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

КУРС МЕДИЦИНСКОЙ И БИОЛОГИЧЕСКОЙ ФИЗИКИ

THE COURSE OF MEDICAL AND BIOLOGICAL PHYSICS

Учебно-методическое пособие

Минск БГМУ 2008
УДК 577.3 (811.111) (075.8)
ББК 28.707 (81.2 Англ – 923)

Утверждено Научно-методическим советом университета в качестве учебно-методического пособия 25.06.2008 г., протокол № 11

Авторы: В. Г. Лещенко, М. В. Гольцев, А. Д. Король, Г. К. Ильич, Л. В. Кухаренко, О. В. Недзведь

Перевод с русского языка В. Г. Лещенко, М. В. Гольцева, А. Д. Короля, Г. К. Ильича, Л. В. Кухаренко, О. В. Недзведь

Рецензенты: проф. каф. физики твердого тела Белорусского государственного университета, д-р физ.-мат. наук В. Г. Шепелевич; доц. каф. медицинской реабилитации и физиотерапии Белорусского государственного медицинского университета, канд. мед. наук Л. А. Малькевич

КУРС МЕДИЦИНСКОЙ И БИОЛОГИЧЕСКОЙ ФИЗИКИ
THE COURSE OF MEDICAL AND BIOLOGICAL PHYSICS

Учебное издание

Лещенко Вячеслав Григорьевич
Гольцев Михаил Всеволодович
Король Андрей Дмитриевич и др.

КУРС МЕДИЦИНСКОЙ И БИОЛОГИЧЕСКОЙ ФИЗИКИ
THE COURSE OF MEDICAL AND BIOLOGICAL PHYSICS

Учебно-методическое пособие

Ответственный за выпуск В. Г. Лещенко
В авторской редакции
Компьютерная верстка Н. М. Федорцовой

Подписано в печать 26.06.08. Формат 60×84/16. Бумага писчая «Кюм Люкс».
Печать офсетная. Гарнитура «Times».
Усл. печ. л. 15,58. Уч.-изд. л. 12,98. Тираж 50 экз. Заказ 467.
Издатель и полиграфическое исполнение – Белорусский государственный медицинский университет.
220030, г. Минск, Ленинградская, 6.

PART 1. Mathematical description of medico-biological processes and medical data processing

1.1. The derivative and differential of a function

The higher mathematics operates with a different values, some of them are independent and the other one depends on it. If value \( y \) depends on value \( x \) that means the \( y \) is a function of \( x \): \( y = f(x) \).

In many cases it’s important to determine how quickly the magnitude \( y \) is changed with \( x \). The \textit{derivative} of the function gives us this information.

1.1.1. The derivative of a function

Let’s note on the fig. 1.1.1 some values of argument \( x_0 \) and \( x \), the difference between which is an increment of argument: \( \Delta x = x - x_0 \). The increment of function will be: \( \Delta y = y - y_0 \). For continuous functions, if \( \Delta x \to 0 \), then \( \Delta y \to 0 \).

But it is impossible to foretell the value the attitude \( \frac{\Delta y}{\Delta x} \) aspires at unlimited decrease \( \Delta x \), because it depends on a concrete kind of function \( y(x) \).

\[ \text{Fig. 1.1.1} \]

\textbf{Definition:} Derivative of function \( y(x) \) in the given point \( x_0 \) is a limit of the attitude of an increment of function to an increment of argument at its unlimited decrease. The derivative of function of one argument is designated: \( y' \) or \( \frac{dy}{dx} \). Thus:

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

(1.1.1)

Derivative of function has simple \textit{geometrical sense}. 

3
From fig. 1.1.1 it is visible, that
\[ \frac{\Delta y}{\Delta x} = \frac{\gamma}{AC} = \tan \beta, \]
(1.1.2)

Where \( \alpha \) — a corner of secant \( AB \) slope to an axis \( OX \).

If \( \Delta x \to 0 \) the secant degenerates in a tangent to the curve of function in a point \( A \), having a corner \( \beta \) of an inclination to an axis \( OX \)
\[ y' = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \tan \beta = \tan \beta. \]
(1.1.3)

Thus, derivative of function is numerically equal to the tangent of a corner between the tangent lead to the curve of function in the given point, and an axis OX, — that is the geometrical sense of a derivative.

The physical sense of a derivative of a way on time is a speed of body movement:
\[ v = s' = \frac{ds}{dt} \]
(1.1.4)

Thus, the derivative of any function on time refers to as a speed of the function changing in time.

\[ \text{grad } C = \frac{dC}{dx} \]

\[ \text{Molecules} \]

\[ \text{X} \]

Fig. 1.1.2

The derivative of a function on spatial coordinate refers to as a gradient of this function.

For example, some substance is non-uniformly distributed along coordinate \( x \), i.e. its concentration \( C \) is some function of \( x \) (fig. 1.1.2). Then speed of concentration change along OX is defined by a derivative, and named the gradient of concentration.

Similarly there are gradients of pressure, temperatures and other values.

In the physics and mathematics the gradient \( \frac{df}{dx} \) is a vector which is directed aside the fastest increase of value \( f \).

The derivative of a function at a certain point is a measure of the rate at which that function is changing as an argument undergoes change. A derivative is the computation of the instantaneous slopes of \( f(x) \) at every point \( x \). This
corresponds to the slopes of the tangents to the graph of said function at said point; the slopes of such tangents can be approximated by a secant. Derivatives can also be used to compute concavity. The functions don’t have derivatives at points where they have either a vertical tangent or a discontinuity.

To find the derivative \( f'(x) \) of the function \( f(x) = x^2 \), let \( dx \) be an infinitesimal. Then \( f'(x) = (f(x + dx) - f(x))/dx = (x^2 + 2x*dx + dx^2 - x^2)/dx = 2x + dx = 2x, \) since \( dx \) is infinitesimally small.

It is important to remember that the limit of the infinitesimal is zero when \( \Delta x \to 0 \).

Hence, in the definition of derivatives we can read that a tangent is a limit position of a secant with \( \Delta x \to 0 \).

**What are the ways for finding the derivative of a suitable function?**
- The derivatives of elementary functions have been found already and given in table 1.1.

<table>
<thead>
<tr>
<th>Function</th>
<th>Derivative</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x^a )</td>
<td>( ax^{a-1} )</td>
</tr>
<tr>
<td>C (constant)</td>
<td>0</td>
</tr>
<tr>
<td>( e^x )</td>
<td>( e^x )</td>
</tr>
<tr>
<td>( \ln x )</td>
<td>( \frac{1}{x} )</td>
</tr>
<tr>
<td>( \sin x )</td>
<td>( \cos x )</td>
</tr>
<tr>
<td>( \cos x )</td>
<td>( -\sin x )</td>
</tr>
<tr>
<td>( \tan x )</td>
<td>( \frac{1}{\cos^2 x} )</td>
</tr>
<tr>
<td>( \cot x )</td>
<td>( -\frac{1}{\sin^2 x} )</td>
</tr>
<tr>
<td>( \arctan x )</td>
<td>( \frac{1}{1+x^2} )</td>
</tr>
</tbody>
</table>

**Table 1.1**

There are certain rules for taking a derivative of the combinations of elementary functions:
- **Constant Rule**: The derivative of any constant \( C \) is zero: \( C' = 0 \).
- **Constant Multiple Rule**: If \( C \) is some real number; then the derivative of \( C \cdot f(x) \) equals \( C \) multiplied by the derivative of \( f(x) \): \( (C \cdot f(x))' = C \cdot f'(x) \).
- **Sum Rule**: \( (f + g)' = f' + g' \).
- **Product Rule**: \( (f \cdot g)' = f' \cdot g + f \cdot g' \).
- **Quotient Rule**: \( \left( \frac{f}{g} \right)' = \frac{f' \cdot g - f \cdot g'}{g^2} \).
- **Chain Rule for composite function**: If \( f(x) = h(g(x)) \),
then $f'(x) = h'[g(x)] \cdot g'(x)$

– **What does a composite function mean?**
– This is a function obtained through the combination of two or more functions. The two given functions $f(x)$ and $g(x)$ of the composite function $f(g(x))$ are obtained by replacing each occurrence of $x$ in $f(x)$ by $g(x)$. A composite function is a function with variable as an argument which in its turn is a function of their variables: $f(g(x)) = f(u)$, where $u = g(x)$.

– **How can we calculate a composite function derivative?**
It is necessary to calculate a derivative of each function which is an argument of another function being a part of a composite function. Then, it is necessary to multiply it by each other:

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

*Example 1*: $y = \sin x$, and $x = (t^2 + 3t)$. Then $y = \sin (t^2 + 3t)$ is a composite function of $x$.
It’s derivative is $y' = (\sin x)'(t^2+3t)' = \cos x \cdot (2t + 3) = (2t + 3) \cdot \cos (t^2 + 3t)$.

*Example 2*: $y = \sin^3(tg(x^2))$. So to find the derivative of this composite function it is necessary to make a number of the conditional steps-mental operations:
- **To determine the number of functions-arguments which are the parts of a composite function.**
  - The power function with power equals 3 — the derivative is $3\sin^2(tg(x^2))$
  - $\sin(tg(x^2))$ — derivative is equal to $\cos(tg(x^2))$
  - $tg(x^2)$ — derivative is $\frac{1}{\cos^2 x^2}$
  - $x^2$ — derivative is equal $2x$

- **To calculate the derivative of a composite function:**

$$y'_x = 3\sin^2 (tg(x^2)) \cdot \cos(tg(x^2)) \cdot \frac{1}{\cos^2 (x^2)} 2x$$

- **Verification** (it is necessary to calculate the number of the multipliers and to compare them with the number of functions-arguments (see the example with the matreshka).

So we have selected 4 functions-arguments that go hand in hand with the number of multipliers being included in the derivative of a composite function.

### 1.1.2. The Multiple Derivatives

When the derivative $f''(x)$ of a function of $x$ has been found, the result being also a function of $x$, may be also differentiated which in its turn gives
the derivative of the derivative or the *second* derivative \( f''(x) \). Similarly, the derivative of the second derivative is called the *third* derivative \( f'''(x) \) and so on.

In order to avoid such “cumbersome” notation, the following options are often preferred:

\[
\frac{df}{dx} ; \frac{d}{dx} \left( \frac{df}{dx} \right) = \frac{d^2 f}{dx^2} ; \ldots \frac{d^n f}{dx^n} , \text{ or alternately, } f'(x); f''(x); \ldots f^{(n)}(x).
\]

**What is a derivation applied in biology and medicine for?**

Arguably the most important application of calculus, to physics, is the concept of the “time derivative” — the rate of change over the time — which is required for the precise definition of several important concepts. In particular, the time derivatives of an object's position are significant in Newtonian physics:

- Velocity (instantaneous velocity; the concept of average velocity predates calculus) is the derivative (with respect to the time) of an object's position.

The *distinction between constant velocity and varying velocity is important.* People often say velocity is equal:

\[
v = \frac{\text{total distance}}{\text{total time}}
\]

That's certainly true about the *average velocity* over the entire trip. But unless you have cruise control, you are probably going faster than the average velocity some of the time and slower than the average velocity other parts of the time.

**What is the instantaneous velocity?**

We can think of it as the average velocity over a very short interval of time. So suppose \( f(t) \) tells us our position at time \( t \), where we think of our position as a value along a line. At a particular time \( t \), we would like to find our instantaneous velocity \( v(t) \). Well, let \( \Delta t \) be a very small interval of time. Then \( f(t) \) is where we are at time \( t \) and \( f(t+\Delta t) \) is where we are at time \( t+\Delta t \). We have traveled a distance \( f(t+\Delta t) - f(t) \) over the time interval \( \Delta t \). Therefore, our average velocity over the interval of time \( \Delta t \) is \( \frac{f(t+\Delta t) - f(t)}{\Delta t} \).

Now, we don't want the average velocity over a little time interval, WE WANT TO KNOW THE VELOCITY RIGHT NOW, AT THIS VERY INSTANT. We just take the average velocities as the length of the time interval shrinks to nothing. Thus,

\[
v(t) = \lim_{\Delta t \to 0} \frac{f(t+\Delta t) - f(t)}{\Delta t} . \hspace{1cm} (1.1.5)
\]

- Acceleration is the derivative (with respect to time) of an object's velocity.
For example, if an object's position $p(t) = -16t^2 + 16t + 32$; then, the object's velocity is $p'(t) = -32t + 16$; the object's acceleration is $p''(t) = -32$; and the object's jerk is $p'''(t) = 0$.

### 1.1.3. THE SEARCHING OF THE FUNCTION EXTREMMUMS

The function extremum is a local minimum or a local maximum of the function. The points on the function graph where the derivative equals to zero are called “stationary points” and function extremums may occur at these points. To find these points it is necessary to solve the algebraic equation:

$$f'(x) = 0.$$  

The roots $x_1, x_2, x_3 \ldots$ of this equation are the stationary points.

To learn the type of extremum in this points it is necessary to calculate the second derivative $f''(x)$ and definite its sign at the 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 are 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 on the left side ($x<x_3$) and on the right one ($x>x_3$) from the $x_3$. If the sign on the left side is “−” and on the right one is “+” there is a local minimum at the point $x_3$. If the sign on the left side is “+” and on the right one is “−” there is a local maximum at the point $x_3$. And if the sign not changes there is no extremum at this point.

### 1.1.4. PARTIAL DERIVATIVES

The concept of a derivative has been entered for function of one variable. But there are many processes depending on several variables. For example, the condition of an organism can be considered as some function $U (x, y, z \ldots t)$ which value depends on a lot of parameters $x, y, z \ldots t$. Here $x$ can mean a body temperature, $y$ — sistolic pressure, $z$ — the maintenance of blood hemoglobin, etc. How to define a derivative of this function? In this cases, for functions of many variables, one enters concept of a partial derivative $Z'_x$ which characterizes speed of function change only on certain independent variable $x$ while the other variables are considered not changing:

$$\frac{\partial z}{\partial x} = \lim_{\Delta x \to 0} \frac{D_x z}{\Delta x} = \lim_{\Delta x \to 0} \frac{f(x_0 + \Delta x, y_0) - f(x_0, y_0)}{\Delta x}.$$  \hspace{1cm} (1.1.6)

$$\frac{\partial z}{\partial y} = \lim_{\Delta y \to 0} \frac{D_y z}{\Delta y} = \lim_{\Delta y \to 0} \frac{f(x_0, y_0 + \Delta y) - f(x_0, y_0)}{\Delta y}.$$

- **What does the partial derivative mean?**
  - Where a function depends on more than one variable, the concept of a partial derivative is used. Partial derivatives can be thought of informally as taking the derivative of the function with all but one variable held temporarily
constant near a point. Partial derivatives are represented as $\partial / \partial x$ (where $\partial$ is a rounded “d” known as the “partial derivative symbol”). Mathematicians tend to speak about the partial derivative symbol as “der” rather than the “dee” used for the standard derivative symbol, “d”.

Differentiating a function of more than one variable is more complicated than differentiating a function of one variable.

With introduction of partial derivative with respect to $x$, for instance, the rest of the variables were fixed. This definition coincides with that of the derivative of a function of one variable. So, a partial derivative can be found when all the variables have been fixed apart from one and when they are considered to be constant. A derivative is found in the same way as the derivative of a function of one variable, i. e.:

$$z' = \frac{dz}{dx} = \frac{d}{dx}(f(x, y_0));$$

(1.1.7)

All the rules and formulas being true to the derivative of a function of one variable are true to a partial derivative.

Example 1. $z(x, y) = 3x^3y^2 - 7xy^8 + 5$

We find the partial derivatives of $z$ with respect to $x$ and $y$. Initially we will not specify the values of $x$ and $y$ when we take the derivatives; we will just remember which one we are going to hold is constant while taking the derivative.

First, hold $y$ fixed and find the partial derivative of $f$ with respect to $x$:

$$\frac{\partial z}{\partial x} = 9x^2y^2 - 7y^8;$$

Secondly, hold $x$ fixed and find the partial derivative of $f$ with respect to $y$:

$$\frac{\partial z}{\partial y} = 6x^3y - 56y^7.$$

Example 2. Let's study the function

$$z = f(x, y) = x^4 + x^3 - 18x^2 - 16x + 32 - y^2$$

which is shown as a surface in xyz space below.
We find the partial derivatives of $z$ with respect to $x$ and $y$:

$$z'_x = 4x^3 + 3x^2 - 36x - 16,$$
$$z'_y = -2y.$$

### 1.1.5. Differential of Function

- Differential of argument $dx$ is equal to an increment of argument $\Delta x$, i.e. $dx = \Delta x$.
- Differential of function $dy$ is the product of derivative function on an increment (or differential) argument:

$$dy = y' \cdot \Delta x = y' \cdot dx.$$  \hspace{1cm} (1.1.8)

Differential $dy$ of function is not equal to its increment $\Delta y$ but represents its main part: $\Delta y \approx dy = y' \cdot dx$.

- Will anything change if “limit” is absent from the formula? Or: Is there any difference between the equality and tendency which limit expresses?
- That's right. We can write the derivatives of a function in the following way:

$$\frac{\Delta y}{\Delta x} = f''(x) + o(\Delta x)$$  \hspace{1cm} (1.1.9)

where $o(\Delta x)$ — *infinitesimal*. This value is the difference between the position of points A and B (we have spoken about it).

But, it is possible to write the definition of a function's derivative in the following way:

$$\Delta y = f'(x) \Delta x + o(\Delta x) \Delta x$$  \hspace{1cm} (1.1.10)

We can see that it consists of two items. What is the main item from the given 2?

- The first item is the differential $dy$ of function and it is the main.
- That's right. But why?
Because the multiplication of the infinitesimal parts results in more less infinitesimal.

The first item called as the *differential of a function* (abbreviated $df$ and $dy$) is a linear dependence $\Delta y$ from $\Delta x$, where $f'(x)$ is the coefficient of proportionality. It is important to remember the following definition of the function differential:

$$dy = f'(x) \cdot \Delta x \equiv f'(x) \cdot dx,$$  

(1.1.11)

So, it is important to remember:

$$f'(x) = \frac{dy}{dx}$$

1.1.6. **Partial differentials. Total differential of function**

Consider function of several (for simplicity — two) variables: $U(x, y)$. Private differential of this function on $x$ (a designation $d_x U$) is a product of a partial derivative $U'_x = \frac{\partial U}{\partial x}$ on differential of argument $dx$:

$$d_x U = U'_x \cdot dx = \frac{\partial U}{\partial x} dx.$$  

(1.1.12)

Similarly define private differential on $y$:

$$d_y U = U'_y \cdot dy = \frac{\partial U}{\partial y} dy.$$  

(1.1.13)

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

$$dU = U'_x \cdot dx + U'_y \cdot dy, \text{ or } dU = \frac{\partial U}{\partial x} dx + \frac{\partial U}{\partial y} dy.$$  

(1.1.14)

**Example.** Let’s consider a parallelepiped with lengths of edges: $x = 1m$, $y = 2m$, $z = 3m$. We shall admit the first two edges have increased on $\Delta x = 0,01m$ and $\Delta y = 0,02m$, and the third one has decreased on $\Delta z = -0,01m$. What size $\Delta V$ the volume of a parallelepiped has changed?

The volume of a parallelepiped is equal $V = x \cdot y \cdot z$.

It’s private derivatives are:

$$V'_x = yz, \quad V'_y = xz, \quad V'_z = xy.$$

Now we calculate the total differential of the volume:

$$dV = y \cdot z \cdot \Delta x + x \cdot z \cdot \Delta y + x \cdot y \cdot \Delta z = 2 \cdot 3 \cdot 0,01 + 1 \cdot 3 \cdot 0,02 - 1 \cdot 2 \cdot 0,01 = 0,10 \text{ (m}^3).$$
1.2. Antiderivative function. Indefinite and definite integral

As far as we know the equation of a curve, say \( F(x) \) then, by differentiation we can find a function which gives the derivative of a function (the gradient of the curve), namely \( F'(x) \). What about the reverse situation where we are given a function for the gradient of a curve \( F'(x) \) and we want to find the equation for the curve itself?

The essence of derivatives of a function is closely related to the differentiation i.e. fragmentation, revelation of instantaneous value. But very often it is necessary to find such function \( F(x) \) the derivative of which is equal to the initial function \( f(x) \). The function \( F(x) \) is called as antiderivative function and is the original image of the initial function \( f(x) \).

**Example 1.** Suppose that we are given \( F'(x) = 4x \).

In order to find \( F(x) \) we have to find the function that differentiates to give 4x. It's clear, \( 2x^2 \) is such a function, however there are many others e.g. \( 2x^2 + 1 \), \( 2x^2 + 2 \), \( 2x^2 + 2004 \), etc. In fact the function, \( F(x) = 2x^2 + C \), where \( C \) is an arbitrary constant, will differentiate to yield \( F'(x) = \frac{dy}{dx} = 4x \).

So there is a whole family of curves \( x^4 + C \) which gradient function is given by \( F'(x) = 4x \).

**Example 2.** \( F(x) = \frac{x^3}{3} \) is an antiderivative of \( f(x) = x^2 \). As the derivative of a constant is zero, \( x^2 \) will have an infinite number of antiderivatives such as \((x^3/3) + 0 \) and \((x^3/3) + 7 \) and \((x^3/3) – 36 \) etc.; the antiderivative family of \( x^2 \) is collectively referred to \( F(x) = (x^3/3) + C \); where \( C \) is any constant. Essentially, related antiderivatives are vertical translation of each other; the location of each graph depending upon the value of \( C \).

This process of finding the family of curves \( y = F(x) + C \) which has a given gradient function \( f(x) = F'(x) \) is called the finding of the **indefinite integral** or antiderivative of the function \( f(x) \).

So, the term “integral” means the following: if \( F(x) \) is a function, the derivative of which is given as:

\[
  f(x) = (F(x))'
\]  

Then we say that \( F(x) \) is the **indefinite integral** or antiderivative of \( f(x) \). The symbol \( \int f(x)\,dx \) is used to indicate the indefinite integral of \( f(x) \). The indefinite integral of any given function is not unique and can differ by up to a constant. Thus we write,

\[
  \int f(x)\,dx = F(x) + C \]  

(1.2.2)
where $C$ is an arbitrary constant known as the *constant of integration*, $dx$ is the infinitesimal.

There is an infinite difference between derivatives and integrals as inverse operations, which may be showed in more expressive way. For example, let’s study the given function $y = x^5$:

<table>
<thead>
<tr>
<th>$f(x) = x^5$</th>
<th>$F(x) = \frac{x^6}{6}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f'(x) = 5x^4$</td>
<td>$x^7$</td>
</tr>
<tr>
<td>$f'' = 5 \cdot 4 x^3$</td>
<td>$\frac{1}{6 \cdot 7} \cdot x^8$</td>
</tr>
<tr>
<td>$f''' = 5 \cdot 4 \cdot 3 \cdot x^2$</td>
<td>$\frac{1}{6 \cdot 7 \cdot 8} \cdot x^9$</td>
</tr>
<tr>
<td>$f'''' = 5 \cdot 4 \cdot 3 \cdot 2 \cdot x^1$</td>
<td>$\frac{1}{6 \cdot 7 \cdot 8 \cdot 9} \cdot x^{10}$</td>
</tr>
<tr>
<td>$f'''''' = 120x^0 = 120$</td>
<td></td>
</tr>
</tbody>
</table>

While differentiating, the exponent of the function decreases in series by 1, whereas the integration leads to the increase of the exponent in series by 1.

Listen to a funny example for better understanding. Let's suppose that at first you have found the derivatives of a function, and then the second derivative. It is similar to going down two steps. Then to get back to your initial position you need to go up two steps up. The latter is referred to the operation of integration.

Hence differentiation and integration are reverse (inverse) operations.

– Why does the term “integral” always include the infinitesimal $dx$?
– Because derivatives and integrals are inverse operations, and the derivation may be expressed as:

\[
\frac{dy}{dx} = f''(x) \\
\int f''(x)dx = \int dy
\]

It means that, $\int f''(x)dx = \int dy$.

So, we have basic rules for the definition of the integral, which demonstrate the connection with the derivation of a function:

\[
(\int f(x)dx)' = f(x) \\
d(\int f(x)dx) = f(x)dx
\]
\[ \int d(F(x)) = \int f(x)dx = F(x) + C \]

Two signs both located on the left and right sides mean the signs of the integral and differential correspondingly.

For example, \( dy = \sin x \, dx \). Then \( y = \int dy = \int \sin x \, dx = -\cos x + C \)

Besides, as well as in the case with calculating the derivative of the functions there are the following rules to calculate the integrals:

\[ \int (y_1 \pm y_2) \, dx = \int y_1 \, dx \pm \int y_2 \, dx \]
\[ \int kydx = k \int ydx \]

– What is the “definite integral”?
– Let’s ask ourselves the following question: “What is the exact way to calculate the length of a curve”?

It is necessary to divide a curve into such small curves so that the distance between its two points are segments. It is only possible if the number of the segments is infinite, i.e. the elementary lengths of such segments are infinitesimal!

Thus, the length \( l \) of a curve will be equal to the sum of the lengths of the segment \( \Delta l \), if the number of such segments indefinitely increases. It follows:

\[ l = \lim_{n \to \infty} \sum_{i=1}^{n} \Delta l_i \]

Have a look at the graph below, where we see that \( y = f(x) \) is a positive, continuous function in the interval \([a, b]\). There is an area bounded by the graph of \( f(x) \), the x axis, and the lines \( x = a \) and \( x = b \) which we have labeled as \( A \).

![Fig. 1.2.1](image-url)
The question which we want to ask is: how can we find the value of the area $A$?

We can get an approximate value for $A$ by dividing it into strips, and estimating the area of each strip as that of a rectangle, and then adding up these results.

Let's divide or, using the mathematical term, *partition* the interval $[a,b]$ into $n$ *closed subintervals* by specifying a set of points $x_k$ ($k = 0, 1, 2, \ldots n$), so $a = x_0 < x_1 < x_2 < \ldots < x_{n-1} < x_n = b$.

We can conveniently label each of our subintervals as $I_1 = [x_0, x_1]$, $I_2 = [x_1, x_2]$, ..., $I_n = [x_{n-1}, x_n]$. We use the symbol $\Delta x_k = x_k - x_{k-1}$ for the length of the $k$-th subinterval $I_k$, as shown in the graph below. The largest of the subintervals $\Delta x_k$ is known as the *mesh* of the partition and the length of the mesh is called the *size* or *norm* of the partition.

![Fig. 1.2.2](image.png)

Having divided the area $A$ into $n$ sub-areas, now we approximate each of these smaller areas by a rectangle. We already have a value for the width of each of these rectangles, namely $\Delta x_k$, now we need to specify what we are going to take as their height. To this end, let's define a sampling point $x^*_k$ for each subinterval $\Delta x_k$. Then we define the height of each rectangle to be $f(x^*_k)$.

The area of a little subinterval under the curve can be thought of as the width of the strip multiplied by the height of the strip.
Hence our approximate value for the area of the $k$-th subinterval is $f(x_k^*) \cdot \Delta x_k$. Adding up all the approximate areas of all the subintervals gives us our estimation of the area $A$:

$$A \approx \sum_{k=1}^{n} f(x_k^*) \Delta x_k$$

(1.2.3)

We have seen how we can approximate the area under a curve by the use of a the sum. We would expect to get a better approximation for the area under a curve by increasing the number of subintervals $n$. In fact, if we were to increase indefinitely the number of subdivisions so that the width of each approaches zero, then we would hope to get the value of the area $A$. Expressing this in mathematical language we say that

$$A = \lim_{n \to \infty} \sum_{k=1}^{n} A_k = \lim_{n \to \infty} \sum_{k=1}^{n} f(x_k) dx_k.$$  

The right part of this equation is the definite integral:

$$\int_{a}^{b} f(x) dx = \lim_{n \to \infty} \sum_{i=1}^{n} f(x_i) dx_i$$

(1.2.4)

If this limit of this sum exists, and does not depend on how the interval $[a,b]$ is partitioned or on the position of the sampling point in each subinterval, then it is known as the definite integral of $f(x)$ with respect to $x$ between $a$ and $b$, and is represented by $\int_{a}^{b} f(x) dx$.

The integral is also regarded as the limiting value of the sum of great numbers of differentials, when the magnitude of the differentials decreases and their number increases indefinitely.

So, the integral gives you a mathematical way of drawing an infinite number of blocks and may be defined as the limit of a sum of strips. Considered so, integration allows us to calculate the area under a curve and the surface area and volume of solids such as spheres and cones.

What is the geometrical difference between the derivatives and integrals?

---

**Derivative**

$$\frac{df(x)}{dx}$$

The derivative of the function $f(x)$ evaluated at $x=a$ gives the slope of the curve at $x=a$.

**Integral**

$$\int f(x) dx$$

The integral of the function $f(x)$ over the range $x=b$ to $x=c$ gives the area under the curve between those points.
If the derivative of a function can be geometrically interpreted as the slope of a curve of the mathematical function \( f(x) \) plotted as a function of \( x \), the integral of a function is the area under the curve of the mathematical function \( f(x) \) plotted as a function of \( x \).

- What is the difference between the definite and indefinite integrals?
- The value of a definite integral is the number, whereas the value of the indefinite integral is the function.

\[
\int f(x)\,dx = F(x) + C \quad (1.2.5)
\]

\[
\int_a^b f(x)\,dx = F(b) - F(a) = \text{number} \quad (1.2.6)
\]

**It's necessary to think over the following!**

If the **DEFINITE integral** geometrically refers to the area of the square, i.e. it symbolizes the **NUMBER**, what does the **indefinite integral** (which refers to a **function**) geometrically mean?

---

**Techniques of integration**
- How is the integral calculated?
- If derivatives and integrals are inverse operations, we can draw the table of integrals on the basis of the table of the derivative of a function.

**Table 1.3**

<table>
<thead>
<tr>
<th>Function</th>
<th>Derivatives</th>
<th>Integrals</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x^a )</td>
<td>( ax^{a-1} )</td>
<td>( \int x^a , dx = \frac{x^{a+1}}{a+1} + C )</td>
</tr>
<tr>
<td>( k ) (constant)</td>
<td>0</td>
<td>( \int k \cdot dx = k \cdot x + C )</td>
</tr>
<tr>
<td>( a^x )</td>
<td>( a^x \ln a )</td>
<td>( \int a^x , dx = \frac{a^x}{\ln a} + C )</td>
</tr>
<tr>
<td>( e^x )</td>
<td>( e^x )</td>
<td>( \int e^x , dx = e^x + C )</td>
</tr>
<tr>
<td>( \ln x )</td>
<td>( \frac{1}{x} )</td>
<td>( \int \frac{1}{x} , dx = \ln</td>
</tr>
<tr>
<td>( \sin x )</td>
<td>( \cos x )</td>
<td>( \int \sin x , dx = -\cos x + C )</td>
</tr>
<tr>
<td>( \cos x )</td>
<td>( -\sin x )</td>
<td>( \int \cos x , dx = \sin x + C )</td>
</tr>
<tr>
<td>( \tan x )</td>
<td>( \frac{1}{\cos^2 x} )</td>
<td>( \int \tan x , dx = -\ln</td>
</tr>
</tbody>
</table>
\[
\begin{array}{|c|c|}
\hline
ctg x & -\frac{1}{\sin^2 x} \\
\hline
arctg x & \frac{1}{1 + x^2} \\
\hline
\end{array}
\]

\[
\int ctg x dx = \ln|\sin x| + C
\]

\[
\int \frac{1}{1 + x^2} dx = arctg x + C
\]

– What are the other ways to calculate the indefinite integral?
– To find antiderivatives is more difficult than to find derivatives.

