic axis is a direction in the crystal not one particular line. Crystals with two optic axes are called biaxial. Crystals with one optic axis is called uniaxial ones. Only uniaxial crystals will be discussed here.
A plane passing through an optical axis is called a principal plane of the crystal. Customarily, the principal plane passing through the light ray is used.

The ordinary and extraordinary rays are both completely polarized in mutually perpendicular directions. The plane of oscillations of the ordinary ray is perpendicular to a principal plane of the crystal. In the extraordinary ray, the oscillations of the light vector $E$ occur in a plane coinciding with a principal plane.

**18.5. Polarization in Double Refraction**

Many crystals are anisotropic to light and exhibit properties such as birefringence. **Birefringence** is defined as the double refraction of light in a transparent, molecularly ordered material, which is manifested by the existence of orientation-dependent differences in refractive index.

When a light ray normally incident on a birefringent crystalline surface it will be divided into two rays (ordinary and extraordinary) at the boundary according to the refraction law because $n_o \neq n_e$ (fig. 18.9).

$$\sin \beta_o = \frac{\sin \alpha}{n_o}; \quad \sin \beta_e = \frac{\sin \alpha}{n_e}. \quad (18.7)$$

![Fig. 18.9. Double refraction phenomenon](image)

The index of refraction for the extraordinary ray $n_e$ is a continuous function of direction. The index of refraction for the ordinary ray $n_o$ is independent of direction. The two indices of refraction are equal only in the direction of an optic axis. The extraordinary ray will deviate from the incident direction while the ordinary ray will not. The ordinary ray index $n_o$ and the most extreme extraordinary ray index $n_e$ are together known as the principal indices of refraction of the material. Birefringent crystals are used in many polarization devices. In some devices the difference in the refractive index is used to separate the rays and eliminate one of the polarization states, as in the Nicol prism.
18.6. The Nicol Prism

*The Nicol prism* is an optical device used to generate a beam of polarized light. It was the first type of polarizing prism to be invented, in 1828 by William Nicol. It consists of a rhombohedral crystal of calcite (Iceland spar) that has been cut at a 68° angle, split diagonally, and then joined again using Canada balsam (fig. 18.10).

Unpolarized light enters one end of the crystal and is split into two polarized rays by birefringence. One of these rays (the *ordinary* or *o*-ray) experiences a refractive index of \( n_o = 1.658 \) and at the balsam layer (refractive index \( n_b = 1.55 \)) undergoes total internal reflection at the interface since \( n_o > n_b \), and is reflected to the side of the prism. Then ordinary ray is absorbed by black mounting material in the prism housing. The other ray (the *extraordinary* or *e*-ray) experiences a lower refractive index (\( n_e = 1.486 \)), is not reflected at the interface, and leaves through the second half of the prism as plane polarized light.

Fig. 18.10. The Nicol polarizing prism

18.7. Phenomenon of Dichroism

When light passes through an absorbing material with the absorption coefficient \( k \) its intensity \( I \) decreases exponentially with increase in medium thickness \( x \):

\[
I = I_0 e^{-kx}.
\]

For anisotropic crystals the absorption coefficient \( k \) depends on the polarization of wave. This anisotropy in absorption is called *dichroism* (selective absorption). If the absorption in a material is different for different linear states of polarization, the material is termed *linear dichroic*. In linear dichroic crystals one of the rays (ordinary or extraordinary) is absorbed to a greater extent than the other as illustrated in fig. 18.11.

Phenomenon of dichroism has been taken for manufacturing a polarizing device called a *polaroid*. A polaroid (polarizer) is a material that selectively absorbs light depending on polarization. The mineral tourmaline is the best known of natural dichroic materials. An ordinary ray is virtually completely absorbed in it over a distance of 1 mm. However, this crystal is seldom used as a polarizer, since the dichroic effect is strongly wavelength dependent and
the crystal appears coloured. In crystals of herapathite (iodoquinine sulphate), one of the rays is absorbed over a path of about 0.1 mm. Herapathite is a celluloid film into which a great number of identically oriented minute crystals of iodoquinine sulphate have been introduced. Polaroid filters absorb one component of polarization while transmitting the perpendicular components. The intensity of transmitted light depends on the relative orientation between the polarization direction of the incoming light and the polarization axis of the polarizer and is described quantitatively by Malus intensity Law.

**18.8. Polarized Light Transition through a Polarizer. Malus’s Law**

Plane-polarized light can be obtained from unpolarized light using polarizers. Polarizers freely transmit oscillations parallel to the plane which is called the **plane of polarization (transmission)** and completely or partly retain the oscillations perpendicular to this plane.

Assume that a beam of plane-polarized light of amplitude $E_0$ and intensity $I_0$ strikes a polarizer whose transmission axis is at an angle $\phi$ to the incident polarization direction (fig. 18.12).

**Fig. 18.12.** Light transmission through linear polarizer. Vertical polarizer transmits only the vertical component of a wave (electric field $E_{II} = E_0\cos\phi$) incident upon it.
An amplitude $E_0$ can be resolved into two components: $E_{ii} = E_0 \cos \phi$ (parallel to the transmission axis of polarizer) and $E_{\perp} = E_0 \sin \phi$ (perpendicular to the transmission axis of polarizer).

Only the component of the electric field vector that is along the transmission axis can pass through the polarizer. Therefore the projection of $E_0$ along the transmission axis ($E_{ii}$) should be considered. As light passes through the polarizer, the amplitude of the electric field vector is given by $E_{ii} = E_0 \cos \phi$. The intensity $I$ of the light can be shown to be proportional to $E_{ii}^2$. Hence, the intensity $I$ of the light after it passes through the polarizer is proportional to $E_0^2 \cos^2 \phi$. Equivalently, the transmitted intensity $I$ is related to the incident intensity $I_0$ by the following equation:

$$I = I_0 \cos^2 \phi,$$

(18.8)

where $I_0$ is the incoming intensity and $\phi$ is the angle between the polarizer transmission axis and the plane of polarization of the incoming wave. This equation (18.8) is known as Malus’s Law.

It is obvious from equation (18.8) that if $\phi = 90$ (or 270) degrees, then $I = 0$. One can conclude that if the polarizer is rotated so that no light can pass through, then the polarizer's transmission axis must be perpendicular to the polarization of the incident light. On the other hand, if the transmission axis is aligned with the polarization direction, $\phi = 0$ (or 180) degrees and the transmission is maximum ($I = I_0$).

When unpolarized light is incident on an ideal polarizer, the intensity of the transmitted light is one-half that of the incident light. This can be explained if we resolve the electric fields of the incident waves into components parallel and perpendicular to the polarizing axis. Because the incident light is a random mixture of all states of polarization, these components will, on average, be equal (all the values of $\phi$ are equally probable). Since the polarizer transmits only the component parallel to the axis of polarization, one-half of the incident intensity is transmitted $I = I_0/2$.

**Fig. 18.13.** The crossed polarizers. Their transmission axes are perpendicular to one another — unpolarized light can be entirely stopped.

172
If two polarizers are placed one after another (the second polarizer is generally called an analyzer), the mutual angle between their polarizing axes gives the value of $\varphi$ in Malus’ Law. If their polarizing axes are perpendicular ($\varphi = 90^\circ$), the polarizers are crossed and no light is transmitted (fig. 18.13). If the polarizers axes are parallel ($\varphi = 0$) one-half of the incident intensity $I = I_0/2$ is transmitted.

**18.9. OPTICAL ACTIVITY**

The phenomenon of optical activity is the rotation of plain of polarization of linearly polarized light as it travels through certain materials (fig. 18.14). A material is said to be optically active if it rotates the plane of polarization of linearly polarized light transmitted through it. Such substances include crystalline bodies (for example, quartz, cinnabar), pure liquid (turpentine, nicotine), and solutions of optically active substances in inactive solvents (aqueous solutions of sugar, tartaric acid).

![Fig. 18.14. A rotation of linearly polarized light as it travels through optically active substance](image)

Crystalline substances rotate the plane of polarization to the greatest extent when the light propagates along the optical axis of crystal. The angle of rotation $\alpha$ is proportional to the path $L$ traveled by a ray in the crystal:

$$\alpha = \alpha_0 L,$$  \hspace{1cm} (18.9)

The coefficient $\alpha_0$ is called the rotation constant. It depends on the wavelength (dispersion of the ability to rotate).

In solutions, the angle of rotation of the plane of polarization proportional to the path $L$ of the light in the solution and to the concentration of the active substance $c$:

$$\alpha = \alpha_0 c L,$$  \hspace{1cm} (18.10)

Here $\alpha_0$ is a quantity called the specific rotational constant.

Depending on the direction of rotation of the polarization plane optically active substances are divided into right-hand and left-hand ones. There exist right-hand and left-hand quartz, right-hand and left-hand sugar, etc. If we place
an optically active substance (a crystal of quartz, a transparent tray with a sugar solution, etc.) between two crossed polarizers, then the field of vision becomes bright. To get darkness again, one of the polarizers has to be rotated through the angle \( \alpha \) determined by expression (18.9) or (18.10). When a solution is used, we can determine its concentration \( c \) by equation (18.10) if we know the specific rotational constant \( \alpha_0 \) of the given substance and the length \( L \) and have measured the angle of rotation \( \alpha \). This way of determining the concentration is used in the production of various substances; in particular in the sugar industry (the corresponding instrument is called a saccharimeter). It is used in the sugar industry to measure syrup concentration, in optics to manipulate polarization, in chemistry to characterize substances in solution, and is being developed as a method to measure blood sugar concentration in diabetic people.

**Questions:**

1. What is an electromagnetic wave? What does absolute refractive index characterize?
2. Specify the electromagnetic spectrum ranges.
3. What types of polarization are known? What is the degree of polarization?
5. What is the double refraction phenomenon? Describe properties of ordinary wave and extraordinary one.
6. Explain the Nicol prism construction and a light propagation through it.
7. Explain phenomenon of dichroism. What are the polarizers?
8. Write Malus’s Law.
9. What is the optical activity? How to determine a concentration of optically active substance by polarizer?

**Chapter 19. THERMAL RADIATION**

**19.1. BASIC CHARACTERISTICS OF THERMAL RADIATION**

All bodies in nature are the sources of electromagnetic radiation, which depends on the body temperature. This electromagnetic radiation is called thermal radiation. Let’s consider the characteristics of thermal radiation.

The intensity of thermal radiation is characterized by the magnitude of the energy flux measured in watts (\( W \)) and use the symbol \( \Phi \). The energy flux \( \Phi \) is the total rate of emitted energy and can be written in the form:

\[
\Phi = \frac{E}{t},
\]  

(19.1)

where \( E \) is the radiation energy, \( t \) is the time.

The second important characteristic of thermal radiation is the radiant emittance. The radiant emittance \( R \) of the body is the energy flux emitted by unit surface area of a radiating body in all directions. The radiant emittance \( R \) has units of \( \text{Wm}^{-2} \) and is given by formula:
\[ R = \frac{\Phi}{S} \, , \]  

where \( \Phi \) is the energy flux, \( S \) is a surface area.

Radiation consists of electromagnetic waves having different wavelengths \( \lambda \). Let \( dR_\lambda \) be the energy flux emitted by unit surface area of a body on wavelength \( \lambda \) within the wavelength interval \( d\lambda \). When the interval \( d\lambda \) is small, the flux \( dR_\lambda \) will be proportional to \( d\lambda \):

\[ dR_\lambda = r_\lambda d\lambda . \]  

(19.3)

The \( r_\lambda \) is called the emissivity of a body and \( r_\lambda \) is the spectral radiance of the body. The emissivity \( r_\lambda \) has units of \( \text{Wm}^{-3} \). Like the radiant emittance \( R \), the emissivity \( r_\lambda \) depends greatly on the temperature \( T \) of a body. Thus, \( r_\lambda \) is a function of the wavelength \( \lambda \) and temperature \( T \).

The radiant emittance \( R \) and the emissivity \( r_\lambda \) are related by the formula:

\[ R_T = \int_0^\infty dR_T = \int_0^\infty r_\lambda T d\lambda . \]  

(19.4)

(to stress that the radiant emittance \( R \) and the emissivity \( r_\lambda \) depend on the temperature, we have provided them with the subscript \( T \)).

Assume that the flux of radiant energy \( d\Phi_\lambda \) due to electromagnetic waves whose wavelength \( \lambda \) is within the interval from \( \lambda \) to \( \lambda + d\lambda \) falls on an elementary area of a body’s surface. A part of this flux \( d\Phi_\lambda(\text{abs}) \) will be absorbed by the body. The dimensionless quantity \( \alpha_{\lambda T} \)

\[ \alpha_{\lambda T} = \frac{d\Phi_\lambda(\text{abs})}{d\Phi_\lambda} \]  

(19.5)

is called the absorptivity of a body. The absorptivity \( \alpha_{\lambda T} \) of a body is a function of the wavelength \( \lambda \) and temperature \( T \).

By definition, the absorptivity \( \alpha_{\lambda T} \) cannot be greater than unit. There are three type of the bodies with a different absorptivity: a blackbody, the gray bodies and the all other bodies. Fig. 19.1 shows dependence of the absorptivity \( \alpha_\lambda \) on wavelength \( \lambda \) for a blackbody, the gray bodies and all other bodies.

![Fig. 19.1. Dependence of the absorptivity \( \alpha_\lambda \) on wavelength \( \lambda \): 1 — a blackbody; 2 — the gray bodies; 3 — the all other bodies](image_url)
1. A blackbody completely absorbs the radiation of all wavelengths falling on it. The absorptivity $\alpha_T \equiv 1$ at the all wavelengths.

2. The gray bodies have the same absorptivity at the all wavelengths too, but it is smaller than the unit: $\alpha_{\lambda T} \equiv \alpha_T = \text{const} < 1$.

3. The absorptivity $\alpha_{\lambda T}$ of the all other bodies is not constant and depends on wavelength $\alpha_{\lambda T} = f(\lambda)$.

Blackbodies do not exist in nature. Carbon black and platinum black have an absorptivity $\alpha_{\lambda T}$ close to unit only within a limited range of wavelengths. It is difficult if not impossible to make a surface that is completely absorbing; the absorption can be improved by making a completely enclosed cavity provided with a small hole, as in fig. 19.2. The radiation penetrating in the cavity through the hole will undergo multifold reflections, part of the energy is absorbed upon each reflection and as a result virtually the entire radiation of any frequency is absorbed by such a cavity.

![Fig. 19.2. The model of a blackbody](image)

The radiation entering the hole in the cavity bounce from the walls many times before chancing to pass out through the hole again, and they therefore have a greater chance of being absorbed.

Such a hole in a cavity is a better approximation to a blackbody (model of a blackbody) than is the absorbing material lining the cavity.

The blackbody radiant emittance $R_b$ and its emissivity $\varepsilon_\lambda$ are related by the formula:

$$R_b = \int_0^\infty dR_b = \int_0^\infty \varepsilon_\lambda d\lambda.$$  \hspace{1cm} (19.6)

Much work was done on the properties of blackbody or thermal radiation. While some properties could be explained by classical physics, others could not. The description of the emissivity function $\varepsilon_{\lambda}(\lambda, T)$ by Max Planck is one of the foundations of quantum mechanics. Max Planck has made an assumption absolutely alien to classical notions, namely, to assume that electromagnetic radiation is emitted in the form of separate portion of energy (quanta) whose magnitude is proportional to the frequency of radiation:

$$E = h\nu,$$  \hspace{1cm} (19.7)
where the constant of proportionality \( h \) was subsequently named Planck’s constant. Moreover Planck has obtained formula for the emissivity function \( \varepsilon_{\lambda}(\lambda, T) \):

\[
\varepsilon_{\lambda} = \frac{2\pi hc^2}{\lambda^5} \frac{1}{e^{\frac{hc}{\lambda kT}} - 1},
\]

where \( k \) is the Boltzmann’s constant; \( h \) is the Planck’s constant; \( \lambda \) is the wavelength; \( T \) is the temperature.

**19.2. THERMAL RADIATION LAWS**

Let’s consider the main thermal radiation laws.

**Kirchhoff Law**

There is a definite relation between the emissivity and absorptivity of any body. We can convince ourselves that this is true by considering the following experiment. Assume that several bodies are confined in an enclosure maintained at a constant temperature \( T \) (fig. 19.3).

![Fig. 19.3. Several bodies are confined in an enclosure maintained at a constant temperature \( T \)](image)

The cavity inside the enclosure is evacuated so that the bodies can exchange energy with one another and with the enclosure only by emitting and absorbing electromagnetic waves. Experiments show that such a system will arrive at a state of thermal equilibrium after a certain time elapses — all the bodies will acquire the same temperature \( T \) equal to that of the enclosure. In this state, a body having a greater emissivity \( r_{\lambda T} \) loses more energy from unit surface area in unit time than a body whose emissivity \( r_{\lambda T} \) is lower. Since the temperature (and, consequently, the energy) of the bodies does not change, then the body emitting more energy must absorb more, i.e. have a greater \( a_{\lambda T} \). Thus, the greater the emissivity \( r_{\lambda T} \) of a body, the greater is its absorptivity \( a_{\lambda T} \). Hence follows the relation

\[
\left( \frac{r_{\lambda}}{a_{\lambda}} \right)_1 = \left( \frac{r_{\lambda}}{a_{\lambda}} \right)_2 = \ldots = \frac{\varepsilon_{\lambda}}{1} = \varepsilon_{\lambda},
\]

where the subscripts 1, 2 etc. relate to different bodies. This relation expresses the following law established by the German physicist Gustav Kirchhoff: at the stage of thermal equilibrium the ratio of the emissivity \( r_{\lambda T} \) to the absorptivity \( a_{\lambda T} \) does not depend on the nature of a body, it is the same
The universal function of the wavelength (frequency) and temperature for all bodies and equals the emissivity $\varepsilon_\lambda$ of a blackbody.

The quantities $r_{\lambda T}$ and $a_{\lambda T}$ can vary exceedingly greatly for different bodies. Their ratio, however, is identical for all bodies and equals the emissivity $\varepsilon_\lambda$ of a blackbody. This signifies that a body which absorbs certain rays to a greater extent will emit these rays to a greater extent too.

**Stefan–Boltzmann Law**

For a long time, attempts to obtain the form of the function $\varepsilon_\lambda(\lambda, T)$ theoretically did not provide a general solution of the problem. In 1879 the Austrian physicist Joseph Stefan analysing experimental data, arrived at the conclusion that the radiant emittance $R$ of any gray body is proportional to the fourth power of the absolute temperature. But subsequent more accurate measurements, however, showed that his conclusions were erroneous. In 1884 the Austrian physicist Ludwig Boltzmann, on the basis of thermodynamic considerations, obtained theoretically the following value for the radiant emittance $R_b$ of a blackbody:

$$R_b = \sigma T^4,$$

(19.10)

where $\sigma$ is a constant quantity, and $T$ is the absolute temperature.

Relation (19.10) between the radiant emittance $R_b$ of a blackbody and its absolute temperature $T$ was named the **Stefan–Boltzmann Law**. The constant $\sigma$ is called the **Stefan–Boltzmann constant**. Its experimental value is $\sigma = 5.67 \cdot 10^{-8}\text{ W}\cdot\text{m}^{-2}\cdot\text{K}^{-4}$.

The temperature dependence of the gray body radiant emittance $R_g$ is similar:

$$R_g = a\sigma T^4,$$

(19.11)

where $a$ is the gray body absorptivity.

**Wien’s Displacement Law**

At first the wavelength dependence of the black body emissivity has been established experimentally. The value of $\varepsilon_\lambda$ is plotted for several different temperatures in fig. 19.4.

The German physicist Wilhelm Wien has established the relation between the wavelength $\lambda_{\text{max}}$ corresponding to the maximum of the black body emissivity function $\varepsilon_\lambda(\lambda, T)$ and the temperature $T$, that is known as Wien’s displacement Law:

$$\lambda_{\text{max}} = \frac{b}{T},$$

(19.12)

where the experimental value of the Wien’s constant $b$ is: $b = 2900\text{ μm}\cdot\text{K}$. Wien’s displacement law is severe true only for the black and gray bodies.

This relationship is useful for the determining the temperatures of any hot radiant objects whose temperature is far above that of its surroundings. Thus, when the temperature of a blackbody increases, the overall radiated energy...
increases and the peak of the radiation curve moves to shorter wavelengths (fig. 19.4). The value of $\varepsilon_\lambda$ is plotted for several different temperatures in fig. 19.4. As the blackbody become hotter, the spectrum shifts toward shorter wavelengths.

![Graph of blackbody radiation function for several temperatures.](image)

*Fig. 19.4. The blackbody radiation function for several temperatures. The visible spectrum is marked by $\nu$*

**19.3. Heat Transfer Mechanisms in Cooling the Human Body**

Fig. 19.5 gives a simplified model of the process by which the human body gives off heat.