Working out the indefinite integral of any given function \( g(x) \) is not always an easy matter (in fact sometimes it is impossible). Unlike differentiation there are no general rules which can always be applied. Instead we have various techniques which will often lead us to an answer. These techniques range from the easy “by inspection” or “seen that sort of thing before” through to incredibly complicated manipulations involving all sorts of fancy mathematics.

We have various methods at our disposal:
– the linearity of integration allows us to break complicated integrals into simpler ones;
– integration by substitution;
– integration by parts to integrate products of functions.

1) In calculus, linearity is a fundamental property of the integral that follows from the sum rule in integration and the constant factor rule in integration.

Let \( f \) and \( g \) be functions. Now consider:

\[
\int a f(x) + b g(x) \, dx
\]

By the sum rule in integration, this is:

\[
\int (a f(x) \, dx) + \int (b g(x) \, dx)
\]

By the constant factor rule in integration, this refers to:

\[
a \int f(x) \, dx + b \int g(x) \, dx
\]

Hence we have:

\[
\int (a f(x) + b g(x)) \, dx = a \int f(x) \, dx + b \int g(x) \, dx
\]

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

– Is there any difference between the addition or subtraction of the constant \( C \)?

No, there isn't, because an arbitrary constant \( C \) expresses any real number which may have a negative sign: \( -C = +(-C) \).

2) In calculus, the substitution rule is an important tool for finding antiderivatives and integrals. It is the counterpart to the chain rule for differentiation.

Study the integral \( \int_0^2 t \cos(t^2 + 1) \, dt \).

By using the substitution \( x = t^2 + 1 \) we obtain \( dx = 2t \, dt \) and \( t \, dt = dx/2 \).

If \( t = 0 \), the \( x = 0^2 + 1 = 1 \), if \( t = 2 \), then \( x = 2^2 + 1 = 5 \), so
\[ \int_0^2 t \cos(t^2 + 1) dt = \frac{1}{2} \int_0^2 \cos(t^2 + 1) 2t \, dt = \frac{1}{2} \int_1^5 \cos(x) \, dx = \frac{1}{2} (\sin(5) - \sin(1)). \]

**When can we substitute one variable for the other one?**

If there is a “differential” relation between these functions. For example, in the given integral \( \int \frac{x^2 \, dx}{x^3 + 6} \), the algebraic function \( x^2 \) is the derivative of the function \( x^3 + 6 \): \( (x^3 + 6)' = 3x^2 \), so \( d(x^3 + 6) = 3x^2 \, dx \).

The given relation can be easily explored if we remember that the operation of the differentiation of the power function results in the decreasing of the exponent by 1!

Hence, the numerator of the given fraction can be represented as a function which is found in the denominator:

\[ \int \frac{x^2 \, dx}{x^3 + 6} = \frac{1}{3} \int \frac{d(x^3 + 6)}{(x^3 + 6)}. \]

**Why have we used the multiplier \( \frac{1}{3} \) before the integral?**

- Please, have a look at the following equation: \( d(x^3 + 6) = 3x^2 \, dx \), whereas \( x^2 \, dx \) is in the numerator of the initial integral.

It means that for easier calculation of this integral we have expressed the given numerator as the denominator, i.e. we have used one function instead of two initial functions. At the same time the value of the ratio increases threefold and for the elimination of this increase, it is necessary to divide the given ratio by 3 or multiplicate it by \( 1/3 \).

Thus, the possibility to substitute the function for the new variable, for instance, it follows:

\[ \int \frac{x^2 \, dx}{x^3 + 6} = [x^3 + 6 = t, \ dt = 3x^2 \, dx] = \frac{1}{3} \int \frac{dt}{t} = \frac{1}{3} \ln|t| + C = \frac{1}{3} \ln|x^3 + 6| + C = \ln\sqrt[3]{x^3 + 6} + C \]

- So, what is the sequence to calculate the integral by substitution?

Consider, for example, the integral \( \int \frac{\ln^5 x}{x} \, dx \).

- The basic steps of our mental operations are the following:
  - We determine what types of the functions are under the sign of the integral, i.e. it is necessary to compare the functions under the integral. We must answer the question: Is it possible to substitute one function for another one?

  \( \ln x \) and \( \frac{1}{x} \) \( \Rightarrow \ (\ln x)' = \frac{1}{x} \).

So, the substitution is possible in our example: \( t = \ln x, \ dt = \frac{1}{x} \, dx \).
\[
\int \frac{\ln^5 x}{x} \, dx = [\ln x = t] = \int t^5 \, dt = \frac{t^6}{6} + C = \frac{\ln^6}{6} + C.
\]

1.3. Differential equations

Definition: The differential equation is an equation that connects argument \(x\), its function \(y\) and their derivatives \(y', y'', ..., y^{(n)}\) of various orders.

In a common case the differential equation can be written down as

\[F(x, y, y', y'', ..., y^{(n)}) = 0. \tag{1.3.1}\]

The order of the differential equation is defined by the highest order of a derivative entering into it.

An example of the differential equation is the second law of the Newton, defining force \(F\) as product of weight of a body \(m\) on acceleration that have been got under action of the force: \(F = mv'\), or \(F = mx''\).

The solution of the differential equation is function which turns this equation into identity.

There are general and partial (particular) solutions of the differential equation.

The general solution contains the arbitrary constants \(C_1, C_2, ..., \) and their number is equal to the order of the differential equation.

If the constant \(C\) accepts any concrete value the solution will be partial.

There are infinite set of partial solutions. To choose the necessary partial solution, one uses some conditions (initial or boundary).

– How are a number and a function related to the differential equations?

– We can consider a number as a solution of algebraic equation, but a function — as a solution of differential one.

A differential equation explains the conversion of static process to dynamic one.

– What do mathematical statics and dynamics mean?

– Functions are abstractly defined as certain relations. Formally, a function \(f\) from a set \(X\) of input values to a set \(Y\) of possibly output values (written as \(f: X \rightarrow Y\)) is a relation between \(X\) and \(Y\).

If a function may be represented as a dynamic process, a number may be shown as a static one.

– What is the essence of differential equation?

– In mathematics, and more specifically calculus, a differential equation is an equation that describes the relationship between an unknown function and its derivatives.
Differential equations are used to construct mathematical models of physical phenomena such as fluid dynamics or celestial mechanics. Therefore, the study of differential equations is a wide field in both pure and applied mathematics.

**What is the solution to a differential equation?**

The problem of solving a differential equation is to find function $y$ and its derivatives satisfy the equation. For example, the differential equation $y'' + y = 0$ has the general solution $y = A\cos x + B\sin x$, where $A$, $B$ are constants. For application to a physical problem, the constants must be determined by forcing the solution to fit physical boundary conditions. Once a general solution is formed and then forced to fit the physical boundary conditions, one can be confident that it is the unique solution to the problem, as guaranteed by the uniqueness theorem.

For example:

<table>
<thead>
<tr>
<th>Algebraic equations</th>
<th>Differential equations</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y^2 - 7y + 12 = 0$</td>
<td>$y'' - 7y' + 12y = 0$</td>
</tr>
<tr>
<td>$y_1 = 4, y_2 = 3$</td>
<td>$y = C_1 e^{-3t} + C_2 e^{3t}$</td>
</tr>
</tbody>
</table>

**What are the types of differential equations?**

There are many types of differential equations. The simplest one is the separable equation. Those equations are considered to be the equations, which are the easiest for comprehension.

The differential equation of the form $\frac{dy}{dx} = f(x, y)$ is called *separable*, if $f(x, y) = h(x) g(y)$; that is,

$$\frac{dy}{dx} = h(x)g(y),$$ (1.3.2)

**Why should the values be separated?**

To integrate and to obtain the functions. We have already known, that differentiation and integration are inverse operations.

So, in order to solve the equation of this type, perform the following steps:

1. Rewrite the equation (1.3.2) as:

$$\frac{dy}{g(y)} = h(x)dx,$$

and, then, integrate $\int \frac{1}{g(y)} dy = \int h(x)dx$

to obtain the solution;

$$G(y) = H(x) + C.$$
2. The constant $C$ is found from the initial condition and then we find the particular solution.

**Example.**

$x dy = y dx$

**Solution.**

It is necessary to:

1) separate $\frac{dy}{y} = \frac{dx}{x}$,

2) and, then, integrate $\int \frac{dy}{y} = \int \frac{dx}{x}$

3) to obtain the general solution: $ln|y| = ln|x| + C$.

**Attention!**

Sometimes it is better to substitute the constant $C$ for the constant $lnC$:

\[
ln|y| = ln|x| + lnC \\
\downarrow \\
ln|y| = ln|xC| \\
\downarrow \\
y = xC
\]

– Why can’t we separate the variables in the other way?

\[
\frac{x}{dx} = \frac{y}{dy}
\]

– Because the differential as a part of integral can’t be represented as a divisor.

– Is there the rule for the separation of values?

– Yes, of course. There is the following algorithm to solve a separate equation.

To define which variable is *basic*, around which the other variable will be placed. This variable is always under the sign of differential. In the given example, on the left side it is $dy$, on the right side it is $dx$.

The other variable, which is different from the basic one, and which can be displaced to the other side, can be underlined. In our example, $x$ (the basic variable — $y$) is unnecessary on the left side, $y$ is unnecessary on the right side, as the basic variable of this side is $x$.

\[
xdy = ydx
\]  \hspace{1cm} (1.3.3)

Such underlining is used so that both sides of an equation can be divided by the multiplication of underlined multipliers:
\[ \frac{xy}{dy} = \frac{ydx}{xy} \Rightarrow \frac{dy}{y} = \frac{dx}{x} \]

– Does a differential equation have more than one solution? If so, how can we find a solution which satisfies the particular initial conditions?

– A problem in which we are looking for the unknown function of a differential equation where the values of the unknown function and its derivatives at some point are known is called an initial value problem (in short IVP).

– What are the initial conditions?

– They are to specify the state of the system at a given time (usually taken to be \( t = 0 \)).

If we have the following initial conditions: \( x = 3, \ y = 6 \), then, the general solution \( y = xC \) will be represented in the form \( 6 = 3C \Rightarrow C = 2 \).

So, the particular solution is as follows: \( y = 2x \).

If no initial conditions are given, we call the description of all solutions to the differential equation the general solution.

– What is the process of working out a differential equation?

– One of the most difficult problems that a scientist deals with in his everyday research is: “How do I translate a physical phenomenon into a set of equations which describes it?”

It is usually impossible to describe a phenomenon totally, so one usually strives for a set of equations which describes the physical system approximately and adequately.

In general, once we have built a set of equations, we compare the data generated by the equations with real data collected from the system (by measurement). If the two sets of data “agree” (or are close), then we gain confidence that the set of equations will lead to a good description of the real-world system. For example, we may use the equations to make predictions about the long-term behavior of the system. It is also important to keep in mind that the set of equations stays only “valid” as long as the two sets of data are close. If a prediction from the equations leads to some conclusions which are by no means close to the real-world future behavior, then we should modify and “correct” the underlying equations. As you can see, the problem of generating “good” equations is not an easy exercise.

Note that the set of equations is called a Model for the system.

– How do we build a Model?

The basic steps in building a model are:

**Step 1:** Clearly state the assumptions on which the model will be based. These assumptions should describe the relationships among the quantities to be studied.

**Step 2:** Completely describe the parameters and variables to be used in the model.
**Step 3:** Use the assumptions (from Step 1) to derive mathematical equations relating to the parameters and variables (from Step 2).

The best example of mathematical modeling is the one related to the decrease of *light intensity* which is sent to an obstacle.

Let’s give an example of making out a differential equation

**First step.**

The more is the thickness of an obstacle, the more is the decrease of light intensity $\uparrow dx \Rightarrow \uparrow dl$, but as $dl = I - I_0$, that is $dl \leq 0$, the increase of the positive value of thickness $dx$ results in the increase of the negative value $dl$.

Hence, one can write that: $dl \sim -dx$

**Second step.**

The change of intensity of photon stream (decrease) will depend on the value of the intensity itself: $dl \sim I$

For example, the twofold increase of light intensity will results in greater absorption of light by a substance. Combining the two conditions and changing the sign of proportionality into the coefficient of proportionality, we have got the equation: $dl = -kldx$

**Initial conditions**

Let’s analyze the similar case: if an obstacle is absent ($x = 0$), the intensity of light $I$ (passed through the obstacle) will be the same as the intensity of the initial $I_0$ (falling on the obstacle): $I = I_0$.

The solution of this DE is the function, that is the dependence $I = I(x)$. To solve the equation, that is to eliminate the differentials, it is necessary to integrate the left and right sides. However, it is necessary to separate the variables first. Variable $I$ should be on the left side, and variable $x$ — on the right side:

$$\frac{dl}{l} = -kdx$$

$$\int \frac{dy}{y} = -k \int dx$$

(In the given example the coefficient of proportionality is called as the coefficient of monochromatic light absorption by a substance. And according to the constant rule, it is replaced under the sign of integral)

$$ln I = -kx + ln C$$

$$ln I - ln C = -kx$$

$$ln \frac{l}{C} = -kx$$

$$\frac{l}{C} = e^{-kx}$$
We have received a general solution. Now, if we use the initial conditions: \( x = 0, I = I_0 \), we will receive constant \( C \) and will find a particular solution:

\[
I_0 = Ce^{-k_0}
\]

\[
I_0 = C.
\]

So, the final solution: \( I = I_0e^{-kx} \)

**PROOF!**

*The particular solution to a differential equation can be expressed as the solution to a definite integral.*

For example: \( \frac{dl}{I} = -kdx \)

\[
\int_{l_0}^{l} \frac{dl}{I} = -k \int_{0}^{x} dx
\]

\[
ln I - ln I_0 = -kl
\]

\[
ln \frac{I}{I_0} = -kl
\]

\[
\frac{I}{I_0} = e^{-kl}
\]

\[
I = I_0e^{-kl}
\]

1.4. **Probability theory**

*Probability theory* is the mathematical study of accidental (casual) events. It is the mathematical device intended for the quantitative description of casual events which at realization of a complex of conditions can occur or not occur. A quantitative measure of occurrence of casual event is the probability. The probability theory has an impotent rule for the analysis of medical and biologic processes as often their course depends from numerous, practically not considered random factors.

**Classical definition of probability**

It is admissible, some event \( A \) occurs in \( m \) cases from \( n \) possible. Then classical probability \( P \) of the event \( A \) is equal:

\[
P(A) = \frac{m}{n}.
\]  (1.4.1)
Example: let some box contains 20 black and 80 white spheres, then the probability of extraction at random of a black sphere from the box will make
\[ P = \frac{20}{100} = 0.2. \]

From definition and the resulted examples follows, that the classical probability of event accepts value between zero and unit:
\[ 0 \leq P(A) \leq 1. \]

If \( P(A) = 1 \), event \( A \) is authentic, if \( P(A) = 0 \), it is impossible. In other cases \( A \) is a casual event.

The methods of probability assist us in understanding randomness and therefore provide us tools for defining the measures of the unpredictability or uncertainty. The probability theory is a mathematical representation of random phenomena.

For example, a single card is picked from an ordinary 52 card deck. **What is the probability that we will pick an ace?**

\[ N = 52, \ m = \text{the accepted event is any ace in the deck, that is equal to 4.} \]

\[ P(A) = \frac{m}{N} = \frac{4}{52} \]

Another example, two fair dice are tossed simultaneously.

**Statistical definition of probability**

There is other approach to probability in math statistic, where the value \( \frac{m}{N} \) is a relative frequency of event \( A \), and we think of probability as a limit of relative frequency:

\[ P(A) = \lim_{N \to \infty} \frac{m}{N}, \quad \text{if} \quad N \to \infty. \]  

Thus, statistical probability of event is a limit to which relative frequency of event aspires at unlimited increase of the general number of tests.

The relative frequency approach to probability states that if an experiment is conducted a large number of times, the probability of that event will become stable and equal to the relative frequency.

The relative frequency is always a number between 0 and 1, as \( 0 \leq m \leq N \).

**For example**, a fair dice was tossed 100,000 times during an experiment. It was found that 50,097 times an even number (2, 4, 6) appeared. **What is the relative frequency \( f \) of tossing an even number?**

**Solution:**

\[ f(\text{even number}) = \frac{50097}{100000} = 0.50097, \quad \to \quad P = \lim f = 0.5. \]

1.4.1. **THE MAIN LAWS OF PROBABILITY THEORY**

There are 3 main kinds of random events:

- Disjoint events;
- Independent events;
• Dependent events;

1. **Disjoint events** — they cannot simultaneously occur, as appearance of one of them excludes appearance of another.

   **Definition:** Events $A_1, A_2, A_3, \ldots$ form the *full group of events* if at any tests there can be only one of them, and can’t be any other events.

   If events $A_1, A_2, A_3, \ldots$ form the *full group of events* the sum of their probabilities must be equal to one:

   \[
   \sum_{i=1}^{n} P(M_i) = 1. \tag{1.4.3}
   \]

2. **Independent events** — they can occur simultaneously and independently from each other.

3. **Dependent events** — the probability of appearance of one of them depends on the second event has occurred or not.

   — *What can we say about the laws of probability theory for those events?*

**1.4.2. ADDITIVE RULE FOR PROBABILITIES OF DISJOINT EVENTS**

Generally the theorem of addition of probabilities can be formulated as follows. Disjoint casual events $A$ and $B$ have an appearance probabilities $P(A)$, $P(B)$ and cannot occur simultaneously.

- The probability of realization of one of two disjoint events $A$ and $B$ (it is indifferent what event):

  \[
  P(A \text{ or } B) = P(A) + P(B). \quad (1.4.4)
  \]

- The probability of appearance of one (it is indifferent what) events from several disjoint events $A_1, A_2, A_3, \ldots A_n$ is equal to the sum of their probabilities:

  \[
  P(A_1 \text{ or } A_2 \text{ or } A_3 \text{ or } \ldots A_n) = P(A_1) + P(A_2) + P(A_3) + \ldots + P(A_n).
  \]

- If the events $A_1, A_2, A_3, \ldots .A_n$ form the *full group* of events the sum of their probabilities must be equal to one: $\sum_{i=1}^{n} P(A_i) = 1.$

**1.4.3. THE MULTIPLICATION RULE FOR PROBABILITIES OF INDEPENDENT EVENTS**

The probability of joint appearance of several *independent events* is equal to product of their probabilities:

*For 2 events:*

\[
P(A \text{ and } B) = P(A) \cdot P(B). \quad (1.4.5)
\]

*For many events:*

\[
P(A_1 \text{ and } A_2 \text{ and } A_3 \text{ and } \ldots A_n) = P(A_1)\cdot P(A_2)\cdot P(A_3) \cdot \ldots \cdot P(A_n).
\]

*For example*, a coin is tossed six times in succession. **What is the probability that at least one head occurs?**
Each toss can result in either head (H) or tails (T), their probabilities are equal: $P(H) = P(T) = \frac{1}{2}$. If A is the event when no head occurs, this can only happen in one way when all tosses result in tails:

$$P(A) = P(T \text{ 6 times}) = [P(T)]^6 = \left(\frac{1}{2}\right)^6 = \frac{1}{64}.$$  

$$P(\text{at least one head}) = P(E^C) = 1 - P(E) = 1 - \frac{1}{64} = \frac{63}{64}.$$  

The probability that at least one head occurs is $P = 1 - P(A) = 1 - \frac{1}{64} = \frac{63}{64}$.

### 1.4.4. The Multiplication Rule for Probabilities of the Dependent Events

Sometimes, if additional information about an event is available, the probability of the event changes.

The are two kinds of probabilities for dependent events A and B.

If event A appears the first, when event B has not occurred, its probability $P(A)$ is unconditional one. But if the event A appears when event B has occurred, its probability designates $P(A/B)$ and is named conditional probability. It is true for event B too. The theorem of multiplication of probabilities becomes complicated a little for dependent events.

So the probability of joint appearance of two dependent events A and B is equal to product of the unconditional probability of first event on conditional probability of another one:

$$P(A \text{ and } B) = P(A) \cdot P(B/A),$$  \hspace{1cm} (1.4.6)  

or

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

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

Comparing these two cases, we can find conditional probability of event A:

$$P(A/B) = \frac{P(A) \cdot P(B \mid A)}{P(B)}. \hspace{1cm} (1.4.7)$$  

Expansion of this formula on a case of many events $B_k$ forming full system of events $(B_k)$, is Bayes formula:

$$P(A/B_k) = \frac{P(A) \cdot P(B_k/A)}{\sum_{k=1}^{n} P(A) \cdot P(B_k/A)};$$  \hspace{1cm} (1.4.8)

This formula is very useful in many cases.
Example. Let’s consider that some diseases \( M_1, M_2, \ldots, M_n \) form the full system of events, so \( \sum_{i=1}^{n} P(M_i) = 1 \). If symptom \( S_j \) is present, the conditional probability \( P(M/S_j) \) of disease \( M \) is defined under Bayes formula:

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

– How do I find the probability of independent and dependent events?

Can you explain the difference?

First let’s talk a little about the sample space. The sample space is the set of all possible outcomes for an event or experiment. For example, the sample space of a die (one of a pair of dice) is six: \( S = \{1, 2, 3, 4, 5, 6\} \). Each number is what you would see on each side of the die. The sample space for a coin is two: \( S = \{H, T\} \), \( H \) for heads and \( T \) for tails. Now let’s see if we can understand the idea of an independent event first. Informally speaking, we say that two events \( A \) and \( B \) are independent if when one of them happens, it doesn't affect the other one happening or not. Let's use a real life example. Let's say that you have a coin and a die (one of a pair of dice). You want to find the probability of tossing the coin, getting heads one time, and then tossing the die and getting a five one time. We'll call the coin toss event \( A \). The plain old probablity of tossing a coin and getting heads is \( 1/2 \). It follows that:

\[
P(A) = \frac{\text{The number of favorable outcomes}}{\text{Total possible outcomes}} = \frac{1}{2}.
\]

The probability of getting a five when you toss the die will be event \( B \) and that is:

\[
P(B) = \frac{\text{The number of favorable outcomes}}{\text{Total possible outcomes}} = \frac{1}{6}.
\]

Now for the independent part. Does your chance of getting a five when you toss the die have anything to do with whether you get heads or tails when you toss the coin? It does not. That's why they are independent. The probability of independent events occurring is found by multiplying the probability of the first event occurring by the probability of the second event occurring. In general, it looks like this:

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

In our example it looks like this:

\[
P(H \text{ and } 5) = P(A) \cdot P(B) = 1/2 \cdot 1/6 = 1/12.
\]

Now for dependent events. A dependent event is one where the outcome of the second event is influenced by the outcome of the first event. For example, let's say we have a box with 6 marbles: 3 red, 1 blue, 1 green and 1 yellow. What's the probability of picking a yellow marble? We know that probability is 1/6. What's the probability of picking a blue marble? Can it also be 1/6? Well, it
could be if we put back the first marble we picked. But if we don't put back the first marble, our sample space will have changed. We started with six marbles, picked one, and now we only have five marbles in the sample space, so the probability of picking a blue marble is now 1/5. And in such a case we have dependent events, because something about the first one changed the second one. The probability of two dependent events occurring, one right after the other, is still found by using the same formula:

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

The big difference is that the individual probabilities won't have the same sample spaces. So from our example, what is the probability of picking a yellow marble and then a blue marble, without putting the first marble back?

- \[ P(\text{Yellow}) = 1/6, \]
- \[ P(\text{Blue/}} \text{Yellow}) = 1/5, \]
- \[ P(Y \text{ and } B) = 1/6 \cdot 1/5 = 1/30. \]

This is a very different number from what we would get if the events were independent, that is if the sample space remained the same because we put the first marble we picked back into the box. Then:

- \[ P(\text{Yellow}) = 1/6, \]
- \[ P(\text{Blue}) = 1/6, \]
- \[ P(Y \text{ and } B) = 1/6 \cdot 1/6 = 1/36. \]

**How is probability related to genetics?**

- Here's the basic idea: You have two copies of every gene. One of them came from your mom and one came from your dad. (Each of them also had two copies of each gene, but randomly gave one of each to you.) The two copies you have are not necessarily identical. Brown eye color is a good example. Everyone has a gene that says one of two things:
  1) “Make brown eyes”.
  2) “Don't make brown eyes”.

So if both copies of your brown eye gene say “Make brown eyes”, then your eyes will be brown. On the other hand, if both copies say “Don't make brown eyes”, then your eyes will be some other color (which other color depends on other genes).

- **What happens if one copy says “Make brown eyes” and the other copy says “Don't make brown eyes”?**

  You might think that you'd end up with something in between. But that's actually not what happens. What you end up with is just simply brown eyes. So if you have one copy of your brown eyes gene (say, the one you got from mom) saying “Make brown eyes” and another copy (the one from dad) saying “Don't make brown eyes” then your eyes will be brown. In fact, they will be just as brown as someone who has BOTH copies saying “Make brown eyes”. Biologists say that brown eyes are “dominant”. You can think of it like this.
The copy that says “Make brown eyes” is a really big, loud bully of a gene. Whenever it gets into a cell, it pushes aside anything else that's there and makes sure it gets its way. On the other hand, the copy that says “Don't make brown eyes” is a little, quiet gene that lets the “Make brown eyes” gene walk all over it. The only time it gets heard at all is if there are two copies of it and no one else around to bully it. As a convention, the two copies of a gene are written using letters. Capital letters stand for dominant genes, so the “Make brown eyes” copy would be written B, and lower case letters stand for “recessive” (not dominant) genes, so the “Don't make brown eyes” copy would be written b. Since you have two copies, you get two letters. So an individual who got a B from mom and a B from dad would be written BB, someone who got a b from both parents would be written bb, and someone who got a B from one parent and a b from the other would be written Bb. Notice that BB and Bb individuals BOTH look exactly the same (they both have brown eyes), but their genes are different so their children might look different. Also, keep in mind that the only individuals who do not have brown eyes are bb individuals.

**Is this related to mathematics?**

You were right that the big link is probability. If we know what genes parents have, we can figure out the probabilities that their children will have different genes. Let's say that there's a woman who has brown eyes and her genes are Bb (that is, she has one “Make brown eyes” copy and one “Don't make brown eyes” copy). And she marries a man with blue eyes (since he doesn't have brown eyes, the only possibility is that he is bb: that is, he has two copies of the “Don't make brown eyes” gene). What color eyes will their kids have? Remember, each kid will get one copy from mom and one copy from dad.

**So what could a child possibly get from his dad?**

Well, dad is bb and he has to give one or the other of his copies to his kid. But his two copies are identical, so all he really has is that he can possibly give is b. So the kid has a 100% chance of getting a b from dad. What about mom? Well, she is Bb, so she has two “choices” — she can give her kid either a B or a b. Since the sorting is random, there is an equal chance of either event happening, so the kid has a 50% chance of getting a B from mom and a 50% of getting a b from mom. So what are the possible outcomes? Well, we have to multiply the probabilities, so there is a 50% chance that the kid will get a B from mom, times a 100% chance of getting a b from dad, equals a 50% chance of the kid ending up Bb (with brown eyes). On the other hand, there is a 50% chance that the kid will get a b from mom, times a 100% chance of getting a b from dad, equals a 50% chance of the kid ending up bb (with eyes that aren't brown). Let's try one more example. Let's say there's another woman who has
brown eyes and her genes are Bb, and she marries a man who also has brown
eyes and whose genes are also Bb.

- **What color eyes will their kids have?**
  - Again, each kid gets one copy from mom and one copy from dad.

- **So what could a kid possibly get from his mom?**
  - Well, mom is Bb, so she has a 50% chance of giving her kid a B and
    a 50% chance of giving her kid a b.

- **What could a kid get from his dad?**
  Dad is also Bb, so he has a 50% chance of giving his kid a B and a 50%
  chance of giving his kid a b. So what will the kids have? Again, we have
to multiply the probabilities, so there is a 50% chance that the kid will get a B
from mom, times a 50% chance that the kid will get a B from dad, equals
a 25% chance that the kid will be BB. And there is a 50% chance that the kid
will get a b from mom, times a 50% chance that the kid will get a b from dad,
equals a 25% chance that the kid will be bb. And there is a 50% chance that
the kid will get a B from mom, times a 50% chance that the kid will get a B
from dad, equals a 25% chance that the kid will be Bb. And there is a 50%
chance that the kid will get a b from mom, times a 50% chance that the kid
will get a B from dad, equals a 25% chance that the kid will be Bb. So there are
actually two ways the kid can end up Bb: he or she can get a B from mom and
a b from dad, or get a b from mom and a B from dad. So the total probability of
the kid ending up Bb is 25% + 25%, which equals 50%. So there is a 25%
chance of being BB, a 50% chance of being Bb, and a 25% chance of being bb.
What color eyes will the kids have? Well, the 25% that are BB plus the 50%
that are Bb will all have brown eyes, so 75% of the kids will have brown eyes,
and 25% of the kids will be bb and have non-brown eyes. Wow! So it is
actually possible for two people with brown eyes to have a kid with non-brown
eyes! That can happen if both parents have a recessive b hiding in
the background. Try a few more marriages on your own. For example, what will
happen if a woman with green eyes (bb) marries a man who is BB (brown
eyes)? (Be careful when you do this that you keep in mind that one copy in
every child has to come from mom and one copy has to come from dad.)
So that's the link between math (probability) and genetics.

**1.4.5. Practical Tasks**

**Example 1.**

For example, suppose we have a jar with 4 red marbles and 6 blue. We want
to find the probability of drawing a red one at random. So our event is
“drawing a red marble”. The probability of this is:
number of red marbles (the chances of our event)/total marbles in jar
(the number of total chances)

In our example, this is 4/10 which is 2/5, reduced. So the probability of
drawing a red marble is 2/5. This is because all the outcomes are equally likely.
It follows that any individual marble has the same chance of being drawn.

– If we numbered all the marbles, what is the probability of picking
out # 5?
– Well, there is only 1 number “5” marble, and still 10 marbles in the jar,
so the answer is 1/10. Now suppose we have 2 events. Let's say that Niki is
going to draw 1 marble, and then Tom is going to draw one from the remaining
marbles.

– What is the probability that Niki gets a blue one? What is
the probability that Tom gets a red one?
– Again, we use our fraction. When Niki draws, there are 10 marbles in
the jar, of which 6 are blue, so her probability of drawing a blue is 6/10 or 3/5.
After she draws, it is Tom's turn. But now there are only 9 marbles left. 4 of
these are red, so his probability of drawing a red marble is 4/9. Now, it is
important to distinguish in any probability problem how many events you have.
Here we have figured the probability for two events. The first is that Niki draws
a blue marble. The second is that Tom draws a red one after Niki has drawn.
But, suppose we want to know the probability of the one event: “Niki draws
a blue marble and Tom draws a red one”. It seems like the same question, but it
isn't. The reason is that now we have more than one way this could happen.
These are the only 4 possibilities:

(1) Niki draws a blue, then Tom draws a blue
(2) Niki draws a blue, then Tom draws a red
(3) Niki draws a red, then Tom draws a blue
(4) Niki draws a red, then Tom draws a red

They are not all equally likely, however.

So, the probability of (2) above, is:

\[
P(2) = \frac{3}{5} \cdot \frac{4}{9} = \frac{4}{15}.
\]

– How about the probability of (1) possibility?
– Well, we already figured the probability of Niki drawing blue; it's 3/5.
How about Tom drawing blue also? Well, after Niki draws blue, there are
9 marbles left, and 5 blue, so its 5/9.

\[
P(1) = \frac{3}{5} \cdot \frac{5}{9} = \frac{1}{3}.
\]
Example 2.

- What is the probability that of any seven persons exactly one was born on Monday?

1) Let’s begin solving a simpler problem. Let’s suppose you just want the first person to be born on Monday and the other six to be born on any other day. Then the probability of this happening is equal to the probability of being born on Monday \( \left( \frac{52}{365} = 0.14 \right) \) times the probability of “not” being born on Monday (six times for the other six persons):

\[
0.14 \cdot (1-0.14)^6 = 0.14 \cdot 0.40 = 0.0566.
\]

But if you want any one of the seven persons (not just the first one) to be born on a Monday, then you want also the second, or third or ... person to be born on a Monday (and the others on any other day). So you have a total of seven possibilities each with the same probability just calculated. So the total probability is \( 7 \cdot 0.14 \cdot (1-0.14)^6 = 7 \cdot 0.0566 = 0.39647 = 0.40 \), and the answer is 40 percent.

1.5. Random variables

In the mathematician the size is the general name of various quantitative characteristics of subjects and the phenomena. Length, the area, temperature, pressure, etc. — examples of various sizes. The size which numerical values depends on casual circumstances, refers to as a random variable. Examples of random variables: 1) number of patients on reception at the doctor, 2) growth and weight of the person, etc.

Random variables happen discrete and continuous.

The discrete random variable accepts only the certain values separated from each other which can be established and listed. Examples:

1) the number of students in an audience — can be only the whole positive number: 0, 1, 2, 3, 4 ... 20 ...

2) relative frequency of hit in the purpose at 10 shots — its values: 0; 0.1; 0.2; 0.3 .... 1.

Continuous random variable can accept any values inside of some interval. For example, an atmospheric pressure, temperature of air, weight and growth of people, the sizes of blood elements, etc. concern to continuous random variables.

1.5.1. The low of distribution of the discrete random variables
To have full representation about casual discrete size, it is necessary to list all its possible values and to specify probabilities corresponding them. Conformity between values of a discrete random variable and their probabilities refers to as the law of distribution of this size. The law of distribution of a discrete random variable can be set in three ways: in the form of the table, the schedule or the formula. Thus possible values of random variable $X$ can be designated through $x_i$, and corresponding them probabilities — through $p_i$.

1. In the table, possible values of discrete random variable $X$ and probabilities $P$ corresponding these values $(X)$ are listed all:

$$
\begin{array}{cccccccc}
X & x_1 & x_2 & \ldots & x_i & \ldots & x_n \\
P(X) & p_1 & p_2 & \ldots & p_i & \ldots & p_n \\
\end{array}
$$

thus the normalization condition necessarily should be satisfied: the sum of all probabilities $p_i$ should be equal to unit

$$
\sum_{i=1}^{n} p_i = p_1 + p_2 + \ldots + p_n = 1. \quad (1.5.1)
$$

Graphically distribution of a random variable is set by a broken line which can be named a polygon of distribution (fig. 1.5.1).

![Fig. 1.5.1](image)

Here on a horizontal axis postpone all possible values of a random variable $X_i$, and on a vertical axis — probabilities corresponding them $P_i$. The broken line connecting received points also is a polygon of distribution.

Analytically distribution of a random variable is set in the form of the formula. For example, if the probability of hit in the purpose at one shot is equal $p$ the probability of a miss at one shot is equal $q = 1 - p$, and the probability of defeat of the purpose only once at $n$ shots is given by formula $P(n) = q^{n-1} \cdot p$. 

35
1.5.2. THE DISTRIBUTION LAW OF THE CONTINUOUS RANDOM VARIABLES.

DENSITY OF PROBABILITY DISTRIBUTION

It is impossible to make the same table for continuous random variables, because of infinite a plenty of possible values of a random variable. Besides the probability of its any concrete value is very small (is close to 0), and various intervals of its values usually are not equally probable. Let possible values of continuous random variable $X$ entirely fill some interval $[a, b]$ (fig. 1.5.2).

![Fig. 1.5.2](image)

Then small probability $dp$ that random variable $X$ will accept any value from a small interval $(x, x + \Delta x)$, will be proportional to size of this interval $\Delta x$: $dp \sim \Delta x$, or, entering factor of proportionality $f$, which itself can depend from $x$, we receive:

$$dp = f(x) \cdot \Delta x.$$  \hfill (1.5.2)

The function $f(x)$ entered here refers to as density of probability distribution of random variable $X$ or, simply, density of probability (density of distribution). The equation (1.5.2) can be considered as the differential equation. If we’ll integrate it within the limits of from $x_1$ up to $x_2$, we shall find probability of random variable $X$ hit in this interval $(x_1, x_2)$:

$$P(x_1 < X < x_2) = \int_{x_1}^{x_2} f(x)dx.$$  \hfill (1.5.3)

Graphically this probability $P(x_1 < X < x_2)$ is equal to the area of the curvilinear trapeze limited by an axis absciss, a curve $f(x)$ and the ordinates lead in points $x_1$ and $x_2$ (see fig. 1.5.3) as follows from geometrical sense of the certain integral.

The curve $f(x)$ thus refers to as a curve of distribution. Function of probability distribution $f(x)$ completely defines the distribution law of continuous random variables because knowing it, it is possible to find probability of hit of random variable $X$ in any interesting interval.

The condition

$$\int_{a}^{b} f(x)dx = 1, \quad \int_{-\infty}^{+\infty} f(x)dx = 1,$$  \hfill (1.5.4)

must be occur for the probability distribution density $f(x)$.

From geometrical sense of the certain integrals (1.5.4) follows the area under the distribution curve $f(x)$ and an axis OX is always equal to 1.
1.5.3. NUMERICAL CHARACTERISTICS OF RANDOM VARIABLE DISTRIBUTIONS

Full representation about distribution of a discrete or continuous random variable is given with their laws of distribution.

To express in the compressed form the most essential features of distributions one use numerical characteristics (parameters) of distributions of random variables. These parameters are concrete numbers which can be found from the distribution law or from the received experimental data.

The basic types are characteristics of random variable position and characteristics of its dispersion

1. Characteristics of position:
   • The Population mean \( M(X) \) of a random variable is calculated under formulas:

   for a discrete random variable
   \[
   M(X) = x_1p_1 + x_2p_2 + \ldots + x_np_n = \sum_{i=1}^{n} x_i p_i = \overline{X}, \quad (1.5.5)
   \]

   and for a continuous random variable
   \[
   M(X) = \int_a^b x \cdot f(x)dx, \quad \text{or} \quad M(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx, \quad (1.5.6)
   \]

   We shall note, that the population mean of a random variables is a limit to which to aspire average arithmetic value of a random variable at unlimited increase of number of experiences. Therefore the term “population mean” is convenient to be replaced by more clear term “average arithmetic value”.

   Example. Calculate the average arithmetic value of a continuous random variable, uniformly distributed on a piece \([a, b]\).

   The decision: The probability density of uniform distribution on an interval \([a, b]\) is constant, i. e. \( f(x) = f_0 = \text{const} \), and it is equal to zero outside of \([a, b]\). From the condition (1.4.3) we shall find value \( f_0 \):

   \[
   1 = \int_a^b f_0 dx = f_0 \int_a^b dx = f_0 \cdot x \bigg|_a^b = (b-a)f_0, \quad \text{hence} \quad f_0 = \frac{1}{b-a}.
   \]

   Therefore

   \[
   M(X) = \int_a^b \frac{x}{b-a} dx = \frac{1}{b-a} \int_a^b x dx = \frac{1}{2(b-a)} \cdot x^2 \bigg|_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{1}{2} (a+b).
   \]

   Thus, in this case the population mean \( M(X) \), defining an average arithmetic value \(<X>\) of a random variable, coincides with the middle of an interval \([a, b]\), i. e.

   \[
   <X> = M(X) = \frac{1}{2} (a+b).
   \]
- **Fashion Mo(\(X\))** of discrete random variable is a variable value whose probability has maximal value, a fashion of continuous random variable is value \(X\) at which the probability density is maximal (fig. 1.5.3).

- **Median Me(\(X\))** of distributions is the following characteristic of position. \(Me(\(X\))\) of a random variable distributions is such value \(X\) which divides all distribution to two equiprobable parts. In other words for a random variables the probability to accept values less \(Me(\(X\))\) or more \(Me(\(X\))\) is equal:

\[
P(\(X < Me\)) = P(\(X > Me\)) = \frac{1}{2}.
\]

Therefore the median can be calculated from a parity:

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

Graphically the median is a value of a random variable which ordinate divides the area limited by a distribution curve, half-and-half (\(S1 = S2\)) (fig. 1.5.3).

In case of discrete sizes it is necessary do the rising set of data and then a variant \(X\), average under number in this set will be a median \(Me(\(X\))\).

If the \(M(\(X\)), Mo(\(X\))\) and \(Me(\(X\))\) coincide, the distribution of a random variable names **symmetric**, otherwise — **asymmetric**.

2. **Characteristics of dispersion** are a dispersion and a standard deviation. (an average square-law deviation)

- **Dispersion D(\(X\))** of random variable \(X\) is defined as a population mean of a square of a deviation casual \(X\) from its population mean \(M(\(X\))\):

\[
D(\(X\)) = M[\(X - M - (X)\]^2, \tag{1.5.8}
\]

or as

\[
D(\(X\)) = M(\(X^2\)) - [M(\(X\))]^2.
\]

So for a discrete random variable it is calculated under formulas:

\[
D(\(X\)) = \sum_{i=1}^{N} [x_i - M(\(X\))]^2 p_i, \quad \text{or} \quad D(\(X\)) = \sum_{i=1}^{N} x_i^2 p_i - [M(\(X\))]^2. \tag{1.5.9}
\]

and for the continuous variable distributed in an interval \([a, b]\):

\[
D(\(X\)) = \int_a^b [x - M(\(X\))]^2 f(x)dx, \quad \text{or} \quad D(\(X\)) = \int_a^b x^2 f(x)dx - [M(\(X\))]^2. \tag{1.5.10}
\]
The dispersion characterizes average scattering of values of $X$ concerning its population mean. The word “dispersion” means “scattering”. But dispersion $D(X)$ has dimension of a square of a random variable $X$ that is rather inconvenient at an estimation in physical, biological, medical, etc. appendices. Therefore usually use other parameter, which dimension coincides with $X$ dimension — an average square-law (differently — standard) deviation.

- This **average square-law (differently — standard) a deviation** of random variable $X$ which designate $\sigma(X)$:

$$\sigma(X) = \sqrt{D(X)} \quad (1.5.11)$$

So, the population mean, the fashion, the median, the dispersion and the average square-law deviation are most often used numerical characteristics of distributions of random variables. Each of them expresses the certain property of the distribution.

### 1.5.4. The Normal Law of Random Variable Distribution

Strictly speaking, each random variable has the owner law of distribution. But it is established, that at all these laws are reduced only to several basic types (the binomial law, Poisson’s law, Gaussian law). Exclusively important role in probability theory is played the normal law of distribution (Gaussian law).

First, the normal law of distribution meets in practice most often. Secondly, it is the limiting law in the sense that the other laws of distribution converge to it under certain conditions.

![Fig. 1.5.4](image)

The normal law of distribution of probability density of a random variable $X$ is given by the formula:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-M(X))^2}{2\sigma^2}} \quad (1.5.12)$$

Here $x$ — current values of random variable $X$. A population mean $M(X)$ and a standard deviation $\sigma(X)$ completely define this function $f(x)$.
So if the random variable is distributed under the normal law it is necessary to know only two numerical parameters $M(X)$ and $\sigma(X)$ that completely define its distribution law (1.4.10). The function (1.4.10) refers to as a normal curve of distribution (Gaussian curve). It is a symmetric curve concerning a population mean, i.e. concerning ordinate $x = M(X)$. The maximal density of probability equal to $\frac{1}{\sqrt{2\pi}}$ and corresponds to a population mean $M(X)$. As $X$ removes from $M(X)$, the probability density decreases and gradually comes to zero (fig. 1.5.4). A root-mean-square deviation $\sigma(X)$ characterizes average size of scattering and the width of a distribution curve (see fig. 1.5.5). It is natural, that at any values of $M(X)$ and $\sigma(X)$ the area under a normal curve is equal to 1 (a condition (1.5.4):

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

A fashion and a median of this distribution coincide with $M(X)$, therefore the normal distribution is symmetric: $M(X) = Mo(X) = Me(X)$.

If value $M(X)$ changes the form of a normal curve does not vary, and the curve shifts along an axis OX only. But with changing of size of a root-mean-square deviation $\sigma(X)$ the form of a curve of distribution changes: with $\sigma(X)$ increasing value of a curve maximum decreases and the curve becomes more flat, being stretched along an axis OX. On the contrary at reduction $\sigma(X)$ the maximum of a curve grows and the curve is narrowed (fig. 1.4.5).

1.5.5. THE NORMAL DISTRIBUTION’S PROPERTIES

It is known the probabilities of size $X$ hits in certain intervals are equal (fig. 1.5.6):

\[
P(M(X) - \sigma < X < M(X) + \sigma) = 0,6827 = 68,27 \%.
\]

\[
P(M(X) - 2\sigma < X < M(X) + 2\sigma) = 0,9545 = 95,45 \%.
\]
From (1.5.14) follows: if random variable $X$ is distributed under the normal law, practically all (99.73%), its values lay in an interval of $[\mu(X) - 3\sigma, \mu(X) + 3\sigma]$.

This way of an estimation of possible values $X$ is known as “a rule of three standard deviation $\sigma$”.

Example. It is known the person pH blood is normally distributed size with average value (population mean) 7.4 and a standard deviation 0.2. Define a range of pH values.

The decision: We shall take advantage “of a rule of three standard deviation $\sigma$”. 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.

1.5.6. PROBLEMS TO PRACTICAL EMPLOYMENT

1. Dialing the number of phone, the subscriber has forgotten one figure and has typed it at random. Find probability He is typed necessary one.

2. At test of a party of devices relative frequency of suitable devices has appeared equal 0.9. Find number of suitable gauges if all 200 pieces have been checked up.

3. About one child from 700 is born with a Down syndrome. In hospital of the big city 2500 children are born in a year. Define expected number of newborns with a Down syndrome.

4. Operation on skin transplantation leads to success in 40% of all cases. What probability the skin transplantation will successful at the fourth attempt only?

5. Discrete random variable $X$ has two possible values only: $X_1$ and $X_2$. The probability of the $X_1$ occurring is $p_1 = 0.6$. Write down the law of distribution of this size.

6. The distribution law of discrete random variable $X$ is set:
Construct a polygon of distribution. Find a population mean of $\overline{M(X)}$, an average arithmetic, fashion $\overline{M_0}$, dispersion $D(X)$ and a standard deviation for this random variable.

7. Distribution of continuous random variable $X$ is described by density of probability $f(x) = a \cdot \sin x$ in an interval $(0, \pi/3)$ and $f(x) = 0$ outside of this interval. Find probability that $X$ will accept value in an interval $(\pi/6, \pi/4)$.

8. The distribution density of Random variable $X$ is $f(x) = 2x$ in an interval $(0,1)$; outside of this interval $f(x) = 0$. Find $M(X)$ and $\sigma(X)$.

9. Normally distributed random variable $X$ is set by density of probability $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. Find a population mean and dispersion $X$.

### 1.6. The elements of math statistics

**Deviation** is the difference of a quantity from some fixed value, usually the “correct” or “expected” one.

**Mean**

A mean is a *homogeneous function* that has the property that a mean $\mu$ of a set of numbers $x_i$ satisfies $\min(x_1, \ldots, x_n) \leq \mu \leq \max(x_1, \ldots, x_n)$

There are several statistical quantities called means, e. g., arithmetic-geometric mean, geometric mean, harmonic mean, quadratic mean, root-mean-square. The quantity commonly referred to as “the” mean is the arithmetic mean, also called the average. The mean (or, more specifically, the population mean) of a probability function $P(x)$ for a continuous distribution defined by

$$m = \int xP(x)dx$$

where the integral is taken over the domain of $P(x)$, and for a discrete distribution, is given by the sum

$$m = \sum x_i P(x_i)$$

**Absolute deviation**

Let $\mu$ denote the mean of a set of quantities $U_i$, then the absolute deviation is defined by

$$D_{u_i} = |u_i - \mu|.$$
Relative error

Let the true value of a quantity be \( x \) and the measured or inferred value \( x_0 \). Then the relative error is defined by

\[
\Delta x = \frac{x_0 - x}{x} = \frac{x_0}{x} - 1,
\]

(1.6.4)

where \( \Delta x \) is the absolute error. The relative error of the quotient or product of a number of quantities is less than or equal to the sum of their relative errors. The percentage error is 100 % times the relative error.

Statistical dispersion

\((\Delta u)^2_i = (u_i - \bar{u})^2\), where \( \bar{u} \) is the average of \( \{u_i\} \).

There are 2 ways to calculate the dispersion:

\[
D(x) = m(x - m)^2 = \sum_{i=1}^{n} (x_i - m)^2 p_i = \sum_{i=1}^{n} (x_i - m)^2 m_i/n
\]

\[
D(x) = m(x^2) - m^2
\]

(1.6.5)

For example, the experience of tossing a fair dice. It's necessary to find:

- Population mean \( \mu \);
- Dispersion \( D(x) \);

In this case: \( N = 6 \), \( p_i = \frac{1}{6} \)

1) \( m = \sum_{i=1}^{N} x_i p_i = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = \frac{21}{6} = 3.5 \)

2) \( D(x) = \sum_{i=1}^{N} (x_i - m)^2 p_i = \frac{1}{6}((1-3.5)^2 + (2-3.5)^2 + (3-3.5)^2 + ... + (6-3.5)^2) \)

But, \( (1-3.5)^2 = (6 - 3.5)^2 \); \( (2-3.5)^2 = (5 - 3.5)^2 \); \( (3-3.5)^2 = (4 - 3.5)^2 \)

So, we obtain:

\( D(x) = \frac{1}{6}((1-3.5)^2 2 + (2-3.5)^2 2 + (3-3.5)^2 2) = \frac{1}{3}(6,25 + 2,25 + 0,25) = \frac{8,75}{3} = 2.91 \)

If \( N \leq 30 \), we must multiplicate the dispersion (from a sample of \( N \) elements) by the so-called estimator defined by \( \frac{N}{N-1} \):

\[
D(x)_{\text{est}} = \sum_{i=1}^{N} (x_i - m)^2 p_i \frac{N}{N-1} = \frac{\sum_{i=1}^{N} (x_i - m)^2 m_i}{N} \cdot \frac{N}{N-1} = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - m)^2 m_i
\]

So, \( D(x)_{\text{est}} = 2.91 * \frac{6}{5} = 3.5 \)

Standard deviation
The standard deviation $\sigma$ of a probability distribution is defined as the square root of the variance $\sigma^2$,
\[ y = \sqrt{\langle x^2 \rangle - \langle x \rangle^2} = \sqrt{M_2 - M^2}, \]
where $M = \bar{x} = \langle x \rangle$ is the mean, $M_2 = \langle x^2 \rangle$ is the second raw moment, and $\langle f \rangle$ denotes an expectation value. The variance $\sigma^2$ is therefore equal to the second central moment (i.e., moment about the mean), $\sigma^2 = \mu_2$.

\[ S(x) = \sqrt{D(x)} \] (1.6.7)

In our case: $S(x) = \sqrt{3.5}$.

The sample standard deviation distribution is a slightly complicated, though well-studied and well-understood, function.

However, consistent with widespread inconsistent and ambiguous terminology, the square root of the bias-corrected variance is sometimes also known as the standard deviation,

\[ S_{N-1} = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2} \] (1.6.8)

\[ S_x = \frac{S_{N-1}}{\sqrt{N}} = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - M)^2 m_i} = \sqrt{\frac{\sum_{i=1}^{N} (x_i - M)^2 m_i}{N(N-1)}} \] (1.6.9)

\[ S_x = \sqrt{\frac{3.5}{6}} = \sqrt{0.58} \approx 0.7 \]

**Discrete distribution**

Discrete distribution includes the following types of the distribution: Bernoulli Distribution, Binomial Distribution, Poisson Distribution.

A discrete distribution with probability function $P(x_k)$ defined over $k = 1, 2, ..., N$ has distribution function.

\[ D(x_k) = \sum_{k=1}^{N} P(x_k) \] (1.6.10)

and population mean.

\[ M = \frac{1}{N} \sum_{k=1}^{N} P(x_k). \] (1.6.11)

The square root of the sample variance of a set of $N$ values is the sample standard deviation.
Normal Distribution

A normal distribution in a variate $X$ with mean $\mu$ and variance $\sigma^2$ has probability function

$$P(x) = \frac{1}{\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$$
on the domain $x \in (-\infty, \infty)$. While statisticians and mathematicians uniformly use the term “normal distribution” for this distribution, physicists sometimes call it a Gaussian distribution and, because of its curved flaring shape, social scientists refer it to the “bell curve”. The so-called “standard normal distribution” is given by taking $\mu = 0$ and $\sigma^2 = 1$ in a general normal distribution. An arbitrary normal distribution can be converted to a standard normal distribution by changing variables to $Z \equiv (X - \mu)/\sigma$, so $dz = dx/\sigma$, yielding

$$P(z)dz = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}dz.$$
The normal distribution is the limiting case of a discrete binomial distribution \( P_p(n|N) \) as the sample size \( N \) becomes large, in which case \( P_p(n|N) \) is normal with mean and variance
\[
\mu = Np \\
\sigma^2 = Npq,
\]
with \( q = 1 - p \)

The distribution \( P(x) \) is properly normalized since
\[
\int_{-\infty}^{\infty} P(x)dx = 1. \quad (1.6.13)
\]

A distribution of values of a discrete variate represented graphically by plotting points \((x_1, f_1), (x_2, f_2), \ldots, (x_k, f_k)\), and drawing a set of straight line segments connecting adjacent points. It is usually preferable to use a histogram for grouped distributions.

**Student's t-distributions**

A statistical distribution was published by William Gosset in 1908. His employer, Guinness Breweries, required him to publish under a pseudonym, so he chose “Student”.

Given \( n \) independent measurements \( x_i \), let
\[
t = \frac{\bar{x} - \mu}{s / \sqrt{n}},
\]
where \( \mu \) is the population mean, \( \bar{x} \) is the sample mean, and \( s \) is the estimator for population standard deviation (i.e., the sample variance)

Student's t-distribution is defined as the distribution of the random variable \( t \) which is (very loosely) the “best” that we can do not knowing \( \sigma \).
If $\sigma = s$, $t = z$ and the distribution becomes the normal distribution. As $N$ increases, Student's $t$-distribution approaches the normal distribution.

**Confidential Interval**

A confidence interval is an interval in which a measurement or trial falls corresponding to a given probability. Usually, the confidence interval of interest is symmetrically placed around the mean, so a 50% confidence interval for a symmetric probability function would be the interval $[-a, a]$ such that

$$\frac{1}{2} = \int_{-a}^{a} P(x)dx.$$

For a normal distribution, the probability that a measurement falls within $n$ standard deviations ($n\sigma$) of the mean $\mu$ (i.e., within the interval $[\mu - n\sigma, \mu + n\sigma]$) is given by

$$P(\mu - ny < x < \mu + ny) = \frac{2}{\gamma \sqrt{2p}} \int_{-n\gamma}^{n\gamma} e^{-(x-\mu)^2/(2\gamma^2)} dx. \quad (1.6.14)$$

Now let $e = S_x t_{\gamma,n}$, so, obtain:

$$M - e \leq M \leq M + e$$

The coefficient of the Student $t_{\gamma,N}$ depends on:

- independent measurements
- confidence probability $\gamma$.

The so-called term “number of degrees of freedom” $f$ is very often used $f = N - 1$.

There are 3 prevalent values of the confidence probability: 0.95 %, 0.99 %, 0.999 %, where $\gamma$ — confidence probability.

For $N = 6$ $\Rightarrow$ $t_{0.95} = 2.571$. 

47
So, $3,5 - 0,7 \cdot 2,571 \leq M \leq 3,5 + 0,7 \cdot 2,571$
$1,7 \leq M \leq 5,3$
Sound and hearing

2.1. Harmonic oscillations

The equality of the oppositions defines harmonicity of the forms of transition of one opposition into another. The situation “the more the one – the less the other” means harmony: under certain condition it can take the form “everything is for one opposition – nothing is for the other”. If one opposition smoothly transfers into another, they will be equal! However, a single whole consisting of oppositions always remains constant.

Such situation can be demonstrated by means of fig. 2.1.1, in which the projections of the constant radius of a circumference on axis (OX, OY) will be the changing opposition.

If $\delta = 0 \Rightarrow \{ \cos \delta = 1(\max), \sin \delta = 0 \}$

$\uparrow \delta \Rightarrow \{ \downarrow \cos \delta, \uparrow \sin \delta \}$

If $\delta = 90^\circ \Rightarrow \{ \cos \delta = 0, \sin \delta = 1(\max) \}$

\[
\sin^2 x + \cos^2 x = 1
\]

Over the determined equal time periods and through the angular distance the situation is repeated.

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.

Frequency is a measurement of the number of cycles repeated per event in a period. To compute the frequency, one fixes a time interval, counts the number of occurrences of the event, and divides this count by the length of the time interval. The result is presented in units of hertz (Hz) after German physicist Heinrich Rudolph Hertz, where 1 Hz is an event that occurs once per second. Alternatively, one can measure the time between two occurrences of the event (the period) and then compute the frequency as the reciprocal of this time,
\[ v = \frac{1}{T}, \]  
\[ (2.1.1) \]

where \( T \) is the period.

In physics (specifically mechanics, \textit{angular frequency} \( \omega \) (sometimes called \textit{angular velocity}) is a measure of rotation rate, almost invariably given in units of radians per second, or simply \( s^{-1} \) since radians are dimensionless. One revolution is equal to \( 2\pi \) radians, hence

\[ \omega = 2\pi \nu = \frac{2\pi}{T}, \]
\[ (2.1.2) \]

where \( \nu \) is the \textit{frequency}.

The oscillation repeats every \( 2\pi/\omega \) seconds. In other words, there are \( \omega/2\pi \) oscillations per second.

The similar periodic situations in nature (and not only!) are repeated very often. For example, in mechanics, \textit{displacement} \( x \) (the coordinate) and \textit{velocity}, as well as kinetic and potential energy (being the sort of \textit{OPPOSITIONS}), form the whole \textit{constant unity}, as, in the case of \textit{radius} to circumferences.

\[ \vartheta = 0 \]
\[ x = X_{\text{max}} \]
\[ \vartheta = \vartheta_{\text{max}} \]
\[ x = X_{\text{max}} \]

\[ \Theta \]
\[ X \]

\textbf{Fig. 2.1.2. The simple harmonic oscillator}

The \textit{simple harmonic oscillator} (SHO) is a mass connected to some elastic object of negligible mass that is fixed at the other end and constrained so that it may only move in one dimension. This simplified model approximates many systems that vibrate or oscillate: drum heads, guitar strings, the quantum mechanical descriptions of an atom, etc.

Let's study the change of these parameters as a result of movement of a body (fig. 2.1.2). The \textit{first} stage will be considered to be the extreme \textit{right} position of the body, in which we have placed it primarily, having stretched the spring to the maximum distance (the amplitude) from the position of balance.

The further you move the object (larger \( x \)) the stronger is the force pulling it back (imagine a rubber band or a spring). If there wasn't a minus sign in front of \( k \), the force would be pulling the object still farther away, instead of back to
the origin (there are forces like that in nature, but they don't lead to oscillations). We can see the oscillations of displacement and velocity at the table below (fig. 2.1.3).

<table>
<thead>
<tr>
<th>№</th>
<th>$x_0$</th>
<th>$m \varrho_l^2$</th>
<th>$kx_0^2$</th>
<th>Essence of nature harmony</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>max</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1*</td>
<td>↑</td>
<td>↓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>max</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2*</td>
<td>↓</td>
<td>↑</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>max</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3*</td>
<td>↑</td>
<td>↓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>max</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4*</td>
<td>↓</td>
<td>↑</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>max</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sinx</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cosx</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 2.1.3. Displacement and velocity are two opposite things

Thus we can see that the fifth stage is equal to the first one. It means we’ve finished a complete cyclic action. Over the determined equal period of time and through angular distance the situation is repeated.

Then we can see the fluent transfer from one opposition (the displacement $x$) into another (the velocity). At that moment when, for example, the velocity is maximum, the displacement is zero, and on the contrary.

Such relations between the displacement $x$ and velocity refer to the relation between $\cos x$ and $\sin x$: $\cos x$ function leads to $\sin x$ by $90^\circ$ — when $\cos x$ is maximum $\sin x$ is equal to zero.

When the coordinate begins to increase, the velocity decreases. However, the full mechanical energy of a pendulum (the friction is absent) is not changed:

$$\frac{m}{2} \varrho_l^2 + \frac{k}{2} x^2 = \text{const}$$

(2.1.3)

where $x$ — displacement, $k$ — spring constant, $m$ — mass, $\varrho_l$ — velocity.

If we move a piece of paper upwards, toward the perpendicular movement of the pendulum, we will get $\sinusoid$ on it (fig. 2.1.4).
“There are places on the land, where the spirit dies and will be born the truth as its direct negation”.

The words of A. Kamu can be referred to the harmonic oscillation. The absence of the spirit means the presence of the truth, the notions of “spirit” and “truth” are the oppositions, according to the words mentioned above.

In fig. 2.1.5 two sinusoids are represented, which do not coincide with each other. They differ by the mutual location — an offset on value, which is equal to \( \frac{\pi}{2} \) radian, or 90°. The definitions by some sinusoids and cosinusoids are explained in this way.

\[
sin(\delta + \frac{\pi}{2}) = \cos \delta
\]

The angle at 90° always separates two oppositions! For example, the hot and cool can be represented in the following way (fig. 2.1.6).

When dissipative forces such as friction and air resistance are ignored, the net force will be directly proportional to the displacement of mass from the equilibrium position of a system and it will point in the opposite direction (a condition known as Hooke's law).
In case of mass hanging on a spring, Newton's Laws, combined with Hooke's law for the behavior of a spring, states the following: \( m \ddot{x} = \vec{F} \), or in projection on horizontal axis \( x \): \( m \ddot{x} = -kx \), where \( k \) is the spring constant, \( m \) is the mass, \( x \) is the position of the mass, and \( a \) is its acceleration.

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

\[
\dot{v} = \frac{dx}{dt}
\]

The acceleration, or the rate of change of the velocity, is

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

So, we've obtained the second order linear differential equation. And its solution gives us the displacement of the mass as a function of time. This DE describes the free harmonic oscillation:

\[
m \frac{d^2 x}{dt^2} = -kx
\]

\[
\downarrow
\]

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

\[
\downarrow
\]

\[
\frac{d^2 x}{dt^2} + \omega_0^2 x = 0
\]

(2.1.4)

where

\[
\omega_0^2 = \frac{k}{m}.
\]

(2.1.5)

The solution of this equation is harmonic function of type:

\[
x = A \sin(\omega_0 t + \varphi),
\]

(2.1.6)

or

\[
x = A \cos(\omega_0 t + \varphi).
\]

So, it is the general solution of the free harmonic oscillations, where \( A \) is the amplitude, \( \omega_0 \) — the frequency and \( \varphi \) — initial phase of oscillation, \( t \) — running time.