![Simplified model of heat transfer](image)

*Fig. 19.5. A simplified model of the process by which the human body gives off heat*

Even when inactive, an adult male must lose heat at a rate of about 90 watts as a result of his basal metabolism. One implication of the model is that radiation is the most important heat transfer mechanism at ordinary room temperatures and is $\sim 50\%$ from total heat transfer mechanism. When the body
temperature is above ambient temperature the net radiation loss rate takes the form:

\[ P = a_{gb} \sigma \cdot S(T_b^4 - T_s^4), \]  

(19.1)

where \( P \) is net radiated power, \( S \) is the area of the human body (a typical body area according to physiology texts is equal \( 1.5–2 \text{ m}^2 \)) and \( a_{gb} \) is the absorptivity of the skin (human skin is a near blackbody radiator in the infrared, \( a_{gb} = 0.97 \)), \( \sigma = 5.67 \times 10^{-8} \text{ W m}^{-2} \text{K}^{-4} \) is Stefan-Boltzmann constant. This model indicates that an unclothed person at rest in a room temperature of 23 °Celsius \( (T_s) \) would be uncomfortably cool. The skin temperature of 34 °C \( (T_b) \) is a typical skin temperature taken from physiology texts, compared to the normal core body temperature of 37 °C. In this case the temperatures must be in Kelvins and in this case \( P = 133 \) Watts. This suggests that radiation alone is more than adequate for body under these conditions.

As one of the basic heat transfer mechanisms, convection involves the transport of energy by means of the motion of the heat transfer medium, in this case the air surrounding the body.

Another basic heat transfer mechanism is heat conduction. In estimating the effect of convection on the cooling of the body, it is lumped in with conduction. Together, they are not generally adequate for cooling.

This becomes a problem when the ambient temperature is above body temperature, because all three standard heat transfer mechanisms work against this heat loss by transferring heat into the body. Since there must be a net outward heat transfer, the only mechanisms left under those conditions are the evaporation of perspiration from the skin and the evaporative cooling from exhaled moisture.

19.4. INFRARED RADIATION FROM THE HUMAN BODY

The principle of infrared thermography is based on the physical phenomenon that any body of a temperature above absolute zero \((-273 °C)\) emits electromagnetic radiation. There is clear correlation between the surface of a body and the intensity and spectral composition of its emitted radiation. By determining its radiation intensity the temperature of an object can thereby be determined in a non-contact way. Researches on thermal processes developed inside the human body and on the quantity of heat emitted by the body in its environment allowed obtaining important information about the equilibrium between the human body and its environment and about the body’s biological activity and state of health. Methods like thermography and thermovision, involving measuring human body’s temperature, are presently used at present as medical diagnose methods for diseases even in their early stages of development. A very accurate piece of information about the thermal processes that are developing inside the human body can be obtained from direct measurements of the heat emitted by the body’s surface using thermal flux sensors of a thermoelectric type. Thermoelectric effects occurring in anisotrope and
inhomogeneous media are involved in functioning of this type of sensors that can detect heat fluxes up to $10^{-8} \text{ W/cm}^2$. Thermographic cameras detect radiation in the infrared range of the electromagnetic spectrum (roughly 0.9–14 µm) and produce images of that radiation. Since infrared radiation is emitted by all objects based on their temperatures, according to the black body radiation law, thermography makes it possible to «see» one’s environment with or without visible illumination.

The human body radiates energy in the infrared, and this is a significant source of energy loss. According to Wien’s displacement Law one can estimate the wavelength $\lambda_{\text{max}}$ corresponding to the human body maximum emissivity considering the temperature of skin is $T = 34 ^\circ \text{C} = 273 + 34 = 307 \text{ K}$:

$$\lambda_{\text{max}} = \frac{b}{T} = \frac{2900 \mu\text{mK}}{307 \text{ K}} = 9.5 \mu\text{m}. \quad (19.14)$$

In the infrared region in which the human body radiates, the skin is very nearly a blackbody. Measurements of the absorptivity of human skin have shown that for $5 \mu\text{m} < \lambda \leq 25 \mu\text{m}$, $a_{xt} = 0.98 \pm 0.01$. This value was found for white, black, and burned skin.

Two types of infrared imaging are used.

1. In near infrared photography the subject is illuminated by an external source with wavelengths from 0.8 to 25 µm. The difference in absorption between oxygenated and nonoxygenated hemoglobin allows one to view veins lying within 2 or 3 mm of the skin. Either infrared film or a solid-state camera can detect the reflected radiation. Thermal imaging detects thermal radiation from the skin surface.

2. Significant emission from human skin occurs in the range 4–30 µm, with a peak at 9 µm (fig. 19.6). The detectors typically respond to wavelengths below 6–12 µm.

Thermography began about 1957 with a report that skin temperature over a breast cancer was slightly elevated. There was great hope that thermography would provide an inexpensive way to screen for breast cancer, but there have been too many technical problems. Normal breasts have more variability in vascular patterns than was first realized, so that differences of temperature at corresponding points in each breast are not an accurate diagnostic criterion.

Advantages of thermography are: it shows a visual picture so that can help compare temperatures over a large area; it is capable of catching moving targets in real time; measurement in areas inaccessible or hazardous for other methods; it is a non-destructive test method. There are some limitations and disadvantages of thermography: quality cameras are expensive and are easily damaged; images can be hard to interpret accurately even with experience; accurate temperature measurements are very hard to make because of emissivities; most cameras have ±2 % or worse accuracy (not as accurate as contact); ability to only measure
surface body areas. IR detector allows to determine temperature of internal organs with accuracy up to 0.1–0.2 °C.

Fig. 19.6. The blackbody radiation function $\varepsilon(\lambda, T)$ for $T = 310$ K and $T = 312$ K

Thermography has also been proposed to detect and to diagnose various circulatory problems. Clinical applications of thermography are phlebology — vein thrombosis, vascular cancer, ischemia of the limbs. Spectron thermography for use also include: adjunctive diagnostic screening for the detection of the breast cancer, neuromusculoskeletal disorders, extracranial cerebral and facial vascular disease, thyroid gland abnormalities, and various other neoplastic, metabolic and inflammatory conditions.

Questions:
1. What basic characteristics of thermal radiation are known? Specify relationship between the characteristics.
2. What is an absorptivity of a body? What three types of the bodies with a different absorptivity are known?
4. Describe human body infrared radiation, its spectrum and peak emission wavelength.
5. Explain heat transfer mechanisms in cooling the human body.
6. Describe thermography fundamentals. Explain advantages of this method.
Chapter 20. OPTICAL SPECTRA OF ATOMS AND MOLECULES

20.1. LIGHT ABSORPTION

When light travels through a medium, it interacts with the medium. The important interactions are absorption and scattering. Absorption is a transfer of energy from the electromagnetic wave to the atoms or molecules of the medium resulting in decreasing of incident light intensity (fig. 20.1).

![Diagram of absorption of a beam of light as it travels through a cuvette of width x. I is the transmitted light, I₀ is the incident light.](image)

The lose of energy depends on the path length of light in the medium, properties of the material and the light wavelength. The absorption is described by the empirical expression called Bouguer’s Law of absorption. Bouguer described how intensity changes with distance in an absorbing medium. For solids for a parallel beam of light passing through a homogeneous absorbing material (fig. 20.2) the lose of light intensity ΔI is proportional to the path length through the material Δx and the initial light intensity I:

\[ \Delta I = -kI \Delta x \quad \text{or} \quad dI = -kIdx, \quad (20.1) \]

where the coefficient of proportionality k is called the coefficient of light absorption or the linear decay constant. The distance x light traveled through the medium is called the path length. The solution of the differential equation (20.1) is the function \( I = I(x) \). To obtain the general solution of the differential equation (20.1) it is necessary to separate the variables first. Variable I should be on the left side, and variable x — on the right side:

\[ \frac{dI}{I} = -kdx. \]

Then it is necessary to integrate left and right sides:

\[ \int \frac{dI}{I} = -k \int dx \]

\[ \ln I = -kx + \ln C \]

\[ I = Ce^{-kx} \]
Taking into account the initial conditions: if the medium is absent \( x = 0 \), the light intensity \( I \) passed through the absorbing medium will be the same as the intensity \( I_0 \) of incident light \( I = I_0 \), one can find constant \( C = I_0 \) and obtain a particular solution of the differential equation (20.1):

\[
I = I_0 e^{-kx},
\]

(20.2)

where \( I \) is intensity of transmitted light; \( I_0 \) is intensity of incident light; \( k \) is the linear decay constant; \( x \) is path length of the light absorbing sample.

The equation (20.1) is the Bouguer’s Law of absorption. When light passes through a uniform absorbing medium with the linear decay constant \( k \) its intensity \( I_0 \) decreases exponentially with increase of medium thickness \( x \) (fig. 20.3).

The linear decay constant \( k \) is a characteristic of the medium and depends on wavelength \( \lambda \). The unit of the linear decay constant \( k \) is \( \text{m}^{-1} \). The linear decay constant \( k \) is the path length over which the intensity \( I \) is attenuated to \( 1/e \). The dependence of \( k(\lambda) \) or \( k(\nu) \) is individual for each substance and determines its absorption spectrum.

Beer found that linear decay constant \( k \) for a solution of an absorbing substance is linearly related to its concentration \( C \):

\[
k = \alpha C,
\]

(20.3)

where \( \alpha \) is molar extinction coefficient depending on the \( k \) and wavelength \( \lambda \).

Beer’s Law is valid at low concentrations, but breaks down at higher concentrations. Bouguer’s and Beer’s Laws are combined to describe
the attenuation of light by a solution (fig. 20.4). Substituting equation (20.3) in equation (20.2) the Lambert–Beer–Bouguer Law can be obtained.

The **Lambert–Beer–Bouguer Law** states that the amount of light $I$ absorbed by a solution is an exponential function of the concentration $C$ of the absorbing substance present and of the length of the path $x$ of the light through the sample:

$$I = I_0 e^{-aCx}. \quad (20.4)$$

Typical units are: $[k] = \text{cm}^{-1}$; $[C] = \text{M}$ or (moles/liter); $[a] = \text{M}^{-1}\text{cm}^{-1}$.

Let’s introduce two characteristics of light absorbance by substance: **transmittance $T$** and **optical density $D$**.

$T$ is called the **transmittance** and defined as the ratio of the intensity $I$ of transmitted light to intensity $I_0$ of incident light:

$$T = \frac{I}{I_0} = \frac{I_0 e^{-aCx}}{I_0} = e^{-aCx}. \quad (20.5)$$

Transmittance $T$ is a measure of the light that passes through the sample. $T$ is expressed in percent (%). 100 % transmittance means no light is absorbed by the solution so that incident light is 100 % transmitted. Transmittance $T$ exponentially depends on the concentration $C$ of the colored solution (fig. 20.5).

**Optical density $D$** is the logarithmic ratio of intensity $I_0$ of incident light to that $I$ of transmitted light. $D$ can be defined as the base-ten logarithm of the reciprocal of the transmittance:
\[ D = \log \frac{I_0}{I} = \log \frac{1}{T} = -\log T = k_1 x = \alpha_1 C x, \quad (20.6) \]

where \( k_1 = k \lg e = 0.43 k, \quad \alpha_1 = \alpha \lg e = 0.43 \alpha. \)

Optical density \( D \) represents the amount of light absorbed by the sample. \( D \) is directly proportional to concentration \( C \) of the colored solution (fig. 20.6).

Optical density \( D \) depends on wavelength \( \lambda \) of incident light. Moreover the dependence \( D(\lambda) \sim k(\lambda) \sim \alpha(\lambda) \) determines the substance absorption spectrum. \( D \) has not unit (numerical number only).

**Colorimetry** is the method of definition of the colored solution concentration \( C \). Colorimetry is based on the Lambert–Beer–Bouguer Law. The colorimeter is used to quantitatively measure and record the light absorption (\( D \)) and transmission (\( T \)) of a colored solution at a specific wavelength (fig. 20.7).

![Fig. 20.6. Dependence of the optical density \( D \) on the concentration \( C \) of the solution](image)

![Fig. 20.7. Basic design of the colorimeter](image)

Schematic diagram of a single-beam colorimeter is shown in fig. 20.8.

A colorimeter measures the intensity of light passing through a colored solution compared to the intensity of light passing into a reference solution of the same solvent. A detector measures the transmittance \( T \) (% of light passing through) of the solution. This is mathematically converted to optical density \( D \) (\( D = -\log T \)) (absorbance). The optical density \( D \) is directly proportional to the solution concentration \( C \). Higher optical density (absorbance) \( D \) means
a higher concentration $C$ and a more intense colour. By measuring the light absorbed one can determine the concentration of a solution $C$.

*Fig. 20.8. Schematic diagram of a single-beam colorimeter:*

1 — source of light (visible light has a range of wavelength of 400–760 nm); 2 — lens (for creating parallel beam of light); 3 — wavelength selector (Coloured filter) is used to remove all but a narrow band (specific wavelength) of visible radiation; 4 — a sample cell (light is passed (transmitted) thought the sample contained within a little glass (plastic) cuvette. A solution containing an absorbing material (4) is compared to a reference solution (8) of the same solvent and non-absorbing materials. The transmittance of the reference solution is set to 100 % (Abs = 0), then the relative transmittance of the solution is measured; 6 — photoelectric cell (light that penetrates hits the photoelectric cell. The current developed by the photoelectric cell is translated into percent transmission or absorbance through a detector; 7 — detector (records transmittance $T$ and optical density $D$)

Colorimetry is used to measure haemoglobin content of blood, glucose in blood, cholesterol.

**20.2. LIGHT SCATTERING**

Any material whose refractive index is different from that of the surrounding medium (optically inhomogeneous) scatters light. **Scattering** is the process by which particles suspended in a medium of a different index of refraction diffuse a portion of the incident radiation in all directions (fig. 20.9).

Difference between scattering and absorption is:
- both scattering and absorption remove flux from an incident wave;
- during scattering process flux is not lost from the incident beam but is redistributed over the total solid angle centered around the scatterer and it does not change the internal energy states of the molecules;
- absorption changes the internal energy states of the molecules;
- absorption is spectrally selective, scattering is not;
- scattering depends on the ratio of particle size to wavelength of light.
When light passes through a scattering medium its intensity $I$ decreases exponentially with increase in medium thickness $x$.

$$I = I_0 e^{-\sigma x},$$  \hspace{1cm} (20.7)

where $\sigma$ is the scattering coefficient, $I_0$ is the intensity of the incident light. Scattering coefficient $\sigma$ is a characteristic of the medium and depends on wavelength $\lambda$ of light; $[\sigma] = m^{-1}$.

When light passes through a medium which is absorbing and scattering simultaneously, its intensity $I$ decreases exponentially with increase in medium thickness $x$:

$$I = I_0 e^{-(\sigma + k)x},$$ \hspace{1cm} (20.8)

where $I$ is intensity of the transmitted light; $I_0$ is intensity of the incident light; $\sigma$ is the scattering coefficient; $k$ is the linear decay constant; $x$ is path length of the medium.

The strength of scattering depends on the wavelength of the light and also the size of the particle which cause scattering. When a light penetrates into a medium composed of particles whose sizes are much smaller than the wavelength of the incident light ($d < 0.2\lambda$) the scattered intensity on both forward and backward directions are equal (fig. 20.10), the scattering process is elastic and is called Rayleigh scattering. In this scattering process, the energy (and therefore the wavelength) of the incident light is conserved and only its direction is changed. The scattering phase function, or phase function, gives the angular distribution of light intensity scattered by a particle at a given wavelength (fig. 20.11). The scattering function is given by $I_{\text{scatt}}/I_0 = (1 + \cos^2 \varphi)$ for the Rayleigh scattering.

$$\text{Fig. 20.10. The scattered intensity on both forward and backward directions are equal for Rayleigh scattering}$$
The intensity $I_{\text{scatt}}$ of the scattered light is inversely proportional to the fourth power of the wavelength $\lambda$. This is known as Rayleigh scattering Law:

$$I_{\text{scatt}} \sim \frac{1}{\lambda^4}.$$  \hspace{1cm} (20.9)

Hence, the shorter wavelengths are scattered much more than the longer wavelengths. The blue appearance of sky is due to scattering of sunlight by the atmosphere. According to Rayleigh’s scattering Law, blue light is scattered to a greater extent than red light. This scattered radiation causes the sky to appear blue. At sunrise and sunset the rays from the sun have to travel a larger part of the atmosphere than at noon. Therefore most of the blue light is scattered away and only the red light which is least scattered reaches the observer. Hence, sun appears reddish at sunrise and sunset.

When light passes through a colloidal solution it is scattered by the particles of solution. When a light penetrates into a medium composed of particles whose sizes are comparable to wavelength of radiation (aerosols, water vapour) $d > 0.2\lambda$, the scattered intensity is more in the forward direction relative to the backward direction (fig. 20.12). The scattering of light by the colloidal particles is called Tyndal scattering.

$$I_{\text{scatt}} \sim \frac{1}{\lambda^2}.$$  \hspace{1cm} (20.10)

In case of Tyndal scattering the intensity $I_{\text{scatt}}$ of the scattered light is inversely proportional to the second power of the wavelength $\lambda$.
For Tyndal scattering the longer wavelengths are scattered more than shorter ones.

When scatterers are large \((d \gg \lambda)\) the dependence of intensity \(I_{\text{scatt}}\) on \(\lambda\) practically disappears and angular distribution of scattered intensity becomes more complex. Clouds are white as all wavelengths are scattered equally.

### 20.3. Types of Spectrum

Atomic spectroscopy is the determination of elemental composition of a test sample and also the relative concentration of the several composition compounds by its absorption or emission spectrums. An absorption spectrum occurs when light passes through a cold, dilute gas and gas atoms absorb light at characteristic frequencies. This gives rise to dark lines (absence of light) in the spectrum (fig. 20.13). An element's emission spectrum is the relative intensity of electromagnetic radiation of each frequency it emits when it is excited (fig. 20.14).

![Fig. 20.13. Hydrogen absorption spectrum](image1)

![Fig. 20.14. Hydrogen emission spectrum](image2)

Each element emits a characteristic set of discrete wavelengths according to its electronic energy structure, by observing these wavelengths the sample elemental composition can be determined. With the use of emission spectroscopy in the late 19\(^{th}\) century, it was found that the radiation from hydrogen, as well as other atoms, was emitted at specific quantized frequencies. It was the effort to explain this radiation that led to the first successful quantum theory of atomic structure, developed by Niels Bohr.

**The substance absorption spectrum** \(k(\lambda)\) or \(D(\lambda)\) is the dependence of the absorption coefficient (decay constant) \(k\) or the substance optical density \(D\) on the frequency \(k(\nu)\), \(D(\nu)\) or the wavelength \(k(\lambda)\), \(D(\lambda)\) of the incident radiation. Each substance has its own individual absorption spectrum.

**The substance emission spectrum** is the dependence of electromagnetic radiation intensity \(I\) of the substance on the frequency \(I(\nu)\) or wavelength \(I(\lambda)\). Each substance has its own individual emission spectrum.
20.4. BOHR’S THEORY OF THE HYDROGEN ATOM

The simplest system that can emit or absorb light is an isolated atom. In the early part of the 20th century, experiments by Ernest Rutherford and others had established that atoms consisted of a diffuse cloud of negatively charged electrons surrounding a small, dense, positively charged nucleus. The planetary model of the atom still had shortcomings. Firstly, a moving electric charge emits electromagnetic waves; according to classical electromagnetism, an orbiting charge would steadily lose energy and spiral towards the nucleus, colliding with it in a tiny fraction of a second ($10^{-9}$ s). Secondly, the model did not explain why excited atoms emit light only in certain spectrum. Quantum theory revolutionized physics at the beginning of the 20th century when Max Planck and then Albert Einstein postulated that light energy is emitted or absorbed in fixed amounts known as quanta. In 1913, Niels Bohr used this idea in his model of the atom, in which the electrons could only orbit the nucleus in particular circular orbits with fixed angular momentum and energy (fig. 20.15). They were not allowed to spiral into the nucleus, because they could not lose energy in a continuous manner; they could only make quantum leaps between fixed energy levels.

![Fig. 20.15. The Bohr atom model](image)

**Bohr’s theory** of the hydrogenic (one-electron) atom is based on the following postulates:

1. An atom can exist in certain allowed or **stationary states**, with each state having a definite value for its total energy $E_1, E_2, E_3 \ldots E_n$ (fig. 20.16). When the atom is in one of these states it is stable and does not radiate energy.

2. An atom emits or absorbs energy only when an electron moves from one stable state with energy $E_n$ to another stable state with energy $E_k$. In a transition from its initial state to its final state, a photon is either **emitted** (if $E_n > E_k$) or **absorbed** (if $E_n < E_k$) and the energy $h\nu$ of the photon is equal to the difference in the energy of the two states:

$$h\nu = |E_n - E_k|.$$  \hspace{1cm} (20.11)
Example of the light emission is illustrated in fig. 20.17. The electron jumps from an stationary orbit of higher energy $E_2$ to an stationary orbit of lower energy $E_1$ and a photon of energy $h\nu = E_2 - E_1$ is emitted.

![Energy level diagram](image)

**Fig. 20.16.** The Bohr atom (a). Energy level diagram (b)

The total energy of an orbiting electron is quantized such that the electron’s angular momentum has a set of discrete values given by equation:

$$m\nu r = \frac{nh}{2\pi}, \quad n = 1, 2, 3... \quad (20.12)$$

i.e. electron revolves in orbits where the angular momentum of electron is an integral multiple of $\hbar/2\pi$, where $m$ is the mass of the electron, $\nu$ is the electron orbital velocity, and $r$ is the orbital radius, $\hbar$ is Planck's constant.

The electron can absorb energy from some source and jump from a lower energy level to a higher energy level and then emits energy jumping from a higher energy level to a lower energy level as shown in the following fig. 20.18.

Thus from the Bohr model of the atom follows that electrons exist only in the certain energy levels within an atom. The electron energy in these levels has well defined values and electrons jumping between them must absorb or emit the energy equal to the difference between them. In optical spectroscopy, the energy absorbed by electron to move it to a higher energy level and/or
the energy emitted as the electron moves to a lower energy level is in the form of a photon (a particle of light). Because this energy is well-defined, an atom’s identity can be found by the energy of this transition. The wavelength $\lambda$ of the emitted light can be related to its energy $h \nu = \frac{hc}{\lambda} = \Delta E$. It is usually easier to measure the wavelength of light than to directly measure its energy.

\[ h \nu = \frac{hc}{\lambda} = \Delta E. \]

![Diagram of energy levels and transitions](image)

Fig. 20.18. The various ways of an energy absorption by electron from some source (jump from a lower energy level to a higher energy level) and then energy emission (jump from a higher energy level to a lower energy level)

### 20.5. ENERGY STATES OF A HYDROGEN ATOM

The above postulates can be used to calculate allowed energies of the atom for different allowed orbits of the electron. The theory developed should be applicable to hydrogen atoms and ions having just one electron. Thus, within the Bohr atom framework, it is valid for He$^+$, Li$^{++}$, Be$^{3+}$ etc.

Let us consider allowed energies of the hydrogen atom for different allowed orbits of the electron. An electron in a circular orbit of radius $r$ would have a centripetal acceleration $a = \frac{v^2}{r}$ produced by the electrical force of attraction between the negative electron and the positive nucleus. This force is given by Coulomb’s law:

\[ F = \frac{(Ze)(e)}{4\pi\varepsilon_0 r^2} = \frac{Ze^2}{4\pi\varepsilon_0 r^2}. \]  

(20.13)

The charge on the nucleus is $+Ze$, where $Z$ is the number of positive charges (i.e., protons). For the hydrogen atom, $Z = +1$. In Newton’s second Law, $F = ma$, we substitute Coulomb’s Law for $F$, and $a = \frac{v^2}{r}$ for a particular allowed orbit of radius $r$, and obtain:

\[ F = ma; \]

\[ \frac{Ze^2}{4\pi\varepsilon_0 r^2} = \frac{mv^2}{r}. \]  

(20.14)
Using Bohr’s angular momentum quantization rule (20.12) for the value \( n \), the principal quantum number, we obtain both the velocity \( \upsilon \) and the radii of all possible orbits \( r \) as:

\[
\upsilon = \frac{Ze^2}{2\varepsilon_0 \hbar n}, \quad r = \frac{\varepsilon_0 \hbar^2 n^2}{m \pi Ze^2}.
\] (20.15)

It is seen that the allowed radii are proportional to \( n^2 \). The smallest orbit is for \( n = 1 \), and for hydrogen (\( Z = 1 \)) has the value \( r_1 = 0.0529 \text{ nm} \). \( r_1 \) is called the Bohr radius and is a convenient unit for measuring lengths in atomic physics. It is generally denoted by the symbol \( a_0 \). The second allowed radius is \( 4a_0 \) and the third allowed radius is \( 9a_0 \) and so on. In general, the radius of the \( n_{th} \) orbit is

\[
r_n = n^2 a_0.
\] (20.16)

In each of its possible orbits, the electron would have a definite energy, as the following calculation shows. The total energy equals the sum of the kinetic and potential energies. The potential energy of the electron is given by

\[
E_p = -e\varphi = -\frac{Ze^2}{4\pi \varepsilon_0 r},
\] (20.17)

where \( \varphi \) is the potential due to a point charge \( +Ze \) as given:

\[
\varphi = \frac{Ze}{4\pi \varepsilon_0 r}.
\] (20.18)

When we substitute \( r \) from equation (20.15) into equation (20.17), we obtain the potential energy of the electron \( E_p \):

\[
E_p = -e\varphi = -\frac{Ze^2}{4\pi \varepsilon_0 r} = -\frac{mZ^2 e^4}{4\varepsilon_0^2 \hbar^2 n^2},
\] (20.19)

The kinetic energy of the electron in the \( n_{th} \) orbit is:

\[
E_k = \frac{m\upsilon^2}{2} = \frac{mZ^2 e^4}{8\varepsilon_0^2 \hbar^2 n^2}.
\] (20.20)

The total energy \( E \) for an electron in the \( n_{th} \) orbit of radius \( r \) is the sum of the kinetic and potential energies:

\[
E = E_p + E_k = -\frac{mZ^2 e^4}{8\varepsilon_0^2 \hbar^2 n^2} < 0, \quad n = 1, 2, 3 \ldots.
\] (20.21)

If we evaluate the constant term in equation (20.21) and convert it to electron volts, one can obtain:

\[
E_n = -(13.6 \text{ eV})\frac{Z^2}{n^2}, \quad n = 1, 2, 3 \ldots.
\] (20.22)
The lowest energy level \((n = 1)\) for hydrogen \((Z = 1)\) is \(E_0 = -13.6\) eV. Since \(n^2\) appears in the denominator of equation (20.22), the energies of the larger orbits in hydrogen \((Z = 1)\) are given by

\[
E_n = \frac{E_0}{n^2} = \frac{-13.6\text{ eV}}{n^2}.
\]

(20.23)

For example,

\[
E_2 = \frac{-E_0}{2^2} = \frac{-13.6\text{ eV}}{4}.
\]

(20.24)

\[
E_3 = \frac{-E_0}{3^2} = \frac{-13.6\text{ eV}}{9}.
\]

(20.25)

The quantum number \(n\) that labels the orbit radii also labels the energy levels. The lowest energy level or energy state has energy \(E_1\), and is called the ground state. The higher states, \(E_2, E_3,\) and so on, are called excited states (fig. 20.16). The fixed energy levels are also called stationary states.

Notice that although the energy for the larger orbits has a smaller numerical value, all the energies are less than zero. Thus, \(E_2 = -3.4\) eV is a higher energy than \(E_1 = -13.6\) eV. Hence the orbit closest to the nucleus \((r_1)\) has the lowest energy \(E_1 = -13.6\) eV. If an electron is free and has kinetic energy, then the total energy \(E > 0\). Since \(E > 0\) for a free electron, then an electron bound to an atom needs to have \(E < 0\). The minimum energy required to remove an electron from an atom initially in the ground state is called the binding energy or ionization energy. The ionization energy for hydrogen has been measured to be 13.6 eV, and this corresponds precisely to removing an electron from the lowest state, \(E_1 = -13.6\) eV, up to \(E = 0\) where it can be free.

It is useful to show the various possible energy values as horizontal lines on an energy-level diagram. This is shown for hydrogen in fig. 20.19. The electron in a hydrogen atom can be in any one of these levels according to Bohr’s theory.

The radiation of atoms that do not interact with one another consists of separate spectral lines. The emission spectrum of atoms is accordingly called a line spectrum. The atomic spectra show the energy structure of atoms therefore the studying of these spectra served as a key to cognition of the structure of atoms. It was noted first of all that the lines in the spectra of atoms are arranged not chaotically, but are combined into groups or, as they are called, series of lines. One can predict frequencies of the spectral lines emitted by combining equation (20.11) with equation (20.23):

\[
\nu = \frac{E_k - E_n}{h} = \frac{E_0}{h} \left( \frac{1}{n^2} - \frac{1}{k^2} \right),
\]

(20.26)

where \(n = 1, 2, 3, 4, \ldots; k = n + 1, n + 2, n + 3 \ldots\). Equation (20.26) is the generalized Balmer formula.
Once in an excited state, an atom’s electron can jump down to a lower state, and give off a photon in the process. This is, according to the Bohr model, the origin of the emission spectra. The vertical arrows in fig. 20.18 represent the transitions or jumps that correspond to the various observed spectral lines.

The group of spectral lines that corresponds to transitions from any higher energy levels to certain low level forms spectral series. There are some spectral series in hydrogen emission spectrum:

1. **The Lyman series** of lines corresponds to transitions or «jumps» that bring the electron down to the ground state \( E_1 \) \((n = 1)\) from any exited energy levels \( k \geq 2 \) (where \( n \) and \( k \) are the principal quantum numbers of the states). The lines of the Lyman series are located in the ultraviolet range of the spectrum. The frequencies of the Lyman series are obtained from formula (20.26) if \( n = 1 \) and \( k = 2, 3, 4, 5, \ldots \):

\[
v = \frac{E_0}{h} \left( 1 - \frac{1}{k^2} \right), \text{ where } k = 2, 3, 4, 5 \ldots \tag{20.27}
\]
2. The Balmer series is characterized by the electron transitions from any excited energy levels $k \geq 3$ to the second energy level $E_2 (n = 2)$, where $n$ and $k$ are the principal quantum numbers of the states. The spectral lines associated with this series are located in the visible part of the electromagnetic spectrum. The frequencies of the Balmer series can be represented in the form:

$$v = \frac{E_k - E_2}{h} = \frac{E_0}{h} \left( \frac{1}{4} - \frac{1}{k^2} \right), \text{ where } k = 3, 4, 5, 6 \ldots . \quad (20.28)$$

3. The Paschen series is the emission lines corresponding to an electron transitions from $k \geq 4$ to the third energy level $E_3 (n = 3)$. The lines of the Paschen series are located in the near infrared range of the spectrum. The frequencies of the Paschen series are given by formula:

$$v = \frac{E_0}{h} \left( \frac{1}{9} - \frac{1}{k^2} \right), \text{ where } k = 4, 5, 6, 7 \ldots . \quad (20.29)$$

20.6. Molecular spectrum

The internal energy of a molecule $E_{mol}$ includes the electronic energy $E_{el}$, the vibrational energy of nuclei $E_{vib}$, and the rotational energy of the whole molecule $E_{rot}$:

$$E_{mol} = E_{el} + E_{vib} + E_{rot}. \quad (20.30)$$

The every type of the internal molecule energy is quantized. In a molecular system every electronic state includes some vibrational levels and a lot of rotational ones as shown in fig. 20.20. The rotational motion of molecule is quantized and described by the rotational quantum number $j$ also giving a ladder of unequally spaced energy levels. Separations of rotational energy levels ($h\nu = E_{j2} - E_{j1}$) correspond to the microwave region of the electromagnetic spectrum. The vibration motion of nuclei is also quantized and described by vibrational quantum number $v$.

Vibration and rotational energy levels are very closely spaced while the energy spacing of electronic levels are much larger:

$$E_{el} >> E_{vib} >> E_{rot}.$$ Absorption of a photon results in a change of the electronic energy accompanied by changes in the vibrational and rotational energies. If the molecule is initially at the ground state $E_1 (n = 1)$, the molecule will remain in this state unless got excited. When an radiation of resonant frequency $\nu$ is incident on the species the molecule will absorb the incident energy and jump from its ground electronic state $E_1$ to one of the various vibrational-rotational states of excited electronic level $E_2 (n = 2)$ or $E_3 (n = 3)$. These transitions between closed located quantized energy states result in the absorption spectral lines formation. Absorption spectral series is composed
of more than one wavelength $\lambda$ of light and this spectral line is broadening, thus resulting in the formation of the first and the second absorption bands. Collisions with other molecules cause the excited molecule to lose vibrational energy until it reaches the lowest vibration state of the excited electronic state. So, in $\tau \sim 10^{-11} - 10^{-12}$ seconds molecule turns from the higher vibrational-rotational state of electronic level $E_2 (n = 2)$ to the lowest vibrational-rotational state of the excited electronic state $E_2 (n = 2)$, losing energy by non-radiative means, such as transfer of energy as heat to another molecules. This phenomenon is called internal conversion. After that a molecule falls back down to the any vibrational-rotational energy levels of the ground electronic state $E_1 (n = 1)$ and leaves the excite state. These transitions result in the emission spectral lines formation. Note however that the emission extends over a range of frequencies, thus spectral lines are broadening. Wavelength band (absorption band) appears in emission spectrum. The emitted photon due to the internal conversion has less frequency $\nu$ than the absorbed photon. This frequency difference is known as the Stokes shift: emitted light always has a longer wavelength than the absorbed light due to the internal conversion.

Fig. 20.20. Molecular energy diagram illustrating an absorption and emission band appearance
Thus molecules have various states referred to energy levels. If the frequency of the radiation matches the vibrational frequency \((h\nu = E_{v2} - E_{v1})\) of the molecule then radiation will be absorbed, causing a change in the amplitude of molecular vibration. The energy of a vibrational mode depends on molecular structure and environment. **Infrared spectroscopy** is the measurement of the wavelength and intensity of the absorption of mid-infrared light by a sample. Mid-infrared is energetic enough to excite molecular vibrations to higher vibrational-rotational energy levels.

The wavelength of infrared absorption bands is characteristic of specific types of chemical bonds, and infrared spectroscopy finds its greatest utility for identification of organic and organometallic molecules. Infrared and microwave probes are used extensively in the laboratory. Since the vibrational and rotational levels depend on the masses, separations, and forces between the various atoms bound in a molecule, it is not surprising that spectroscopy can be used to identify specific bonds. This is a useful technique in chemistry. Biological applications are difficult because the absorption coefficients are large; thin samples must be used, particularly in an aqueous environment.

**20.7. THE SPECTRAL DEVICES**

Spectral devices are optical instruments used for decomposition of light into the monochromatic components. Such devices are used for qualitative and quantitative spectral analysis of the light emitted, absorbed, reflected or scattered by a sample. These studies define the properties of a sample, its chemical composition and the nature of physical processes associated with the optical radiation and its interaction with a sample.

There are prism and grating spectrometer designs. The prism spectrometer configuration is shown in fig. 20.21.

![Prism spectrometer schematic](attachment:image.png)

*Fig. 20.21. Prism spectrometer schematic:*  
1 — a light source; 2 — a substance under investigation; 3 — a slit; 4 — a collimating lens; 5 — a prism; 6 — a camera lens; 7 — a camera lens focal plane
Light entered a slit and a collimating lens transformed the light into a thin beam of parallel rays. The light was then passed through a prism that refracted the beam into a spectrum because different wavelengths were refracted different amounts due to dispersion (the index of refraction $n$ dependence on the wavelength $n(\lambda)$). In the spectroscope this image was then viewed through a tube with a scale that was transposed upon the spectral image, enabling its direct measurement. Spectrograph has a camera in place of the viewing tube. In recent years the electronic circuits built around the photomultiplier tube have replaced the camera, allowing real-time spectrographic analysis with far greater accuracy. In the grating spectrometer a diffraction grating is used in place of the prism. The grating spectrometer splits light into a spectrum because of the phenomenon of light diffraction, $\varphi(\lambda)$ dependence.

The spectral analysis has become an important scientific tool for analyzing the composition of material in biophysical and analytical chemistry for the identification of substances through the spectrum emitted from or absorbed by them.

### 20.8. Luminescence

Luminescence is light produced using energy sources other than heat. Sometimes luminescence is called «cold light», because it can occur at room temperature and cooler temperatures. To produce luminescence, energy is absorbed by an electron of an atom or molecule, causing it to become excited, but unstable. When the electron returns to a lower energy state the energy is released in the form of a photon (light). The energy of the photon $h\nu$ determines its wavelength or color ($\lambda = \frac{C}{V}$).

There are different ways of atom and molecule exciting and the following types of luminescence are known:

1. **Photoluminescence** is a process in which a substance absorbs photons (electromagnetic radiation, usually ultraviolet or visible range) and then emits photons. This can be described as an excitation to a higher energy state and then a return to a lower energy state accompanied by the emission of a photon (luminescence).

2. **Cathodoluminescence** is an optical phenomenon where the atomic excitation is produced by a beam of electrons which is generated by an electron gun (e. g. cathode ray tube) and then impacts on a luminescent material, causing the material to emit visible light.

3. **Electroluminescence** is an optical phenomenon where a material emits light in response to an electric current passed through it, or to a strong electric field.

4. **Chemiluminescence** is the emission of light due to excitation in the result of a chemical reaction: $A + B \rightarrow AB^* \rightarrow AB + h\nu$. 

200
5. **Bioluminescence** is chemiluminescence which takes place in numerous living organisms. For example, the American firefly is a widely studied case of bioluminescence. The firefly reaction has the highest known quantum efficiency $\sim 88\%$.

6. **Roentgenluminescence** is the optical luminescence produced by X-rays.

7. **Radioluminescence** is the phenomenon by which luminescence is produced in a material by the bombardment of ionizing radiation such as alpha and beta particles.

The main characteristics of luminescence are the following:

- **Luminescence spectrum** is a dependence of luminescence intensity $I_{\text{lum}}$ on luminescence wavelength $\lambda$.

- **Exciting radiation spectrum** is a dependence of photoluminescence intensity $I_{\text{lum}}$ on exciting radiation wavelength $\lambda$.

- **The luminescence quantum yield** $\gamma$ gives the efficiency of the luminescence process. It is defined as the ratio of the number of photons emitted to the number of photons absorbed:

$$\gamma = \frac{n_{\text{emitted}}}{n_{\text{absorbed}}}.$$  \hspace{1cm} (20.31)

The maximum luminescence quantum yield is 1.0 (100\%); every photon absorbed results in a photon emitted. Compounds with quantum yields of 0.10 are still considered quite luminescent.

- **The delay time** $\tau$ is the time during that the intensity of the luminescence decreases in $e = 2.7$ times.

When the radiation causing luminescence has been stopped the intensity of the luminescence decays exponentially (fig. 20.22) with time and is described by the formula:

$$I_{\text{lum}} = I_0 e^{-t/\tau},$$  \hspace{1cm} (20.32)

where $I_0$ is the intensity of the stationary luminescence; $t$ is the time; $\tau$ is the delay time during that the intensity of the luminescence decreases in $e = 2.7$ times.

![Fig. 20.22. The luminescence intensity decay with time](image)
There are two principal varieties of luminescence: fluorescence and phosphorescence, distinguished by the delay time $\tau$. If the luminescence decays in very short time ($\sim 10^{-9} - 10^{-7}$ s) it is known as fluorescence (in this case $\tau < 10^{-7}$ s). If $\tau > 10^{-4}$ s, then luminescence is known as phosphorescence ($\tau > 10^{-4}$ s). Fluorescence appearances if the electron selection rules are satisfied, the transition is fairly rapid (typically $10^{-8}$ s). Sometimes the electron becomes trapped in a state where it cannot decay according to the electronic selection rules. It may then have a lifetime up to several seconds before decaying and in this case phosphorescence occurs.

Fluorescence spectroscopy is primarily concerned with electronic and vibrational states. Tryptophan is an important intrinsic fluorescent probe (amino acid), which can be used to estimate the nature of microenvironment of the tryptophan. When performing experiments with denaturants, surfactants or other amphiphilic molecules, the microenvironment of the tryptophan might change. For example, if a protein containing a single tryptophan in its «hydrophobic» core is denatured with increasing temperature, a red-shift emission spectrum will appear. This is due to the exposure of the tryptophan to an aqueous environment as opposed to a hydrophobic protein interior. In contrast, the addition of a surfactant to a protein which contains a tryptophan which is exposed to the aqueous solvent will cause a blue shifted emission spectrum if the tryptophan is embedded in the surfactant vesicle or micelle.
Proteins that lack tryptophan may be coupled to a fluorophore. At 295 nm, the tryptophan emission spectrum is dominant over the weaker tyrosine and phenylalanine fluorescence.

There are several laws that deal with molecular luminescence:

1. **The Stokes Law**: the wavelength of the luminescence light is always greater than that of the exciting radiation (fig. 20.23).

2. **The Kasha–Vavilov Law**: luminescence spectrum and the quantum yield of luminescence are independent of the wavelength of exciting radiation.

Both of these laws are explained by internal conversion in molecules.

**Questions**:

1. What is the light absorption? Describe fundamental laws of the light absorption. What is the light absorption spectrum?
2. Write Lambert-Beer-Bouguer Law. What does molar extinction coefficient depend on?
3. What are transmittance and optical density? Describe their dependence on the wavelength and solution concentration.
5. What do Bohr’s postulates describe? Specify these postulates.
6. Explain appearance of emission spectra and absorption one.
7. Give the energy states of a hydrogen atom. Explain spectral lines formation for a hydrogen atom.
9. What is a luminescence? Which types of luminescence are known?
10. Write a formula for luminescence intensity decay with time.
11. Explain the Stokes Law and the Kasha–Vavilov one.