The choice between \( \sin x \) and \( \cos x \) functions as the solutions of DE depend on the initial conditions of the oscillatory process. If a pendulum begins the oscillations with \( x(0) = A \) (from the maximum amplitude position), one should choose the function of the \( \cos x \), because \( \cos 0 = 1 \). In case of the beginning of oscillations of a pendulum from the position of equilibrium \( (x(0) = 0, \ A = 0) \), one should choose \( \sin x \) dependency, because \( \sin 0 = 0 \). Thus, the choice of the needed function depends upon the initial phase of oscillation:

\[
sin \omega_0 t = \cos \left( \omega_0 t - \frac{\varphi}{2} \right).
\]
We have studied that displacement leads velocity by 90°. It means the following:

\[ x = X_{\text{max}} \cos \omega_0 t \]
\[ \theta = \theta_{\text{max}} \sin \omega_0 t \]

\( A \left( X_{\text{max}}, V_{\text{max}} \right) \) is called the amplitude and \( \omega_0 t \) is called the phase \( \varphi \), \( \omega_0 \) — cyclic frequency of an oscillator. The general form of equation of oscillation is the following:

\[ x = A \cos(\omega_0 t + \varphi) . \]

2.2. Damped harmonic oscillation

Let us find the law, according to which the coordinate changes in the course of time, when a body makes the oscillations overcoming the power of friction.

For this reason, it is necessary to form and solve the equation, where the resultant of all forces (according to Newton's second law) is a single force acting along the horizontal axis — the force of friction.

– In physics friction is the resistive force, the physical deformation and the heat buildup that occurs when two surfaces travel along each other whilst forced together.

The force of friction is always directed opposite a moving body, that is to say: \( F_{fr} \sim \theta \Rightarrow F_{fr} = -r \theta \), where \( r \) — the coefficient of proportionality.

Newton's second law relates to the mass and velocity of a particle to a vector quantity known as the force. Suppose \( m \) is the mass of a particle and \( F \) is the vector sum of all applied forces (i. e. the net applied force). Beginning with Newton's second law of motion, we can derive an algebraic equation:

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

In the projections on horizontal axis OX along which an oscillatory process occurs, the equation will take the type:

\[ m \ddot{a} = -kx - r \theta \]

Or, in the differential form:

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

If \( \frac{r}{m} = 2\beta, \frac{k}{m} = \omega_0^2 \), obtain:

\[
\frac{d^2 x}{dt^2} + 2\beta \frac{dx}{dt} + \omega_0^2 x = 0 \tag{2.2.1}
\]

The general solution of his DE is the function:

\[
x = Ae^{-\beta t} \cos \omega t (t + \phi_0) \tag{2.2.2}
\]

\( \beta \) — damping factor; \( r \) — coefficient of drag; \( \omega \) — damping frequency

The given value shows how different the amplitudes of the oscillations are:

\[
T \approx \ln A e^\gamma = \ln e^\beta T \tag{2.2.3}
\]

\[
\lambda = b T \ln e = b T \tag{2.2.4}
\]

With damped harmonic oscillations the amplitude is changed. The period of the damped harmonic oscillations depend on the power of resistance of the ambience.

It is known, that the frequency of damped harmonic oscillations depends on the factor of fading:

\[ \omega^2 = \omega_0^2 - b^2 \] \tag{2.2.5}

If the period of the free harmonic oscillations is defined as \( T = \frac{2p}{\omega_0} \), then, for damped harmonic oscillations the period will be as follows:

\[ T = \frac{2p}{\sqrt{\omega_0^2 - b^2}}, \text{ or } \omega = \sqrt{\omega_0^2 - b^2}. \tag{2.2.6} \]

### 2.3. The driven harmonic oscillations

Another common mechanical problem arises when a damped harmonic oscillator is driven by some time-dependent external applied force: the *driven harmonic oscillator*. The most important case is that of a force that oscillates in a sinusoidal manner:

\[
F(t) = F_0 \cos(\omega t + \phi_0) \tag{2.3.1}
\]

\[
m \frac{d^2 x}{dt^2} + r \frac{dx}{dt} + kx = F_0 \cos(\omega t + \phi_0) \tag{2.3.2}
\]

The solution:
\[ x = Ae^{-\omega t} \cos(\omega_0 t + \varphi_0) + \frac{F_0 / m}{\sqrt{\left(\omega_0^2 - \omega^2\right)^2 + 4B^2 \omega^2}} \sin(\omega + \varphi_0 + \gamma) \]

\( F_0 = \) maximum driving force; \( \varphi_0 = \) driving phase; \( \omega = \) driving frequency

\[ \gamma = \tan^{-1} \frac{\omega^2 - \omega_0^2}{2B^2} \]

As the damping factor approaches zero, the steady state amplitude approaches the infinity. This illustrates the importance of damping in structures susceptible to vibration such as suspension bridges and steel framed buildings.

Without damping, a structure could shake itself to pieces from a tiny external force with just the right frequency.

\[ E_k = \frac{m \vartheta^2}{2} \]

\[ x = A \sin \omega_0 t \Rightarrow \vartheta = \frac{dx}{dt} \Rightarrow \vartheta = A \omega_0 \cos \omega_0 t \]

\[ E_k = \frac{m A^2 \omega_0^2}{2} \cos^2 \omega_0 t \]

\[ E_p = \frac{k \vartheta^2}{2} = \frac{k A^2}{2} \sin^2 \omega_0 t \]

\[ E = E_k + E_p = \frac{1}{2} m A^2 \omega_0^2 \left( \sin^2 \omega_0 t + \cos^2 \omega_0 t \right) = \frac{1}{2} m A^2 \omega_0^2 \] \quad (2.3.3)

The full energy of the free harmonic oscillator is in direct proportion to the square of the amplitude and to the square of frequency of the oscillations.

### 2.4. Characteristics of a wave. Sound

The wave is characterized by the transfer of energy without the transfer of matter. In case of propagation of a wave, we notice some effect of dominoes: each molecule transfers the impulse and energy to the neighbouring one, and the latter, in turn, transfers them to their neighbours. Thus, the molecules continue to oscillate around the center of their position in a substance. So, waves are a means of transferring energy, without actually transporting matter.

The basic characteristics of a wave are the length of a wave and speed of its propagation in a substance.

The interrelation of these characteristics to each other is easy to remember if to use the known formula for the definition of a displacement with the known values of speed and time.

\[ d = \vartheta t, \text{ where } d \text{ means "distance", } \vartheta \text{ — velocity, } t \text{ — time.} \]

The following correlation is true for a wave:

\[ \lambda = cT, \] \quad (2.4.1)
where \( \lambda \) — wavelength, \( c \) — speed of light in vacuum, \( T \) — period of oscillations.

Or:

\[
\lambda = \frac{c}{f},
\]

(2.4.2)

where \( f \) — frequency.

The waves are defined to the *longitudinal* and *transverse* waves. In a longitudinal wave the particle displacement is parallel to the direction of wave propagation.

In a *transverse* wave the particle displacement is perpendicular to the direction of wave propagation. The particles do not move along with the wave; they simply oscillate up and down about their individual equilibrium positions as the wave passes by. Pick a single particle and watch its motion.

*Sound* is defined as mechanical compression or longitudinal waves that propagate through a medium (solid, liquid or gas). Most sounds are combinations of signals, but a theoretically pure sound can be described as having a given speed of oscillation or frequency measured in hertz (Hz). The range of sound audible to the human ear falls roughly between 20 Hz and 20 kHz at typical amplitudes with wide variations in response curves. Above and below this range are ultrasound and infrasound, respectively.

Humans and several animals perceive sounds with their ears, but low frequency sounds can also be felt by other parts of the body. Sounds are used in several ways, most notably for communication through speech or, for example, music. Sound perception can also be used for acquiring information about the surrounding environment in properties such as spatial characteristic and presence of other animals or objects. For example, bats use one sort of echo-location for flying. Ships and submarines use sonar. Humans acquire and use spatial information perceived in sounds.

The perceived “amount” of sound (energy of pressure wave) is defined as loudness and measured in sones. The human ear is most sensitive to sound in the middle of the audible frequency range.

The hearing range varies between different animals: bats have an unusually large one, while frogs have a narrow one. Typical frequencies (in hertz) are listed below:

- Bat: 100–100,000 Hz
- Dog: 10–35,000 Hz
- Elephant: 1–20,000 Hz
- Frog: 100–2,500 Hz
- Human: 20–20,000 Hz

It is reasonable to state that a sound wave in the air and in the other surroundings (for instance, in the endolymph) propagates with the different velocity. It means that this wave influences the surroundings with the different pressure.
Then, the energy of a wave must be decreased with its transfer from the middle ear (in the air) into the inner one (in the liquid). But in the middle ear a special system (in the form of the lever) of the ossicles (the malleus (hammer), the incus (anvil), and the stapes (stirrup)) acts what creates an additive force, i.e. an additional pressure on the perilymph.

![Diagram of the ear](image)

**Fig. 2.4.1**

The ossicles form a chain of levers that links the external ear to the inner ear. They are composed of compact bone and joined to each other by synovial joints. The ossicles are covered by epithelium that is continuous with the epithelium of the tympanic cavity. The malleus is attached to the inner aspect of the tympanic membrane. The stapes is attached to the oval window of the inner ear by a fibrous ligament. The incus joins the malleus and the stapes. The three bones are associated with 2 small skeletal muscles, the Tensor Tympani (attached to the malleus), and the Stapedius (attached to the stapes). These muscles adjust the tension of the tympanic membrane in order to convey vibrations to the inner ear. During very loud noises, they also serve to protect the delicate receptors of the inner ear.

Here is the structure of this lever.
Fig. 2.4.2. The special system of the middle ear in the form of the lever

The *malleus* is connected with the *tympanic membrane* (TM), and the *stapes* with the membrane of the *oval window* (OW) (see fig. 2.4.1, 2.4.3). It follows that on the *tympanic membrane* (TM) the pressure of a sound in the air $p_1 = \frac{F_1}{A_1}$ acts, and on the OW — the pressure $p_2 = \frac{F_2}{A_2}$. So, for finding the ratio of the pressure $\frac{p_2}{p_1} = \frac{F_2}{A_2} \cdot \frac{A_1}{F_1} \Rightarrow \frac{p_2}{p_1} = \frac{F_2}{F_1} \cdot \frac{A_1}{A_2}$,

$\frac{A_1}{A_2} = \frac{64mm^2}{3mm^2} \approx 22\text{(times)}$

According to the law of the equilibrium of a lever, the multiplication of force $F$ by the corresponding arm $l$ (if force is perpendicular to arm) is constant, i. e.:

$\frac{F_1}{F_2} = \frac{l_2}{l_1}$

So, $l_1 > l_2$ is as much as $F_1 < F_2$.

Thus, in this case:

$\frac{F_2}{F_1} = \frac{l_1}{l_2} \approx 1,3$

Then:

$\frac{p_2}{p_1} = \frac{F_2}{F_1} \cdot \frac{A_1}{A_2} = 1,3 \cdot 22 \approx 26$

So, the lever in the middle ear increases the pressure 26 fold!

Let us study the way of sound signal propagation in an internal ear, and its structure.

The internal ear consists of 2 channels — *scala vestibuli* (SV) and *tympanic canal of cochlea* (TC) which are filled by *perilymph* (PL) and are connected with each other — *helicotrema* (H).

*Endolymph* (EL) is between these two channels. The sound signal reaches PL and contributes to the basilar membrane deformation in the following way (fig. 2.4.3).

59
Now let us study the structure of the inner ear.

The inner ear can be thought of as two organs: the semicircular canals which serve as the body's balance organ and the cochlea which serves as the body's microphone, converting sound pressure impulses from the outer ear into electrical impulses which are passed on to the brain via the auditory nerve.

The basilar membrane of the inner ear plays a critical role in the perception of pitch according to the place theory.
The organ of Corti is the sensitive element in the inner ear and can be thought of as the body's microphone. It is situated on the basilar membrane in one of the three compartments of the Cochlea. It contains four rows of hair cells which protrude from its surface. Above them is the tectorial membrane which can move in response to pressure variations in the fluid-filled tympanic and vestibular canals. There are some 16,000–20,000 of the hair cells distributed along the basilar membrane which follows the spiral of the cochlea. The place along the basilar membrane where maximum excitation of the hair cells occurs determines the perception of pitch according to the place theory. This theory of pitch perception suggests that pitch is determined by the place along this collection at which excitation occurs.

The sensitive hair cells of the organ of Corti may have about 100 tiny stereocilia which in the resting state are leaning on each other in a conical bundle. In response to the pressure variations in the Cochlea produced by sound, the stereocilia may dance about wildly and send electrical impulses to the brain.

The deformation of the basilar membrane (BM) contributes to the reception of a nervous impulse.

In fact, in the BM there is a transformation of the mechanical energy into electromagnetic one. If then a nervous impulse is transformed back into the mechanical fluctuations (as in case of speakers), we will hear the same signal as in the inner ear.

The nervous impulse or active potential is a “splash” of the electromagnetic field caused by the change of concentration of the ions across the membrane cell.

What causes this “splash”?

The biophysical principles of the change of concentration of the ions as well as the electric charge of cytoplasm are based on the change of the membrane permeability. It consists of the channels (their size is 2–5 nanometers). Their walls cover the albuminous polypeptide molecules.

It means that the change of the sizes of the ion channels influences the permeability of a membrane.

First, with the pressure on the BM, the hair cells also undergo the deformation, and it means, the mechanical change of the geometrical sizes of
the channels (see fig. 2.4.3). Second, with the deformation of the BM and tectorial membrane, the hormone of acetylcholine (AH) (which is a mediator in the synapse of involuntary nervous system (INS) and which action is always directed to the increase of permeability of the cellular membranes) is secreted.

It means that the increase of AH concentration intensifies the metabolism.

| ↑ AH →↑ Permeability of membranes → relaxation of muscles → decrease in arterial and peripheral pressure. |
| Notes: |
| Symbol “↑” means increasing, “↓” – decreasing, “→” — “leads to…” |

The increase of permeability of the membranes contributes to the relaxation of muscles.

Thus, the transformation of mechanical energy into electromagnetic one in the inner ear is caused by two things:
1) mechanical — the change of the geometrical sizes of ion channels;
2) biochemical — the increase of AH concentration and that of the membrane permeability.

### 2.5. The weber-fechner law

It is necessary to disprove the statement, that all people are capable to perceive the sound waves with the same loudness.

The Weber-Fechner's law represents the relation between the magnitude of physical stimulus and the magnitude of psychological sense in human being to the utility function.

There are three main characteristics of sound perception: pitch, loudness, and sound quality (timbre).

“Timbre” describes those characteristics of sound which allow the ear to distinguish sounds which have the same pitch and loudness. Timbre is then a general term for the distinguishable characteristics of a tone. Timbre is mainly determined by the harmonic content of sound and the dynamic characteristics of sound.

The perceived pitch of a sound is just the ear's response to frequency. There are circumstances in which a constant frequency sound can be perceived to be changing in pitch.

Sound loudness is a subjective term describing the strength of the ear's perception of a sound. It is intimately related to sound intensity but can by no means be considered identical to intensity. The sound intensity must be factored by the ear's sensitivity to the particular frequencies contained in the sound.
A bel (symbol B) is a unit of measure of ratios of power levels, i.e., relative power levels. With 1000 Hz threshold of hearing $I_0$ is equal to $10^{-12} \frac{W}{m^2}$ (in this case the sound pressure is equal to $p_0 = 2 \cdot 10^{-5} Pa$). The threshold of pain $I$ is equal to $10 \frac{W}{m^2}$ ($p = 63 Pa$). So:

$$\frac{I}{I_0} = \frac{10}{10^{-12}} = 10^{13}$$

(2.5.1)

The number of bels for a given ratio of power levels is calculated by taking the logarithm, to the base 10, of the ratio. Therefore, one bel corresponds to a ratio of 10:1. Mathematically, the number of bels is calculated as:

$$\lg \frac{I}{I_0} = \lg 10^{13} = 13$$

(2.5.2)

The measured threshold of hearing curve shows that the sound intensity required to be heard is quite different for different frequencies. The standard threshold of hearing at 1000 Hz is nominally taken to be 0 dB, but the actual curves show the measured threshold at 1000 Hz to be about 4 dB. There is marked discrimination against low frequencies so that about 60 dB is required to be heard at 30 Hz. The maximum sensitivity at about 3500 to 4000 Hz is related to the resonance of the auditory canal.

Fig. 2.5.1
The given scale from 13 levels (units) of intensities $L$ is designated and measured in bel (b). However usually we use 10 times the smaller unit of measure:

$$L_{dB} = 10 \log \frac{I}{I_0}$$

(2.5.3)

The bel is too large for everyday use, so the decibel (dB) equal to 0.1 B, is more commonly used. One decibel is equivalent to a ratio of about 1.259:1. It is defined as $10 \log_{10}(P_1/P_2)$, where $P_1$ and $P_2$ are powers.

The decibel is not an SI unit, although the CIPM has recommended its inclusion in the SI system.

The decibel is often used in acoustics to quantify sound levels relative to some 0 dB reference. The reference may be defined as a sound pressure level (SPL), commonly 20 micropascal (20 μPa). To avoid confusion with other decibel measures, the term dB(SPL) is used for this. The reference can also be defined as the sound intensity at the threshold of human hearing, which is conventionally taken to be one picowatt per square metre (1 pW/m²), roughly the sound of a mosquito flying 10 feet (3 m) away.

It is possible to receive the scale of sound pressure using the logarithm.

Perceived loudness is directly related to the pressure amplitude of the sound wave reaching our ear. Though “loudness” is perceived, however, it is hard to measure. Pressure amplitude on the other hand is easy to measure [microphones do this by converting pressure waves into mechanical motion, into
an electrical signal that can be measured]. Due to the range of human hearing, the range of pressures we need to be able to measure is vast and we use a logarithmic scale, based on a unit called decibel. The sound wave intensity $I$ is proportional to the second degree of sound pressure $p$:

$$I \sim p^2,$$

so the sound intensity level in decibels is

$$L_{dB} = 10 \log \frac{I}{I_0} = 10 \log \left( \frac{p}{p_0} \right)^2 = 20 \log \frac{p}{p_0}.$$

Let us find the relation between the sound stimulus and its perception.

Weber established, that the change of the stimulus $\Delta I$ which corresponds to the change of the loudness $\Delta E$, depends on the initial intensity $I$ under the condition $\frac{\Delta I}{I} = \text{const}$.

This kind of relationship can be described by the differential equation:

$$dE \sim \frac{dI}{I} \Rightarrow dE = k \frac{dI}{I}$$

$$\downarrow$$

$$E = \int I \frac{dI}{I} = k \ln \frac{I}{I_0}$$

The Weber–Fechner law describes the human perception of various physical stimuli. E. H. Weber (1795–1878) was one of the first people to quantitatively study the human response to physical stimuli. In one of his classic experiments, Weber gradually increased the weight that a blindfolded man was holding and asked him to respond when he first felt the increase. Weber found that the response was proportional to a relative increase in the weight. That is to say, if the weight is 1 kg, an increase of a few grams will not be noticed. Rather, when the mass is increased by a certain factor, an increase in weight is perceived. If the mass is doubled, the threshold is also doubled.

In other words, you have to multiply the sound intensity by the same factor to have the same increase in loudness. That is why the numbers around the volume control dial on a typical audio amplifier are related not to the absolute power amplification, but to its logarithm.

This logarithmic relationship means that if the perception is altered in an arithmetic progression the corresponding stimulus varies as a geometric
The point is that this logarithmic relationship is valid for not just the sensation of weight alone, but for other stimuli as well.

\[
1 + \left( \frac{1}{2} \right)^1 + \left( \frac{1}{2} \right)^2 + \ldots + \left( \frac{1}{2} \right)^n \quad \text{— geometric progression}
\]

\[
\ln \left[ \left( \frac{1}{2} \right)^1 + \left( \frac{1}{2} \right)^2 + \ldots + \left( \frac{1}{2} \right)^n \right] = \ln \left( \frac{1}{2} \right) + 2 \ln \left( \frac{1}{2} \right) + \ldots + n \ln \left( \frac{1}{2} \right) \quad \text{— arithmetic progression}
\]

All aforesaid also means, that in connection with features of the structure of an ear each person is capable to perceive subjectively different loudness at the fixed frequency and intensity. Weber's findings were later popularized by G. T. Fechner (1801–1887) and hence the name.

Sounds above 85 dB are considered harmful, while 120 dB is unsafe and 150 dB causes physical damage to the human body. Windows break at 163 dB. Jet airplanes are 165 dB. Eardrums pop at 190 to 198 dB. Shock waves and sonic booms are 194 dB. Sounds around 200 dB can cause death to humans and are generated near bomb explosions. The space shuttle is around 215 dB. Nuclear bombs are 240 dB to 258 dB. Even louder are earthquakes, tornados, hurricanes and volcanoes.

The Weber–Fechner law states that the amount of change needed for sensory detection to occur increases with the initial intensity of stimulus, and is proportional to it. The change in stimulus that will be just noticeable is a constant ratio of the original stimulus. Applied to vision, the amount of change in brightness detected by the visual senses is dependant on and proportional to the present brightness.

**Some Weber-Fechner values for Different Senses**

<table>
<thead>
<tr>
<th>Sense</th>
<th>Increase in Intensity/Background Intensity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vision (brightness of white light)</td>
<td>1/60</td>
</tr>
<tr>
<td>Kinesthesis (lifted weights)</td>
<td>1/50</td>
</tr>
<tr>
<td>Pain (heat on skin)</td>
<td>1/30</td>
</tr>
<tr>
<td>Hearing (tone of mid pitch, loudness)</td>
<td>1/10</td>
</tr>
<tr>
<td>Pressure (touch spot on skin)</td>
<td>1/7</td>
</tr>
<tr>
<td>Smell (odour of India rubber)</td>
<td>1/4</td>
</tr>
<tr>
<td>Taste (table salt)</td>
<td>1/3</td>
</tr>
</tbody>
</table>

The Relative Loudness levels are important insofar as they demonstrate that a 10-deciibel increase will be perceived as twice as loud as the pervious level or conversely, a decrease of 50 % from the previous higher level.
2.6. Ultrasound, its characteristics and applying in medicine

What is the ultrasound? What does the term “ultra” mean? Ultrasound is sound or vibrations with frequencies too high to be audible by the human ear, above approximately 20 kilohertz. Some sound that is ultrasound to the human ear can be heard by animals, such as dogs. Ultrasound has industrial and medical applications. Medical ultrasonography can visualise muscle and soft tissue, making them useful for scanning the organs. However, they will not penetrate bone so they are not a substitute for conventional X-rays in many cases.

Ultrasonic cleaners, sometimes called supersonic cleaners, are used at frequencies from 20–40 kHz for jewellery, lenses and other optical parts, dental instruments, surgical instruments, and industrial parts.

The influence of the ultrasound on biological objects consists of three points: mechanical, heat and chemical.

Drug delivery to diseased tissue is certainly enhanced under irradiation with ultrasound. However, the interaction between an ultrasound wave and the tissue can be attended with deleterious biological effects, especially at high wave intensities. It is very important to find the conditions at which ultrasound passed through the blood vessel containing microbubbles does not pose threats. The dissipation of acoustic energy and acoustic cavitation are two major reasons for bioeffects of ultrasound. When passing through a human body the energy of an ultrasonic signal is scattered by regions of different acoustic impedance and absorbed locally through viscous shearing and relaxation processes. The absorption results in heating of blood and tissue and, together with the scattering, produces radiation pressure and acoustic streaming within blood. Radiation pressure forces in turn push up blood cells against vessel walls and may force the cells to aggregate. Acoustic streaming is known to increase the shear stresses, and therefore the likelihood of cell lysis, though the destruction of cells rarely occurs in the blood free of gas inclusions or cavitation bubbles.

The most deleterious nonthermal bioeffects are produced by acoustic cavitation. Pressure waves of high intensity are capable to rupture blood and tissue, giving rise to the formation and oscillation of cavities, i.e. gas or vapor bubbles. It is precisely these cavitation bubbles that are responsible for hemolysis, hemorrhage, and DNA fragmentation. Cavitation inception is a threshold process and significantly depends on the presence of “nuclei” in the liquid. The nuclei are small pockets of gas which are trapped in conical pits of solid impurities or covered with organic or surface-active molecules. The more is the quantity of “nuclei” in the liquid, the smaller is the wave amplitude needed for cavitation inception. Like any natural liquid, blood has
cavitation nucleation agents, and therefore sufficiently high-amplitude ultrasound pulses cavitate it. There is also emerging a body of data identifying cavitation thresholds for tissue. Encapsulated microbubbles, especially of very small size, can be considered as additional “man-made” cavitation nuclei. The administration of microbubbles in blood should therefore decrease its cavitation thresholds and enhance the risk of biological damage.

Cavitation may be stable or inertial. Stable cavitation refers to bubbles that pulsate about some equilibrium radius and may exist for some time. The inhomogeneous cyclic field around each of stably oscillating bubbles causes steady blood flow of small size in a process known as microstreaming. The microstreaming creates extremely high shear stresses near the bubble surface and actually deforms or even damages neighbor cells in vivo. At the onset of inertial cavitation, bubbles oscillate violently, undergoing initial explosive growth and subsequent rapid collapse. In the final phase of collapse, pressure and temperature inside the bubble can reach thousands of bars and degrees kelvin. That high temperatures lead to the emission of light (sonoluminescence) and can cause bond dissociation in molecules, producing free radicals capable to react with biomedical species in the same way as those produced by ionizing radiation. Immediately after the bubble rebound, the high-pressure shock wave emanates from the bubble location and causes mechanical damage to the surrounding fluid. It is very likely that most of the bubbles collapse asymmetrically in blood vessels because there is limited “space” for them to undergo expansion. Hence, the greatest destruction of human tissue is made by violently collapsing bubbles. The American Institute of Ultrasound in Medicine accepted the Mechanical Index (MI), which is a dimensionless quantity proportional to the ultrasound rarefaction pressure and inversely proportional to the square root of the center frequency of the ultrasound wave, as a predictor of possible biological responses to inertial cavitation activity. It is agreed that if MI is less than 0.7, then the probability for inertial cavitation is negligible.

On the other hand, acoustic cavitation may have its good points. In particular, exposure from a 3.5 MHz diagnostic ultrasound scanner has been shown to produce sonoporation, i. e. transient membrane permeabilization of monolayer cells situated at the vicinity of microbubbles. Consequently, ultrasound is capable to enhance gene/drug delivery from the microbubble surface into cells and tissue. Again, the sonoporation is a therapeutic effect only at low dosage of ultrasound. An increase in the ultrasound intensity results in killing the cells and other detrimental “side effects” to the patient. The risk of biological damage is high under the condition of inertial cavitation, i. e. jet-induced sonoporation may lose its therapeutic effect and even become dangerous.
2.7. The applying of ultrasound in medicine.  
The doppler’s effect

For example, the Doppler’s effect in an ultrasonic pulse probe detects the reflected sound from moving blood. For best understanding let us study the theory of this effect.

For example, the origin of some signal (for instance, a speaking teacher) and receiver (as a student) are in the unmoving state — they are keeping the zero velocity. The origin of wave generates the signal with certain frequency, and the receiver obtains this signal with the same frequency. It is very important to note, that distance separating both origin and receiver does not change. It follows that wave’s density does not change too on this distance! But, if somebody (teacher or student) begins to move with certain velocity, the receiver will obtain signal with distortion. The distortions degree depends on the velocity of distance’s change between origin and receiver like at figures below.
Let us study the first situation from two possible ones: the distance between persons is decreasing (wave’s density is increasing).

Case A.

\[ f'_{\text{rec}} = \frac{\varrho}{\varrho - \varrho_{\text{origin}}} f_{\text{origin}} \]

Case B.
The similar situation: the distance between teacher and pupil is decreasing in the following way:

\[ f_{\text{rec}} = \frac{\varrho + \varrho_{\text{rec}}}{\varrho} f_{\text{origin}} \]

Case C.

\[ f_{\text{rec}} = \frac{\varrho + \varrho_{\text{rec}}}{\varrho - \varrho_{\text{origin}}} f_{\text{origin}} \]
Now let us study the “opposite” case: the distance between persons is increasing (the wave’s density is decreasing).

Case A.

The student is going out with $g_{\text{rec}}$

$$f_{\text{rec}} = \frac{g - g_{\text{rec}}}{g} f_{\text{origin}}$$

Case B.

The teacher is going out from student

$$f'_{\text{rec}} = \frac{g}{g + g_{\text{origin}}} f_{\text{origin}}$$

Case C.

Both teacher and student are going out from each other

$$f_{\text{rec}} = \frac{g - g_{\text{rec}}}{g + g_{\text{origin}}} f_{\text{origin}}$$
The particular case:

$$f'_{\text{rec}} = \frac{\mathcal{G}}{\mathcal{G} - g_{\text{origin}}} f_{\text{origin}}$$

$$f'_{\text{rec}} = f_{\text{origin}}$$

The received signal without distortions

The particular case:

$$f'_{\text{rec}} = \frac{\mathcal{G}}{\mathcal{G} - g_{\text{origin}}} f_{\text{origin}}$$

$$f'_{\text{rec}} = f_{\text{origin}}$$

The received signal without distortions
Fig. 2.7.1

The particular case:

\[ f'_{\text{rec}} = \frac{f_{\text{origin}}}{g_{\text{origin}} + g_{\text{origin}}} \]

\[ f_{\text{rec}} = f_{\text{origin}} \]

The received signal without distortions

The particular case:

\[ f'_{\text{rec}} = \frac{f_{\text{origin}}}{g_{\text{origin}} + g_{\text{origin}}} \]

\[ f_{\text{rec}} = f_{\text{origin}} \]

The received signal without distortions
Let us put the following question: in what way is the Doppler’s effect used in medicine?

The ultrasound techniques can measure the blood flow velocity by using of the Doppler’s effect. The reflected ultrasound is shifted in frequency from the frequency of the source, and that difference in frequency can be accurately measured by detecting the *beat frequency* between the incident and reflected waves. The beat frequency is directly proportional to the velocity of flow, so continuous recording of the beat frequencies from the different parts of the artery gives you an image of the velocity profile of the blood flow as a function of time.

High frequency ultrasound in the 7–12 MHz region is used for high resolution imaging of arteries which lie close to the surface of the body, such as the carotid arteries. Using a nominal sound velocity of 1540 m/s in tissue, the sound wavelength in tissue for a 7 MHz sound wave can be obtained from the wave relationship: \( \vartheta = fn \).
Part 3. Biophysical principles of biorheology and hemodynamics

3.1. Bernoulli’s equation

The more is the area $A$ of the cross-section of a tube, i.e. the broader is the area of the tube, the slower are the liquid flows in this area provided the liquid immediately runs through the whole tube with different width, i.e. with various $A$.

Fig. 3.1.1

The liquid will flow without points of break if in the same time $t$:


What is the way to calculate the given volume?
To answer this question we need to recall the equation from geometry:
Volume of cylinder = height multiply by cross-section of area $A$.
Or:

\[ V = hA \]  \hspace{1cm} (3.1.1)

In our case $h$ refers to the length of the way of moving fluid with velocity $\vartheta$, or:

\[ V = A \cdot \vartheta \cdot t. \]  \hspace{1cm} (3.1.2)

So, the increases $A$ cause the decreases of the velocity $\vartheta$ at the same volume.