**Chapter 21. STIMULATED EMISSION. LASER**

Light Amplification by Stimulated Emission of Radiation, commonly referred to as «Laser» describes a wide range of devices. The lasers can function as oscillators (sources of light) and as amplifiers. Lasers have revolutionized various fields of science and technology, and are being used in a wide range of applications in medicine, communications, defense, measurement, and as a precise light source in many scientific investigations.

**21.1. PROCESSES OF ABSORPTION AND EMISSION IN ATOMIC SYSTEM**

The principle of operation remains the same though there is a wide range of lasers. Laser action occurs in three stages: population inversion creation, spontaneous emission, and stimulated emission. We must begin with an account of how light (photon) can interact with individual atoms within an amplifying medium («atoms» will be used to include molecules and ions). Energy levels associated with molecules, atoms and nuclei are in general discrete, quantized
energy levels and transitions between those levels typically involve the absorption or emission of photons. Electron energy levels have been used as the example here, but quantized energy levels for molecular vibration and rotation also exist. Transitions between vibrational quantum states typically occur in the infrared and transitions between rotational quantum states are typically in the microwave region of the electromagnetic spectrum.

Atoms consist of a positively charged core (nucleus) which is surrounded by negatively charged electrons. According to the quantum mechanical description of an atom, the energy of an atomic electron can have only certain values and these are represented by energy levels \( E_1 < E_2 < E_3 \ldots \). The electrons can be thought of as orbiting the nucleus, those with the largest energy orbiting at greater distances from the nuclear core. There are many energy levels that an electron within an atom can occupy, but here we will consider only two. Also, we will consider only the electrons in the outer orbits of the atom as these can most easily be raised to higher unfilled energy states.

A photon of light is absorbed by an atom in which one of the outer electrons is initially in a low energy state. The energy of the atom is raised to the upper energy level, and remains in this excited state for a period of time that is typically less than \( 10^{-7} \) second. It then spontaneously returns to the lower state, with the emission of a photon of light. Absorption is referred to as a resonant process because the energy of the absorbed photon must be equal to the difference in energy between the levels. This means that only photons of a particular frequency (or wavelength) will be absorbed. Similarly, the photon emitted will have energy equal to the difference in energy between the two energy levels. These common processes of absorption and spontaneous emission cannot give rise to the amplification of light. Spontaneous emissions are random and isotropic in nature. The best that can be achieved is that for every photon absorbed, another is emitted. The above processes are represented in fig. 21.1, where \( E_1 \) is the ground-state or lower energy level and \( E_2 \) is the excited-state or higher energy level. The particle of the material, which undergoes the process of excitation, might be an atom, molecule, or ion depending on the laser material.

![Fig. 21.1. The processes of absorption and spontaneous emission](image)

\[ E_2 \text{ (Excited State)} \]
\[ E_1 \text{ (Ground State)} \]

Absorption can occur only when

\[ E_{\text{photon}} = h\nu = E_2 - E_1 \]

\[ E_{\text{photon}} = h\nu = E_2 - E_1 \]

A downward transition involves emission of a photon of energy

\[ E_{\text{photon}} = h\nu = E_2 - E_1 \]
Above it was stated that an atom in a high energy, or excited, state can return to the lower state spontaneously. However, if a photon of light interacts with the excited atom, it can stimulate a return to the lower state (fig. 21.2). One photon interacting with an excited atom results in two photons being emitted.

![Diagram: Energy Levels](image)

*The energy of the incoming photon of light must match the difference in energy between the two energy levels*

\[ E_{\text{photon}} = h \nu = E_2 - E_1 \]

**Fig. 21.2. A process of stimulated emission**

Furthermore, the two emitted photons are said to be in phase, i.e. thinking of them as waves, the crest of the wave associated with one photon occurs at the same time as on the wave associated with the other. This feature ensures that there is a fixed phase relationship between light radiated from different atoms in the amplifying medium and results in the laser beam produced having the property of coherence.

Stimulated emission has very important properties. The direction of its propagation exactly coincides with the direction of propagation of the stimulating radiation, i.e. of the external radiation producing a transition. The same relates to the frequency, phase, and polarization of the stimulated emission and stimulating radiation. Consequently, the stimulated emission and the radiation stimulating it are strictly coherent. This feature of stimulated emission underlies the action of light amplifiers and generators known as lasers. Stimulated emission is the process that can give rise to the amplification of light. As with absorption, it is a resonant process; the energy of the incoming photon of light must match the difference in energy between the two energy levels. Furthermore, if we consider a photon of light interacting with a single atom, stimulated emission is just as likely as absorption; which process occurs depends upon whether the atom is initially in the lower or the upper energy level. However, under most conditions, stimulated emission does not occur to a significant extent. The reason is that, under most conditions, that is, under conditions of thermal equilibrium, there will be far more atoms in the lower energy level, than in the upper level \( n_1 > n_2 \), so that absorption will be much more common than stimulated emission. If stimulated emission is to predominate, we must have more atoms in the higher energy state than in the lower one. This unusual condition is referred to as a population inversion and it is necessary to create a population inversion for laser action to occur.
When the number of particles in the excited state is greater than the number of particles in the ground state \( n_2 > n_1 \), the material is in a state of «population inversion» (fig. 21.3).

![Image of population inversion](image)

*Fig. 21.3. A condition of thermal equilibrium (a) and a state of «Population Inversion» between 2 and 1 levels (b)*

Population inversion is a prerequisite for laser action. Energy can be transferred into a laser medium to achieve population inversion by several mechanisms including absorption of photon, collision between electrons (or sometimes ions) and species in the active medium, collisions among atoms and molecules in the active medium, recombination of free electrons with ionized atoms, recombination of current carriers in a semiconductor, chemical reactions producing excited species, and acceleration of electrons. If during the process of stimulated emission, the population inversion is maintained by continuous pumping of energy, the laser action continues indefinitely and the result is a continuous wave laser. On the other hand, if the pumping cannot be maintained the output is a pulsed laser.

**21.2. CONSTRUCTION OF A LASER**

A laser consists of an amplifying medium (the gain medium), a source of excitation energy, and a resonator or feedback mechanism to perform the three stages of laser action. The general construction of a laser is shown in fig. 21.4.

![Image of laser construction](image)

*Fig. 21.4. Laser components*
**Amplifying medium:** All lasers contain an energized substance that can increase the intensity of light passing through it. This substance is called the amplifying medium or, sometimes, the gain medium, and it can be a solid, a liquid or a gas. Whatever its physical form, the amplifying medium must contain atoms, molecules or ions, a high proportion of which can store energy that is subsequently released as light.

In a neodymium YAG (Nd:YAG) laser, the amplifying medium is a rod of yttrium aluminium garnate (YAG) containing ions of the lanthanide metal neodymium (Nd). In a dye laser, it is a solution of a fluorescent dye in a solvent such as methanol. In a helium-neon laser, it is a mixture of the gases helium and neon. In a laser diode, it is a thin layer of semiconductor material sandwiched between other semiconductor layers. The factor by which the intensity of the light is increased by the amplifying medium is known as the gain. The gain is not a constant for a particular type of medium. It’s magnitude depends upon the wavelength of the incoming light, the intensity of the incoming light, the length of the amplifying medium and also upon the extent to which the amplifying medium has been energized. An amplifying medium is one in which population inversion is possible (fig. 21.3).

The downward transition from the excited to the normal state is triggered by stimulated emission. The lasers are classified depending on the number of energy levels used for the excitation and the stimulated emission process. Commercial lasers are three-level and four-level systems (fig. 21.5), while the simple two-level system is not used in practice, as it is difficult to achieve population inversion in a two-level system.

![Fig. 21.5. Three-level and four-level laser energy diagrams](image)

**Excitation Source:** Population inversion is achieved by «pumping energy» from an external source. Depending on the external source, the excitation process is called as optical pumping or electrical pumping. In electrical
pumping, an AC or DC electrical discharge is used for excitation. Gas lasers and semiconductor lasers are usually excited using electrical pumping. In optical pumping, light is the source of energy and is used for most of the solid-state and dye lasers.

**Resonator:** A resonator consists of a pair of parallel mirrors. The high degree of collimation arises from the fact that the cavity of the laser has very nearly parallel front and back mirrors which constrain the final laser beam to a path which is perpendicular to those mirrors. The back mirror is made almost perfectly reflecting while the front mirror reflecting is $R < 1$, letting out about $(1 - R) \%$ of the beam. This $(1 - R) \%$ is the output beam which one sees. But the light has passed back and forth between the mirrors many times in order to gain intensity by the stimulated emission of more photons at the same wavelength. If the light is the slightest bit off axis, it will be lost from the beam.

### 21.3. Characteristics of Laser Light

1. **Coherent:** different parts of the laser beam are related to each other in phase due to stimulated emission properties. These phase relationships are maintained over long enough time so that interference effects may be seen or recorded photographically. This coherence property is what makes holograms possible.

2. **Monochromatic:** laser light consists of essentially one wavelength, having its origin in stimulated emission between two of atomic or molecular energy levels.

3. **Collimated:** because of bouncing back between mirrored ends of a laser cavity, those paths which sustain amplification must pass between the mirrors many times and be very nearly perpendicular to the mirrors. As a result, laser beams are very narrow and do not spread very much.

### 21.4. The Ruby Laser

The ruby laser takes its place in history by being the first working laser to be demonstrated. Theodore Maiman, working at Hughes Labs. in the USA, showed the first working optical laser to the world in 1960. The active medium is a cylindrical crystal of synthetic sapphire ($\text{Al}_2\text{O}_3$) doped with roughly 0.05 %, by weight, of chromium ions ($\text{Cr}^{3+}$).

Fig. 21.6 gives a diagram of the energy levels of the chromium ion $\text{Cr}^{3+}$ (level $E_3$ is a band formed by a collection of closely arranged levels).

A xenon lamp is rolled over ruby rod and is used for pumping ions to excited state. When a flash of light falls on ruby rod, radiations are absorbed by $\text{Cr}^{3+}$ which are pumped to $E_3$.

The excitation of the ions as a result of pumping is depicted by arrow from $E_1$ to $E_3$. The lifetime for level $E_3$ is very small ($\sim 10^{-8}$ s). The $\text{Cr}^{3+}$ ions after giving a part of their energy to crystal lattice pass to level $E_2$, which is
metastable with a lifetime of about $3 \times 10^{-3}$ s (compared to $10^{-8}$ s for ordinary levels). With strong pumping action, more ions can be found in the $E_2$ state than are in the $E_1$ state. Consequently, levels $E_1$ and $E_2$ become inverted and the population inversion occurs.

![Diagram of energy levels of chromium ion Cr$^{3+}$ in the ruby laser](image)

*Fig. 21.6. A diagram of the energy levels of the chromium ion Cr$^{3+}$ in the ruby laser*

Arrow from $E_2$ to $E_1$ depicts a spontaneous transition from the metastable level $E_2$ to the ground one $E_1$. As soon as a few ions in the $E_2$ state jump down to $E_1$, they emit photons that produce stimulated emission of the other ions, and the lasing action begins. A ruby laser thus emits a beam whose photons have energy 1.8 eV and a wavelength of 694.3 nm (or «ruby-red» light).

### 21.5. Types of Lasers

There are many types of lasers available for research, medical, industrial, and commercial uses. Lasers are often described by the kind of lasing medium they use—solid state, gas, excimer, dye, or semiconductor:

1. Gas lasers (helium and helium-neon, HeNe, are the most common gas lasers) have an output of different spectral range. CO$_2$ lasers emit energy in the far-infrared, 10.6 micrometers, and are used for cutting hard materials.

2. Excimer lasers (the name is derived from the terms *excited* and *dimers*) use reactive gases such as chlorine and fluorine mixed with inert gases such as argon, krypton, or xenon. When electrically stimulated, a pseudomolecule or dimer is produced and when lased, produces light in the ultraviolet range.

3. Dye lasers use complex organic dyes like rhodamine 6G in liquid solution or suspension as lasing media. They are tunable over a broad range of wavelengths.
4. Solid state lasers have lasing material distributed in a solid matrix, e. g., the ruby or neodymium-YAG (yttrium aluminum garnet) lasers. The neodymium-YAG laser emits infrared light at 1.064 micrometers.

5. Semiconductor lasers, sometimes called diode lasers, are not solid-state lasers. These electronic devices are generally very small and use low power. They may be built into larger arrays, e. g., the writing source in some laser printers or compact disk players.

Lasers are also characterized by the duration of laser emission — continuous wave or pulsed laser.

1. Continuous wave (CW) lasers operate with a stable average beam power. In most higher power systems, one is able to adjust the power. In low power gas lasers, such as HeNe, the power level is fixed by design and performance usually degrades with long term use. The direction of a CW laser can be scanned rapidly using optical scanning systems to produce the equivalent of a repetitively pulsed output at a given location.

2. Single pulsed (normal mode) lasers generally have pulse durations of a few hundred microseconds to a few milliseconds. This mode of operation is sometimes referred to as long pulse or normal mode.

3. Repetitively pulsed or scanning lasers generally involve the operation of pulsed laser performance operating at a fixed (or variable) pulse rates which may range from a few pulses per second to as high as 20,000 pulses per second.

21.6. LASER MEDICAL APPLICATIONS

There is a wide range of medical applications. Often these relate to the outer parts of the human body, which are easily reached with light; examples are eye surgery and vision correction (LASIK), dentistry, dermatology (e. g. photodynamic therapy of cancer), and various kinds of cosmetic treatment such as tattoo removal or hair removal. Lasers are also used for surgery (e. g. of the prostate), exploiting the possibility to cut tissues while causing only a low amount of bleeding.

Very different types of lasers are required for medical applications, depending on the wavelength, output power, pulse format, etc. In many cases, the laser wavelength is chosen so that certain substances (e. g. pigments in tattoos or caries in teeth) absorb light more strongly than surrounding tissue, so that they can be more precisely targeted. Fig. 21.7 shows ruby laser use for epidermal hyperpigmentation treatment. Medical lasers are not always used for therapy. Some of them rather assist the diagnosis e. g. via methods of laser microscopy or spectroscopy.
Chapter 22. EYE VISION

22.1. EYE STRUCTURE

The eye is a very complex organ that sends a huge amount of information to the brain. It has a very specific design to capture and analyze light. In its simplest description, the eye is a spherical box, with a lens to focus the light that enters it, and cells to process the light.

The human eye is roughly spherical in shape (fig. 22.1). It is bounded by three distinct layers of tissue. The outer layer, sclera, is extremely tough. It is white in color except in the front. Here it forms the transparent cornea, which admits light into the interior of the eye and bends the light rays so that they can be brought to a focus. The surface of the cornea is kept moist and dust-free by the secretion from the tear glands.

The middle coat of the eye, choroid, is deeply pigmented with melanin and well supplied with blood vessels. It serves the very useful function of stopping the reflection of stray light rays within the eye. This is the same function that is accomplished by the dull black paint within a camera.

In the front of the eye, the choroid forms the iris. This may be pigmented and is responsible for the color of the eye. An opening, the pupil, is present in the center of the iris. The size of this opening is variable and under automatic control. In dim light (or times of danger) the pupil enlarges, letting more light
into the eye. In bright light, the pupil closes down. This produces clearer vision, because a smaller opening, or aperture, creates a sharper image.

Behind the pupil and iris are the crystalline lens and the ciliary body. The ciliary body contains muscles that support the lens and changes its shape. The lens is a colorless, nearly transparent double convex structure, similar to an ordinary magnifying glass. Its only function is to focus light rays onto the retina. By changing its curvature, the lens can focus on objects at different distances from it. This process is called accommodation.

The lens of the eye is bathed on one side by the aqueous humor and supported on the other side by the vitreous humor. Aqueous humor is located in the anterior chamber of the eye, the space between the lens and the cornea. It maintains the intraocular pressure and inflates the globe of the eye. The vitreous is the transparent, colorless, gelatinous mass that fills the space between the lens of the eye and the retina and occupying about 80% of the volume of the eyeball.

The inner coat of the eye is the retina. It contains photoreceptors that translate light energy into electrical signals and sends them to the brain through the optic nerve. The center area of retina, called the macula, is used for fine central vision and color vision. The fovea is located in the center of the macula. It is responsible for sharp central vision, which is necessary in humans for reading, watching, driving, and any activity where visual detail is of primary importance. The blind spot lacks photoreceptors; it is located where the optic nerve fibers leave the eye.
22.2. IMAGE FORMATION BY THE EYE OPTICAL SYSTEM

The eye function is to collect light emitted or reflected by a distant object and form an image of object for presentation to the brain. The eye can see the object under study clearly if precise real optic image of this object is built on the retina. This problem is solved by an optical system of the eye. It consists mainly of the cornea and the lens, and to a lesser extent of other structures.

Most of the refraction occurs at the cornea (40–43 diopters). The cornea has an index of refraction of 1.38. The index of refraction of the cornea is significantly greater than the index of refraction of the surrounding air. This difference in optical density combined with the fact that the cornea has the shape of a converging lens is what explains the ability of the cornea to do most of the refracting of incoming light rays. The refractive index of the aqueous humor — 1.33; the crystalline lens (on average) — 1.41; and the vitreous humor — 1.34. Total refractive power $D$ of the eye is varied from 60 to 73 diopters; the lens refractive power $D$ is in the range 20–30 diopters.

If all the refractive surfaces of the eye are algebraically added together and then considered to be one single lens, the optics of the normal eye may be simplified and represented schematically as a «reduced eye». This is useful in simple calculations. In the reduced eye, a single refractive surface is considered to exist, with its central point 17 millimeters in front of the retina and a total refractive power of 60 diopters when the lens is accommodated for distant vision.

Eye optical system forms on the retina a **real inverted diminished** image of the distant object (fig. 22.2).

![Fig. 22.2. The formation scheme of a real inverted diminished image on retina](image-url)

22.3. ACCOMMODATION

**Accommodation** is the process by which the eye changes optical power to maintain a clear image as object distance varies.
Eye optical system makes clear image on retina, if thin-lens equation is satisfied:

$$\frac{1}{d} + \frac{1}{f} = \frac{1}{F},$$  \hspace{1cm} (22.1)

where $d$ is the object to lens distance (fig. 22.3), $f = 17 \text{ mm}$ is the image distance (lens to retina distance), $F$ is the eye focal length.

Light rays from distant objects ($d \rightarrow \infty$) are nearly parallel and do not need as much refraction to bring them to a focus ($f = F$).

Let’s compare the operation of a camera and an eye. In both cases the instrument (eye or camera) must make an adjustment to put clear images on the retina or film of objects that are a variety of distances away. In a camera, the focal length of the lens $F$ is fixed and the image distance $f$ is adjusted (lens to film distance is adjusted). In the eye the lens to retina distance (the image distance) $f$ is fixed and the eye adjusts its focal length to place clear images on the retina. Eye accommodation is carried out by the lens curvature change. When the eye is relaxed, the lens has its minimum optical power for distant viewing. As the muscle tension around the ring of muscle is increased, the lens rounds out to its maximum optical power (fig. 22.4).
The ability for accommodation depends on the elasticity of the eye lens and decreases with age (fig. 22.5).

![Graph showing the refractive power range AD and age in years from 0 to 70 with a point labeled "average eye" at 40 years.]

**Fig. 22.5.** Dependence of accommodation on the age

*Far point* $R$ is the point at which an object must be placed along the optical axis for its image to be focused on the retina when the eye is not accommodating. For the normal eye the far point $R$ is located at distance $l_R \geq 20$ m from the eye. *Near point* $P$ is the point nearest the eye at which an object is accurately focused on the retina when the maximum degree of accommodation is employed. For the normal eye near point $P$ at $l_P = 10–12$ cm distance from the eye. **Range of accommodation** $A_{PR}$ is

$$A_{PR} = \frac{1}{l_P} - \frac{1}{l_R}. \quad (22.2)$$

**22.4. THE EYE REFRACTION DEFECTS AND EYESIGN IMPROVEMENT**

*Emmetropic eye* is a condition of the normal eye. It is achieved when the refractive power of the cornea and the axial length of the eye is balanced, and in this case rays are focused exactly on the retina, resulting in perfect vision. An eye in a state of emmetropia requires no correction.

In this case parallel light rays from distant objects are in sharp focus on the retina when the ciliary muscle is completely relaxed (fig. 22.6). This means that the emmetropic eye can see all distant objects clearly with its ciliary muscle
relaxed. However, to focus objects at close range, the eye must contract its ciliary muscle and thereby provide appropriate degrees of accommodation.

**Fig. 22.6.** The back focus location for normal eye

**Hyperopia**, which is also known as farsightedness, is usually due to either an eyeball that is too short or a lens system that is too weak. In this condition, parallel light rays are not bent sufficiently by the relaxed lens system to come to focus by the time they reach the retina (fig. 22.7, a). To overcome this abnormality, the ciliary muscle must contract to increase the strength of the lens. In old age, when the lens becomes **presbyopic**, a farsighted person is often unable to accommodate the lens sufficiently to focus even distant objects. Hyperopia can be corrected by adding refractive power using a convex lens in front of the eye (fig. 22.7, b).