We can mathematically express the continuity equation.

\[ A_1 \vartheta_1 = A_2 \vartheta_2 = A_3 \vartheta_3 = k \]  \hspace{1cm} (3.1.3)

The $k$ is any constant of proportionality. This formula expresses the continuity equation.

The quicker the liquid flows through one or another part of a tube, the more is the absolute value of the kinetic energy of liquids ($\frac{mv^2}{2}$) and, hence, the less is the fluid potential energy ($mgh$), which is defined by the liquid pressure on the walls of a tube. So, the total energy of liquids being equal to the amount of kinetic and potential energy is constant.
But nevertheless, if the liquid is blood, with the increase of fluid flow velocity the pressure on the walls of vessels will increase. It can be explained by the elasticity of blood vessels. The continuity equation is true for absolutely rigid tubes and this equation is also true for a sample liquid, so called Newtonian fluids.

Kinetic energy of a moving fluid is more useful in applications when it is expressed as kinetic energy per unit of volume:

$$ \frac{kinetic\ldots energy}{Volume} = \frac{m\vartheta^2}{2V} = \frac{1}{2} \frac{m}{V} \vartheta^2 = \frac{1}{2} \rho \vartheta^2, $$

(3.1.5)

where $\rho$ — density of liquid, $\vartheta$ — fluid velocity.

What is hydrostatic fluid pressure?

As far as you know, the pressure exerted by a static fluid depends only upon the depth of a fluid, the density of a fluid, and the acceleration of gravity.

The potential energy of a moving fluid is expressed as potential energy per unit of volume

$$ \frac{potential\ldots energy}{Volume} = \frac{mgh}{V} = \rho gh $$

(3.1.6)

$\rho = m/V$ = fluid density, $g$ = acceleration of gravity, $h$ = depth of fluid

The most obvious application is to the hydrostatic fluid pressure of a fluid, where pressure can be used as energy density alongside kinetic energy density and potential energy density.

It is the hydrostatic fluid pressure that characterizes the blood pressure on the walls of vessels.

We can state then, that under conditions of continuous flow the amount of hydrostatic and hydrodynamic pressure is a constant. Bernoulli equation can be
considered to be the statement of the conservation of energy principle appropriate for flowing fluids.

\[
\begin{aligned}
P_1 + \frac{1}{2} \rho v_1^2 + \rho g h_1 &\equiv P_2 + \frac{1}{2} \rho v_2^2 + \rho g h_2 \\
\text{Total pressure (consists of 3 parts) does not change for lamina flow}
\end{aligned}
\]

3.2. The laminar (streamlined) and turbulence flow

What are the laminar (streamlined) and turbulence flow? What is the lamina of a liquid? Do you know that liquid being at rest is uniform, but its molecules are equally distributed throughout the whole bulk occupied by a liquid?

We use the term “lamina of a fluid” concerning the moving fluid. The molecules of the fluid lamina have the same velocity.

NOTES:

- \( \rho \) – density of the liquid
- \( v \) – liquid speed
- \( A \) – cross-section of the area
In this case, the velocity of the flow varies from zero at the walls to a maximum along the centerline of a vessel.

Turbulence or turbulent flow is a flow regime characterized by low momentum diffusion, high momentum convection, and such characteristics of liquid motion as pressure and velocity variation with time. It is opposed by a laminar flow.

Friction force is always directly proportional to the velocity of any moving body and opposite directed to the motion. For example, let us imagine that you try to come in time to lecture. The increases of body velocity will result in the increases of frequency interactions of air molecules with your body, then, the friction force increases.

So, $F_{\text{friction}} \sim \vartheta$, or (in projection on horizontal axis):

$$F_{\text{friction}} = -r \vartheta,$$

(3.2.1)

where $r$ — proportionality coefficient.

Joke example!
Let us imagine, that you are rubbing your hand along another one. What does the friction force depend upon?

\[
\begin{align*}
\text{area of the interacted hands} & \quad F_{fr1} > F_{fr2} \\
F_{fr} \sim A & \quad \text{speed of frictions} \\
F_{fr} \sim g & \\
\end{align*}
\]

Or:

\[
F_{fr} \sim A \ g
\]

Fig. 3.2.2
What does the friction force of your copying of information from screen to notebook depend upon? What is the role of air temperature, quality of pen, etc. in these processes?

The reason for internal friction is explained by the force of interaction between the liquid molecules. The force of friction between two moving lamina of liquid is directly proportional:

1) to the area of lamina $A$;

2) to the velocity gradient $\frac{\partial \vartheta}{\partial x}$ (how fast the change of velocity along the direction perpendicular to lamina of a flow liquid, in our given example — axis $OX$);

Hence, $F_{\text{friction}} \sim A \frac{\partial \vartheta}{\partial x}$.

The constant of the proportionality is called viscosity. The resistance to the flow of a fluid and the resistance to the movement of an object through a fluid are usually stated in terms of the viscosity of the fluid.

$$ F_{\text{friction}} = 3A \frac{\partial \vartheta}{\partial x} \quad (3.2.2) $$

$$ \eta = \frac{F_{\text{friction}}}{A \partial \vartheta} \quad \text{— coefficient of proportionality (viscosity).} $$
The units of viscosity are the following: \( \frac{Nm}{m^2 \cdot m/s} = \frac{N \cdot s}{m^2} = 1Pa \cdot s = 10Poise \)

(The GGS unit is dyne s/cm², which is called Poise).

The viscosity of blood at body temperature is about 0.03 Poise. The \( Pa \cdot sec \) is called a Poisuisse and is equal to 10 Poise. \( 1Poise = 0.1 \frac{Ns}{m^2} \)

What kinds of viscosity are there?

The viscosity of fluids is either given as absolute or dynamic viscosity \( \eta \) or as kinematic viscosity \( \nu \). Both terms are related to each other via the fluid density \( \rho \): \( \nu = \frac{3}{c} \). According to this formula, the units of kinematic viscosity are the following:

\[
\nu = \frac{1Poise}{1kg} = 1Poise m^3 = 1Ns m^3 \]

Inasmuch as \( N = \frac{m}{s^2} \Rightarrow \frac{m s^3}{s^2 m^2} = \frac{m^2}{s} = 10^4 Strokes \)

The old unit for kinematic viscosity is stokes after George Gabriel Stokes (1819–1903): 1 stokes = 1 sm²/s = 0.0001 m²/s.

What are the Newtonian and non-Newtonian fluids?

Newtonian fluid is such fluid in which viscosity \( \eta \) does not depend on velocity gradient (shear rate) \( \frac{\partial \vartheta}{\partial x} \). It is an ideal case. Otherwise it is non-Newtonian fluid.

We can state that the more is the amount of molecules with large molecular mass, the more is the difference between Newtonian and non-Newtonian fluids. With the increase of the amount of big molecules which have large molecular mass the intensity of electromagnetic coupling in a fluid increases what leads to non-Newtonian nature of a liquid, complicates the interaction between the molecules. Blood is also considered to be Non-Newtonian liquid.

Some dynamic viscosities of Newtonian liquids (at 20 °C) are listed below:

- acetone \( 0.326 \times 10^{-3} \) Pa·s
- benzene \( 0.64 \times 10^{-3} \) Pa·s
- castor oil \( 985 \times 10^{-3} \) Pa·s
- glycerol \( 1485 \times 10^{-3} \) Pa·s
- methanol \( 0.59 \times 10^{-3} \) Pa·s
nitrobenzol $2.0 \times 10^{-3}$ Pa·s
olive oil $81 \times 10^{-3}$ Pa·s
water $1.025 \times 10^{-3}$ Pa·s

Contrary to many assertions, glass is an *amorphous solid*, not a liquid, and it does not flow, but still we can talk about its viscosity. Many fluids such as honey have a wide range of viscosity.

**Can we prove, that a turbulence flow appears with heating?**

Of course, but not only with heating. The main reason for transfer of laminar flow into turbulence is based on the increase of kinetic energy of fluid molecules.

The Reynolds number characterizes whether flow conditions lead to *laminar* or *turbulent flow*. The Reynolds number is the most important dimensionless number in fluid dynamics providing a criterion for dynamic similarity. It is named after Osborne Reynolds (1842–1912). Typically it is given as follows:

$$Re = \frac{\rho v d}{\eta}$$  \hspace{1cm} (3.2.3)

where $Re$ is the *Reynolds number* and $\rho$ fluid density, $d$ — diameter of the vessel, $v_s$ — mean fluid velocity, $L$ — characteristic length (equal to diameter $2r$ if a cross-section is circular), $\eta$ — (absolute) dynamic fluid viscosity.

Laminar flow within pipes, for example, will occur when the Reynolds number is below the critical Reynolds number of $Re_{crit, \text{pipe}} = 2300$ and turbulent flow will happen when it is above 2300 where the Reynolds number is based on the pipe diameter and the mean velocity $v_s$ 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.

We can consider, that the increase of blood temperature ($\uparrow T$) results in the increase of erythrocyte sedimentation rate (ESR).

Does the increase of temperature $T$ lead to the increase of the velocity of blood elements including erythrocytes? Can we state then that the increase of body temperature results in the decrease of blood viscosity? If the blood viscosity decreases (that is the increase of $Re$ (according to the formula)) must the turbulence of blood flow also be increased? The increase of blood turbulence leads to the intensification of chaotic movement of blood corpuscles, doesn't it?

Will more time be required for their sedimentation?

It means that the increase of body temperature leads to the increase of ESR.

$\uparrow T \Rightarrow \uparrow \vartheta \Rightarrow \downarrow \eta \Rightarrow \uparrow Re \Rightarrow \uparrow \text{ESR}$
3.3. Hagen-poiseuille law

The similarity of Hagen-Poiseuille's law and Ohm's law allows to determine the resistance of blood vessels and blood pressure in arterioles. According to the parallel location of arterioles in cardiovascular system, we must obtain the findings of the low resistance of these vessels, though it is not true. How can we explain this?

Poiseuille's law (or Hagen–Poiseuille's law also named after Gotthilf Heinrich Ludwig Hagen (1797–1884) is the physical law concerning the voluminal laminar stationary flow $\Phi$ of incompressible uniform viscous liquid through a cylindrical tube with the constant circular cross-section, experimentally derived in 1838, formulated and published in 1840 and 1846 by Jean Louis Marie Poiseulle (1797–1869).

$$\vartheta = \frac{P_1 - P_2}{4l\eta} (R^2 - r^2) \Rightarrow dQ = \vartheta dA = \vartheta \cdot 2\pi r dr \quad dQ = \frac{P_1 - P_2}{2l\eta} (R^2 - r^2) r dr$$

Volume flowrate is equal to:

$$Q = \pi \frac{P_1 - P_2}{2l\eta} \int_0^R (R^2 - r^2) dr ,$$

$$Q = \frac{\pi R^4}{8\eta} \frac{P_1 - P_2}{l} \quad (3.3.1)$$

where $\frac{P_1 - P_2}{l}$ or $\frac{dP}{dx}$ — pressure gradient, $V$ is a volume of the liquid, poured in the time unit $t$, $v_s$ median fluid velocity along the axial cylindrical coordinate $z$, $r$ internal radius of the tube, $\Delta P$ the pressure drop at the two ends (pressure difference), $\eta$ dynamic fluid viscosity and $l$ characteristic length along $z$, a linear dimension in a cross-section (in non-cylindrical tube).

It is important to note, that a 19 % increase in radius will double the volume flowrate!
The law is also very important, especially in hemorheology and hemodynamics, they both being fields of physiology.

**Suppose you also have an emergency requirement for a five-fold increase in blood volume flowrate. How does your body supply it?**

For example, blood pressure increases from 120 mmHg to 600 mmHg. This is not a realizable pressure. But the body has a much more potent method for the increasing volume flowrate in the vasodilatation of the arterioles (small vessels). A 50% dilation of the vessels is within the body's control limits (if \( r \Rightarrow 1.5r \)). For constant pressure, this increases the volume of flowrate:

\[
(1.5)^4 = 5.06
\]

![Diagram showing vasodilatation and vasoconstriction](image)

*Fig. 3.3.2. The effect of vessel radius is just as dramatic in reduction*

**We can show the similarity between the flow of the liquid molecules through a vessel and the flow of the electric charge carriers in a conductor.**

The *electric current* may be represented as the movement of multiple microscopic charge carriers with a drift velocity in a conductor under the action of voltage.

\[
V_{AB} = \varphi_B - \varphi_A = \Delta \varphi = \frac{W}{q_{(B \rightarrow A)}}
\]

According to Ohm's Law the electric current \((I)\) is directly proportional for the voltage \((V)\):

\[
I \sim V \Rightarrow I = \frac{1}{R} V,
\]

where \( \frac{1}{R} = G \) — the coefficient of proportionality (conductivity), which depends on the «quality» of the conductors of electricity (shape and size of atoms) and «quantity» (length \(l\) and area \(S\)).
According to Hagen–Poiseuille's law:

\[ V = \frac{\pi r^4}{8\eta l} \Delta p t \]  
(3.3.2)

or

\[ Q = \frac{(P_1 - P_2)}{R_{hydr}}, \]  
(3.3.3)

where \( \Delta p = P_1 - P_2 \), \( Q = \frac{V}{t} \), \( R_{hydr} = \frac{\pi r^4}{8\eta l} \).

The voltage seems to be a difference of potentials \( \Delta \phi \) similar to difference of pressure \( \Delta p \) on the both sides of a tube, which creates liquid flows.

Electric current is directly proportional to electric charge concentrations, modulus of electric charge carrier, and it's velocity in the material. So, the more is the total electric charge (through the area for less time), the more is the electric current.

The Ohm's Law for electric current flow and Hagen-Poiseuille's law for the smooth flow of fluids are similar.

This similarity may be represented by the figure below:

\[ I = \frac{|\phi_1 - \phi_2|}{R_{el.}} \]  
Ohm's law

\[ R_{el.} = \rho_0 \frac{1}{A} \]  
active resistance (in electricity)

Fig. 3.3.3. The whole base of both electricity and fluid mechanics

The resistance of blood vessels is inversely proportional to its four power radius.

What are the other reasons for the resistance of blood vessels?

The resistance of blood vessels also depends on:

1) velocity of blood flow;
2) positional relationship of the vessels (parallel connection and concatenation).
3) viscosity of blood.
The first reason. Just as in the analogy given above, the drop of pressure in a blood vessel is directly proportional to the blood flow rate.

\[ P_0 - P = IR \]

where \( R \) — hydrodinamic resistance.

The given equality was called as the Hagen's equation. Hence, blood pressure \( P \) in any part of the vessel:

\[ P = P_0 - IR \]

The second reason. Just as electric resistance, the general flow resistance of concatenated blood vessels is equal to the amount of the resistance of these vessels.

\[ R_{gen} = R_1 + R_2 + \ldots + R_n \]  \hspace{1cm} (3.3.4)

In case of parallel connection (arterioles, capillaries, veins) the general flow resistance decreases:

\[ \frac{1}{R_{gen}} = \frac{1}{R_1} + \frac{1}{R_2} + \ldots + \frac{1}{R_n} \]  \hspace{1cm} (3.3.5)

That is why, for instance, arterioles are of the greatest resistance: the influence of small radius (the first reason of resistance increase) predominates over the influence of parallel arterioles in the cardiovascular systems.

The third reason. 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 (the third reason) 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.

3.4. Surface tension. Capillarity

What does the term “surface tension” mean? What is the reason for surface tension?

It is the force of interaction of the molecules with each other. Surface tension is caused by the attraction between the molecules of the liquid. In the bulk of the liquid each molecule is pulled equally in all directions by neighbouring molecules, resulting in a net force of zero. At the surface of the liquid, the molecules are pulled inwards by other molecules deeper inside the liquid, but there are no liquid molecules on the outside to balance these forces, so the surface molecules are subject to a net inward force.
We can calculate the surface tension.

The surface tension can be considered as surface energy. If a surface with surface tension is expanded by a unit of area, then the increase in the surface's stored energy is also equal to $y$. It follows then that the surface energy of a material is equal to one half of its energy of cohesion.

$$y = \frac{W}{A} \quad (3.4.1)$$

The units of surface tension are joules per square metre ($\text{J} \text{m}^{-2}$). Under the condition that $W = Fl$, we can rewrite this equation in the following way:

$$y = \frac{Fl}{l^2} = F/l \quad (3.4.2)$$

Surface tension is measured in newton per metre ($\text{Nm}^{-1}$), is represented by the symbol $y$ and is defined as the force along a line of unit of length perpendicular to the surface.

There is a relation between the coefficient of liquid surface tension and interaction between its molecules and the molecules of air and solid surface.

Molecules of liquid state experience strong intermolecular attractive forces. When those forces are between similar molecules, they are referred to as cohesive forces.

When the attractive forces are between dissimilar molecules, they are said to be adhesive forces. The adhesive forces between water molecules and the walls of a glass tube are stronger than the cohesive forces and lead to an upward turning meniscus at the walls of the vessel and contribute to capillary action.

The attractive forces between molecules in a liquid can be viewed as residual electrostatic forces and are sometimes called van der Waals forces or van der Waals bonds.

What forces will influence fluid ability to wet a surface?

Interaction between molecules of three mediums (solid, liquid, gas)!
Fig. 3.4.2

$\vec{F}_{fs}$ — the forces of interaction between fluid and solid mediums.

$\vec{F}_{sg}$ — the forces of interaction between solid and gas mediums (on the border «solid body – gas»).

$\vec{F}_{fg}$ — the forces of interaction between fluid and gas mediums (on the border «fluid – gas»).

Contact angle $\Theta$ (formed by tangent line to fluid surfaces) quantity refers a fluid ability to wet some solid surfaces.

$$\cos \Theta = \frac{F_{fs} - F_{sg}}{F_{fg}}$$ (3.4.3)

Fig. 3.4.3

If $\Theta < \frac{\pi}{2}$ ⇒ wetting. Solid surface is hydrophilic.

$\Theta > \frac{\pi}{2}$ ⇒ nonwetting. Solid surface is hydrophobic.

Fig. 3.4.4
The influence of surface tension forces leads to the curvature of fluid surface lamina in narrow tubes. It is followed by additional pressure appearance on fluid.

\[ \Delta p = \frac{F}{A} = \frac{y \cdot l}{pr^2} = \frac{y \cdot 2pr}{pr^2} = \frac{2y}{r} \]  \hspace{1cm} (3.4.4)

This equation reflects Laplace's pressure. Surface has spherical form, area projection is \( pr^2 \).

**The curved liquid tension force is called meniscus.**

So, we have studied the phenomena of cohesion and adhesion.

The cohesive forces between molecules down into a liquid are shared with all neighboring atoms. Those on the surface have no neighboring atoms above, and exhibit stronger attractive forces upon their nearest neighbors on the surface. This enhancement of the intermolecular attractive forces at the surface is called surface tension.

Can we state that if the temperature increases (and the molecules of a liquid move faster), the interactions between the molecules have less effect on their motion and the surface tension decreases?

The temperature is related to the energy of molecular motion in a material. The surface tension of a particular liquid usually decreases as temperature increases.

**When taking blood, it is better to use a capillary tube instead of wide one in the laboratories.**

1. Is it reasonable to postulate that if the blood goes upward along the tube contrary to the effect of gravity, it is explained by the net forces directed upward?

![Fig. 3.4.5](image)
2. Can we state that the adhesion of the fluid (for instance, blood) in a capillary tube results in surface tension directed upward or downward according to the calculations given below?

\[
\cos \Theta = \frac{(y_{32} - y_{31})}{y_{21}}
\]

\[
\Delta p = 2y/r
\]

\[
\Delta p = 2y \cos \Theta / R
\]

\[
\rho gh = 2y \cos \Theta / R,
\]

or:

\[
h = 2y \cos \Theta / R \rho g
\]

3. Is it true that the less is the radius \( R \) of a narrow tube, the more is the value of the force pressure of the surface layer of blood \( F \) directed downward?

4. Can we say that according to the third Newton's law the resistance force will be directed upward and will be the very force which makes the blood go upward along the narrow tubes?

### 3.5. Surfactant

Surface tension coefficient can be changed due to the surfactant (surface active agent) action. Under the condition, that about 80 % of body are fluids, surfactant role in nature is greatly important. For instance, surfactant is surface active agent on the boarder “air – alveoli lungs”. Surfactant consists of phospholipids (negative charge), phosphatidyl-aetanolamins (positive charge), and other proteins and carbons. 50 mm³ of surfactant are distributed on 1 m² of alveoli surfaces.

**Surfactant role is to decrease the surface tension coefficient.**

Surfactants decrease the surface tension of fluid by adsorbing at the liquid-gas interface. They also reduce the interfacial tension between oil and water by adsorbing at the liquid-liquid interface.

Some pathological factors (dust, microorganisms, etc.) can be negatively influenced on the surfactant. It can be followed *atelectasis* or pulmonary collapse.

*Atelectasis* (from Greek “ateles” — incomplete, defective, “ektasis” — extension) means collapse of the expanded lung, defective expansion of the pulmonary alveoli at birth.

We can explain the surfactant role in lungs diseases according to Hook’s law for elastic tissue:

\[
\text{stress} = k \cdot \text{strain}
\]

But for biological tissues we can’t apply this law:

\[
\Delta V \neq k \cdot \Delta P,
\]

where \( \Delta P \) — mechanical stress at breath, \( \Delta V \) — volume change.
The oxygen exchange in the lungs takes place across the membranes of small balloon-like structures called alveoli. These alveoli change their volume with inhalation and exhalation. The behavior of the alveoli is explained by Laplace's law and surface tension. For example, to increase the alveoli lungs volume on \( \Delta V_1 \) it will be necessary to apply non-proportionally \( \Delta P_1 \).

After that the larger change of volume \( \Delta V_2 \) will be caused by non-proportionally smaller change of pressure \( \Delta P_2 \).

According to Laplace law, pressure is inversely proportional to radius of sphere. I. e, with small radius of alveoli lungs (while exhalation), their resistance is great. It means to apply an additional pressure to remove it. And, inversely, with large radius of alveoli lungs their volume changes very slowly.

Thus, while exhalation, the decreases of area of alveoli surfaces lead to the increases of the density of surfactant molecule. It is followed by reduction of surface tension.

The baby's first breath depends on this surfactant concentration. The difficulty of the baby's breath is explained by the incomplete formation of the surfactant. It is very important because all the balloons must be inflated from a collapsed state.
Part 4. The principles of biomechanics

4.1. Deformation and stress

What does the term “deformation” mean?
For each kind of deformation we will introduce a quantity called stress that characterizes the strength of the forces causing the stretch, squeeze, and twist. Stress is a reason, which changes the distance between the molecules of the body, that results in the change of the shape of body, and means its deformation.

We can represent the consequences of the stress as a quantity, which is called a strain, and the latter describes the resulting deformation.

What are the ways to deformate any physical body?
There are some kinds of the stress:
- Tensile stress;
- Compressive stress;
- Bulk (a volume) stress;
- Shear stress.

Between the stress (the reason of deformation) and strain (the consequences) (when stress and strain are small enough) there is a direct proportionality called as an elastic modulus.

\[
\frac{\text{Stress}}{\text{Strain}} = \text{const}
\]  

(4.1.1)

What does the tensile stress mean?
The tensile stress at the cross section can be defined as the ratio of the force \( F \) to the cross section \( A \):

\[
\sigma = \frac{F}{A}
\]  

(4.1.2)

This is scalar quantity because \( F \) is the magnitude of the force.

Hence the SI unit of stress is Newton per square meter: \( \frac{1N}{1m^2} = 1 \text{Pascal} \) (abbreviated Pa)

\[\text{Fig. 4.1.1. Tensile stress}\]

We can estimate the quantity of mechanical deformation.
The simple way can be represented as an elongation \( \Delta l \) which occurs under the stress. The elongation does not occur only at the ends; every part of the bar
is stretched in the same proportion. Hence more effective way is described as the ratio of the elongation $\Delta l$ to the original length $l_0$:

$$\varepsilon = \frac{l-l_0}{l_0} \quad (4.1.3)$$

This ratio is called a tensile strain.

If mechanical stress is the reason for the deformation of a body, the relative elongation of the body is the consequence of such effect. Hence $\varepsilon$ is the *dimensionless quantity* because numerator and denominator are the length and are measured in meters which are cancelled like terms.

**How is the reason of mechanical deformation related to its consequences?**

$$\sigma \sim \varepsilon \Rightarrow y = E\varepsilon, \quad (4.1.4)$$

where $E$ — proportionality constant, as called Young's modulus or elastic modulus.

So, for small deformations:

*Reason = $E \cdot $consequences*.

**What are the units of measurement of elastic modulus?**

If (in projection on horizontal axis):

$$\frac{F}{A} = E \frac{\Delta l}{l_0} = \frac{Fl_0}{A\Delta l}. \text{ Hence, } [E] = \frac{N \cdot m}{m^2} = \frac{N}{m^2} = Pa$$

**What does the compressive stress mean?**

When the forces on the end of a bar are pushes rather than plus, the bar is in compression, and the stress is a compressive stress (fig. 4.1.3). The compressive strain of a bar in compression is defined in the same way as a tensile strain, but $\Delta l$ has opposite direction.

**What does the bulk stress mean?**

If we choose an arbitrary cross section within a fluid (liquid or gas) at rest, the force acting on each side of the section is always perpendicular to it. If we tried to exert a force parallel to a section, the fluid would slip sideways to counteract the effort. When a solid is immersed in a fluid and both are at rest, the forces that the fluid exerts on the surface of the solid are always perpendicular to the surface at each point. The force $F$ per unit of area $S$ on such a surface is called the pressure in the fluid.
What does the shear stress mean?
It is possible to define the shear stress as the force tangent to a material surface, divided by the area on which it acts. On this figure the deformed block has been shifted and rotated so that its bottom edge coincides with that of the underformed block to show the angle clearly.

We can show the graphical relations between the stress and strain.
Part 5. Physics of cell membrane

5.1. PHYSICAL PROPERTIES OF CELL MEMBRANE

As the functions of proteins determine the properties of the cellular membrane, the vital activity of the cell itself and the amount of proteins in the cell must be great in most cases.

5.1.1. BASIC STRUCTURAL PARTS OF THE CELL

All known cells (except for a few specialized cell types) have certain basic components: cellular membrane, cytoplasm, DNA, proteins. They also have several functions.

These functions and abilities are expressed in the cell cycle: the “birth”, growth, reproduction and “death” of individual cells.

The importance of cell membrane can hardly be overestimated, as the cell is a structural free-running unit of any living organism, and for its normal functioning both protection from the adverse effects of the environment and metabolism are essential. The cell membrane has these two major functions (protective and metabolic).

For example, under stress conditions, corticoliberine is secreted in the axons of hypothalamus. Then it causes the production of adrenocorticotropic hormone which decreases the membrane permeability.
Getting into a bad mood also influences the somatic actions, because the central nerve system influences the vegetative (involuntary) nervous system. The mediator of parasympathetic system acetylcholine increases the cell membrane permeability, hence the relaxation of smooth muscles and vasodilatation take place. It is opposite to noradrenaline action.

Death can occur as a result of complete congestion of membrane.

What is the structure, physical and chemical properties of the membrane cell?

According to the modern investigations the cell membrane (or plasma membrane) is a thin, structural layer (very often — bilayer) of lipid and protein molecules that completely envelopes the cell, separates cytoplasm from the surroundings and strictly controls a transport across membrane. The plasma membrane is only about 10 nm thick and may be discerned only faintly with a transmission electron microscope (fig. 5.1.2).

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.

Some attractive notes!

- Total square of all cellular membranes reaches approximately about ten thousand square meters!
- The plasma membrane consists of protein molecules (about 70 % of dry membrane weight), and phospholipid molecules (30 % of dry weight approximately).
**What is the chemical composition of phospholipids?**

The lipid bilayer is formed by the amphiphilic molecules of phospholipids and sphingomyelin in the aqueous phase.

Amphiphilic molecules consist of a water-soluble head (a positively charged \((\text{polar})\) group), linked to two water-insoluble nonpolar tails (by a negatively charged \((\text{phosphate group})\)). Both tails consist of a \((\text{fatty acid})\), each 14–24 carbon groups long.

The head of a phospholipid is attracted to solvent, for example — water (it is \((\text{hydrophilic})\)), due to its polar nature. The nonpolar head is not attracted to water and is the \((\text{hydrophobic})\). Phospholipids form a bilayer, when they are placed in water, where the hydrophobic tails line up against each other (see table 1 below). This forms a membrane with hydrophillic heads on both sides.

<table>
<thead>
<tr>
<th>Monolayer</th>
<th>Bilayer</th>
<th>Vesicle</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Monolayer" /></td>
<td><img src="image2" alt="Bilayer" /></td>
<td><img src="image3" alt="Vesicle" /></td>
</tr>
<tr>
<td>Non polar tail is perpendicular to the surfaces of solvent</td>
<td>Non polar tails orient to each other</td>
<td>Such structure is without protein molecules</td>
</tr>
</tbody>
</table>

Table 5.1

So, lipids mostly contain phospholipids, sphingomyelin, and cholesterol. For example, erythrocyte membranes of a man correspondingly consist of 36, 30 and 22 % of their mass; and also 12 % are in glycolipidoses.

![Fig. 5.1.4. The structure of amphiphilic molecules](image4)

97
5.1.2. TYPES OF STRUCTURE OF THE CELLULAR MEMBRANES

There are some types of structure of the cellular membrane:

- Fluid Mosaic;
- Micelles;
- Monolayers;
- Vesicles.

**Fluid Mosaic Structure**

The model was first introduced by S. J. Singer (1971) as a lipid protein model and extended to include the fluid character in a publication with G. L. Nicolson in “Science” (1972).

The basic composition and structure of the plasma membrane is the same as that of the membranes that surround organelles and other subcellular compartments. This structure is based on *phospholipid bilayer (similar to sea) filled with flowing protein molecules (as iceberg examples)*. Some of these proteins simply adhere to the membrane (*extrinsic* proteins), while others might be said to reside within it or to span it (*intrinsic* proteins — more at *integral membrane protein*). *Glycoproteins* have carbohydrates attached to their extracellular domains. Cells may vary the variety and the relative amounts of different lipids to maintain the fluidity of their membranes despite changes in temperature. In the bilayer *cholesterol* molecules assist in regulating fluidity.

In fact, not all lipid molecules in the cell membrane are “fluid”, in the sense of free to flow. *Lipid rafts* and *caveolae* are examples of more cohesive membrane regions.

*Synapses* are another example of highly structured membrane.

Other self-organizing structures what lipids assume, depend on their concentration and the type of lipid, include *micelles, monolayers* 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. 5.1.5. A micelle](image)
What are the cellular proteins?

Discovered by Berzelius, in 1838, proteins have since come to be understood as amino-acid chains. Proteins are generally large molecules, with molecular masses which sometimes can approach up to 3,000,000 (the muscle protein titin has a single amino acid chain 27,000 subunits long).

Undoubtedly all functions in the living cell depend upon proteins.

The shape of a protein determines its biological activity. A single protein may have varying structure and more than one function.

Proteins have many different biological functions and can be classified according to their biological roles.