**Fig. 22.7.** The back focus location for far-sighted eye (a) and vision correction with a convex lens (b)

**Myopia** (shortsightedness) is a condition of the eye where the light that comes in does not directly focus on the retina but in front of it (fig. 22.8, a). This is usually due to too long an eyeball. Also it can result from too much refractive power in the lens system of the eye. A myopic person has no mechanism by which to focus distant objects sharply on the retina. However, as an object
moves still closer to the eye, the person can use the mechanism of accommodation to keep the image focused clearly. A myopic person has a definite limiting far point for clear vision.

![Figure 22.8](image)

*Fig. 22.8.* The back focus location for short-sighted eye (a) and vision correction with a concave lens (b)

The corrective lenses have a negative optical power (i.e. are concave) which compensates for the excessive positive diopters of the myopic eye (fig. 22.8, b).

### 22.5. Visual Acuity

**Visual acuity** is a quantitative measure of the ability to identify black symbols on a white background at a standardized distance as the size of the symbols is varied. Visual acuity is related with **visual angle** — the minimum angle at which resolution is just possible. It is the angle $\varphi$ under which object $AB$ is seen from the optical center of the eye (fig. 22.9).

![Figure 22.9](image)

*Fig. 22.9.* Visual angle

The sensation of vision occurs when light is absorbed by the photosensitive rods and cones. To resolve two points, the light from each point must be focused on a different cone and the exited cones must be separated from each other by at least one cone that is not exited. The minimum distance between exited cones is $d \approx 5 \, \mu m$, so the minimum visual angle for which two luminous points (or two...
black points over a white background) are perceived by the eye as separate is about one angular minute:

\[ \varphi_{\text{min}} = \frac{d}{f} = \frac{5 \cdot 10^{-3} \, \mu m}{17 \, mm} = 3 \cdot 10^{-4} \, \text{rad} = 1', \]  

(22.3)

where \( f = 17 \, mm \) is the lens to retina distance.

The eye poorly recognizes the details of an object seen at an angle less than 1'.

The angle 1' is an angle at which a segment having a length of 1 cm is seen at a distance of 34 m from the eye. At an insufficient illumination (in twilight), the minimum angle of resolution becomes larger and may reach 1°.

The minimum visual angle of patient is determined using the special tables. Then visual acuity can be calculated as

\[ \gamma = \frac{1'}{\varphi_{\text{min patient}}}. \]  

(22.4)

For example, if \( \varphi_{\text{min patient}} = 2' \), then the visual acuity for this patient:

\[ \gamma = \frac{1}{2} = 0.5. \]

By bringing an object close to the eye, one increase the angle of view, and hence make it possible to resolve finer details. However, objects cannot be brought too close to the eye since it has a limited capacity for accommodation. The most favorable distance for seeing object with a normal eye is \( d_0 = 25 \, cm \). At this distance the eye recognized details well enough without being tired. This is the distance of normal vision.

The eye resolution limit for the distance of normal vision is equal to:

\[ AB = 1' \cdot d_0 = 3 \cdot 10^{-4} \, \text{rad} \cdot 250 \, mm = 73 \, \mu m. \]  

(22.5)

22.6. RETINA ANATOMY AND FUNCTION

After light passes through the eye lens system and then through the vitreous humor, it enters the retina from the inside. Light passes first through several layers before it finally reaches the layer of rods and cones located on the outer edge of the retina (fig. 22.10). This distance is a thickness of several hundred micrometers; visual acuity is decreased by this passage through such nonhomogeneous tissue. However, in the central foveal region of the retina the inside layers are pulled aside to decrease this loss of acuity.

The retina contains two types of photoreceptors, termed rods and cones. (fig. 22.11). Rods are concentrated at the outer edges of the retina and are used in peripheral vision. There are approximately 125 million rods in the human retina. More sensitive than cones, rods are almost entirely responsible for night vision — vision under low illumination levels.
Rods contain the light-sensitive pigment rhodopsin (visual purple) which undergoes a chemical reaction (the rhodopsin cycle) when exposed to visible light. Rhodopsin consists of a lipoprotein called opsin and a chromophore (a light-absorbing chemical compound called 11-cis-retinal). Rods cannot discriminate different wavelengths of light, and vision under low illumination conditions is essentially «colorblind». More than 100 rods are connected to each
ganglion cell, and the brain cannot discriminate among these photoreceptors to identify the origin of an action potential transmitted along the ganglion.

Cones are responsible for color vision. Cone cells are densely packed in the fovea — a part of the eye, located in the center of the retina and responsible for sharp vision. There are approximately 6 million cones in the retina.

Cones are less sensitive to light than the rods, but allow the perception of color. They are also able to perceive finer detail and more rapid changes in images, because their response times to stimuli are faster than those of rods. Cones are maximally sensitive to light of about 550 nm, in the yellow-green region of the visible spectrum. Cones are much less sensitive than rods to light, but in the fovea there is a 1:1 correspondence between cones and ganglions, so the visual acuity is very high.

Essential components of a photoreceptor (either a rod or a cone) are the outer segment, the inner segment, the nucleus, and the synaptic body. The light-sensitive photochemical is found in the outer segment. In the case of the rods, this is rhodopsin; in the cones, it is one of three «color» photochemicals, usually called simply color pigments, that function almost exactly the same as rhodopsin except for differences in spectral sensitivity.

22.7. RHODOPSIN-RETINAL VISUAL CYCLE

When light energy is absorbed by rhodopsin, the rhodopsin begins to decompose. The cause of this is photoactivation of electrons in the rhodopsin, which instantaneously changes of the cis-form of retinal into an trans-form (fig. 22.12). This form has the same chemical structure as the cis-form but has a different physical structure — a straight molecule rather than an angulated molecule.

![Fig. 22.12. Cis-form and trans-form of retinal](image)

Because the three-dimensional orientation of the reactive sites of the trans-retinal, it begins to divide into opsins and a 11-cis-retinal.

This process excites electrical changes in the rods, and the rods then transmit the exciting on nerve cell and the visual image transmit into the central nervous system in the form of optic nerve action potential.

The excitation of the rod causes increased negativity of the membrane potential, which is a state of hyperpolarization. It means that there is more negativity potential than normal inside the rod membrane. This is exactly opposite to the decreased negativity (the process of «depolarization») that occurs in almost all other sensory receptors.
22.8. LIGHT AND DARK ADAPTATION OF EYE

Light and dark adaptation is the ability of the eye to adjust to various levels of darkness and light.

**Light adaptation** occurs when we move from the dark into bright light. Rods and cones are both stimulated and large amounts of the photopigment are broken down instantaneously, producing signals resulting in the light.

Adaption occurs in two ways:
1. By the pupil constriction, it takes about 0.3 sec.
2. By the decreasing of rhodopsin concentration in rods and iodopsin concentration in cones.

Within about one minute the cones are sufficiently excited by the bright light to take over. Visual accuracy and color vision continue to improve over the next ten minutes. During light adaptation retinal sensitivity is lost.

**Dark adaptation** is essentially the reverse of light adaptation. It occurs when going from a well light area to a dark area. Initially blackness is seen because our cones cease functioning in low intensity light. Also, all the rod pigments have been bleached out due to the bright light and the rods are initially nonfunctional.

Once in the dark, rhodopsin regenerates and the sensitivity of the retina increases over time (maximum sensitivity reaches approximately in hour). During these adaptation processes reflexive changes occur in the pupil size.

The eye is extremely sensitive to small amounts of light. For example, as few as 10 photons can generate a visual stimulus in an area of the retina where the rods are present at high concentration.

Differences in signal intensity that can just be detected by the human observer are known as **just noticeable differences (dI)**. This concept applies to any type of signal, including light that can be sensed by the observer. The smallest difference in signal that can be detected depends on the magnitude of the signal. The JND is directly proportional to the intensity of the signal:

\[ dI \sim I \cdot dS \]

\[ dS \sim \frac{1}{I}, \]

where \( I \) is the intensity of stimulus, \( dS \) is an increment of perception, and \( k \) is a coefficient. The integral form of this expression is known as the **Weber–Fechner Law**:

\[ S = k \log \frac{I}{I_0}. \]  \hspace{1cm} (22.6)

The Weber–Fechner law is similar to the expression for the intensity of sound in decibels.
22.9. Color vision

Different cones are sensitive to different colors of light. Let’s discuss of the mechanisms by which the retina detects the different gradations of color in the visual spectrum.

All theories of color vision are based on the well-known observation that the human eye can detect almost all gradations of colors can be received when only red, green, and blue monochromatic colors are appropriately mixed in different combinations.

The spectral sensitivities of the three types of cones in humans are the same as the light absorption curves for the three types of pigment found in the cones with maximum absorption on 440, 540 and 590 nm respectively. The absorption maximum for rods corresponds to 510 nm (fig. 22.13).

For example, an orange monochromatic light with a wavelength of 580 nm stimulates the red cones; it stimulates the green cones to a less stimulus value, but the blue cones not at all. The nervous system interprets the ratios of stimulation of the three types of cones as the sensation of orange.

About equal stimulation of all the red, green, and blue cones gives one the sensation of seeing white. Yet there is no single wavelength of light corresponding to white; instead, white is a combination of all the wavelengths of the spectrum.

The rods are maximally sensitive to light of about 510 nm, in the blue-green region of the visible spectrum.

Questions:
1. Describe eye structure. What are the eye mediums optical properties?
2. Explain the eye optical system functions. Specify eye refractive power.
3. What is the eye accommodation? Describe mechanism of the accommodation. What is range of accommodation? How does it depend on age?
4. Characterize the main eye refraction defects. How are these defects corrected?
5. How is the visual acuity determined? What is the visual angle? Specify the eye resolution limit.
6. What is the eye retina construction? Describe two type photoreceptors, explain quantity and distribution of photoreceptors.
7. What is the difference between rods and cones? Describe rhodopsin-retinal visual cycle.
8. What is the difference between daylight vision and twilight one? Specify the light absorption spectrum for cones and rods.
9. What is the eye adaptation? Specify basic mechanisms of adaptation.

Chapter 23. X-RAYS

X-rays are a form of electromagnetic radiation with a wavelength in the range of $80$ to $10^{-5}$ nanometers. They are longer than γ-rays but shorter than ultraviolet rays. If X-rays have short wavelength they are called hard X-rays. On the other hand, the long-wave radiation is classified as soft X-rays. X-rays is divided into bremsstrahlung and characteristic X-rays according to the mechanism of their formation.

In many languages, X-radiation is called Röntgen radiation, after Wilhelm Conrad Röntgen, who is credited as its discoverer, and who had named it X-radiation to signify an unknown type of radiation.

23.1. BREMSSTRAHLUNG X-RAYS

According to the electrodynamics laws, electromagnetic radiation is emitted when the moving charge is accelerated or decelerated. If a charged particle, which has a large kinetic energy is suddenly decelerated, the electromagnetic radiation of X-ray range arises. X-rays is produced in a highly evacuated glass bulb, called an X-ray tube (fig. 23.1).

![X-Rays tube diagram](image)

Fig. 23.1. X-Rays tube
It contains two electrodes — a cathode which emits electrons and an anode to collect the electrons made of molibdenum, tungsten. The electrons are produced by thermionic effect from a tungsten filament heated by an electric current. The filament is the cathode of the tube. When high voltage is applied between the cathode and the anode the electrons are accelerated. The focusing electrode directs the electron beam towards the anode. Brehmsstrahlung X-rays is produced when the electrons are suddenly decelerated upon collision with the anode.

The kinetic energy of an electron obtained in the electric field between cathode and anode can be written as:

\[ \frac{m v^2}{2} = eU, \]  

where \( m \) is the mass of an electron, \( v \) is the velocity, \( e \) its charge, \( U \) is the applied electrical potential difference between the cathode and the anode.

In a conventional X-ray tube, only about 1% of the electron energy is converted into X-rays (\( h\nu \)). The remaining 99% of the energy is converted into the internal energy of the target (heating of the target) (\( Q \)):

\[ eU = h\nu + Q. \]  

A relation between summands in right part of this equation (23.2) is random. Therefore the different frequencies are observed in the bremsstrahlung radiation spectrum. Bremsstrahlung X-rays has a continuous spectrum. This spectrum has a definite short wavelength \( \lambda_{\text{min}} \).

The minimum wavelength \( \lambda_{\text{min}} \) corresponds to maximum frequency \( \nu_{\text{max}} \). Let’s assume that \( Q = 0 \) (all kinetic energy of electron is converted into X-radiation). Thus:

\[ \frac{h c}{\lambda_{\text{min}}} = eU \Rightarrow \lambda_{\text{min}} = \frac{h c}{eU}. \]  

where \( c \) is velocity of light.

If the numerical values of constants \( h, c \) and \( e \) will be used in the equation (23.3), the minimum wavelength \( \lambda_{\text{min}} \) can be written as:

\[ \lambda_{\text{min}}(\text{nm}) = \frac{1.23}{U(\text{kV})}. \]  

The curve of bremsstrahlung X-rays spectrum for each voltage starts at a particular minimum wavelength, rises rapidly to a maximum and drops gradually but indefinitely towards the longer wavelengths (fig. 23.2). Minimum wavelength \( \lambda_{\text{min}} \) depends on the voltage \( U \) between cathode and anode. The higher voltage \( U \) is, the smaller value of the minimum wavelength \( \lambda_{\text{min}} \) (23.4) will be. The penetrating power (hardness) of the X-rays is controlled by the voltage \( U \) between the cathode and anode.
The spectral radiant flux of the X-ray radiation $\Phi_\lambda$ is the ratio of the flux energy $d\Phi(\lambda)$ emitted by the X-ray tube in a narrow range of wavelengths from $\lambda$ to $d\lambda$ to the width of the interval $d\lambda$.

A total radiant flux $\Phi$ of X-rays can be written as:

$$\Phi = \int_{\lambda_{\text{min}}}^{\infty} \Phi_\lambda d\lambda.$$ 

A total radiant flux $\Phi$ of X-rays depends on filament current $I$ and voltage $U$ between the anode and cathode in the X-rays tube and is determined by formula:

$$\Phi = k I U^2 Z,$$  \hspace{1cm} (23.5)

where $Z$ is the atomic number, $k$ is the coefficient proportionality $k = 10^{-9} (V^{-1})$.

As seen from fig. 23.3 the minimum wavelength $\lambda_{\text{min}}$ is the same for different values of current ($I_1$ and $I_2$) in the X-rays tube when $U$ is constant. Therefore the radiant flux rises but the radiation hardness remains unchanged.

$Fig. \ 23.2.$ The spectrum of the bremsstrahlung X-rays for different voltages $U_1$ and $U_2$ applied between cathode and anode

$Fig. \ 23.3.$ The spectrum of the bremsstrahlung X-rays for different values of current in the X-rays tube
23.2. **Characteristic X-rays**

If the energy of the electrons is high enough, *characteristic X-ray* radiation is generated, which appears in the spectrum as individual emission lines superimposed on the continuous bremsstrahlung spectrum (fig. 23.4).

![Fig. 23.4. X-rays spectrum](image)

These lines are generated when high-energy electrons penetrate deep into the atomic shells of the anode material and eject electrons from the innermost orbitals. This vacancies created in this process are filled by electrons from the outer orbitals. This process results in a photon (X-ray) with a characteristic energy equal to the difference in the energy of the orbitals. As the energy of an orbital is dependent on the material, the resulting X-ray radiation is *characteristic* of that anode material.

This process produces an emission spectrum of X-rays at a few discrete frequencies, sometimes referred to as the spectral lines. The spectral lines generated depend on the target (anode) element used and thus are called characteristic lines. Usually these are transitions from upper shells into K shell (called K lines), into L shell (called L lines) and so on (fig. 24.5).

![Fig. 23.5. Characteristic X-rays](image)
The frequency $\nu$ of the characteristic X-rays rises as the atomic number $Z$ increases. This relation is known as a **Moseley Law**:

$$\sqrt{\nu} = A(Z - B),$$

(23.6)

where $A$ and $B$ are constant.

**23.3. Interaction between X-rays and matter**

Let’s consider an interaction between quanta of X-rays and atoms and molecules of the matter. Obviously, the result of this interaction depends on the energy of the quantum. There are several different cases.

1. The quantum energy $h\nu$ of X-rays is smaller than the energy of the atomic ionization $A_i$ ($h\nu < A_i$). This interaction is called a **coherent scattering**. It is a process in which the photon is scattered on the entire atom. That is, the internal energy of the atom does not change. In this case the energy of the incident photon equals the energy of the scattered photon. Only soft X-ray experiences a coherent scattering (fig. 23.6, a). This is not an ionizing interaction.

2. The quantum energy $h\nu$ is slightly greater than the energy of the atomic ionization $A_i$ ($h\nu \geq A_i$). In this case, the quantum energy $h\nu$ is spent on atom ionization $A_i$ and kinetic energy of electron $\frac{mv^2}{2}$:

$$h\nu = A_i + \frac{mv^2}{2}.$$  

(23.7)

This phenomenon is known as a photoelectric effect or **photoeffect** (fig. 23.6, b).

3. The quantum energy $h\nu$ is much greater than the energy of the atomic ionization $A_i$ ($h\nu \gg A_i$). This interaction not only changes the incident photon direction but reduces its energy and ionizes the atom as well. A photon interacts with an electron, but in contrast to the photoelectric effect, only a part of the photon energy is transferred to the electron. The photon continues on its way, but with reduced energy $h\nu'$ (i.e., a lower frequency). This effect is called a **Compton scattering** or incoherent scattering (fig. 23.6, c). The electron is still emitted from its shell. In addition the electron obtains a kinetic energy $E_k$:

$$h\nu = A_i + h\nu' + E_k.$$  

(23.8)

If the electron is ejected from interior shells then the characteristic X-rays appears.

Secondary X-rays have energy $h\nu' > A_i$ and can produce the ionization of the matter again. Recoil electrons can also ionize adjacent atoms by means of a collision (fig. 23.6, d).

High-energy photons experience more Compton scattering than low energy photons. Unfortunately, Compton scattering is the major source of background
noise in X-rays images. In addition, Compton scattering is the major source of tissue damage due to X-rays. For these reasons, this phenomenon X-rays is applied in medicine for damage cancer tumors.

Fig. 23.6. Interaction between X-rays and matter

23.4. ATTENUATION OF X-RAYS

When the X-rays pass through the matter, its intensity falls due to its absorption and scattering by the matter (fig. 23.7). The attenuation of X-rays depends on the energy of X-rays, nature and thickness of matter. The attenuation of X-rays is described by the exponential law:

\[ I = I_0 e^{-\mu x}, \]  

where \( I \) is the X-rays intensity after traversing a thickness \( x \); \( I_0 \) is the intensity of incident X-rays; \( x \) is a thickness of matter; \( \mu \) is the linear attenuation coefficient, which is the sum of scattering and absorption coefficients: \( \mu = \mu_{\text{absorption}} + \mu_{\text{scattering}} \) (typically expressed in cm\(^{-1}\)). \( \mu \) depends on material density \( \rho \).

A mass attenuation factor \( \mu_m \) is also used: \( \mu_m = \mu / \rho \). A mass attenuation factor \( \mu_m \) is independent of a density of material.

The beam of X-rays encloses quanta with different energy. They have different penetrating power. Therefore the coefficient \( \mu \) in equation (23.9) is constant only for monoenergetic X-ray photons. For X-rays with different photon energies the effective attenuation coefficient is used.
Let’s estimate the penetrating power of X-rays. In practice a half-value layer is used, which is the thickness required to attenuate the beam intensity by 50 % (fig. 23.8). One can relate the half-value layer to the linear attenuation coefficient analytically. If in equation (23.9) the thickness \( x \) is equal to a half-value layer \( d_{1/2} \) \( (x = d_{1/2}) \), then \( I = I_0/2 \):

\[
I_0/2 = I_0 e^{-\mu d_{1/2}}
\]

\[
e^{+\mu d_{1/2}} = 2
\]

\[
\ln e^{+\mu d_{1/2}} = \ln 2
\]

\[
\mu d_{1/2} = \ln 2 = 0.69
\]

Thus:

\[
d_{1/2} = \frac{\ln 2}{\mu} = 0.69
\]

(23.10)

For example, the half-value layer for X-rays is equal 10 mm of water or 1 mm of aluminium when the applied in X-rays tube voltage \( U \) is 60 kV.
The half-value layer is a function of the energy of X-ray beam. Therefore a spectral composition of X-ray is changed when the beam goes through the half-value layer. The radiation becomes harder because short rays have a big penetrating power. Soft X-rays are absorbed more strongly. This phenomenon is called «beam hardening».