1. **Enzymatic Proteins** (with catalytic activity). The most varied and most highly specialized proteins with catalytic activity are enzymes. Virtually all the chemical reactions of organic biomolecules in cells are catalyzed by enzymes. Thousands of different enzymes, each capable of catalyzing a different kind of chemical reaction, have been discovered in different organisms. Digestive enzymes hydrolyze the polymers in food.

2. **Transport Proteins**. These proteins transport other substances. For example, hemoglobin, the iron-containing protein of blood, transports oxygen from the lungs to other parts of the body.

3. **Structural Proteins**. Structural proteins are parts of muscle tissues and very important for support. Collagen and elastin provide a fibrous framework in animal connective tissues. Keratin is the protein of hair, horns, feathers, quills and other skin appendages.

4. **Storage Proteins**. These proteins store amino acids. Ovalbumin is the protein of egg white, used as an amino acid source for the developing embryo. Casein, the protein of milk, is the major source of amino acids for baby mammals.

5. **Hormonal Proteins**. Hormonal proteins coordinate bodily activities. Insulin, a hormone secreted by the pancreas, helps to regulate the concentration of sugar in the blood.

6. **Receptor Proteins**. Receptor proteins are built into the membrane of a nerve cell and they detect chemical signals released by other nerve cells. They are involved in the cell's response to chemical stimuli.

7. **Contractile Proteins**. These proteins are very important in movement. Actin and myosin are responsible for the movement of muscles.

8. **Protect Proteins**. These proteins protect against diseases. Antibodies combat bacteria and viruses.

5.1.3. MECHANISMS OF PERMEATION ACROSS CELL MEMBRANES

*Ion channels* are present in the membranes that surround all biological cells. The existence of ion channels was hypothesized by the British biophysicists Alan Hodgkin and Andrew Huxley as part of their Nobel Prize-
winning theory of the nerve impulse (published in 1952). The existence of ion channels was confirmed in the 1970s with an electrical recording technique, which led to a Nobel Prize to Erwin Neher and Bert Sakmann, the inventors of the technique.

An ion channel is an integral membrane protein or more typically an assembly of several proteins. 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.

As “voltage-gated” channels underlie the nerve impulse and “transmitter-gated” channels mediate conduction across the synapses, channels are especially prominent components of the nervous system.

- **Voltage-gated** channels sense the transmembrane potential and open or close in response to depolarization or hyperpolarization, respectively. Examples include the sodium and potassium voltage-gated channels of nerve and muscle, and the voltage-gated calcium channels that control neurotransmitter release in pre-synaptic endings.
- **Ligand-gated** channels open in response to a specific ligand molecule on the external face of the membrane in which the channel resides. Examples include the “nicotinic” Acetylcholine receptor, AMPA receptor and other neurotransmitter-gated channels.
- **Cyclic nucleotide-gated** channels, Calcium-activated channels and others open in response to internal solutes and mediate cellular responses to second messengers.
- **Stretch-activated** channels open or close in response to mechanical forces that arise from local stretching or compression of the membrane around them; for example when their cells swell or shrink. Such channels are believed to underlie touch sensation and the transduction of acoustic vibrations into the sensation of sound.
- **G-protein-gated** channels open in response to G protein-activation via its receptor. For example the “muscarinic” Acetylcholine receptor

**Transmembrane receptors** are integral membrane proteins, which reside and operate typically within a cell's plasma membrane, but also in the membranes of some subcellular compartments and organelles. Binding to a signalling molecule or sometimes to a pair of such molecules on one side of the membrane, transmembrane receptors initiate a response on the other side. In this way they play a unique role in cellular communications and signal transduction.

In what way can we explain the permeability of some part of ions and other chemical compounds through ion channels? A moving charge particle (for example, anion or cation) has an electromagnetic field. When this particle approaches the membrane cell, we can state that between the given ion and
protein polypeptide molecules (they line the internal surface of the ion channel and form the size and shape such like the dimensions of an ion channel), the electromagnetic coupling arises. This phenomenon results in the conformational alteration of size and shape of the molecules and correspondingly the change of the size of an ion channel. It means, that the permeability of membrane cell changes. Thus, the permeability of a membrane depends upon diameter of an ion channel. We will call such phenomenon as ECI (electroconformational interaction). That is the conformational alteration of the proteins molecules change:

1. Dimensions of membrane.
2. The distribution of frequency of the electrical and magnetic fields in a membrane.

This refers to the electromechanical way of change of cellular membrane permeability.

Since the ion channels have certain size (0.35–0.8 nm), such molecules and ions (having the same size as ion channels) can only permeate across the membrane cell.

**What are the mechanisms of permeation across cell membranes?**

There are two main ways of permeation of the particles across the membrane cells:

1) by means of solution of species in the phospholipids of membrane cells;
2) by means of passing through the ion channels.

The choice of the way depends upon many factors and the main of them are polarity and non-polarity of a matter. That is electromagnetism which is the basic phenomenon of a particle passing across the membrane. The fact is that the lipids of the membranes are non-polar chemical compounds. Hence, non-polar solvents such as methyl, ethyl, phenyl compounds, are dissolved easily in such non-polar chemical compounds. At the same time the size of non-polar molecules is not important, when their transport takes place across the cell membrane.

Then polar charge chemical compounds (hydroxyl, carboxyl and amine compounds are the centers of their positive and negative electrical charges, and these centers do not correspond to each other because of their non-symmetric chemical links) permeate into the cytoplasm through the ion channels.

Such chemical compounds have a high permittivity.

The transport of ions across the cell membrane can be demonstrated in the given table.

<table>
<thead>
<tr>
<th></th>
<th>Polar charge particles</th>
<th>Non-polar charge particles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Permittivity $\varepsilon$</td>
<td>$\uparrow \varepsilon$</td>
<td>$\downarrow \varepsilon$</td>
</tr>
<tr>
<td>Solubility</td>
<td>$\downarrow$ solubility in lipids</td>
<td>$\uparrow$ solubility in lipids</td>
</tr>
<tr>
<td>The way to permeate across a membrane</td>
<td>ion channels</td>
<td>solubility in lipids</td>
</tr>
</tbody>
</table>
Hence, there are two factors closely related to the membrane permeability:
1. ECF (electrical conformational factor).
2. Chemical. That is dissolving of substances in the lipids of the membranes.

The functions of proteins determine the properties of the cellular membrane in most cases. According to the modern concepts, the cellular membranes are about 70 % of proteins.

5.2. MEMBRANE TRANSPORT

The membrane transport is the mechanism of ions flux across cellular membranes. There are two major types of the membrane transport: passive (which does not require the electrochemical energy of the hydrolysis ATP) and active.

5.2.1. TYPES OF PASSIVE TRANSPORT

The main types of the passive transport are the following:
- Diffusion;
- Osmosis;
- Filtration.

Diffusion is the flux of particles from high to low concentration, i. e. down along a concentration gradient.

Let us consider the container of two types of liquids, separated by a semi-permeable partition. If the partition is removed as in fig. 5.2.1, the liquids will mix. In time a uniform mixture of molecules of two liquids will be produced in the given container.

Fig. 5.2.1. A divided container: solute and water move in opposite directions because of the concentration gradients

Fig. 5.2.2. A person on the sledge rolls down from point B towards point A due to the work of gravity

In the same way, a person on the sledge rolls down from point B towards point A due to the work of gravity (fig. 5.2.2), and the velocity will be directly proportional to the altitude difference
Look at this picture of a cell alongside a blood vessel (fig. 5.2.3).

Fig. 5.2.3. Diffusion allows us to explain how the lungs and blood system work, as well as many forms of transport.

**What is the relation between the intensity of diffusion and the main physical constants?**

Intensity of some flux is the ratio of this flux by area

\[ J = \frac{\text{Flux}}{\text{Area}} \]

(5.2.1)

According to the **law of diffusion**, the transfer of solute atoms can be described by the following equation:

\[ J = -D \frac{dc}{dx} \]

(5.2.2)

where \( J \) is the particle flux, \( C \) is the concentration of the solute, \( D \) is the diffusion coefficient, \( x \) is the distance into the substrate, and \( t \) is the diffusion time. The negative sign indicates that the diffusing mass flows in the direction of decreasing concentration.

\[ D = \frac{1}{3} \frac{\delta^2}{\tau} \]

where \( \delta \) — average distance between the molecules, \( \tau \) — average time of their “rest” state.

Any diffusion processes are described for **biological cellular membranes** by the following equation:

\[ \frac{dm}{dt} = -PA(C_1 - C_2) \]

(5.2.3)

Where \( P \) means the permeability coefficient, \( A \) — area of the membrane’s surface, \( C \) — concentration.

But **Fick’s equation** concerns neither the types of charge nor the **chemical nature** of a substance. What is the relation between the intensity of diffusion and its parameters?

To answer this question, it is necessary to consider such important parameters of a substance as the **chemical potential**. So, the chemical potential of some substance equals:
Page 104

\[ \mu_i = \mu_{0i} + RT \ln \frac{C_i}{C_e} \]  (5.2.4)

For the relation between \( \frac{dc}{dx} \) and \( \mu \), it is necessary to differentiate the left part from the given equation \( \mu_i \):

\[ d\mu_i = RTd \ln C, \text{ where } C = \frac{c_i}{c_e} \Rightarrow d\mu_i = RT \frac{dc}{c} \]

Let's divide the left and right parts of the given equation by \( dx \):

\[ \frac{d\mu_i}{dx} = \frac{RT \frac{dc}{dx}}{c} \]
\[ \frac{dc}{dx} = \frac{c \frac{d\mu_i}{dx}}{RT} \]

It follows:

\[ J = -D \frac{dc}{dx} \Rightarrow J = -\frac{Dc}{RT} \frac{d\mu_i}{dx} \]  (5.2.5)

If the coefficient of diffusion is directly proportional to the temperature, the given ratio \( \frac{D}{RT} \) is determined as the mobility of the particle \( U_m \). Then:

\[ J = -U_m c \frac{d\mu}{dx} \]  (5.2.6)

It is a formula which considers the influence of the electrical factor in the diffusion process. If there is a concentration gradient for the ion, and there is a voltage gradient across the membrane, the \textit{total} flux of the ion will be \textbf{the sum of the diffusional flux and the electrical flux}.

For diffusion (taking into account influence electrical factor) there is Nernst-Plank equation:

\[ J = -(D \frac{dc}{dx} + FZU_m c \frac{d\varphi}{dx}) \]  (5.2.7)
Diffusion — the flow of SOLUTE particles due to the concentration gradient: from high concentration part (I) to low concentration part (II).

Osmosis — the flow of SOLVENT particles due to the concentration gradient: from high concentration part (II) to low concentration part (I). Osmosis — is the diffusion too! But it is the diffusion of solvent molecules exactly (arrow b).

Some conclusions!
- Diffusion can be defined as a flow of particles: a) solute; b) solvent molecules (osmosis) due to the concentration gradient. Such transport does not require the additional energy.
- Diffusion and Osmosis — two opposite directed processes (because the concentration of both solvent molecules and solving ones has opposite symmetrical pattern — the increasing solvent molecules concentration, the decreasing solving ones).

- If two solutions of different concentration are separated by a semi-permeable membrane which is permeable to the smaller solvent molecules but not to the larger solute molecules, then the solvent will tend to diffuse across the membrane from the less concentrated to the more concentrated solution. This process is called osmosis.

Let’s compare two equations below:

<table>
<thead>
<tr>
<th>Diffusion</th>
<th>Osmosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \frac{dm}{dt} = -PA(C_1 - C_2) )</td>
<td>( \frac{dm}{dt} = kA(P_1 - P_2) )</td>
</tr>
</tbody>
</table>
The difference in the sign of “minus” characterizes the opposition in directions of the solute molecules flux and solvent molecules flows.

So, we can define an osmosis from two “sides” — from the side of considering the solute (white) molecules, and from the side of studying the solvent (smaller black) molecules.

First side: osmosis is the solvent molecules flow from low concentration of solute white molecules (right part of vessel) towards high concentration of this molecules (left part of vessel).

Second side: osmosis is the solvent molecules diffusion from high concentration of solvent molecules (right part of vessel) towards low concentration of this molecules (left part).

All depends upon the points of view on molecule’s type — solute or solvent, and their concentrations.

Osmosis has a great importance in biological processes where the solvent is water. The transport of water and other molecules across biological membranes is essential to many processes in living organisms. The energy which drives the process is usually discussed in terms of osmotic pressure.

We can calculate an osmotic pressure using the following equation:

\[ P_{\text{osmotic}} = iCRT, \]

(5.2.8)

where \( i \) — isotonic coefficient of the solution, \( R \) (gas constant) = 8.3145 J/k mol is the normal gas constant.

The osmotic pressure of human blood is about seven atmospheres!

Let us consider the last type of passive transport such as filtration. Filtration can be defined as the flow of liquid according to the main pressure gradient. The filtration speed we can calculate due to Hagen-Poiseuille equation:

\[ \frac{dV}{dt} = \frac{\pi R^4}{8 \eta l} (P_1 - P_2) \]

(5.2.9)

where \( P_1 - P_2 \) — pressure difference between the beginning and the end of membrane pore.

The filtration and osmosis phenomena play very important role in mechanism of the water transport between blood and conjunctive tissue.

The colloids of plasma blood play main role in the water molecules metabolism. The osmotic pressure, which is determined by the high-molecular colloids of plasma is called the oncotic pressure. This pressure has small value in comparison with the osmotic pressure (approximately in 200 times less, or 7 % of osmotic pressure).

There is the hydrostatic pressure gradient on the level with oncotic pressure one. This hydrostatic pressure difference exists due to the heart work (left atrium constriction results in blood filling of the artery).
Hydrostatic pressure in each arterial part of blood vessel prevails oncotic pressure on 10 mm Hg.

In central part of blood vessel we can see the equilibrium state: hydrostatic pressure is equal to oncotic pressure.

In venous part of each blood vessel situation is inverse in comparison with arterial one: oncotic pressure prevails hydrostatic pressure on 10 mm Hg.

Thus, in arterial and venous parts different pressure types have the priority.

It is logically to suppose, that water molecules transport will have direction from plasma to lymph in arterial part of blood vessel according to the filtration (hydrostatic pressure gradient).

But in venous part of vessel we can see opposite directed process of water transport — from lymph to plasma due to the oncotic pressure gradient.

Different pathological influences can break the given water exchange.

It is very important to note the following!

The decreasing of oncotic pressure can be explained by the decreasing of colloids concentration in plasma, because:

\[ P_{\text{o}} = iCRT \]

Such processes result in the decreasing of concentration of the plasma colloids based on the destroying of the cellular membranes. I. e., the permeability of any membrane is increased, and protein molecules will flow outside cytoplasm of cell. It happens in bleeding, in shock, under the radiation sickness.

For example, depolarization of some globular proteins in radiation sickness leads to the increasing of permeability of the capillary walls. It means the high molecular colloids flow from plasma. It is followed by the decreasing of the oncotic pressure. Thus, the hydrostatic pressure increasing will result in the transport of water molecules from plasma to connective tissue and lymph (the edema).
In opposite case, inflammation, muscle work leads to the intensification of glicolysis and proteolysis (sodium and potassium ions are released from connection with protein molecules). It will result in the colloid concentration increasing.

Let us speak about active transport across cellular membrane.

5.2.2. TYPES OF THE ACTIVE TRANSPORT

In some organs, e. g. the kidney, the body has to move solutes such as sodium ions against their concentration gradients, effectively reversing the flow of water.

Active transport typically moves molecules from low concentration to high, or against their concentration gradient. Active transport requires metabolic energy in the form of the active chemical ATP which is produced as a result of respiration.

There are two main types, primary and secondary. Primary transport involves the consumption of metabolic energy (often in the form of ATP) and is directly coupled to movement across a membrane, independent of any other species. Secondary transport concerns the diffusion of one species across a membrane to drive the transport of another. Transporters generally are membrane-spanning or “transmembrane” proteins.

Primary active transport directly uses energy to transport molecules across a membrane. Most of the enzymes that perform this type of transport are transmembrane ATPases. A primary ATPase universal to all cellular life is the sodium-potassium pump, which helps to maintain the cell potential.

In secondary active transport, there is no direct coupling of ATP; instead, the electrochemical potential difference created by pumping ions out of cells is used. The two main forms of this are counter- and co-transport.

The proteins in the membrane can also act as pumps!

- Move ions or small molecules from low concentration to high concentration (i.e. up their gradients).
- Require cellular energy, usually as ATP.
- Saturates when substance reaches high concentrations due to lack of available protein.
- Other pumps include the Ca-ATPase, and the H-ATPase.
5.3. THE CELL MEMBRANE POTENTIALS

5.3.1 RESTING POTENTIAL

- A major characteristic of all living cells is their generation of a standing voltage difference between the inside and outside of the cell.

As carrier of charge, the most abundant ions found on either side of the membrane are Na\(^+\), Cl\(^-\), K\(^+\) and organic anions (A\(^-\)). The specific distribution of the single ions evokes characteristic properties of a cell. Na\(^+\) and Cl\(^-\) — are more concentrated outside the cell and K\(^+\) and A\(^-\) — are more abundant inside the cell.

Charge separation across the membrane of a cell produced by outward and inward movements of ions and molecules leads to thin clouds of positive and negative ions spread over the inner and the outer surfaces of the cell membrane.

To maintain the unequal distribution of the ionic species leading to the membrane potential several principles are applied.

At rest the membrane is slightly more permeable to K\(^+\) than to Na\(^+\). Since the concentration of K\(^+\) is greater inside the cell than outside, this causes K\(^+\) ions to move out of the cell and build up “+” charges on the outside of the plasma membrane. If potassium ions (“+” charge) are higher inside the cell, proteins and amino acids (“−” charge) are normally higher inside the cell. Since we always refer to the charge on the inside of the plasma membrane, the inside of the membrane is slightly negative. The resting potential indicates what is happening with the neuron at rest.
Different neurons have different rest potentials, but for humans it is typically around -70 mV (fig. 5.3.1).

Ca$^{2+}$ and Mg$^{2+}$ are also present in real cells but their contribution to the resting membrane potential is negligible.

In 1949 the first membrane potential ever recorded by Ling and Gerard was obtained from the preparation of frog sartorius muscle fibres. Thus, experimentation with the frog sartorius muscle is an excellent introduction to the subject of membrane potentials gaining insight into applied and practical Neurosciences, focusing on the recording membrane potentials from large and hard cells of muscle.

### Table 5.3

<table>
<thead>
<tr>
<th>Ion</th>
<th>Extracellular concentration, (mmol/L)</th>
<th>Intracellular concentration, (mmol/L)</th>
<th>Relative permeability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Na$^+$</td>
<td>150</td>
<td>15</td>
<td>low</td>
</tr>
<tr>
<td>K$^+$</td>
<td>5</td>
<td>150</td>
<td>high</td>
</tr>
<tr>
<td>Cl$^-$</td>
<td>120</td>
<td>10</td>
<td>low</td>
</tr>
</tbody>
</table>

\[
\frac{C_{K^+}}{C_{K^+}'} = 20 - 40
\]

**Fig. 5.3.2.** The generation of the rest potential
**Why does the diffusion of potassium ions stop?**

The similar flux of potassium ions will have lasted by the net force between the electric forces field and the forces causing diffusion (concentration gradient). The more amount of potassium ions will diffuse outward the more positive charge of the inner side of the membrane and the greater repulsive force acts from outside upon the potassium flux.

The **Nernst equation** allows us to predict the cell potential for voltaic cells under conditions rather than standard conditions of 1M, 1 atmosphere, 25 °C. The effects of different temperatures and concentrations may be tracked in terms of the Gibbs free energy change \( \Delta G \).

This **free energy change** depends upon the temperature and concentrations according to

\[
\Delta G = \Delta G_0 + RT \ln Q
\]

where \( \Delta G_0 \) is the free energy change under **standard conditions** and \( Q \) is the thermodynamic reaction quotient. The free energy change is related to the cell potential \( \phi_{cell} \) by

\[
\Delta G = -ZF \phi_{cell}
\]

Thus, for non-standard conditions:

\[
-zF \phi_{cell} = -zF \phi_{ocell} + RT \ln Q
\]

or

\[
\phi_{cell} = \phi_{ocell} - \frac{RT}{zF} \ln Q
\]

where \( R \) means the gas constant, \( T \) the temperature, \( z \) the valence of the ion and \( F \) the Faraday constant.

So, the conditions for potassium to be in equilibrium across the cell membrane can be described by the Nernst equation, where a concentration gradient is equated to an electrical potential, provided that all the channels are specifically open for K-ions.

\[
\phi = \frac{RT}{Fz} \ln \left[ \frac{[K]}{[K]} \right] = 58 \ln \left[ \frac{[K]}{[K]} \right]
\]

\([K]_e\) and \([K]_i\) are the external and internal potassium ion concentrations.

According to the original **theory of Bernstein** the cell membrane permeability at \( \phi_{rest} \) is minimal to Na\(^+\), Cl\(^-\) and Ca\(^{2+}\) while it is high to K\(^+\). However, notably skeletal muscle cells like the sartorius frog muscle fibres, have also a high resting permeability to Cl\(^-\) with opposing effects leading to following relation:

\[
\frac{RT}{Fz} \ln \left[ \frac{[K]}{[K]} \right] = -\frac{RT}{Fz} \ln \left[ \frac{Cl^-}{Cl^-} \right]
\]
The Goldman-Hodking-Katz (GHK) equation results by implementing concentrations outside \([X]_e\) and inside \([X]_i\), the cell and membrane permeabilities for each ion \(P_x\) (with the approximation for room temperature and conversion into log):

\[
\varphi = -\frac{RT}{F} \ln \left( \frac{P_K \left[ K^+ \right]_e + P_{Na} \left[ Na^+ \right]_e + P_{Cl} \left[ Cl^- \right]_e}{P_K \left[ K^+ \right]_i + P_{Na} \left[ Na^+ \right]_i + P_{Cl} \left[ Cl^- \right]_i} \right) \tag{5.3.7}
\]

In applying the GHK equation to nerve cells following assumption must be made:
- A voltage gradient across the membrane is uniform.
- Overall net current flow across the membrane is zero for each species.
- Membrane must be in steady-state conditions.
- No active transport mechanism occur.
- No interactions between monovalent ions occur.
- Membrane surface charges are ignored.
- No impeding interaction with the membrane occurs.
- Ionic activities can be replaced by their concentrations.

**How long can a cell be at rest?**
For the stimulation of a cell it is necessary to apply some stimulus of a certain force. If the amplitude of such electric influence surpasses the *gate threshold* then the necessary prerequisites for the generation of the action potentials will be created.

### 5.3.2. ACTION POTENTIAL

The transmembrane voltage changes that take place during an action potential result from changes in the permeability of the membrane to specific ions, the internal and external concentrations of which are in imbalance.

An action potential involves a *depolarization*, a *repolarization* and finally a *hyperpolarization* (or “undershoot”). In response to the appropriate stimulus, the cell membrane of a nerve cell goes through a sequence of depolarization from its rest state followed by repolarization to that rest state. In the sequence, it actually reverses its normal polarity for a brief period before reestablishing the rest potential.

The action potential sequence is essential for neural communication. Action potentials are triggered by an initial depolarization to the point of *threshold*. Typically action potential initiation occurs at a synapse, but may occur anywhere along the axon. In his discovery of “animal electricity” Luigi Galvani elicited an action potential through contact of his scalpel with the motor nerve of a frog. He was dissecting, causing one of its legs to kick as in life.

The simplest action in response to thought requires many such action potentials for its communication and performance. For modeling the action
potential for a human nerve cell, a nominal rest potential of -70 mV will be used. The process involves several steps:

1. A stimulus is received by the dendrites of a nerve cell. This causes the Na⁺ channels to open. If the opening is sufficient to drive the interior potential from -70 mV up to -55 mV, the process continues.

2. Having reached the action threshold, more Na⁺ channels (sometimes called voltage-gated channels) are opened. Since the concentration of Na⁺ is greater outside the cell than inside, this causes Na⁺ ions to move into the cell and to build up + charges on the inside of the plasma membrane. The Na⁺ influx drives the interior of the cell membrane up to about +30 mV. The process to this point is called *depolarization* (fig. 5.3.3).

![Sodium ions diffusion](image)

![depolarization](image)

If the net-movement is stopped the sodium equilibrium potential (E_{Na}) is reached. At E_{Na} the potential gradient and the concentration gradient balance one another exactly, evoking a dynamic equilibrium.

3. The Na⁺ channels are closed and the K⁺ channels are opened. Since the K⁺ channels open more slowly, the depolarization has time to be completed. Having both Na⁺ and K⁺ channels open at the same time would drive the system toward neutrality and prevent the creation of the action potential.

4. With the open K⁺ channels, the membrane begins to *repolarize* back toward its rest potential.

5. The *repolarization* (fig. 5.3.4) typically overshoots the rest potential to about -90 mV. This is called hyperpolarization and would seem to be counterproductive, but it is actually important in the transmission of information. Hyperpolarization prevents the neuron from receiving another stimulus during this time, or at least raises the threshold for any new stimulus. Part of the importance of hyperpolarization is in preventing any stimulus already sent up an axon from triggering another action potential in the opposite...
direction. In other words, hyperpolarization assures that the signal is proceeding in one direction.

After hyperpolarization, the Na$^+$/K$^+$ pumps eventually bring the membrane back to its resting state of -70 mV.

Changes in membrane permeability and the onset and cessation of ionic currents reflect the opening and closing of “voltage-gated” ion channels, which provide portals through the membrane for ions. Residing in and spanning the membrane, these enzymes sense and respond to changes in transmembrane potential.

So, in the axon fibers of nerves, depolarization results from the inward rush of sodium ions, while repolarization and hyperpolarization arise from an outward rush of potassium ions. Calcium ions make up most or all of the depolarizing currents at an axon's pre-synaptic terminus, in muscle cells (including the heart's) and in some dendrites.
Action potentials are measured with the recording techniques of electrophysiology. In the case of an archetypal nerve action potential on an oscilloscope, the relatively large swing to a more positive value, followed by the repolarization recovery and undershoot together trace an arc that could be described as a distorted sine wave — or like the blips on hospital EKG machines that can be seen on TV (these EKG waves are a smear of all the action potentials in one heartbeat, so they enact more slowly than any individual “A.P.” and have a somewhat more complicated shape). In an unmyelinated axon that is “firing” an action potential, the transmembrane potential at any instant will vary from point to point along the fiber, with its amplitude depending on whether the A.P. wave has reached that point or passed it, and how long ago. A recording from a single point will show the various stages of the action potential enacted — depolarization, repolarization, hyperpolarization — as the wave passes.

Prototypically, depolarization and repolarization together are complete in about two milliseconds, while undershoots can last hundreds of milliseconds, depending on the cell. In neurons, the exact length of the roughly two-
millisecond delay in repolarization can have a strong effect on the amount of neurotransmitter released at a synapse. The duration of the hyperpolarization determines a nerve's "refractory period" (how long until it may conduct another action potential) and hence the frequency at which it will fire under continuous stimulation. Both of these properties are subject to biological regulation, primarily (among the mechanisms discovered so far) acting on ion channels selective for potassium.

In pacemaker and other cardiac muscle cells, inward calcium currents determine shape and duration of the plateau phase, which in turn controls the strength and duration of contraction.

5.4. A NERVE IMPULSE TRANSMITTING ALONG AN AXON

While using some drugs, for example, psychotropic, the reaction of a man on the exogenous irritant becomes slower.

In the fine fibers of simple (or "unmyelinated") axons, action potentials propagate as waves, which travel at speeds up to 120 meters per second.

The propagation speed of these "impulses" is faster in fatter fibers than in thin ones, other things being equal. In their Nobel prize-winning work uncovering the wave nature and ionic mechanism of action potentials, Alan Hodgkin and Andrew Huxley performed experiments on the "giant fiber" of Atlantic squid. Responsible for initiating flight, this axon is fat enough to be seen without a microscope (100 to 1000 times larger than typical). This is assumed to reflect an adaptation for speed. Indeed, the velocity of nerve impulses in these fibers is among the fastest in nature.

What are the ways to transmit a nerve impulse along an axon?

A nerve cell is like a receiver, transmitter and transmission line with the task of passing a signal along from its dendrites to the axon terminal bundle.
The stimulus triggers an action potential in the cell membrane of the nerve cell, and that action potential provides the stimulus for a neighboring segment of the cell membrane. When the propagating action potential reaches the axon, it proceeds down that “transmission line” by successive excitation of segments of the axon membrane.

Just the successive stimulation of action potentials would result in slow signal transmission down the axon. The propagation speed is considerably increased by the action of the myelin sheath.

Many neurons have insulating sheaths of “myelin” surrounding their axons, which enable action potentials to travel faster than in unmyelinated axons of the same diameter. The myelin sheath around the axon prevents the gates on that part of the axon from opening and exchanging their ions with the outside environment.

There are gaps between the myelin sheath cells known as the nodes of Ranvier. The myelin sheathing normally runs along the axon in sections about 1 mm long, punctuated by unsheathed nodes of Ranvier.

It is important to note that at those uncovered areas of the axon membrane, the ion exchange necessary for the production of an action potential can take place. As the salty cytoplasm of the axon is electrically conductive and the myelin inhibits charge leakage through the membrane, depolarization at one node is sufficient to elevate the voltage at a neighboring node to the threshold for action potential initiation. Thus in myelinated axons, action potentials do not propagate as waves – the action potential at one node is sufficient to excite a response at the next node, so the nerve signal can propagate faster by these discrete jumps than by the continuous propagation of depolarization/repolarization along the membrane. This enhanced signal transmission is called salutatory conduction.

The disease multiple sclerosis (MS) is due to a breakdown of myelin sheathing, and degrades muscle control by destroying axons' ability to conduct action potentials.

The membrane capacitance per unit of length of a myelinated axon is much smaller than in an unmyelinated axon. Therefore, the myelin sheath increases the conduction velocity. The resistance of the axoplasm per unit of length is inversely proportional to the cross-sectional area of the axon and thus to the square of the diameter. The membrane capacitance per unit of length is directly proportional to the diameter. As the time constant formed from the product controls the nodal transmembrane potential, it is reasonable to suppose that the velocity would be inversely proportional to the time constant. On this basis the conduction velocity of the myelinated axon should be directly proportional to the diameter of the axon. This is confirmed in fig. 5.4.2, which shows the conduction velocity in mammalian myelinated axons as linearly
dependent on the diameter. The conduction velocity in myelinated axon has the approximate value shown:

\[ \vartheta = 6d, \text{ where } \vartheta = \text{velocity [m/s]}, d = \text{axon diameter [µm]} \]  

(5.4.1)

It is mean that blockage of the nodes of Ranvier by some psychotropic leads to slow signal transmission along the axon.
Part 6. Electric fields and physical principles of the electrography

6.1. THE GENERAL CHARACTERISTICS OF ELECTRIC FIELDS

The electric force acting on a point charge $q_0$ as a result of the presence of a second point charge $q$ is given by Coulomb's Law:

$$\vec{F} = k \frac{q_0 q}{r^2},$$

(6.1.1)

where $k$ is a proportionality constant (Coulomb’s constant):

$$k = \frac{1}{4 \pi \varepsilon_0} \approx 9 \cdot 10^9 \frac{N \cdot m}{C^2}$$

(6.1.2)

In SI units, $F$ is in newtons (N), $q$ is in coulombs (C), $r$ is in meters (m), and $k = 1/4 \pi \varepsilon_0$, where $\varepsilon_0$ is the permittivity of free space ($8.854 \cdot 10^{-12}$ farads [F]/m), or $k = 9 \cdot 10^9$ m/F.

Coulomb's law is a vector equation and includes the fact that the force acts along the line joining the charges. If a positive charge $q_0$ is fixed at some point in space, any other positive charge $q$ (test charge) which is brought close to it will experience a repulsive force. The repulsive force $F$ is directly proportional to the test charge $q$ (in each point of electric field):

$$F \sim q, \text{ or } \vec{F} = \vec{E} q,$$

(6.1.3)

Thus, the coefficient of proportionality $\vec{E} = \frac{\vec{F}}{q}$ (are vector quantities with definite magnitude and direction) does not depend on the point charge $q_0$ and is called the electric field strength or electric field intensity. So, the electric field strength is defined as the electric force per unit of charge.

If the electric field intensity is a vector, what is the direction of this vector?

The direction of the field is taken to be the direction of the force it would exert on a positive test charge. The electric field is radially outward from a positive charge and radially inward to a negative point charge.

![Figure 6.1.1](image.png)

Fig. 6.1.1. The electric field strength
An electric field line is an imaginary line or curve drawn through a region of space. A streamline is a line or curve whose tangent at any point is in the direction of the velocity of the fluid at that point. Electric field lines show the direction of $\vec{E}$ at each point of space. Their spacing gives a general idea of the magnitude of $\vec{E}$ at each point. Where $\vec{E}$ is strong, we draw lines bunched closely together. At any particular point, the electric field has a unique direction, so only one field line can pass through each point of the field: field lines never intersect. The direction of the total electric field at every point is along the tangent to the electric field line passing through the point.

Because $\vec{F} = \frac{kq_0q}{r^2}$, and $\vec{E} = \frac{\vec{F}}{q}$, thus

$$\vec{E} = \frac{kq_0}{r^2}$$ (6.1.4)

For more than two electrical charge carriers, the force on one charge, $q$, from all others is obtained by the summation of the separate forces (the principle of superposition). For example, given three point charges in free space ($q$, $q_2$, $q_3$), the force on $q$ becomes:

$$\vec{F} = \left[k(\frac{qq_2}{R_2^2})\vec{a}_{R_2} + k(\frac{qq_3}{R_3^2})\vec{a}_{R_3}\right]$$

where $\vec{a}_R$ is the unit vector in the direction of $R$. For example, if $q$ and $q_2$ lie on the x-axis, the unit vector is $\langle 1,0,0 \rangle$ in 3D Cartesian coordinates.

What is the quantitative factor of the relation between the two point charges?

On the other hand, if a positive charge $q$ is fixed at some point in space, any other positive charge which is brought close to it will have the potential energy.

The electric potential is defined as the work done in moving a charge unit from point A to point B in an electric field.

To calculate the point charge potential it is necessary to calculate the work to displace the charge from point A towards point B as below:

$$W = \int_{AB}FdW = \int_{A}^{B}\frac{q dq}{r}dr = q_0q\left[\frac{1}{2} - \frac{1}{r_2}\right] = q_0q\left[\frac{1}{r_2} - \frac{1}{r_1}\right] = q\left(\frac{q_0}{r_1} - \frac{q_0}{r_2}\right)$$ (6.1.5)

The work done does not depend on the path of motion from point A to point B.

The ratio

$$\frac{q_0}{r} = \varphi$$ (6.1.6)

is the potential of the point charge.

Then:
Electric potential is measured in volts, where 1 volt = 1 joule/coulomb.

If \( Q_2 = 0 \), we can write \( \phi_1 = \frac{W}{q} \). (Setting a zero potential at some infinite point B from point A, the electric potential at A is given by \( V_A = \frac{W}{q} \), where \( W \) is the work done by an external agent in moving the particle from infinity to A).

**Electric potential is a magnitude only** (it has no direction) and it is defined in relation to positive charge. A positive electric potential measures the work done in moving a positive charge toward another positive charge (pushing the charge against a repulsion force) and a negative electric potential measures the work done in moving a positive charge away from a positive charge (releasing the charge into a repulsion force).

\[
W = q \left( \frac{Q_0}{r_1} - \frac{Q_0}{r_2} \right) = qV. \tag{6.1.7}
\]

**Fig. 6.1.2. Force and energy characteristics of electrical field**

**What does the term "grad" mean?**

Over a surface \( (S) \) the relationship between electric potential (voltage, \( \phi \)) and electric field strength (\( \vec{E} \)) is given by:

\[
\phi = -\int EdS
\]

Solving for \( \vec{E} \), using the inverse relationship between integrals and derivatives, we have:
Given measures of $\phi$ from scalp electrodes in 3D Cartesian coordinates, $\phi = \phi(x,y,z)$, we can calculate the $x$, $y$, $z$ components of $\vec{E}$ from the partial derivatives ($\partial$) of $\phi(x)$, $\phi(y)$, $\phi(z)$:

$$E_x = -\frac{\partial \phi}{\partial x}; \quad E_y = -\frac{\partial \phi}{\partial y}; \quad E_z = -\frac{\partial \phi}{\partial z} \quad (6.1.10)$$

These equations give the first partial derivatives of a scalar field with respect to the $x$, $y$, $z$ axes, which are known as the directional derivatives, and they measure the rate of change or gradient in the direction of $x$, $y$, $z$. Thus, in terms of the Cartesian unit vectors $\vec{i} = /1,0,0/$, $\vec{j} = /0,1,0/$, $\vec{k} = /0,0,1/$ or $|\vec{i}| = |\vec{j}| = |\vec{k}| = 1$, the electric field strength is given by:

$$\vec{E} = E_x\vec{i} + E_y\vec{j} + E_z\vec{k} = -((\partial \phi / \partial x)\vec{i} + (\partial \phi / \partial y)\vec{j} + (\partial \phi / \partial z)\vec{k}) \quad (6.1.11)$$

In compact vector notation, we have:

$$\vec{E} = -\nabla \phi = -\text{grad}(\phi) \quad (6.1.12)$$

where $\nabla$ ("del") is the vector partial differential operator.

Mathematically, it is true that a “level surface” for a function is always perpendicular to the gradient.

### 6.2. The Einthoven's Theory of Electrocardiographic

**When myocardial fiber can be represented as an electric dipole, then it is possible to estimate the work of cardiac muscles in the quantitative aspect.**

**What is the Heart's Electrical Sequence?**

Previously we have considered the formation biopotentials of cells. Hence, the sum of biopotentials of cells is the biopotentials of the adequate organs. The rhythmic contractions of the heart which pump the life-giving blood occur in response to periodic electrical control pulse sequences.

In the upper part of the right atrium of the heart is a specialized bundle of neurons known as the sinoatrial node (SA node). Acting as the heart's natural pacemaker, the SA node “fires” at regular intervals to cause the heart of beat with a rhythm of about 60 to 70 beats per minute for a healthy, resting heart. The electrical impulse from the SA node triggers a sequence of electrical events in the heart to control the orderly sequence of muscle contractions that pump the blood out of the heart. So, the synchronized electrical sequence of the heart is initiated by the SA node, the heart's natural pacemaker. The firing of the SA node sends out an electrical impulse via its neurons to the right atrium, left...
atrium, and AV node simultaneously. Since the right atrium is closer to the SA node, it depolarizes first, resulting in pumping action by the right atrium before the left atrium. At the AV node, the impulse is delayed to allow for the ventricles to fill up with blood. After the delay, the AV node sends the impulse to the Bundle of His and the Purkinje fibers. This triggers the contraction of the ventricles to send blood either to the lungs or out to the body.

Fig. 6.2.1. The Heart's cross-section

The depolarization and repolarization of the SA node and the other elements of the heart's electrical system produce a strong pattern of voltage change which can be measured with electrodes on the skin. Voltage measurements on the skin of the chest are called an electrocardiogram or ECG.

Fig. 6.2.2
While it is the norm for nerve cells that they require a stimulus to fire, the SA node can be considered to be “self-firing”. It repetitively goes through a depolarizing discharge and then repolarizes to fire again. *This action is analogous to a relaxation oscillator in electronics.* Such an oscillator is routinely used to produce a periodic flash from light. The circuit involves a capacitor which is charged by the energy of a battery (roles played by the membranes of the SA node and the ion transport processes which repolarize it and a resistor which controls the rate of flashing of the light.

The electrical potentials (voltages) that are generated in the body have their origin in membrane potentials where *differences in the concentrations* of positive and negative ions give a localized separation of charges. This charge separation is called *polarization*. Changes in voltage occur when some event triggers a depolarization of a membrane, and also upon the repolarization of the membrane.

At rest the charge on the inside of the plasma membrane is slightly *negative*, whereas the process of exciting (Na$^+$ ions movement into the cell) builds up positive charges on the inside of the plasma membrane. The voltage as the difference between the potentials of the inside and the outside of the plasma membrane is an electric dipole (fig. 6.2.3).

![Fig. 6.2.3. Myocardium as an electric dipole](image)

**What is the electric dipole?**

An electric dipole is a *pair* of point charges with equal magnitude and opposite sign (a positive charge and negative charge). The electric field of an *electric dipole* can be constructed as a vector sum of the *point charge fields* of the two charges.

**What does the term “electric dipole moment” mean?**

The *electric dipole moment* for a pair of opposite charges of magnitude $q$ is defined as the magnitude of the charge times the distance between them.

$$\vec{p} = q\vec{d}$$  \hspace{1cm} (6.2.1)
What is the direction of the dipole moment?

The direction of the dipole moment is from "−" toward the "+". So, we can say that the dipole moment and the electrical field intensity are opposite directed.

The potential of an electric dipole can be found by superposing the point charge potentials of the two charges:

\[
\phi = k q \left[ \frac{1}{r_+} - \frac{1}{r_-} \right] = k q \left[ \frac{r_+ - r_-}{r_+ r_-} \right]
\]

(6.2.2)

For cases \( r \gg d \):

\[
\phi = \frac{k p \cos \theta}{r^2},
\]

(6.2.3)

where \( \vec{p} = k \vec{d} \) can be defined as the dipole moment.
It means, that the potential of an electrical dipole is directly proportional to its magnitude on the so-called radius-vector (the distance between two point charges).

**What does the theory of Einthoven mean?**

In Einthoven's electrocardiographic model the cardiac source is a two-dimensional dipole in a fixed location within a volume conductor that is either infinite and homogeneous or a homogeneous sphere with the dipole source at its center.

The curves (a) and (b) represent the recorded positive and negative isopotential lines, respectively. These indicate that the heart is a dipolar source having the positive and negative poles at (A) and (B), respectively. The curves (c) represent the assumed current flow lines.

![Electric field of the heart on the surface of the thorax, recorded by Augustus Waller (1887)](image)

Einthoven first recognized that as the limbs are generally long and thin, no significant electrocardiographic currents from the torso would be expected to enter them. Accordingly, Einthoven realized that the potential at the wrist was the same as at the upper arm, while that at the ankle was the same as at the upper thigh. Einthoven consequently assumed *that the functional position of*
the measurement sites of the right and left arm and the left leg corresponded to points on the torso which, in turn, bore a geometric relationship approximating the apices of an equilateral triangle. He further assumed that the heart generator could be approximated as a single dipole whose position is fixed, but whose magnitude and orientation could vary. The location of the heart dipole relative to the leads was chosen, for simplicity, to be at the center of the equilateral triangle. The signals were obtained from the two arms and left leg (modern Lead I). To enhance conduction, hands and foot were bathed in saline solution with the tubs wired to the input of the electrocardiograph.

Noticing a recurring pattern of movement, Einthoven named the prominent waves alphabetically, P, Q, R, S and T: the P-Wave, representing the impulse across the atria to the A/V Node; the QRS representing the impulse as it travels across the ventricles; the T-Wave, representing the repolarization of the ventricles. Why PQRST and not ABCDE? The four deflections prior to the correction formula were labelled ABCD and the 5 derived deflections were labelled PQRST. The choice of P is a mathematical convention by using letters from the second half of the alphabet. N has other meanings in mathematics and O is used for the origin of the Cartesian coordinates. In fact Einthoven used O ... X to mark the timeline on his diagrams. P is simply the next letter.

**What is the ECG?**

The depolarization and repolarization of the SA node and the other elements of the heart's electrical system produce a strong pattern of voltage change which can be measured with electrodes on the skin. Voltage measurements on the skin of the chest are called an electrocardiogram or ECG.
The electrocardiogram, or ECG (also known as EKG, abbreviated from the German word), is a surface measurement of the electrical potential generated by electrical activity in cardiac tissue. Current flow, in the form of ions, signals contraction of cardiac muscle fibers leading to the heart's pumping action.

You can see on the figure below the "secret" of the ECG forming.
What are the components of the electrical sequence in association in the heart?

P Wave — Firing of the SA node and depolarization of the atria.

PR Interval — Delay of the electrical impulse at the AV node and the depolarization of the atrium.

QRS Complex — Ventricular depolarization
- Q-wave = first negative deflection
- R-wave = first positive deflection
- S-wave = second negative deflection

ST Segment — The beginning of ventricular repolarization.

T Wave — Ventricular repolarization

What is the diastolic depolarization?
At the cellular level cardiac pacemaking which sets the rate and rhythm of the heartbeat is produced by the slow diastolic depolarization. Diastolic depolarization underlies automaticity of the sinoatrial (SA) node and starts when net ionic membrane current changes from outward to inward. Membrane current
is composed of an outward current, mainly carried by the delayed rectifier current ($I_K$), and several inward currents: the hyperpolarization-activated current ($I_h$); two calcium currents, the T-type calcium current ($I_{Ca,T}$) and the L-type calcium current ($I_{Ca,L}$); and a background current ($I_b$), which is inward during diastole.
Part 7. Conductivity of the biological tissues for direct current. Galvanization and electrophoresis

What is a conductor of electrical current?
The electrical charges in pure and alloy metals are carried by electrons, therefore, the conduction is electronic in nature. Electrical conduction in solutions is ionic because charges are carried by ions and dependent on dissociation of the solute. Technically pure water is a non-conductor, to carry a charge it must have a conductive ion from a salt, acid or base.

\[ R = \rho_0 \frac{L}{A}, \]  

(7.1)

where \( R \) is resistance, \( \rho_0 \) is resistivity (Ohm·m), \( L \) is length, \( A \) is cross-section area.

\[ A = \pi R^2 \]

The degree of dissociation or conduction is dependent on the solution concentration and temperature. Conduction in biological substances, such as blood, is ionic and subject to laws of physical chemistry. The human body and its fluids, tissues, organs and fat may be treated as an ionic conductor, where skin and fat are less conductive than fluids and tissues. Pure fat is like oil and has little conductivity, it is an insulator.

The electrical properties of biological tissues has been an area of scientific interest for many years. The discovery of the electric battery by Galvani in the late 17\(^{th}\) century is considered to be the first observation of bioelectric phenomena. Two dissimilar metals in contact with the leg muscle of a frog were shown to cause stimulation due to the generation of electrical current. Many beneficial clinical uses have resulted, of which the most familiar is probably the electrocardiogram (ECG), where the electrical activity of heart is recorded.

What is the electrical conductivity?
The electrical conductivity of an electrical conductor is defined as

\[ \sigma = \frac{1}{\rho_0} \]  

(7.2)

where \( \rho_0 \) is the resistivity. For a body of length \( L \), resistance \( R \), and cross sectional area \( A \), the electrical conductivity is given by

\[ \sigma = \frac{L}{RA} \]  

(7.3)

Let us reveal the dependence of electrical conductivity of the cytoplasm and intercellular fluid which are the electrolytes.

The current, \( I \), is a coarse, average quantity that tells what is happening in an entire wire. If we want to describe in more detail how the charges move through a conductor, we use the concept of the current flux density \( J \):
\[ j = qn \vartheta, \]  
(7.4)

where \( n \) is the number of free charged particles per unit of volume, \( q \) is their charge, and \( \vartheta \) is their average drift velocity. Since it is a vector quantity, it can indicate the direction of the current flow.

So, \( J \) is the current per unit of area.

The electrolyte current flux density for ions of both signs is represented below:

\[ J_+ = qn_+ \vartheta_+ \quad J_- = qn_- \vartheta_- \]  
(7.5)

The total electrolyte current flux density is equal:

\[ J = J_+ + J_- = q(n_+ + n_-) \]  
(7.6)

The concentration of the cations and anions in the case of dissociation is equal (each molecule dissociates only into one positive charge ion and one negative charge ion). It follows:

\[ n_+ = n_- = \alpha n, \]  
(7.7)

where \( \alpha \) — the coefficient of the dissociation.

In the absence of a resisting force ions would accelerate in an electrical field. When a constant velocity is achieved, the accelerating force is equal to the resisting force. The resisting force is a function of size, shape, and viscosity in accordance with the Stokes equation. Larger molecules migrate slower, while the greater the charge the faster the migration.

As far as you know the driving force of a molecule in an electrical field is equal to electrical field strength (\( \vec{E} \)) times the charge of the molecule (\( q \)). At this case (when a constant velocity) the resisting force directly proportional to the velocity of an ion. Therefore (in projection on horizontal axis):

\[ qE \sim \vartheta \Rightarrow qE = k \vartheta \]  
(7.8)

(in projection on the horizontal axis), where \( k \) — coefficient of the proportionality.

The rate of migration of an ion would be directly proportional to the electric field strength.

\[ \vartheta \sim E \Rightarrow \vartheta = bE \]  
(7.9)

\[ b = \frac{\vartheta}{E}, \]  
(7.10)

where \( b \) — mobility of the ions.

The mobility of an ion in an electrical field is defined as the velocity of the ion per electrical field strength. In a fixed field strength the mobility of an ion is proportional to the charge and inversely proportional to the size of the ion and the viscosity of the buffer. When there is a variation of temperature, mobility increases as temperature increases.

Comparing the formulas \( qE = k \vartheta \) and \( \vartheta = bE \) one can see the mobility of an ion is also equal:
\[ b = \frac{q}{k} \quad (7.11) \]

For the ions with different signs: \( \vartheta_+ = b_+ E, \vartheta_- = b_-.E \)

\[ J = qn\alpha(\vartheta_+ + \vartheta_-) = qn\alpha E(b_+ + b_-) \quad (7.12) \]

We have to calculate the electrolyte conductivity. So let us find out the similarity between the given equation and Ohm’s law \((I = GV)\).

If \( J = \frac{I}{A} \) and \( \vec{E} = \frac{V}{l} \), let us multiply left and right parts of the given equation above by area of the electrode \( A \), obtain:

\[ I = qn\alpha \frac{V}{l} A(b_+ + b_-) \quad (7.13) \]

Hence, according to the Ohm’s law \((I = GV)\) conductance \( G \) will be equal:

\[ G = qn\alpha(b_+ + b_-) \frac{A}{l} \quad (7.14) \]

Knowing that \( G = \frac{1}{\rho_0 l} A, G = \gamma_0 \frac{A}{l} \), we can obtain the electrolyte conductivity per unit of volume (volume conductivity):

\[ \gamma_0 = qn\alpha(b_+ + b_-) \quad (7.15) \]

So, the volume conductivity depends on the value of the ions’ charge, concentration, coefficient of the dissociation, mobility of the ions.

**It is very important to remember that atoms and ions of an resistor’s lattice converts electrical energy into heat.**

Apart the resistor the tissues of the living organism are characterized by the reactive impedance of capacitors. As far as you know a capacitor consists of two conductors separated by a dielectric (an insulator) (fig. 7.1).

Biologically the cell membrane functions as a permeable barrier separating the intracellular (cytoplasm) and extracellular components. Hence, there is a special combination of the cells and their supporting environment: two electrolytes (electric flux conductors) separated by a cellular membrane (a dielectric) (fig. 7.1).

![Fig. 7.1. Capacitor](image)

**At rest**

<table>
<thead>
<tr>
<th>extracellular liquid</th>
</tr>
</thead>
<tbody>
<tr>
<td>has positive charge</td>
</tr>
<tr>
<td>+</td>
</tr>
<tr>
<td>Membrane cell</td>
</tr>
<tr>
<td>dielectric</td>
</tr>
<tr>
<td>cytoplasm</td>
</tr>
<tr>
<td>has negative charge</td>
</tr>
<tr>
<td>-</td>
</tr>
</tbody>
</table>

![Fig. 7.2. Capacitance of cells](image)
All biologic fluids (cytoplasm, intercellular liquid, blood etc.) consist of the polypeptide molecules high rate. What is the influence of that fact on the conductivity?

The considerable decreasing of the electric current through organs and tissues (if a voltage was constant) has been studied during the experiments (fig. 7.3).

![Figure 7.3](image)

*Fig. 7.3. The decreasing of the electric current with time (voltage is constant) in living tissue*

This fact can be explained by the phenomenon called *polarization*.

The phenomenon of polarization is closely related to the terms of free charges and bound charge. The free charge (for example, the electrons) moves under the external reason (voltage) from one potential toward another and this is a reason of electric current, whereas the bound charges are characterized by the low value of mobility. As a rule a bond charge refers an *electric dipole*. All neutral molecules contain equal amounts of positive and negative charge. If a material contains polar molecules, they will generally be in random orientations when no electric field is applied. An applied electric field will polarize the material by orienting the dipole moments of polar molecules. This decreases the effective electric field. So, in the presence of an external electric field the charges become separated in space. The resulting dipole is termed *induced dipole*.

As far as you know the direction of the electric dipole moment toward positive charge is *opposite* to the electric field direction (fig. 7.4).

![Figure 7.4](image)

*Fig. 7.4. Polarisation of a dielectric*
Polarization means the separation of positive and negative charges due to the Coulomb force that acts in opposite directions for negative and positive charges.

So the polarization stands up against electric current in living tissue.

Galvanization can be defined as stimulation tissues and organs with direct current. It means the work of covering with metal by the use of galvanic current or the coating of iron with zinc to protect it from rusting.

The transport of small electric current (around 15 mA) through the skin is called ioniophoresis. Such “injecting” of electrically charged ions into the skin is similar to the antiperspirant actions.

So, ioniophoresis is a convenient simple technique for introducing drugs into the skin by small electric currents. Ions convey the electrical current as they enter the skin. The delivered dose depends on the current flowing and its duration.

Electrophoresis can be defined as the process of charged particle migrating in an electrical field. Biological macromolecules (with high molecular mass), like proteins or nucleic acids, have a certain electrical charge. Those charged molecules will therefore migrate in an electrical field in a velocity and direction depending on their electrophoretic mobility.

The velocity of a particle can be calculated in the following way:

\[ v = \frac{\varepsilon E \varepsilon}{4\pi \eta}, \]  

(7.16)

where \( \varepsilon \) — relative permittivity, \( E \) — electrical field strength, \( \eta \) — viscosity of dispersional medium.

Thus, the electrophoretic mobility is directly proportional to net charge and inversely proportional to the size of the molecule and to the solution viscosity.

It is necessary to add, that mobility is sometimes expressed with a negative sign, because migration of the solutes or particles generally occurs in the direction opposite to the electrophoretic field.

To obtain a uniform electric field with a constant magnitude and direction over a specified volume of space, it is necessary to use two flat conductive plates set up parallel to each other. When the voltage V will be applied to these plates, a uniform electric field E is produced between the plates. Outside of the plates and near the ends, the field is not uniform.
Part 8. The study of the alternative current circuit

8.1. The major types of the resistance. Impedance

Why do we use three major types of the resistance exactly? Maybe it can be explained by the special function in the alternative current circuit. Let’s ask ourselves: “Why?” Can we state that moving electrical substance possess both electrical and magnetic field? So in general the electric current possesses electromagnetic field.

What is the graphical relation between the magnetic and electric fields?

Fig. 8.1.1. Magnetic and electric fields

Scottish physicist James Clerk Maxwell (1831–1879) discovered that electric and magnetic fields are intrinsically related to one another (in four relatively tidy equations). The main ideas of this relation are:

A changing electric field produces a perpendicular magnetic field, and
A changing magnetic field produces a perpendicular electric field

Fig. 8.1.2. Graphical expression of alternative current over time
Very often two waveforms with the same amplitude and frequency are not synchronized. They are out of step with each other. In technical terms, this is called a phase shift as illustrated below:

**Phase shift examples**

**The amount of phase shift between two waves (expressed in terms of degrees)**

- Phase shift equal to 180 degrees.
- Phase shift equal to 0 degrees.
- Phase shift equal to 90 degrees (A leads B, or B lags A)

**Fig. 8.1.3. Phase shift examples**

A *leading* waveform is defined as one waveform that is ahead of another in its evolution. A *lagging* waveform is one that is behind another.  

**What's a resistor?**

Active resistance (ohmic resistance) deprives its name because at any time it resists the electrical flux (due to the mechanical and the electromagnetic reasons). That is to say due to interaction between the flux and resistance of the lattice the power of the current is wasted at every moment and it results in heating the conductor. At the same time the **DIRECTIONS OF THE CURRENT AND THE VOLTAGE ARE COINCIDED (IN PHASE).**

**Fig. 8.1.4. Current and voltage across a resistor in phase**
It has been known since the time of George S. Ohm (1787–1854) that the electrical resistance of a geometrical structure, such as a cylinder, is dependent on its physical dimensions and volume resistivity (Ohms/cm³).

\[ R = \rho_0 \frac{l}{A} \]  

(8.1.1)

All substances have some obstruction to electrical current and are said to have resistance, or are resistors. Resistance is the inverse of conduction, substances which have a high conduction have low resistance.

\[ I = \frac{V}{R}; I = GV \]  

(8.1.2)

A resistive circuit that absorbs one amp at one volt would have one ohm of resistance.

As far as we know there is a similarity between flow of water and electric current. Water flowing through a hose is a hydraulic example of resistance. A hose with a large diameter would have less resistance to flow than a hose with a smaller diameter.

**What is the nature of the active resistance?**

1) Mechanical collision of the electric current charges and atoms of the lattice of the conductor.

2) Electromagnetic repulsion of the electric current charges and atoms of the lattice of the conductor.

**What's a capacitor?**

By definition, a capacitor consists of two or more conducting plates separated from one another by an insulating, or non-conductive, material (dielectric). Capacitors will store a charge of electrons for a period of time depending on the resistance of the dielectric. The amount of charge a capacitor will hold is determined by dielectric resistivity, the area of the charged plates, and the distance between the plates.

The capacitance of flat, parallel metallic plates of area \( S \) and separation \( d \) is shown by the given expression:

\[ C = \frac{\varepsilon \varepsilon_0 S}{d} \]  

(8.1.3)

\( \varepsilon_0 = 8.85 \cdot 10^{-12} \text{ F/m} \) means the permittivity of space, \( \varepsilon \) — relative permittivity of the dielectric material between the plates. For instance, \( \varepsilon = 1 \) for free space, \( \varepsilon > 1 \) for all media. Thus:

**What's an inductor?**

The electrical device based on the inductance’s principle. Inductance is typified by the behavior of a coil of wire in resisting any change of electric current through the coil.
Arising from Faraday’s law, the inductance \( L \) may be defined in terms of the emf generated to oppose a given change in current:

\[
Emf = -L \frac{\Delta I}{\Delta t}
\]  

(8.1.4)

Unit of measurement: \([\text{emf}] = \text{1 henry}\)

**Why do we only use the capacitor and inductor?**

A capacitor and an inductor are two major electrical devices, both of them accumulate the energy: between the capacitor plates the electrical field energy is accumulated while inside of an inductor — the magnetic field energy is. These electrical devices are characterized by the reactance.

**What does the reactance mean?**

The meaning of the reactance (reaction always differs from action by time) is the following. There is a change of “the structure” of the current — splitting of the electric current and voltage directions by \( \frac{\pi}{2} \) in phase.

Let’s analyse a quality factor in the relations between two opposite electrical devices — a capacitor and an inductor.

<table>
<thead>
<tr>
<th>ACTIVE RESISTANCE</th>
<th>REACTANCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resistor</td>
<td>Capacitor</td>
</tr>
<tr>
<td><img src="image1" alt="Electric field" /></td>
<td><img src="image2" alt="Electric field" /></td>
</tr>
<tr>
<td><img src="image4" alt="Magnetic field" /></td>
<td><img src="image5" alt="Electric field" /></td>
</tr>
<tr>
<td><img src="image7" alt="Voltage and current in phase" /></td>
<td><img src="image8" alt="Current leads voltage by 90°" /></td>
</tr>
</tbody>
</table>

**Fig. 8.1.5. Symmetry in the nature: active resistance and reactance**

The certain “championships” between electrical and magnetical variables of the AC results in the perpendicular positions of this characteristics on the graph above. Of course, if the capacitor is of the electrical field origin, and, at the same time, the inductor stores the magnetic field energy, we will obtain the “inverse” character of the diagrams.
What is the reason for “splitting” the voltage and current directions?
Let us try to answer this question by studying an electrical device consisting of two elements: capacitor and inductor.

In the parallel LC circuit (oscillatory circuit) the energy's transformation takes place: from electric energy to magnetic one.

Let us suggest that at the initial moment \( t = 0 \) capacitor charge is maximum that is to say the total energy of the circuit is the electrical field energy (concentrated between the capacitor plates). The electrical flux appears when the circuit is closed. With the capacitor charge decrease (the electrical field \( \frac{q^2}{2C} \) also decreases) the electric flux increases (magnetic field's energy \( \frac{L}{2i} \) also increases). When electric flux \( i \) reaches maximum the capacitor charge is equal to zero.

Then we can observe the inverse situation: the capacitor charge increases — its plates acquire charge which is opposite to the initial state. This leads to decrease of an electrical flux. Such oscillations of an electrical flux and capacitor charge we can see at the table below.

<table>
<thead>
<tr>
<th>#</th>
<th>Capacitor C</th>
<th>Inductor L</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( q = q_{\text{max}} )</td>
<td>( i = 0 )</td>
</tr>
<tr>
<td>1'</td>
<td>( q \downarrow ) (decreasing)</td>
<td>( i \uparrow ) (increasing)</td>
</tr>
<tr>
<td>2'</td>
<td>( q = 0 )</td>
<td>( i = i_{\text{max}} )</td>
</tr>
<tr>
<td>2'</td>
<td>( q \uparrow ) (recharge)</td>
<td>( i \downarrow )</td>
</tr>
<tr>
<td>3</td>
<td>( q = q_{\text{max}} )</td>
<td>( i = 0 )</td>
</tr>
<tr>
<td>3'</td>
<td>( q \downarrow )</td>
<td>( i \uparrow )</td>
</tr>
<tr>
<td>4</td>
<td>( q = 0 )</td>
<td>( i = i_{\text{max}} )</td>
</tr>
<tr>
<td>4'</td>
<td>( q \uparrow )</td>
<td>( i \downarrow )</td>
</tr>
<tr>
<td>5</td>
<td>( q = q_{\text{max}} )</td>
<td>( i = 0 )</td>
</tr>
</tbody>
</table>

So, we can state that electrical flux and capacitor charge are two opposite things!

Such relations between the electrical flux and capacitor charge refer to the relation between \( \cos x \) and \( \sin x \): \( \cos x \) function leads to \( \sin x \) by 90º — when \( \cos x \) is maximum \( \sin x \) is equal to zero.
Thus we can see that the fifth stage is equal to the first stage.
The total energy for an undamped oscillator is the sum of its kinetic energy and potential energy, which is constant. Hence, the energy's transformation takes place: from kinetic energy $\frac{mg^2}{2}$ to potential one $mgh$.