23.5. PHYSICAL PRINCIPLES OF THE X-RAY DIAGNOSTICS

The mass attenuation coefficient $\mu_m$ of X-rays depends on the matter composition and the wavelength:

$$\mu_m = k \lambda^3 Z^3,$$

(23.11)

where $k$ is a coefficient of proportionality; $Z$ is an atomic number of the material; $\lambda$ is a wavelength.

From equation (23.11) one can see that the mass attenuation coefficient $\mu_m$ increases with the increasing of the atomic number $Z$ and depends on the photon energy. This is a basis of the medical X-rays diagnostics. The purpose of this diagnostic is to measure features of the internal anatomy of a patient through differences in the attenuation of X-rays passing through different parts of the body.

As different tissues and organs in human body have different absorbing abilities of x-rays, the homogeneous intensity of x-rays will be not homogeneous after penetrating human body. If the non-homogeneous x-rays are projected onto fluorescent screen, the image of the organs can be formed on the screen. This is called x-ray fluoroscopy. If the transmitted x-rays irradiate on a negative film, the picture can be seen after development. The technique is called x-ray photography.

A simplified model of the human body consists of three different body tissues: fat, muscle, and bone. Air is also present in the lungs, sinuses and gastrointestinal tract. A contrast agent — a material with high $Z$ number — may be used to accentuate the attenuation of X-rays in a particular region.

X-rays interact in fat and other soft tissues predominantly by photoelectric interactions. Low-energy X-rays are used to accentuate subtle differences in soft tissues (e. g., fat, mussels and other soft tissues) in applications such as breast imaging (mamography) where the object (the breast) provides little intrinsic contrast. When images of structures with high intrinsic contrast are desired (e. g., the chest where bone, soft tissue, and air are present), higher-energy X-rays are used. These X-rays suppress X-ray attenuation in bone which otherwise would create shadows in the image that could hide underlying soft-tissue pathology.

In comparison with muscles and bones, fat has a higher concentration of hydrogen ($\sim 11 \%$) and carbon ($\sim 57 \%$) and a lower concentration of nitrogen ($\sim 1 \%$), oxygen (30 \%) and high-Z trace elements ($< 1 \%$). Hence, the effective atomic number of fat ($Z_{\text{eff}} = 5.9$ to 6.3) is less than that of soft tissues ($Z_{\text{eff}} = 7.4$).
or bones ($Z_{\text{eff}} = 11.6$ to 13.8). Because of its lower $Z_{\text{eff}}$, low-energy photons are attenuated less rapidly in fat than in an equal mass of soft tissues or bones.

The effective atomic number and physical density are greater for bones than for soft tissues. Hence, X-rays are attenuated more rapidly in bone than in an equal volume (not necessarily mass) of soft tissue.

There are many X-ray based procedures used in medical diagnosis, for example, fluoroscopy, mammography, X-rays computer tomography. Spiral computer tomography provides images can be displayed in three dimensions. The X-rays tomography allows receiving a layerwise image when a difference between attenuation coefficients is equal 0.1 %.

**Questions:**

1. Describe the bremsstrahlung X-rays appearance mechanism. Why does it have continuous spectrum? How to determine the minimum wavelength?
2. How to control the intensity and the hardness of radiation in the X-rays tube? Write the formula for bremsstrahlung X-rays radiant flux?
3. Compare the thermal radiation spectrum with the X-rays one. Discuss their similarity and difference.
4. Explain the difference between the formation mechanisms of optical spectrum and characteristic X-rays one.
5. Describe the mechanisms of interaction between X-rays and matter. Why is the hard X-rays more harmful for an organism than the soft X-rays?
6. Write exponential law for X-rays attenuation in matter. What is the linear attenuation coefficient? Describe its relation with the half-value layer.
7. Compare the physical principles of ultrasound and X-rays diagnostics.

**Chapter 24. RADIOACTIVITY**

Radioactive decay is a spontaneous process in which nucleons are emitted from or transformed within the nucleus, resulting in a change in the identity of the nucleus, and usually accompanied by the emission of one or more types of radiation from the nucleus and/or atom.

This decay results in an atom of one type, called the parent one is transformed to an atom of a different type, called the daughter one.

**24.1. CHARACTERISTICS OF NUCLEUS**

An atom consists of a positively charged nucleus surrounded by a cloud of negatively charged electrons. Nuclei consist of positively charged protons, and electrically neutral neutrons held together by the so-called strong or nuclear force. Protons and neutrons are generally called nucleons.

Let’s point basic properties of the nuclear force. The nuclear force is related to a strong interaction. It is a short-range force, its range is limited to distances about $10^{-15}$ meters. The nuclear force is independent of whether
the nucleons are neutrons or protons. This property is called **charge independence**. Every nucleon interacts with a limited number of adjacent nucleons (**property of saturation**).

The symbol of an atomic nucleus is \( ^A_Z X \). It consists of three parts: the symbol of the element \( X \), the atomic number of the element \( Z \) and the mass number of the specific isotope \( A \).

The number of protons in the nucleus, \( Z \), is called the **atomic number**. The **nuclear charge** \( q \) is equal to \( Ze \). The number of neutrons in the nucleus is marked by \( N \). The given element can have many different isotopes, which differ from each other by the number of neutrons contained in the nuclei. The **atomic mass number** of the nucleus can be written as: \( A = Z + N \). Each nuclear species with a given \( Z \) and \( A \) is called a **nuclide**.

The sizes of nuclei grow through the periodic table. The nuclear radius \( R \) and the atomic mass number \( A \) are related by formula:

\[
R = 1.5 \cdot 10^{-15} \cdot \frac{3}{\sqrt[3]{A}} \text{ (m).} \quad (24.1)
\]

**Nuclear stability** depends on the atomic number \( Z \) and on the number of neutrons \( N \). The light atomic nuclei contain practically as many neutrons as protons \( (N/Z = 1) \). They are the most stable. Nuclear stability weakens and the probability of spontaneous decay increases with increase of the total number of nucleons and at excess of the number of neutrons over the number of protons \( N/Z > 1.6 \).

**Nuclear binding energy** is the energy required to break up the nucleus into its separate nucleons or this can be expressed as the energy released when the nucleus is formed from separate nucleons. The unit of energy commonly used in atomic and nuclear physics is **the electron volt (eV)**. An **electron volt** is the energy acquired when an electron falls through a potential difference of 1 Volt.

\[
1 \text{ eV} = 1.6 \cdot 10^{-19} \text{ C} \cdot 1 \text{V} = 1.6 \cdot 10^{-19} \text{ J}.
\]

Thus:

\[
1 \text{ KeV} = 1000 \text{ eV} = 1.6 \cdot 10^{-16} \text{ J}
\]

\[
1 \text{ MeV} = 10^6 \text{ eV} = 1.6 \cdot 10^{-13} \text{ J}.
\]

**24.2. Modes of radioactive decay**

Unstable atoms undergo a radioactive decay in order to have a more stable configuration. For any radioactive decay or nuclear reaction, there are four conservation laws which must be obeyed: the law of conservation of electric charge; the law of conservation of mass number \( A \); the law of conservation of mass energy; the law of conservation of momentum.
Nuclei can spontaneously transform to stable nuclei by one of the following processes:

- **α-decay**;
- **β-decay**.

**Alpha-decay (α-decay)** is a type of radioactive decay in which an atomic nucleus \( X \) emits an alpha particle \( ^4_2\alpha \) (a helium nucleus \( ^4_2\text{He} \) ) and is transformed into a new nucleus \( Y \) called the daughter nucleus which has an atomic number 2 less and an atomic mass 4 less than the parent nucleus \( X \).

Alpha-decay proceeds according to the following scheme:

\[
^{A}_{Z}X \rightarrow ^{A-4}_{Z-2}Y + ^{4}_{2}\alpha + \gamma. \tag{24.2}
\]

Gamma radiation often accompanies alpha decay, for example:

\[
^{239}_{94}\text{Pu} \rightarrow ^{235}_{92}\text{U} + ^{4}_{2}\alpha + \gamma.
\]

After a decay reaction, the daughter nucleus is often in an «excited» state. A gamma radiation is emitted when daughter nucleus makes a transition from excited energy level to the ground one.

**Beta-decay** is accompanied by the interconversion between neutrons and protons inside a nucleus. There are three types of beta-decay:

1. **Electron or beta-minus β-decay.**

**Beta-minus decay** occurs because the nucleus has too many neutrons relative to protons. One of these neutrons decays into a proton and an electron. The proton remains in the nucleus but the electron \( ^{0}_{-1}\beta \) and an antineutrino \( ^{0}_{-1}\nu \) are emitted:

\[
^{1}_{0}\bar{n} \rightarrow ^{1}_{1}p + ^{0}_{-1}\beta + ^{0}_{-1}\bar{\nu}. \tag{24.3}
\]

Beta-minus decay proceeds according to the following scheme:

\[
^{A}_{Z}X \rightarrow ^{A}_{Z+1}Y + ^{0}_{-1}\beta + ^{0}_{-1}\bar{\nu} + \gamma. \tag{24.4}
\]

The daughter nucleus \( Y \) has an atomic number 1 more and an atomic mass \( A \) the same as the parent nucleus \( X \).

Antineutrino \( ^{0}_{-1}\nu \) is an elementary particle that travels with the speed, which is approximately equal to the speed of light, having no electric charge, little or no mass. And thus it is extremely difficult to detect. The difference between neutrino \( ^{0}_{0}\nu \) and antineutrino \( ^{0}_{0}\bar{\nu} \) is the opposite direction of spins.

Energy emitted by the β-decay is distributed randomly between an electron and an antineutrino. Therefore the kinetic energy of emitted β-particles takes all possible values, from 0 to \( E_{\text{max}} \). Thus the kinetic energy of the β-particles has a continuous spectrum (fig. 24.1).
There are examples of the beta-minus decay:
\[ ^{137}_{55}\text{Cs} \rightarrow ^{137}_{56}\text{Ba} + ^0_{-1}\beta + ^0_0\nu + \gamma \]
\[ ^{131}_{53}\text{I} \rightarrow ^{131}_{54}\text{Xe} + ^0_{-1}\beta + ^0_0\nu + \gamma \]
\[ ^{90}_{38}\text{Sr} \rightarrow ^{90}_{39}\text{Y} + ^0_{-1}\beta + ^0_0\nu \]
\[ ^{90}_{39}\text{Y} \rightarrow ^{90}_{40}\text{Zr} + ^0_{-1}\beta + ^0_0\nu \]

The gamma radiation is normally emitted soon after the daughter nucleus is created in an excited state.

2. Positron or beta-plus $\beta^+$ decay

The beta-plus particle $^0_{+1}\beta$ is emitted by nucleus that has too many protons. One of these protons decays into a neutron and a positron $^0_{+1}\beta$. The neutron remains in the nucleus but the positron $^0_{+1}\beta$ and a neutrino $^0_0\nu$ are emitted:
\[ ^1_1\text{P} \rightarrow ^0_0\nu + ^0_{+1}\beta + ^0_0\nu. \] (24.5)

Thus, in beta-plus decay the daughter nucleus $Y$ has an atomic number 1 less and an atomic mass $A$ the same as the parent nucleus $X$:
\[ ^Z_A\text{X} \rightarrow ^{Z-1}_{A}\text{Y} + ^0_{+1}\beta + ^0_0\nu. \] (24.6)

There is example of the beta-plus decay:
\[ ^{30}_{15}\text{P} \rightarrow ^{30}_{14}\text{Si} + ^0_{+1}\beta + ^0_0\nu. \]
3. Electron capture

In electron capture, one of the inner orbital electrons $^{0}_1$β is captured by nucleus (fig. 24.2) and a nuclear proton $^{1}_1$P is converted into a neutron $^{0}_0$n and a neutrino $^{0}_0$ν is emitted:

$$^{1}_1P + ^{0}_1\beta \rightarrow ^{0}_0n + ^{0}_0\nu.$$  \hspace{1cm} (24.7)

Electron capture results in a new nucleus formation with the same atomic mass $A$ and atomic number $Z$ which decreases by $I$ and emission of a neutrino. Electron capture is accompanied by characteristic X-rays when an electron from the $L$ or $M$ shell «jumps in» to fill the $K$ shell vacancy (fig. 24.2).

An example of electron capture is the transformation of beryllium $^{7}_4$Be into lithium $^{7}_3$Li:

$$^{7}_4\text{Be} + ^{0}_1\beta \rightarrow ^{7}_3\text{Li} + ^{0}_0\nu.$$

24.3. Nuclear reactions

A nuclear reaction is a process in which two nuclei or nuclear particles collide to produce particles different from the initial particles.

The first nuclear reaction was carried by Rutherford, who bombarded nitrogen $^{14}_7$N with alpha particles $^{4}_2$$\alpha$:

$$^{14}_7\text{N} + ^{4}_2\alpha \rightarrow ^{17}_8\text{O} + ^{1}_1\text{P}.$$

It is necessary to accelerate elementary particles up to high energy for this process. Then a charged particle can overcome the electrostatic repulsion force of nuclear protons.
Another method of the realization of nuclear reaction is **neutron activation**. A stable nucleus $^{A\ Z}_{Z}X$ absorbs a neutron $^{1}_{0}n$ and is transformed into a radionuclide $^{A+1\ Z}_{Z}X$ of the same element:

$$^{A\ Z}_{Z}X + ^{1}_{0}n \rightarrow ^{A+1\ Z}_{Z}X.$$  \hspace{1cm} (24.8)

It is possible to obtain radioactive cobalt by this way:

$$^{59}_{27}\text{Co} + ^{1}_{0}n \rightarrow ^{60}_{27}\text{Co}.$$  

Radioactive cobalt $^{59}_{27}\text{Co}$ is subjected to electron decay:

$$^{60}_{27}\text{Co} \rightarrow ^{60}_{28}\text{Ni} + ^{0}_{-1}\beta + ^{0}_{0}\nu + \gamma.$$  

Gamma-radiation which appears in this reaction is used in radiotherapy for the destruction of malignant tumors.

Moderated neutrons are more useful for nuclear reactions because fast neutrons can experience elastic collisions with a nucleus and scatter.

### 24.4. Radioactive decay law

The fundamental law of radioactive decay is based on the fact that the decay, i.e. the transition of a parent nucleus to a daughter nucleus is a purely statistical process. Following to the **radioactive decay law** one can say that the number of undecayed nuclei $N$ decreases exponentially with time $t$:

$$N = N_0 e^{-\lambda t},$$  \hspace{1cm} (24.9)

where $N$ is the number of parent nuclei at given time $t$; $N_0$ is the initial number of parent nuclei at time $t = 0$; $\lambda$ is the **decay constant** which determines a rate of decay (unit: s$^{-1}$). Each radioactive nuclide has a different decay constant $\lambda$.

**Half-life time** $T$ is the time interval, during which the number of radioactive nuclei is reduced to half the original number.

It is determined by the condition:

$$\frac{1}{2} N_0 = N_0 e^{-\lambda T} \Rightarrow 2 = e^{-\lambda T}.$$  

Finally, relation between $T$ and decay constant $\lambda$ is the following:

$$T = \frac{\ln2}{\lambda} \approx \frac{0.69}{\lambda}.$$  \hspace{1cm} (24.10)

The half-life time $T$ can be determined from fig. 24.3.
**Mean life time** \( \tau \) is the time during which the number of undecayed nuclei decreases in \( e = 2.72 \) time. There is following relation between \( \tau, \lambda \) and \( T \):

\[
\tau = \frac{1}{\lambda} = \frac{T}{0.69}.
\]  

**24.5. Radioactive substance activity**

The *activity* of a radioactive source \( A \) is defined as a number of decays per unite time, or other words the *activity* \( A \) is the number of decays per second:

\[
A = -\frac{dN}{dt}.
\]  

The activity \( A \) determines a decay rate.

Let’s put (24.9) in (24.12) and then differentiate \( N \) with respect to time \( t \):

\[
A = -\frac{dN}{dt} = \lambda N = \lambda (N_0 \cdot e^{-\lambda t}) = A_0 e^{-\lambda t}.
\]  

Thus:

\[
A = \lambda N = 0.69 \frac{N}{T}.
\]  

It is follows from equation (24.14), that activity \( A \) is directly proportional to the number of undecayed nuclei \( N \). Moreover, the smaller the halflife time \( T \) of a radionuclide is, the greater its activity \( A \) will be.

The SI unit of the activity \( A \) is the *becquerel (Bq)*: 1 Bq is defined as one decay per second. Another unit of the activity \( A \) is the *curie (Ci)*:

\[
1 \text{ Ci} = 3.7 \cdot 10^{10} \text{ Bq}.
\]

One curie is defined as the activity of 1 gram of pure \( ^{226}\text{Ra} \).
Activity $A$ decreases with time exponentially:

$$A = A_0 e^{-\lambda t}. \quad (24.15)$$

Let’s find the relation between activity $A$ and radionuclide mass $m$. The number of undecayed nuclei $N$ is determined as:

$$N = \frac{m}{m_N},$$

where $m$ is a mass of the radionuclide nuclei; $m_N$ is mass of a single atom of radionuclide.

The mass of a single atom of radionuclide $m_N$ can be found as:

$$m_N = \frac{M}{N_A},$$

where $M$ is the atomic mass of the radionuclide, $N_A$ is Avogadro’s number.

$$A = \frac{0.69 \cdot N}{T} = \frac{0.69 \cdot m \cdot N_A}{TM} = \frac{0.69 \cdot 6.02 \cdot 10^{23} \cdot m}{TM} = 4.17 \cdot 10^{23} \frac{m}{TM}. \quad (24.16)$$

Substituting mass of the radionuclide $m$ (g), its atomic mass $M$ (g · mol$^{-1}$), half-life time $T$ (sec) to equation (24.16), the activity $A$ (Bq) of the radionuclide can be obtained. From the known activity $A$ of the radionuclide its mass $m$ can be determined:

$$m = 2.4 \cdot 10^{-24} A \cdot T \cdot M. \quad (24.17)$$

Consider the specific mass activity $A_m$ to characterize the radioactive contamination of solid foodstuffs. The specific mass activity $A_m$ is defined as the activity per unit mass of radioactive substance:

$$A_m = \frac{A}{m}. \quad (24.18)$$

The specific mass activity $A_m$ is measured in Bq/kg or Ci/kg.

The radionuclide content in liquid or gas is characterized by the specific volume activity $A_V$. The specific volume activity $A_V$ is defined as the activity per unit volume of radioactive substance:

$$A_V = \frac{A}{V}. \quad (24.19)$$

The specific volume activity $A_V$ is measured in Bq/m$^3$, Bq/l, Ci/l.

The specific surface activity $A_S$ is used to characterize the radioactive contamination of surfaces. The specific surface activity $A_S$ is defined as activity per unit area of radioactive substance:

$$A_S = \frac{A}{S}. \quad (24.20)$$

The specific surface activity $A_S$ is measured in Bq/m$^2$, Ci/m$^2$. 

238
24.6. INTERACTION OF THE IONIZING RADIATION WITH THE MATTER

Ionizing radiation produces ions during the interaction with atoms or molecules of the medium it traverses. Ionizing radiation includes γ-rays and X-rays. It also includes all the atomic and subatomic particles, such as alpha-particles $\frac{4}{2}\alpha$, beta-particles $\frac{0}{-1}\beta$, $\frac{0}{+1}\beta$, neutrons $\frac{1}{0}n$ and protons $\frac{1}{1}p$.

24.6.1. Characteristics of the radiation-matter interaction

There are three important parameters characterizing the interaction of ionizing radiation with matter.

**Linear specific ionization** is the total number of ion pairs $dn$ produced per length unit $dl$ of the path of the incident radiation: $i = \frac{dn}{dl}$ (fig. 24.4). Specific ionization increases with decrease of energy of the charged particle because of the increase of probability of interaction at low energies.

**Fig. 24.4. Ion formation under ionizing radiation in the matter**

**Linear energy transfer (LET)** is the energy imparted per length unit of the path by the radiation: $LET = \frac{dE}{dl}$, where $dE$ is the energy imparted when the radiation travels a distance $dl$. Electromagnetic radiation and β-particles have low LETs. In contrast, heavy particles (α-particles) lose energy very rapidly, producing a lot of ions at a short distance, and thus they have high LETs.

**Mean linear range** of a charged particle is an average distance which the ionizing particle passes before its energy will be equal to the mean kinetic energy of the particles of this matter.

Let’s consider features of the interaction of the ionizing radiation with matter.

24.6.2. Features of the interaction of the ionizing radiation with matter

**Alpha-particles** $\frac{4}{2}\alpha$ are easily absorbed by materials because of their charge and large mass, and can travel only a few centimeters in air and in biological tissues — 10–100 µm. They can be absorbed by paper or the outer layers of the human skin and that is why they are not generally dangerous to life unless the source is ingested or inhaled. However, if alpha-radiation does enter
the body, *it is the most destructive form of ionizing radiation due to high LET*. Exposure of alpha-particles produces atoms excitation, ionization, characteristics X-rays appearance, nuclear reactions.

**Beta-particles** $^0_1\beta$, $^0_1\beta$ have an electrical charge and mass less than alpha-particle charge. Beta particles are much more penetrating than alpha particles (10–15 cm in biological tissues), but they have smaller ionizing power. Very high energy beta particles can penetrate to a depth of about a centimeter in tissue. Eye and skin damage is possible if the source is strong. They are, however, relatively easy to deal with by shielding. Exposure of beta-particles produces ionization and bremsstrahlung X-rays appearance.

**Gamma-rays** are a form of electromagnetic radiation of the highest frequency and energy, and also the shortest wavelength (below about $10^{-5}$ nanometer), within the electromagnetic spectrum. A high-energy gamma photon passing near a nucleus sometimes produces an electron and positron pair. Gamma-ray photons lose energy by being scattered from free electrons (Compton effect) or are completely absorbed by ejecting electrons from atoms (photoelectric effect). Thus the photoeffect and incoherent scattering (Compton effect) are the main mechanisms of interaction gamma-rays with matter. Gamma radiation frequently accompanies alpha and beta emissions. Gamma-rays have high penetrating power, they can pass tens and hundreds of meters in air and a few meters in soft tissues. Gamma-rays are much more penetrating than alpha-particles and beta-particles (fig. 24.5).

![Fig. 24.5. Penetrating power of α-, β- and γ-radiation](image)

**Neutrons.** Since free neutrons are electrically neutral, they pass free through the electrical fields within atoms and so constitute a penetrating form of radiation, interacting with matter almost exclusively through collisions with atomic nuclei. The way in which neutrons interact with matter depends on their energies. Neutrons will have a low probability of interaction because of the small size of the nucleus in relation to the atom, and could thus travel considerable distances in matter.
24.7. Principles of Radionuclide Diagnostics Methods

Radionuclide diagnostics is based on the radionuclides incorporation in biological tissue. Incorporated radionuclides are the γ-ray source which is registered by special detectors.

Let us specify physical properties of radiopharmaceuticals. The half-life must be short enough so that a reasonable fraction of the radioactive decays take place during the diagnostic procedure; any decays taking place later give a patient a dose that has no benefit. (This requirement can be diminished if the biological excretion is rapid.) On the other hand, the lifetime must be long enough so that the radiopharmaceutical can be prepared and delivered to a patient. For the diagnostic work, the decay scheme should minimize the amount of radiation which provides a dose to the patient but never reaches the detector. The ideal source then is a γ source, which means that the nucleus is in an excited state (an isomer). Such states are usually very short-lived. If the decay is a β⁻ or β⁺ decay, the product has different chemical properties from the parent and may be taken up selectively by a different organ. If it is also radioactive, it can confuse a diagnosis and give an undesirable dose to the other organ. It is necessary to remove the radioactive isotope from stable isotopes of the same element, because the chemicals are usually toxic.

Methods of the radionuclide diagnostics may be divided into two general types: gamma-scintigraphy and quantitative scintigraphy.

**Gamma scintigraphy** is a radiographic image techniques for visualizing the distribution of an injected radionuclide within the given organ as a means of studying of the anatomic structure of an organ via the introduction of an appropriate short lived gamma emitting radioisotope. The observed distribution can then be correlated with the rate and extent of drug absorption.

Different types of radionuclides tend to concentrate in different organs or tissues. So, the radionuclide used depends on which part of the body is to be scanned. For example, for scanning the thyroid gland radioactive iodine is used. Active parts of the tissue will emit more gamma-rays than less active or inactive parts. The gamma-rays which are emitted from inside the body are detected by the gamma-camera, are converted into an electrical signal, and sent to a computer. The computer builds a picture by converting the differing intensities of radioactivity emitted into different colours or shades of grey (fig. 24.6).

*Fig. 24.6. Different types of thyroid scintigraphy*
Radiography (quantitative scintigraphy) is a quantitative assay techniques for measuring the absorption and retention of a radionuclide within an organ as a means of studying the metabolism of the organ. It is displayed on the dependence of gamma-ray intensity on time. This investigation allows conclude about the blood flow, work of liver, kidneys, and lungs.

Let’s consider radiographic study of kidneys (fig. 24.7). The analysis data supplies detailed information about a kidney activity. It allows to find out a disturbance of an internal process (a rising branch) or an elimination process (a descending branch). One can perform this measurement for each kidney and make a comparative assessment their work.

![Graph of gamma-ray intensity over time](image)

*Fig. 24.7. The radiographic study of kidneys*

24.8. **Physical Basics of the Radiation Therapy**

Radiation therapy makes use of ionizing radiation, deep tissue-penetrating rays which can react physically and chemically with diseased cells to destroy them. Radiation therapy is used for cancer and for blood disorders such as leukemia.

Radiation may be injected to the body by implanting radioactive substances into the tumors or by exposing the body to external sources of high-energy rays that penetrate internally. Both methods have shown good results in the treatment or arrest of cancerous growths; the type of treatment used depends largely on the size of the tumor, its location.

The purpose of such radiation therapy is to destroy cancerous cells with minimal damage to normal healthy tissue or systemic involvement. Let’s consider features of different rays for radiation therapy. *X-rays* are applied for the irradiation of superficial tumors. The intensity of X-ray decrease sharply as depth increases (fig. 24.8, dotted line).

*Gamma-rays* have deep penetration and cause a minimum of surface-tissue irradiation. It allows to destroy deeply located tumors. Also it decreases damage to the skin and healthy tissues. Gamma radiation from $^{60}_{27}$Co has been usually used in cancer therapy.
Electron beams with energy about 25 MeV produce a maximum ionization at depth of 1–3 cm. They are used for irradiation of not deeply situated tumors.

Protons, due to their relatively big size, scatter less easily in the tissue. The beam stays focused on the tumor shape without much lateral damage to the surrounding tissues. All the protons of the given energy pass a certain distance; no proton penetrates beyond that distance. Furthermore, the dosage to tissue is maximum just over the last few millimeters of the particle range. This depth depends on the energy to which the particles were accelerated by the proton accelerator. Therefore it is possible to focus the cell damage due to the proton beam at the very depth in the tissues (11–14 sm) where the tumor is situated; the tissues situated before this area receive some reduced dose, and the tissues situated after the peak receive none.

Alpha-particles because of small linear range in matter may be used via the contact with on organism or on introducing it inside. A radon therapy is a characteristic example of it. Radon water is used for action on the skin (radon bath), the digestive apparatus (drinking), the respiratory apparatus (inhalations).

Questions:
1. Specify atomic nucleus characteristics.
2. What are the basic properties of the nuclear forces?
3. Describe modes of radioactive decay.
4. Explain why some radionuclides decay is accompanied by emitting γ- radiation?
5. Give β-particles kinetic energy spectrum and explain the spectrum.
6. Derive radioactive decay law. What are decay constant, half-life, mean lifetime? Describe relation between them.
7. What is the activity of a radioactive substance? What are the units of activity? Give the relationship between units.
8. What is the changing of the activity with time? How does activity depend on radionuclide mass?

9. Describe the characteristics of ionizing radiation interaction with the matter. What feature for α- and β-particles and γ-radiation interaction with the matter are observed?

10. Explain the principles of methods of radionuclide diagnostics.

11. What are the physical basics of the radiation therapy? Describe the features of the action of α- and β-particles, γ-radiation, neutrons and protons on the living organism.

Chapter 25. RADIATION DOSIMETRY

Ionizing radiations consists of photons (X-ray, γ-ray) and/or moving particles (α- and β-particles, neutrons and protons) that have sufficient energy to knock an electron out of an atom or molecule, thus forming an ion.

Nuclear radiation is potentially harmful to humans because the ionization it produces can significantly alter the structure of molecules within a living cell. The alternations can lead to the death of the cell and even of the organism itself. Despite the potential hazards, however, ionizing radiation is used in medicine for diagnostic and the therapeutic purposes. The hazards can be minimized only if the fundamentals of radiation exposure, including dose and biological effects of radiation, are understood.

25.1. RADIATION DOSES

25.1.1. Exposure

Exposure $X$ is a measure of ionization produced in air by X-rays and γ-rays and is defined by the formula:

$$ X = \frac{dQ}{dm}, $$

(25.1)

where $dQ$ is a total charge of the ions of one sign produced in the air and $dm$ is an mass of the air.

If the charge $Q$ is uniformly distributed in the air mass $m$, the exposure $X$ can be written as:

$$ X = \frac{Q}{m}. $$

(25.2)

In SI the exposure is measured in Coulombs per kilogram of air (C/kg). A special unit of the exposure $X$ is roentgen ($R$). One roentgen is the amount of gamma or X-radiation required to produce approximately $2.08 \cdot 10^9$ ion pairs in $1 \text{ cm}^3$ of dry air at standard temperature and pressure.

When X-rays or γ-rays produce an exposure of one roentgen the total charge $Q$: $Q = n \cdot q = 2.08 \cdot 10^9 \cdot 1.6 \cdot 10^{-19} = 3.33 \cdot 10^{-10} \text{ C}$ (n is the number of ion pairs in $1 \text{ cm}^3$ of dry air, $q$ is the charge of ion) is produced in $1 \text{ cm}^3$ of dry
The mass \( m \) of one cubic centimeter of dry air is \( m = \rho V = 1.29 \cdot 10^{-6} \text{ kg} \) \((\rho = 1.29 \text{ kg/m}^3 \) is the air density, \( V = 1 \text{ cm}^3 = 10^{-6} \text{ m}^3 \) is the volume). Therefore,

\[
1 \text{ R} = \frac{3.33 \cdot 10^{-10}}{1.29 \cdot 10^{-6}} = 2.58 \cdot 10^{-4} \text{ C/kg}.
\]

\[
1 \text{ R} = 2.58 \cdot 10^4 \text{ C/kg} \text{ or } 1 \text{ C/kg} = 3876 \text{ R}.
\]

The exposure \( X \) is used only for air and only for \( X \)-rays or \( \gamma \)-rays. It characterizes the environment ionization by electromagnetic radiation. Thus, exposure measures the electric charge (positive or negative) produced by electromagnetic radiation in a unit mass of air, at standard temperature and pressure.

If the exposure \( X \) is produced per unit of time \( t \), then the exposure rate can be determined by formula:

\[
\dot{X} = \frac{X}{t}. \tag{25.3}
\]

If the exposure \( X \) varies with time, then the exposure rate is determined as a derivative of the exposure dose \( X \) with respect to time \( t \):

\[
\dot{X} = \frac{dX}{dt}. \tag{25.3a}
\]

The exposure rate is measured in \( 1 \text{ A/kg} \) or \( m\text{R/hr}, \mu\text{R/s} \).

Exposure \( X \) is a very useful characteristic of environment ionization because it can easily be measured. It is necessary to measure the charge produced by the radiation in air (or the electrical current, multiplied by time) to find out the strength of the electromagnetic field. The instruments used for this purpose will be considered.

### 25.1.2. Absorbed dose

Since the concept of exposure is defined in terms of the ionizing abilities of \( X \)-rays or \( \gamma \)-rays in air, it does not specify the effect of radiation on living tissue. Due to the interaction between radiation and human tissue ionization occurs in the irradiated tissue. Chemical and biological changes in the tissue exposed to ionizing radiation depend on the energy absorbed by tissue. To measure the interaction of all types of radiation with any kind of material, the term absorbed dose \( D \) is used.

**Absorbed dose** \( D \) is defined as the energy \( dE \) absorbed from any ionizing radiation per unit of mass \( dm \) of the absorbing material:

\[
D = \frac{dE}{dm}.
\]
If energy $E$ absorbed by tissue due to radiation is uniformly distributed in the irradiated material of mass $m$, then

$$D = \frac{E}{m}.$$  
(25.4)

The SI unit of absorbed dose $D$ is $J\cdot kg^{-1}$ and the special name for this unit of absorbed dose is gray (Gy): $1\, \text{Gy} = 1\, J\cdot kg^{-1}$. Another unit is often used for absorbed dose — the rad (radiation absorbed dose). The rad and gray are related by: $1\, \text{Gy} = 100\, \text{rads}$.

The absorbed dose rate is determined as a derivative of the absorbed dose $D$ with respect to time $t$:

$$\dot{D} = \frac{dD}{dt}. $$

Mean absorbed dose rate can be written as:

$$\dot{D} = \frac{D}{t}. $$
(25.5)

The absorbed dose rate is measured in $\text{Gy/s, rad/s}$.

Internal absorbed dose can’t be measured directly. The absorbed dose $D$ can be estimated by the exposure $X$, which can easily be measured. There is an empirical relation between the amount of ionization in air and the absorbed dose for a given photon energy and absorber (body tissue). The absorbed dose $D$ is proportional to the exposure $X$:

$$D = f \, X,$$
(25.6)

where $f$ is a coefficient, which depends on the irradiated material structure and photons energy.

Let us evaluate coefficient $f$ for the air. To produce an ion in air (electric charge of $1.6 \cdot 10^{-19}\, C$) an average energy of $34\, \text{eV} = 34\cdot 1.6 \cdot 10^{-19}\, J$ is required. The mass $m$ of one cubic centimeter of dry air is $1.29 \cdot 10^{-6}\, \text{kg}$. An exposure $X$ of $1\, \text{Roentgen}$ corresponds to the absorbed dose $D$ of $88\cdot 10^{-4}\, \text{Gy}$ in dry air:

$$D = \frac{E}{m} = \frac{34\cdot 1.6 \cdot 10^{-19}\cdot 2.08 \cdot 10^9}{1.29 \cdot 10^{-6}} = 88\cdot 10^{-4}\, \text{Gy} = 0.88\, \text{rad}. $$

Therefore, for air the coefficient $f$ is equal to $0.88\, \text{rad/R}$. For water and soft tissues $f$ is equal to $1.0\, \text{rad/R}$. For bone tissue the coefficient $f$ depends on photon energy and its value varies from $1$ to $0.45\, \text{rad/R}$. The coefficient $f$ decreases with the increase of the quantum energy.

### 25.1.3. Equivalent dose

At equal absorbed doses, different types of ionizing radiation have varying effectiveness in producing radiation damage in a biological system because of the way they deposit their energy. For example, one rad dose of neutrons is far
more likely to produce eye cataracts than one rad dose of X-rays. To compare the damage caused by different types of radiation, the \textit{relative biological effectiveness (RBE)} is used. For certain biological effect the lethal dose LD\textsubscript{50/30} can be used. LD\textsubscript{50/30} means the certain biological effect i.e. 50 \% of the animals exposed will die within 30 days. The relative biological effectiveness of a particular type of radiation is the ratio of the absorbed dose of 180–200 keV X-rays needed to produce a certain biological effect to the absorbed dose of the test radiation needed to produce the same biological effect:

\[
\text{RBE} = \frac{\text{Absorbed dose of a test radiation that produces the same biological effect}}{\text{Absorbed dose of 180–200 keV X-rays that produces a certain biological effect}}. \quad (25.7)
\]

The \textit{RBE} depends on the nature of the ionizing radiation and its energy, as well as on the type of tissue being irradiated. In dosimetry the \textit{RBE} is represented in radiobiological standardization and regulatory law by the \textit{quality factor} \(k\). The value of quality factor \(k\) is equal 1 \((k = 1)\) for X-rays, \(\gamma\)-rays and beta-particles; for neutrons with energy 0.1–10 MeV \(- k = 10\), and for alpha particles with energy less than 10 MeV \(- k = 20\). The value of the quality factor \(k = 1\) for \(\gamma\)-rays and beta-particles indicates that they produce the same biological damage as do 180–200 keV X-rays. The larger \(k\) values for \(\alpha\)-particles and neutrons indicate that they cause substantially more damage. Thus, the damage produced by 20 Gy of X-rays is equal to that from 1 Gy of alpha radiation.

The \textit{equivalent dose} \(H\) is the measure which allows the different radiobiological effectiveness of different types of radiation to be taken into account. The product of the absorbed dose \(D\) and quality factor \(k\) is the equivalent dose \(H\):

\[
H = kD, \quad (25.8)
\]

The SI unit of the equivalent dose \(H\) is Sievert (Sv). Sievert is the product of the absorbed dose \(D\) in gray and quality factor \(k\). One sievert is generally defined as the amount of radiation roughly equivalent in biologic effectiveness to one gray (or 100 rads) of gamma radiation: 1 Sv = 1 J/kg. The sievert is inconveniently large for various applications, and so the millisievert (mSv), which equals to 0.001 sievert, is frequently used instead. Another unit is often used for equivalent dose — the \textit{rem} (roentgen equivalent man). The rem and sievert are related by: 100 rem = 1 sievert.

The equivalent dose is more biologically significant than the absorbed dose. It gives information about the biological effects of different types of radiation to a tissue or organ and helps to compare them, assessing the health risk of radiation exposure.
**Equivalent dose rate** is determined as a derivative of the equivalent dose $H$ with respect to time $t$:

$$\dot{H} = \frac{dH}{dt}. \quad (25.9)$$

If the equivalent dose $H$ doesn’t vary with time $t$, then the equivalent dose rate is determined as:

$$\dot{H} = \frac{H}{t}. \quad (25.9a)$$

The units of the equivalent dose rate are $\text{Sv/s, mZv/hr, rem/s}$.

### 25.1.4. Effective dose equivalent

Recognizing the fact that different tissues have different sensitivities and, therefore, proportionality constant between dose and effect is not the same for all tissues, an **effective dose equivalent** $H_{\text{eff}}$ is defined as the sum of the products of the equivalent dose to the organ $H_i$ and the weighting factor $w_i$ for each organ irradiated:

$$H_{\text{eff}} = \sum_i w_i H_i. \quad (25.10)$$

**Tissue weighting factor** is a factor that indicates the ratio of the risk of stochastic effects attributable to irradiation of a given organ or tissue to the total risk if the whole body is uniformly irradiated. By definition, the sum of individual weighting factors $w_i$ represents the weighting factor of whole body and is equal to one: $\sum_i w_i = 1$. The interpretation of the weighting factor is as follows. Consider the $w_i = 25$ for the gonads. This means that irradiation of the gonads alone would present about one-fourth the risk for stochastic effects expected to appear after uniform irradiation of the whole body at the same dose level. The values of $w_i$ are presented in table 25.1.

<table>
<thead>
<tr>
<th>Organ</th>
<th>$w_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gonads</td>
<td>0.25</td>
</tr>
<tr>
<td>Breast</td>
<td>0.15</td>
</tr>
<tr>
<td>Red bone marrow</td>
<td>0.12</td>
</tr>
<tr>
<td>Lungs</td>
<td>0.12</td>
</tr>
<tr>
<td>Thyroid gland</td>
<td>0.03</td>
</tr>
<tr>
<td>Bone surfaces</td>
<td>0.03</td>
</tr>
<tr>
<td>Remainder</td>
<td>0.30</td>
</tr>
</tbody>
</table>

Organs that have a large tissue weighting factor are those that are susceptible to radiation-induced carcinogenesis (such as the breast or red bone marrow) or to hereditary effects (the gonads).
The effective dose equivalent $H_{\text{eff}}$ is a risk-related parameter, taking relative radiosensitivity of each organ and tissue into account. The SI unit of the effective dose equivalent $H_{\text{eff}}$ is Sievert (Sv) (the weight factor is dimensionless).

### 25.1.5. Collective effective dose

For the purposes of assessing the overall effect of radiation on a large group of people or entire population, the term collective effective dose $S$ is used. The collective effective dose $S$ is defined as the sum of the products of the individual effective doses $H_{i,\text{eff}}$ and the population number exposed $N_i$:

$$S = \sum_i H_{i,\text{eff}} N_i.$$  \hspace{1cm} (25.11)

It is used to predict the magnitude of stochastic effects of radiation on the population. The collective dose is usually measured in units of person-sieverts (person-Sv) or man-sieverts (man-Sv).

### 25.2. Ionizing radiation detectors

The ionizing radiation detector is a device that is sensitive to radiation and can produce a response signal suitable for measurement or analysis. There are different detector types which are based on the effects of interaction between radiation and matter.

**Trace detectors** help to define a particle trajectory and its track length in the matter. A Wilson cloud chamber consists essentially of a closed container filled with a supersaturated vapor, e.g., water in the air. When ionizing radiation passes through the vapor, it leaves a trail of charged particles (ions) that serve as condensation centers for the vapor, which condenses around them. Thus the path of the radiation is indicated by tracks of tiny liquid droplets in the supersaturated vapor.

The tracks of alpha and beta particles have distinctive shapes (for example, alpha particle’s track is broad and straight, while that of an electron is thinner and shows more evidence of deflection). When a vertical magnetic field is applied, positively and negatively charged particles curve in the opposite directions.

One of the disadvantages of the cloud chamber is the relatively low density of the gas, which limits the number of interactions between ionizing radiation and molecules of the gas. For this reason physicists have developed other particle detectors, notably the bubble chamber. In the bubble chamber (fig. 25.1) the particle track is formed in the result of boiling a superheated liquid along the particle trajectory. As charged particles move through the liquid, they knock electrons out of the atoms of the liquid, creating ions. If the liquid is close to its boiling point, the first bubbles are formed around...
these ions. The observable tracks can be photographed and analyzed to measure the behavior of the charged particles.

Fig. 25.1. A bubble chamber tracks of some particles

In the basic type of an ionization detector a number and characteristics of an electric beam produced in the gas by the radiation are measured. For example, Geiger counters are widely used to indicate the presence and intensity of nuclear radiations. When a fast-moving charged particle traverses a Geiger counter, an electrical impulse is produced and can be counted.

A Geiger counter consists of a gas between two electrodes (fig. 25.2). One electrode, usually cylindrical and hollow, is the cathode. The other electrode, stretched along the axis of the cylinder, is the anode. A potential of about 1000 volts is placed on the wire. As particles enter the tube, they create a large avalanche of ionization in the gas, which then discharges, creating a brief electric pulse. The tube produces the same large output pulse for virtually every charged particle that passes through the gas and so it is useful for detecting individual particles. It can therefore indicate lower levels of radiation than is possible in comparison with other types of detectors.

The degree of ionization per volume unit is measured by ionization detectors. X-rays and gamma-rays have a great track length in the gas they rarely cause ionization. Mainly they knock electrons out of tube wall atoms which get into gas and ionize it.

Scintillation detector or a scintillation counter (fig. 25.3) also measures ionizing radiation. The sensor, called a scintillator, consists of a transparent crystal, plastic, or organic liquid that fluoresces when struck by ionizing radiation. A sensitive photomultiplier measures the light from the crystal. It is attached to an electronic amplifier and other electronic equipment to count and
possibly quantify the amplitude of the signals produced by the photomultiplier. In order to direct as much as possible of the light flash to the photosensitive surface, reflecting material is placed between the scintillator and the inside surface of the container.

![Geiger counter diagram](image)

*Fig. 25.2. Geiger counter*

A charged particle, moving through the scintillator, loses energy and leaves a trail of ions and excited atoms and molecules. Rapid interatomic or intermolecular transfer of electronic excitation energy follows, leading eventually to a burst of luminescence characteristic of the scintillator material. When a particle stops in the scintillator, the integral of the resulting light output, called the scintillation response, provides a measure of the particle energy, and
can be calibrated by reference to particle sources of the energy. Scintillation counters may be used to detect the various types of radioactivity (alpha, beta, and gamma rays), cosmic rays, and various elementary particles.

The registration of α-particles is most difficult due to their short path in matter. Alpha-radiation may be registered only from a thin surface layer so special preparation of patterns is necessary. Beta-particles have a longer path in matter so their detection is slightly simpler. The registration of γ-rays is the simplest due to their long path in matter. They may be registered even from a deep-seated object layer (fig. 25.4).

\[ q_1 - q_2 = \Delta q = C(U_1 - U_2). \]

where \( C \) is electrocapasity.

---

**Fig. 25.4.** The detection features of different particles

### 25.3. Radiation Monitoring Instruments

Radiation monitoring instruments are devices for the radiation doses or activity measurement. They are divided into dosimeters and radiometers.

A dosimeter is a device used to measure an exposure dose. A dosimeter consists of a detector and an electronic measuring device, which transform a detector signal into a form useful for registration.

Let us consider a dosimeter is based on ionization camera use. The ionization camera is filled with air under atmospheric or low pressure. Its active volume is \( V \). The ionization chamber consists of two electrodes (fig. 25.5). Before using these electrodes are filled with a potential difference \( U_1 \) and obtain a charge \( q_1 \).

When ionizing radiation penetrates the gas in the camera this radiation liberates electrons from the gas atoms leaving positively charged ions. The current begins to flow in a camera as soon as the electrons and ions begin to separate under the influence of the applied electric field. Therefore the potential difference decreases from the initial value \( U_1 \) to \( U_2 \), and the charge decrease from \( q_1 \) to \( q_2 \). Their changes are related in the following formula:

\[ q_1 - q_2 = \Delta q = C(U_1 - U_2). \]
Therefore the exposure dose is equal to:
\[ X = \frac{\Delta q}{m} = \frac{C(U_1 - U_2)}{\rho \cdot V} = k(U_1 - U_2) = k \Delta U, \]
where \( V \) is a camera volume, \( m \) is air mass, \( \rho \) is the air density. Constants in this formula may be joined in coefficient \( k \) in this formula. This coefficient is determined under device calibration.

An individual radiation dosimeter is a pen-like device that measures the cumulative dose of radiation received by the device. It is usually clipped to somebody clothing to measure his actual exposure to radiation.

Measure of exposure dose rate is based on the current determination in the detector:
\[ \dot{X} = \frac{dX}{dt} = \frac{dQ}{dt \cdot m} = \frac{I}{m}. \]

Exposure dose rate measuring instruments are usually calibrated in \( mR/hr \) or \( \mu R/s \).

So dosimeters help to measure the exposure dose in the air and to control \( \gamma \)-rays and X-rays background level. They can not be used to control radiation pollution degree of foodstuffs and the human organism.

A radiometer is a device for activity measurement. The activity is determined as the number of decay events per time unit. Therefore, radiometers count electrical pulse caused by particles hitting on the detector per time unit.

Let us consider one of the methods for determining a specific volume activity \( A_v \) or mass activity \( A_m \). A radiometer detector and the tested specimen are placed into the lead-wall camera to minimize the influence of the natural radiation background. It is necessary to improve the measurement accuracy. At first initial background activity (without the test specimen) is measured. Let \( N_1 \) be the impulse count of background activity in time \( t_1 \). Then a measuring cell is filled up with the tested product. Let a detector indicate the impulse count \( N_2 \) in time \( t_2 \). Then a specific volume activity \( A_v \) is calculated by the formula:
\[ A_v = \frac{N_2/t_2 - N_1/t_1}{P}, \]

where coefficient \( P \) takes into account a specimen volume and radiometer sensitivity to different types of radiation.

The determination of radioactive substance content in the human organism is a very important task. The internal irradiation radiometry is the most effective for gamma-emitting radionuclides. A special apparatus for internal the radiation dose consist of a steel protective room, a scintillation counter set, a recording system and a chair for a patient. A multichannel analyzer registers gamma-quanta and determines radionuclide type and the concentration of in the organism.

25.4. Background radiation

Background radiation is the ionizing radiation emitted from a variety of natural and artificial radiation sources. Some of this radiation is man-made, such as radiation used in medical applications and some is «natural». Natural sources include cosmic rays, terrestrial, and internal. Man-made radiation includes medical X-rays, medical nuclear procedures, consumer products, industrial sources, and some miscellaneous sources of radiation. The actual background encountered by each individual varies significantly, depending upon where he lives, the food that is consumed, the radon levels in the house, and so on.

Cosmic rays have always bombarded the earth. A typical person receives 0.31 mSv per year from cosmic rays. The earth’s atmosphere provides some shielding from cosmic rays. This shield is reduced at greater heights, and the cosmic ray dose is increased. Inhabitants at heights of 1600 meters receive 0.50 mSv/yr from cosmic rays, while those at the heights of 3200 meters receive 1.25 mSv. The effective equivalent dose, received by a person living at a sea level in the result of cosmic rays equals about 0.31 mSv per year.

Cosmic rays may broadly be divided into two categories, primary and secondary. The cosmic rays that arise in extrasolar astrophysical sources are primary cosmic rays; these primary cosmic rays can interact with interstellar matter to create secondary cosmic rays. The secondary cosmic rays reach the ground surface and contain all known elementary particles. The sun also emits low energy cosmic rays associated with solar flares. The exact composition of primary cosmic rays, outside the Earth’s atmosphere, is dependent on which part of the energy spectrum is observed. However, in general, almost 90 % of all the incoming cosmic rays are protons, about 9 % are helium nuclei (alpha-particles) and about 1 % are electrons.

Terrestrial background originates from radioisotopes that are found everywhere in our surroundings. All elements found in nature have radioactive isotopes, many of which are also present in the environment. The exact
composition of soil influences the local terrestrial background, because the minerals present determine which elements are most abundant. Terrestrial background sources are categorized as «primordial» if their half-lives are the same order of magnitude as the presumed lifetime of the earth \((4.5 \cdot 10^9 \text{ years})\). That is, these sources were present when the earth was formed and there is no way to replenish them in nature. Two isotopes of uranium, \(^{238}\text{U}\) and \(^{235}\text{U}\), and one of thorium \(^{232}\text{T}\), give rise to three different decay series. In each of these series, the radioactive nuclide decays to another stable isotope of bismuth or lead. Seventeen other nuclides are primordial, but are not part of a decay series. Of these nonseries radionuclides, \(^{40}\text{K}\) and \(^{87}\text{Rb}\) make the greatest contribution to the background dose so the dose accrue \(0.65 \text{ mSv per year}\). The mean background exposure dose rate for Belarus in normal stage is \(10-12 \mu\text{R/h}\).

Internal background is the dose imposed by the isotopes contained in our bodies. A small percentage of the potassium in the human body is \(^{40}\text{K}\). This radioactive nuclide emits both locally absorbed beta radiation and more penetrating gamma radiation. Similarly, \(^{14}\text{C}\), which comprises a small percentage of the carbon atoms found in organic molecules throughout our bodies, contributes to the total dose of \(1.35 \text{ mSv/yr}\) from internal background.

Radon, as part of the \(^{238}\text{U}\) decay series, is significant because it is an alpha emitter that exists as an inert gas. Since it is inert, radon generated by decay of \(^{226}\text{Ra}\) at some depth in the soil does not bind chemically with other elements. Instead, it percolates up to the surface to escape into the atmosphere. Being heavier than most constituents of the atmosphere, it tends to remain at lower elevations. Although minable deposits of uranium ore are primarily associated with granite rock formations, uranium is found everywhere in the earth’s crust. Ninety-nine percent of the uranium found in nature is \(^{238}\text{U}\). Thus, the air we breathe anywhere on earth contains some amount of radon.

Radon itself is not particularly hazardous, when inhaled, because it does not react and in most cases is simply exhaled. A more significant concern is that two of the decay products of radon (\(^{218}\text{Po}\) and \(^{214}\text{Po}\)), delivered to the air by decaying radon, are not inert. These products adhere to dust particles in the air, which may then be inhaled into the lung. Therefore the natural source dose is equal to \(2.0 \text{ mSv}\) per year.

Various human activities add to the annual radiation background. At first it is diagnostic radiology (\(0.39 \text{ mSv}\)). Radiation received by patients in radiation therapy is not counted in man-made background, because the intention is to track radiation doses that are associated only with stochastic effects.

Various consumer products emit small amounts of radiation. Some examples include exit signs that contain \(^{3}\text{H}\), a low-energy beta emitter, and smoke detectors that contain \(^{241}\text{Am}\), an alpha emitter. Luminous dials on watches, clocks, and instruments contained \(^{3}\text{H}\) and \(^{147}\text{Pm}\), both of which are low-energy beta emitters. The low-energy beta particles emitted by these substances are
absorbed in the instrument components and provide negligible amounts of the radiation dose to their owners. Increasingly, liquid crystal displays and light-emitting diodes are replacing the use of radioactive materials in luminous displays. Collectively, consumer products are estimated to contribute approximately 0.1 Sv to the yearly dose from man-made background radiation.

**Questions:**
1. What is the exposure dose? What are the units of the exposure dose? Give the relationship between units.
2. What is the absorbed dose and the absorbed dose rate? What are the units of the absorbed dose?
3. Calculate a coefficient which connects the exposure dose and the absorbed dose for the air.
4. Give a definition of the relative biological effectiveness. What does the equivalent dose mean? What are the units of the equivalent dose?
5. Write formula for the effective equivalent dose. What does this dose characterize?
6. Why is alpha-particles registration more difficult than a gamma-rays registration?
7. What is the difference between a dosimeter and a radiometer?
8. What does a background radiation consist of?
**CONTENTS**

Chapter 1. MATHEMATICS FUNDAMENTALS .............................................. 3

1.1. The function derivative ........................................................................ 3

1.2. Maxima and minima of functions.......................................................... 6

1.3. Differential of a function ....................................................................... 7

1.4. Partial derivatives ................................................................................. 8

1.5. Partial differentials, total differential of function .................................... 9

1.6. Antiderivative function, indefinite integral ............................................. 9

1.7. Definite integral .................................................................................... 11

1.8. Differential equations .......................................................................... 13

Chapter 2. PROBABILITY THEORY ............................................................... 15

2.1. Classical (theoretical) and statistical (empirical) probability definition ...... 15

2.2. Types of random events ........................................................................ 16

2.3. Probabilities addition and multiplication rules ....................................... 17

2.4. Bayes formula ....................................................................................... 19

Chapter 3. RANDOM VARIABLES. DISTRIBUTION OF RANDOM VARIABLES ........................................... 20

3.1. Discrete probability distribution ............................................................. 21

3.2. Continuous probability distribution. Probability density function .......... 22

3.3. Random distribution characteristics ..................................................... 23

3.4. Normal distribution .............................................................................. 25

Chapter 4. MATHEMATICAL STATISTICS FUNDAMENTALS ...................... 27

4.1. General population and sample ............................................................. 27

4.2. Statistical series types ........................................................................... 28

4.3. General population parameter estimation ................................................ 31

4.4. Correlation analysis .............................................................................. 32

Chapter 5. BASICS OF BIOMECHANICS .................................................... 34

5.1. Deformation characteristics ................................................................... 34

5.2. Stress-strain diagram ........................................................................... 37

Chapter 6. MECHANICAL OSCILLATIONS AND WAVES ................................ 40

6.1. Harmonic oscillations .......................................................................... 40

6.2. Damped harmonic oscillations ............................................................... 42

6.3. The forced harmonic oscillations ............................................................ 45

6.4. Superposition of harmonic oscillations .................................................. 46

6.5. The Fourier theorem ............................................................................ 47

6.6. Mechanical waves. The wave equation .................................................. 48
6.7. The Doppler effect .............................................................. 49

Chapter 7. ACOUSTIC .......................................................... 51
7.1. Physical and physiological characteristics of sound ...................... 51
7.2. Audition diagram .................................................................. 53
7.3. The Weber-Fechner Law ....................................................... 54
7.4. Reflection and absorption of the acoustic waves ......................... 55
7.5. Ultrasound in diagnostic and therapeutic applications .................. 57

Chapter 8. PROPERTIES OF LIQUIDS. SURFACE FENOMENA ........ 60
8.1. Surface tension .................................................................... 60
8.2. Phenomenon of the wetting and nonwetting of solids by liquids .... 62
8.3. Laplace pressure .................................................................. 64
8.4. Capillarity ........................................................................... 66
8.5. Methods of surface tension measurement ................................. 67

Chapter 9. BIOPHYSICAL PRINCIPLES OF BIORHEOLOGY AND HEMODYNAMICS ............... 70
9.1. Continuity equation .............................................................. 70
9.2. Bernoulli’s equation ............................................................. 73
9.3. Fluid viscosity ..................................................................... 74
9.4. Poiseuille’s equation ............................................................ 76
9.5. Methods of viscosity measurement ......................................... 79
9.6. Factors affecting blood viscosity ............................................ 82
9.7. Laminar and turbulent flow, Reynolds number ......................... 85
9.8. Pulse wave ........................................................................ 86
9.9. Distribution of blood pressure in cardiovascular system ............ 88
9.10. Blood pressure measurement ................................................ 90
9.11. Heart work and heart power ................................................ 91

Chapter 10. PHYSICAL PROPERTIES AND FUNCTIONS OF THE BIOLOGICAL MEMBRANE ...... 93
10.1. Structure and physical properties of the biological membrane .... 93
10.2. Types of lipids and proteins motion in the cell membrane ........ 95
10.3. Transport of molecules and ions through the membrane .......... 95
10.4. Mathematical description of the passive transport .................... 99
10.5. Active transport of ions ....................................................... 101

Chapter 11. MEMBRANE POTENTIALS OF THE CELL ............................................. 103
11.1. The Nernst equation ........................................................... 103
11.2. Resting membrane potential ................................................ 105
11.3. Action potential in excitable cells .......................................... 105
11.4. Propagation of action potential along an unmyelinated axon .......... 108
11.5. Propagation of action potential along a myelinated axon .......... 110

Chapter 12. ELECTRICAL FIELDS OF THE ORGANS AND TISSUES. METHODS OF THEIR REGISTRATION ......................... 112
12.1. Electrical field and its characteristics ........................................... 112
12.2. Electric Dipole and its field ......................................................... 117
12.3. Electrocardiography ..................................................................... 119

Chapter 13. ELECTROCONDUCTIVITY OF TISSUES AND LIQUIDS FOR DIRECT CURRENT .................................................. 126
13.1. Direct current in electrolytes ......................................................... 126
13.2. Features of electrical conductivity of biological tissues ............... 128
13.3. Some therapeutic methods based on the use of direct current ........ 129

Chapter 14. THE ALTERNATING CURRENT. THE ELECTRICAL IMPEDANCE OF LIVING TISSUE .............................................. 130
14.1. Main characteristics of the alternating current ............................... 130
14.2. AC Circuit with resistor ................................................................. 131
14.3. AC Circuit with a capacitor ............................................................. 132
14.4. AC Circuit with an inductor ........................................................... 133
14.5. Resistor, inductor and capacitor in series ...................................... 133
14.6. Electrical impedance of biological tissues for alternating current .... 135

Chapter 15. ELECTROSTIMULATION OF THE TISSUES AND ORGANS ..... 137
15.1. Characteristics of a rectangular pulse ............................................ 138
15.2. Characteristics of an arbitrary pulse .............................................. 139
15.3. Weiss–Lapicque Law .................................................................. 139
15.4. Electrical stimulation of the heart muscle .................................... 141

Chapter 16. HIGH FREQUENCY ELECTROMAGNETIC FIELDS USE IN MEDICINE ............................................................. 143
16.1. Diathermy ................................................................................. 143
16.2. Inductothermy ............................................................................ 145
16.3. Ultra high frequency therapy ....................................................... 146
16.4. The microwave therapy ............................................................... 148
16.5. Darsonvalisation ......................................................................... 149

Chapter 17. BIOPHYSICAL SIGNALS MONITORING ............................ 150
17.1. The Sensors ............................................................................... 150
17.2. Sensors of temperature ............................................................... 151
17.3. Biopotential amplifier ................................................................. 154
17.4. Differential amplifier ................................................................. 158
22.6. Retina anatomy and function ................................................................. 216
22.7. Rhodopsin-retinal visual cycle ............................................................... 218
22.8. Light and dark adaptation of eye .......................................................... 219
22.9. Color vision .......................................................................................... 220
Chapter 23. X-RAYS ...................................................................................... 221
23.1. Bremsstrahlung X-rays ......................................................................... 221
23.2. Characteristic X-rays ........................................................................... 224
23.3. Interaction between X-ray and matter .................................................... 225
23.4. Attenuation of X-rays .......................................................................... 226
23.5. Physical principles of the X-ray diagnostics ......................................... 228
Chapter 24. RADIOACTIVITY ........................................................................ 229
24.1. Characteristics of nucleus .................................................................... 229
24.2. Modes of radioactive decay .................................................................. 230
24.3. Nuclear reactions ................................................................................. 233
24.4. Radioactive decay Law ....................................................................... 234
24.5. Radioactive substance activity .............................................................. 235
24.6. Interaction of the ionizing radiation with the matter ............................. 237
24.7. Principles of radionuclide diagnostics methods .................................... 239
24.8. Physical basics of the radiation therapy .............................................. 240
Chapter 25. RADIATION DOSIMETRY .......................................................... 242
25.1. Radiation doses .................................................................................... 242
25.2. Ionizing radiation detectors .................................................................. 247
25.3. Radiation monitoring instruments ........................................................ 250
25.4. Background radiation .......................................................................... 252
Учебное издание

Кухаренко Людмила Валентиновна
Недзьведь Ольга Валерьевна
Гольцев Михаил Всеволодович
Лещенко Вячеслав Григорьевич

МЕДИЦИНСКАЯ И БИОЛОГИЧЕСКАЯ ФИЗИКА
ДЛЯ СТУДЕНТОВ-МЕДИКОВ

MEDICAL AND BIOLOGICAL PHYSICS
FOR MEDICAL STUDENTS

Учебно-методическое пособие

На английском языке

Ответственный за выпуск В. Г. Лещенко
Переводчики: Л. В. Кухаренко, О. В. Недзьведь
Компьютерная верстка Н. М. Федорцовой

Подписано в печать 17.09.15. Формат 60х84/16. Бумага писчая «Снегурочка».
Ризография. Гарнитура «Times».
Усл. печ. л. 15,11. Уч.-изд. л. 13,9. Тираж 400 экз. Заказ 664.

Издатель и полиграфическое исполнение: учреждение образования «Белорусский государственный медицинский университет».
Свидетельство о государственной регистрации издателя, изготовителя, распространителя печатных изданий № 1/187 от 18.02.2014.
Ул. Ленинградская, 6, 220006, Минск.