Let us study a quantity factor of the given relations.

$$i = \frac{dq}{dt} \quad (8.1.5)$$

**What do the parameters of the reactance of capacitance depend on?**

On the capacitor plates:

$$V = \frac{q}{C} \Rightarrow \frac{d}{dt} = V_{\text{max}} \cos \omega_0 t \Rightarrow q = CV_{\text{max}} \cos \omega_0 t \quad (8.1.6)$$

$$i = \frac{dq}{dt} \Rightarrow i = \frac{d}{dt} (CV_{\text{max}} \cos \omega_0 t) = -CV_{\text{max}} \omega_0 \sin \omega_0 t =$$

$$CV_{\text{max}} \omega_0 \cos (\omega_0 t + \frac{\pi}{2}) = I_{\text{max}} (\cos \omega_0 + \frac{\pi}{2}) \quad (8.1.7)$$

So, from this equation issues that THE ELECTRICAL CURRENT across the capacitor LEADS TO THE VOLTAGE BY $\frac{\pi}{2}$ IN PHASE (fig. 8.1.6).

![Diagram](image)

**Fig. 8.1.6.** A current across capacitor leads a voltage by $\frac{\pi}{2}$

$$I_{\text{max}} = CV_{\text{max}} \omega_0 = \frac{V_{\text{max}}}{X_C}, \quad (8.1.7)$$

where

$$X_C = \frac{1}{\omega_0 C}. \quad (8.1.8)$$
This is reactive resistance of an inductor.

**Inductor:**

\[ i = I_{\text{max}} \sin \omega_0 t \]

When:

\[ e = -\frac{d\phi}{dt} = -\frac{d}{dt}(Li) = -L\frac{di}{dt} = -LI_{\text{max}} \omega_0 \cos \omega_0 t \]

\[ e = -V \]

When:

\[ V = LI_{\text{max}} \omega_0 \cos \omega_0 t = LI_{\text{max}} \omega_0 \sin(\omega_0 t + \frac{\pi}{2}) = V_{\text{max}} \sin(\omega_0 t + \frac{\pi}{2}) \]

\[ V_{\text{max}} = LI_{\text{max}} \omega_0 \Rightarrow I_{\text{max}} = \frac{V_{\text{max}}}{L\omega_0} \]

\[ X_L = L\omega_0 \quad (8.1.9) \]

So, the **VOLTAGE across inductor LEADS THE CURRENT** by \( \frac{\pi}{2} \) in phase because the Len’z law behavior resists the buildup of the current, and it takes a finite time for an imposed voltage to force the buildup of current to its maximum.
Fig. 8.1.9. A voltage across inductor leads a current by $\frac{\pi}{2}$.

Symmetry

<table>
<thead>
<tr>
<th>Magnetic field</th>
<th>Electric field</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_L = \omega_0 L$</td>
<td>$X_C = \frac{1}{\omega_0 C}$</td>
</tr>
</tbody>
</table>

Electrical field strength is **perpendicular** to magnetic field variable — splitting by $\frac{\pi}{2}$ the vectors of current and voltage across capacitor and inductor (“mirror reflection” phenomenon) — symmetry in mathematical calculations ($X_C = \frac{1}{\omega_0 C}$, $X$ — inverse proportional to frequency: $X_L = \omega_0 L$ — $X$ — direct proportional to frequency).

We can express such relationship graphically as shown in fig. 8.1.10 below.
Impedance is a complex quantity, having both a resistive and reactive component. Specifically, it is the vector sum of resistance and reactance, where reactance is the Y coordinate and resistance is the X coordinate. The square root of the squared sums of X and Y is impedance. Phase angle is the angle (degrees) between the resistance coordinate and the impedance magnitude (line). If, for example, there were zero resistance and any value of reactance then phase angle would be 90 degrees. On the other hand, if reactance were zero and resistance had any finite value, then the phase angle would be zero degrees. Phase angle is a simple method of expressing the effective ratio of resistance and reactance from 0 to 90 degrees and electrically describes how voltage and current lead or lag each other in any circuit of resistors and capacitors. Impedance and phase angle only exist with alternating current.

The formula below represents an impedance for alternating current.

\[ Z^2 = R^2 + (X_L - X_C)^2 \]  

(8.1.10)

Usually we put active resistance along the axis OX, whereas the reactance (reactive resistance) is along OY.

**What is the Power Factor?**

Convenient expressions for the power dissipated in a resistor can be obtained by the use of Ohm's Law.

\[ P = VI = \frac{V^2}{R} = I^2R \]  

(8.1.11)

These relationships are valid for AC applications also if the voltages and currents are rms or effective values. The resistor is a special case, and the AC
power expression for the general case includes another term called the power factor which accounts for phase differences between the voltage and current.

For a DC circuit the power is \( P = VI \), and this relationship also holds for the instantaneous power in AC circuit. However, the *average power* in AC circuit is expressed in terms of the rms voltage and current. The calculations are the following.

\[
\tilde{P} = i^2 R = [i = I_{\text{max}} \cos \omega_0 t] = I_{\text{max}}^2 R \cos^2 \omega_0 t \quad \text{where} \quad \varphi = \omega_0 t \quad \text{is the phase angle between the voltage and current.}
\]

If \( \cos^2 \alpha = \frac{1 - \cos 2\alpha}{2} \),

\[
\tilde{P} = I_{\text{max}}^2 R \left( \frac{1 - \cos 2\omega_0 t}{2} \right) = \frac{I_{\text{max}}^2 R}{2} (1 - \cos 2\omega_0 t) \quad (8.1.12)
\]

As \( \cos x \) is an even function, after \( 2\pi \), \( \cos 2\omega_0 t = 0 \)

Thus, \( \tilde{P} = \frac{I_{\text{max}}^2}{2} R \) Let us compare the given formula with \( P = I^2 R \). We obtain:

\[
I^2 = I_{\text{max}}^2 \Rightarrow I = \frac{I_{\text{max}}}{\sqrt{2}} \quad (8.1.13)
\]

\[
\cos \omega_0 t = \frac{R}{Z} \quad (8.1.14)
\]

From the phasor diagram for AC impedance it can be seen that the power factor is \( R/Z \). If \( R = Z \), the power factor = 1.

### 8.2. Conductance of Biological Tissues for Alternating Current

We have studied previously, that there are two types of biological tissues and organs resistance: active resistance and reactance of capacitor (the body as a composite volume conductor!).

It can be explained easily, under the condition, that our tissues consist of electrolytes (about 80 %).

On the other hand, the cell can be referred to as capacitor — it consists of two electrolytes (cytoplasm and extracellular liquid), separated by the semipermeable membrane. Thus, total electrical impedance of biological tissues and organs is the following:

\[
Z = \sqrt{R^2 + X_c^2}, \quad (8.2.1)
\]

where, \( X_c = \frac{1}{\omega_0 C} \), \( C = \frac{k\varepsilon_0 A}{d} \), and \( R = \rho_0 \frac{l}{A} \).
With the increase of frequency of the alternating current the tissue and organ impedance decreases. Thus, total electrical impedance of biological tissues and organs has two constituents: active resistance and capacitor reactance. Frequency is inversely proportional to impedance, thus it can be shown below:

Let us study consequently these two constituents of impedance.

8.2.1. Electrical Conduction of Tissues

Conductance of biological tissues can be explained by the flux of ionic rather than electronic charge carriers. Human body consists of 80 % of liquid medium approximately. For example, extracellular liquids, cytoplasm of cells, blood etc. If an electric field is applied, a conduction current $I_c$ will develop due to the movement of mobile ions within the aqueous biological medium. The ion mobility will also be temperature dependent: the conductivity will increase by between 1 % and 3 % per °C rise in tissue temperature.

Bound charges within biological tissues give rise to complex dielectric properties, and thus displacement currents $I_d$. What compounds of tissues possess of dielectric properties? First of all, it is necessary to distinguish phospholipids bilayers of membrane cells. Almost all molecules are with high molecular mass proteins. If a sinusoidal voltage $V$ is applied between the faces of a unit cube of tissue, we can calculate the resulting current flow $I_c$ and $I_d$ like the following:

$$I_c = GV$$

where $G$ is the electrical conductivity (Sm$^{-1}$), $\omega = 2\pi f$ where $f$ is the frequency of the alternative current. $\varepsilon_0$ is the dielectric permittivity of free space ($\sim 8.854 \times 10^{-12}$ Fm$^{-1}$), $\varepsilon$ is the relative permittivity (dimensionless).

Both conductivity and relative permittivity vary widely between different biological tissues. These parameters are also changed with the frequency of the applied field.
Fig. 8.2.2. The dielectric properties of tongue muscle of human

Below 100 Hz the impedance of biological material is mostly resistive. It was found that the tissue characteristics obey linearity with respect to the applied field strength. Pethig (in 1979) states that most dielectrics are linear for electrical field strengths up to $10^5 \frac{V}{m^2}$. However, field strengths can prevail $10^7 \frac{V}{m^2}$ across biological cell membranes. Thus, there is evidence that neural membranes exhibit non-linear behaviour (Barber and Brown, 1984). The assumption of tissue linearity is probably valid, however, as this is not the dominant affect when making bulk tissue measurements and only becomes essential with frequencies above 1MHz.

Table 8.3

<table>
<thead>
<tr>
<th>Material</th>
<th>Species</th>
<th>1kHz</th>
<th>10kHz</th>
<th>100kHz</th>
<th>1MHz</th>
<th>10MHz</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brain (grey matter)</td>
<td>B</td>
<td>0.1</td>
<td>0.13</td>
<td>0.15</td>
<td>0.2</td>
<td>0.3</td>
</tr>
<tr>
<td>Liver</td>
<td>B</td>
<td>0.04</td>
<td>0.05</td>
<td>0.09</td>
<td>0.2</td>
<td>0.3</td>
</tr>
<tr>
<td>Kidney</td>
<td>B</td>
<td>0.12</td>
<td>0.15</td>
<td>0.2</td>
<td>0.3</td>
<td>0.5</td>
</tr>
<tr>
<td>Muscle (across)</td>
<td>B</td>
<td>0.3</td>
<td>0.35</td>
<td>0.4</td>
<td>0.5</td>
<td>0.6</td>
</tr>
<tr>
<td>Muscle (along)</td>
<td>B</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.6</td>
<td>0.7</td>
</tr>
<tr>
<td>Lung (inflated)</td>
<td>B</td>
<td>0.05</td>
<td>0.06</td>
<td>0.08</td>
<td>0.10</td>
<td>0.20</td>
</tr>
<tr>
<td>Uterus</td>
<td>H</td>
<td>0.4</td>
<td>0.4</td>
<td>0.4</td>
<td>0.5</td>
<td>0.6</td>
</tr>
<tr>
<td>Skin†</td>
<td>H</td>
<td>7 x 10^4</td>
<td>4 x 10^3</td>
<td>6 x 10^2</td>
<td>3 x 10^1</td>
<td>4 x 10^1</td>
</tr>
<tr>
<td>Adipose Tissue</td>
<td>H</td>
<td>0.022</td>
<td>0.023</td>
<td>0.023</td>
<td>0.24</td>
<td>0.25</td>
</tr>
</tbody>
</table>
In-vitro measurements of a) conductivity (Sm⁻¹) and b) relative permittivity for biological tissues at various frequencies. Tissues are of (B) bovine origin and (H) Human origin. Data is interpolated from graphs in Gabriel et al (1996b).

Electrical properties of tissue depends upon its anisotropic structure. For instance, skeletal muscle is the best example of this dependence, in which the conductivity can be up to 10 times greater along the length of the muscle fibres compared to the perpendicular orientation (Epstein and Foster, 1983).

**What is the biological effects of electrical current?**

At very low frequencies (less than 0.1 Hz), the major effect is electrolysis. This can occur at current levels as low as 100mA, and the biological effect is the formation of small ulcers in the skin beneath the electrode. It can be explained by the following. With low frequencies electrical impedance is high. In this case the depth of permeability is very low and skin is under strong electrical energy influence.

If the frequency increases, electrolysis becomes reversible and neural stimulation is the dominant effect. As the current amplitude is increased, sensory nerve fibers are initially stimulated. With further increases in the current, motor nerves are activated and muscle contractions occur. If the frequency continues to increase beyond 10 kHz, the level of applied current results in tissue heating (dominant effect).

We have shown the particular case of current influences. Electrical current therapy used almost in all branches of medicine. For example, the current at high trans-thoracic levels can result in fibrillation of the cardiac tissues.

### 8.2.2. CAPACITOR REACTANCE OF TISSUES

We have shown previously that capacitor consists of two conductors separated by a dielectric (an insulator).
Biologically, the cell membrane functions as a permeable barrier separating the intracellular (cytoplasm) and extracellular components. Hence, the special combination of the cells and their supporting environment: two electrolytes (electric flux conductors) separated by a cellular membrane (a dielectric) (fig. 8.2.3).

**At rest**

<table>
<thead>
<tr>
<th>Extracellular liquid</th>
<th>+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Membrane cell</td>
<td></td>
</tr>
<tr>
<td>Cytoplasm</td>
<td>−</td>
</tr>
<tr>
<td>Has positive charge</td>
<td></td>
</tr>
<tr>
<td>Has negative charge</td>
<td></td>
</tr>
</tbody>
</table>

*Fig. 8.2.3. Capacitance of the cells*

The phenomenon of decreasing the impedance with the increasing of the frequency within tissues can be explained by the following effects.

In the tissues and organs which conduct the electrical current there are different electrical dipoles (caused by the different types of the polarization). It means that the value of the tissue impedance depends on the electrical current frequency according to the ion dimensions.

The opposite direction of the external electrical field strength and electrical field dipole strength defines the impedance of the tissue and organs.

The electrical field dipole strength is related to its dimensions, for instance, mass and geometrical sizes.

The greater is the amount of atoms in a molecule, the more is the electrical field strength of dipole. Hence, the sizes of the dipoles influence the impedance Z.

Why does the electrical field frequency influence the impedance of tissues and organs?

The size and mass of the dipole are related to its inactivity. For example, which iron ball — small or large — is easier to move from its place or change the direction of its movement with the same effort? A small one, of course.

Then, the frequency of the small dipole oscillations will correspond to the high frequency of the electrical current. And the large dipole will effectively impede the electrical alternative current only with low frequency.
Dielectric dispersion is therefore associated with biological tissues in which the relative permittivity decreases with increasing frequency. On the other hand, the displacement current is proportional to the applied field frequency. These two opposing factors result in a complicated frequency behaviour. According to Schwan (1957), there are three discrete regions of dispersion:

- $\alpha$ — dispersion (10 Hz to a few kHz) associates with tissue interfaces such as membranes.
- $\beta$ — dispersion (1 kHz to several MHz) explained the polarisation of cellular membranes and polarisation of protein and other organic macromolecules.
- $\gamma$ — dispersion (greater than 10 GHz) closely related to the polarisation of water molecules.

In general, in what way can we estimate tissue processes on the base of electrical parameters?

To answer this question it is necessary to turn back to electrical impedance equation.

$$Z = \sqrt{R^2 + X_C^2}$$

But, $X_C = \frac{1}{\omega_0 C}$, where $C = \frac{k\varepsilon_0 A}{d}$, and $R = \rho_0 \frac{l}{A}$.

Thus,

$$Z = \sqrt{\left(\rho_0 \frac{l}{A}\right)^2 + \left(\frac{d}{\omega\varepsilon_0 A}\right)^2} \quad (8.2.2)$$
Knowing the type of the tissue and organ resistance we can give a precise estimation of the functional and organic disturbances of many internal organs with the method of electrical conductivity measurement.

For example, let us consider similarity between cell and capacitor. The permeability of the cellular membrane refers to the permittivity $\varepsilon$ of capacitor. The dimensions of cell refer to dimensions of capacitor $\frac{d}{A}$.

So, the tissue processes are reflected on the both quality and quantity parameters as shown below:

<table>
<thead>
<tr>
<th>Quality factor (internal characteristics of tissues)</th>
<th>Quantity (external characteristics of tissues, dimensions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active resistance</td>
<td>Resistivity $\rho_0$</td>
</tr>
<tr>
<td>Capacitor reactance</td>
<td>Permittivity $\varepsilon$</td>
</tr>
</tbody>
</table>

With constant frequency, the difference within the impedance values from the impedance of normal tissue can be explained by changing of other electrical parameters, both internal (resistivity $\rho_0$, permittivity $\varepsilon$) and dimensions ($\frac{l}{A}$ — resistor, $\frac{d}{A}$ — capacitor).

Some negative pathological factors can influence the cell and its surroundings. Thus, the influences can be detected by calculating both internal and external electrical characteristics of cell. For example, with the inflammation of any organ:

1) Volume (i.e. dimensions) of cells increases;
2) Coefficient of membrane permeability is changed. It referred to changing of permittivity $\varepsilon$;
3) Viscosity of blood, physical and chemical characteristics of cytoplasm, extracellular liquid are changed. It referred to changing of resistivity $\rho_0$.

Hence, the total electrical impedance will be changed.

Furthermore the mechanical damage of tissues and organs can also be estimated by the method of electrical conductivity measurement. This method allows to estimate elasticity of blood vessels.

Blood filling of any vessel leads to the change of its volume, as well as, the active resistance of the very vessel (the resistivity of blood differs from that of a vessel). So, you can judge about the velocity of the blood filling of a vessel (the functions of its elasticity) on the basis of the velocity of the active resistance change. This method of investigation of the blood vessel elasticity and the state
of all functions of the cardiovascular system was called the method of reography.

The ratio of the impedance with low frequency and the one with high frequency indicates the value of Tarusov’s coefficients.

V. M. Tarusov’s coefficient is the following:

\[
K = \frac{Z(10^4)}{Z(10^7)}
\]  

\[(8.2.3)\]

Fig. 8.2.5. The influence of inflammation on the electrical parameters

The basis to calculate Tarusov’s coefficients

Fig. 8.2.6. The basis to calculate Tarusov’s coefficients
Part 9. Low frequency electrophysical therapy.
Electrical stimulation

9.1. Characteristics of pulsed current

Based on previous information about currents, we can get some conclusions.

The difference in bioeffects while a current influence:

High frequency: ↑f → ↓z ⇒ ↑Permeation in tissues

\[ Z = \sqrt{R^2 + \left(\frac{1}{\omega C}\right)^2} \]

Heat effect

Low frequency: ↓f → ↑z ⇒ ↓Permeation in tissues

Stimulation effect

symbols:

↑ - means «increasing»; ↓ - means «decreasing»; ⇒ - means «leads to»

Fig. 9.1.1

There are some following types of electrotherapeutic currents: direct current, alternating current, pulsed current.

 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.

In general, pulse is a splash of the electromagnetical field at space and time.

Fig. 9.1.2. Pulse waveform at oscillograph screen and active potential

For example, active potential is pulse too.
There are some types of impulse waveform:

- **rectangular**
- **square**
- **triangular**
- **sawtooth**
- **an irregular waveform**
Characteristics of pulsed current are the following:

- **Rise time** refers to the time required for a signal to change from a specified low value (10% of the amplitude) to a specified high value (90% of the amplitude).

- **Decay (fall) time** is the time required for the amplitude of pulse to decrease from a value (90 percent of the amplitude) to another specified value (10 percent of the peak value exclusive of overshoot or undershoot). Both parameters depend on rise and fall times of input signal and on the characteristics of the system.

- **Pulse duration** — the elapsed time between the beginning and the end of all phases in a single pulse; on clinical stimulators may be labeled “pulse width”.

  Pulse duration refers to 50% peak amplitude.

  The term **pulse duration** has the following meanings:

  1. The interval between (a) the time, during the first transition, that the pulse amplitude reaches a specified fraction (level) of its final amplitude, and (b) the time the pulse amplitude drops, on the last transition, to the same level.

  2. **Note:** The interval between the 50% points of the final amplitude is usually used to define pulse duration, and this is understood to be the case unless otherwise specified. Other fractions of the final amplitude, e.g., 90% or 1/e, may also be used, as may the root mean square (rms) value of the pulse amplitude. Another synonyms: **pulse length, pulse width**.

- **Phase duration** — the elapsed time between the beginning and the end of one phase.

- **Interphase interval** — the elapsed time between two successive phases of pulse; also known as the intrapulse interval.
Interpulse interval — the time between the end of one pulse and the beginning of the next pulse in a series.

**Duty cycle** $D$ is defined as the ratio between the pulse duration ($\tau$) and the period ($T$) of a rectangular waveform (fig. 9.1.5):

$$D = \frac{\tau}{T} \quad (9.1.1)$$

**Fig. 9.1.5. Duty cycle**

For a pulse train in which the pulse duration is 1 $\mu$s and the pulse period is 5 $\mu$s, the duty cycle is 0.20. The duty cycle of a square wave is 0.4, or 50%.

**Frequency** — the number of pulses per unit of time for pulsed current expressed as pulses per second (pps); the number of cycles of AC per second expressed in cycles per seconds (cps) or hertz (Hz); often on clinical stimulators the frequency of stimulation control is labeled “rate”.

**Period** — the elapsed time from a reference point on a pulse waveform or cycle of AC to the identical point on the next successive pulse; the reciprocal of frequency (Period = $1/f$). For pulsed current, the period is equal to the pulse duration plus the interpulse interval.

**Amplitude** — measure of the magnitude of current or voltage with reference to the zero-baseline at any one moment in time on a graph.

Peak amplitude — the maximum current or voltage reached in a monophasic pulse or for each phase of a biphasic pulse.

**Fig. 9.1.6. Peak amplitude**
Peak-to-peak amplitude — the maximum current or voltage measured from the peak of the first phase to the peak of the second phase of a biphasic pulse.

![Fig. 9.1.7. Peak-to-peak amplitude](image)

Root-mean-square amplitude — measures average current or voltage per time unit, takes pulse shape into account, more accurately reflects the stimulating power of the waveform than peak amplitude measures.

**What is the way to produce electrical pulse?**

An exponential decay voltage wave is a unidirectional voltage that rises rapidly to a maximum value and decays slowly to zero. The circuit in fig. 9.1.8 may be used to generate an exponential decay waveform.

![Fig. 9.1.8. Electrical circuit for the production of exponential decay waveforms](image)

This pulse waveform can be explained by the following. The charging of the capacitor’s conductive plates occurs in time as shown below.
If the voltage of the gas lamp reaches the maximum value, the gas (as insulator) resistance will be broken by current producing. So, resistance of the lamp will be decreasing, what is followed by the charging up of the capacitor.

Both differentiator and integrator consisting of the capacitor can be used to change the pulse waveform like the following.

**Differentiator:**

\[ V_{output} = RC \frac{dV_{input}}{dt} \]

**Differentiator:**

\[ V_{in} = V_{C} + V_{R} \] (9.1.2)

We consider only low frequencies \( \omega << 1/RC \) \( (\tau << RC) \), so that the capacitor has time to charge up until its voltage almost equals to that of the source.

Thus, \( R << \frac{1}{\omega C} \), so \( V_{C} >> IR \), or: \( V_{in} \approx V_{C} \) (9.1.3)
The output is the voltage across the resistor:

\[ V_{out} = V_R = IR = \frac{d}{dt} R = \frac{d(V_cC)}{dt} R = \frac{dV_c}{dt} RC \]  

(9.1.4)

So,

\[ V_{out} = RC \frac{dV_{in}}{dt} \]  

(9.1.5)

**Integrator**

We consider only high frequencies \( \omega >> 1/RC (\tau >> RC) \), so that the capacitor has insufficient time to charge up. It follows, its voltage is small, so the input voltage approximately equals to the voltage across the resistor:

\[ V_{in} \approx V_R \approx IR \]  

(9.1.6)

---

**Integrator**

\[ V_{output} = \frac{1}{RC} \int V_{input} dt \]

Fig. 9.1.11. Integrator and its role in the waveform changing

The output is the voltage across the capacitor:

\[ V_{out} = V_C \]  

(9.1.7)

\[ V_{out} = V_C = \frac{q}{C} = \frac{1}{C} \int idt \approx \frac{1}{C} \int \frac{V_{in}}{R} dt \]  

(9.1.8)

*Medicine is an art of imitation of nature's therapeutic effect.*

(Hippocrates)

**9.2. PULSED CURRENT AND ELECTRICAL STIMULATION**

Physical therapy is the treatment of physical dysfunction by the use of therapeutic exercises and the application of modalities to restore normal function or development.
**Electrical stimulation** can be defined as the influence of low-frequency electrical current to excite skin and surface muscular tissues, because electrical impedance is inversely proportional to frequency: the decreasing of frequency leads to the increasing of impedance, i.e. decreasing of the depth of permeability in tissues.

\[
Z = \sqrt{R^2 + \left(\frac{1}{\omega C}\right)^2}
\]

All electrical pulse current energy is directed to skin and surface levels of tissues.

The results of such electrical current influences is exciting of cells and accelerating of healing processes.

The body tissues are filled with moving electrical charged ions. So, biological tissues have their own bioelectromagnetical system. This system influences wound healing by changing cell membrane permeability, enhancing cellular secretion through cell membranes etc. A natural current called the “current of injury” is generated between the skin and inner tissues when there is a break in the skin. Healing of the injured tissue will be incomplete if currents no longer flow while the wound is open. It’s important to remember the following: applying electrical stimulation will accelerate the wound healing process.

First of all, both the muscle and nerve tissues are potential targets for external electrical influences. Both myelin and non myelin nerve fibers are like to conductors: the active potentials are propagated along these nerve fibers i.e. covered by electromagnetical field.

Muscle tissue is closely related to nerve tissue and is always ready to answer nerve pulse: membrane potential of plasmalemma reaches to 100 mV! The mechanism of muscle both contraction and relaxation is based on the calcium ions current from SR to cytoplasm of muscle sell, so, muscle fibers are filled with natural electrical charge carriers.

Capacitatively coupled electrical stimulation involves the transfer of electric current through an applied surface electrode pad that is in wet electrolytic contact with the external skin surface and/or wound bed. Electrodes are usually placed over wet conductive medium, in the wound bed and on the skin a distance away from the wound.

When discussing electrical stimulation, it is important to distinguish the waveform. Although there are many waveforms used on electrotherapy equipment, very often high voltage pulsed current is used. The given current is a very safe one due to the short pulse duration (prevents significant changes in both tissue pH and temperature). The width of such current has a range from 20–200 microseconds.
9.2.1. ES AND CURRENT PATHWAY

In conductors, resistance is proportional to the distance between electrodes, and is inversely proportional to conductor cross sectional area. In homogenous conductors, resistivity (resistance specific to conductor material) normally is measured between opposite faces of a one centimeter cube. However, not only are carcasses interrupted by tracts of fat with a high resistance and by bones with a variable resistance, but muscles themselves are electrically anisotropic. Resistivity is inversely proportional to tissue temperature, it tends to be greater across rather than along muscle fibers, and it may show a transient increase post mortem followed by a progressive decline.

Resistance of a whole carcass is modified by factors such as:
- the time lapse between animal exsanguination and muscle stimulation;
- the distance between electrodes;
- electrode surface area in contact with meat or connective tissue and not blocked by fat;
- whether or not the carcass is whole, eviscerated or split, and electrode location.

Electrical stimulation is often referred to a pelvic floor muscle electrical stimulation or functional ES. Applying a low grade electrical current to pelvic floor muscles stimulates the pelvic muscle to contract. Muscle contraction form PFES is a useful addition to pelvic floor exercises in the rehabilitation of weakened pelvic muscles. It is very important for patients who are unable to contract these muscles on brain command as it may teach the correct action.

Even in a simplified laboratory model of carcass stimulation, with a muscle strip and a pair of stimulatory electrodes, the response of necrobic muscle may be quite complex. Although stimulation may accelerate post mortem metabolism, muscle with an already accelerated rate of metabolism may lose its excitability at a faster than normal rate. Thus, animals with intrinsically fast glycolytic rates may be detected by their reduced electrical excitability.

Unless special precautions are taken to the contrary, muscle strips contain intramuscular nerves and neuromuscular junctions among the muscle fibers. Immediately post mortem, all three components may be excitable with their own particular activation thresholds and, as these change post mortem, it is difficult to identify the point between the axon and the muscle fiber that responds first to ES.

The complete final common pathway from the spinal cord to the muscles survives for many minutes in pork carcass and probably longer out in the carcass.

The excitability of muscle strips decreases progressively post mortem so that either a higher voltage and/or a longer duration stimulus is needed to obtain a constant response, and it is likely that the initial loss of excitability is caused by fatigue in the excitation-contraction pathway. If neuromuscular junctions are
pharmacologically blocked in samples taken shortly after animal exsanguination, excitability is decreased. This suggests that the high excitability of muscle strips at this time is caused by intramuscular motor axons and/or their neuromuscular junctions.

It is very interesting to add, that chronic electrical stimulation provides muscle structural and physiological changes in response to the stimulation.

It is known that muscles respond to different electrical input. For example, skeletal muscles that play a postural role, and thus have a high proportion of slow fibers, are physiologically activated at low frequencies. Conversely, muscles with a very high proportion of fast fibers may be activated only intermittently with high frequency bursts of electrical activity. The fact that electrical activity and muscle properties seem to be intimately interrelated, provides an experimental basis for understanding muscle plasticity. The best documented effects of electrical stimulation on skeletal muscle are those that occur after chronic, low-frequency stimulation (similar to the activity of a “slow” muscle) is imposed upon a predominantly “fast” muscle. If the stimulator is activated at a nominal frequency of about 10 Hz and allowed to operate 8–24 hours per day, a well-defined progression of changes is observed whereby the fast muscle first changes its metabolic and then its contractile properties to completely “transform” into a “slow” muscle.

So, the fast fibers actually become slow fibers. If low frequency stimulation is applied 8–24 hours per day, the total transformation process requires about 8 weeks. If the stimulator is active for a shorter period of time, the transformation takes longer or may not fully occur.

The earliest observed changes occur within a few hours after the onset of stimulation where swelling begins to occur in the sarcoplasmic reticulum (SR) membrane network (compare upper panel to lower panel). The significance of this morphological change is not clear but it is routinely observed.

After 14 days, the Z-band begins to increase in width, and a decrease in the amount and activity of calcium ATPase is observed.

![Fig. 9.2.1. Electrical stimulation effect for a muscle](image)

The myosin profile is altered with different myosin monomers incorporating into single filaments. The precise role of myosin light chains are not fully understood, however, the fact that the light chains associated with slow fibers are expressed is consistent with the fast muscle fiber becoming more like
a slow muscle fiber. The heavy chain profile may also be altered at this time. In addition, muscle mass and fiber area are decreased.

The first papers introducing the clinical application of magnetic stimulation were published in 1985.

**What is the essence of magnetic stimulation?**

The origin of the biomagnetic field is the electric activity of biological tissue: permanent ional movement of electrical charged ions. This bioelectric activity produces an electric current in the volume conductor which induces the biomagnetic field. The external changing magnetic field will be coupling with the biomagnetic field of tissues and organs. Therefore, this interaction is a base for stimulation of biological tissue.

Magnetic stimulation is a method for stimulating excitable tissue with an electric current induced by an external time-varying magnetic field. It is important to note that, as in the electric and magnetic detection of the bioelectric activity of excitable tissues, both the electric and the magnetic stimulation methods excite the membrane with electric current.

The reason for using a time-varying magnetic field to induce the stimulating current is, on the one hand, the different distribution of stimulating current and, on the other hand, the fact that the magnetic field penetrates unattenuated through such regions as the electrically insulating skull. This makes it possible to avoid a high density of stimulating current at the scalp in stimulating the central nervous system and thus avoid pain sensation. Also, no physical contact of the stimulating coil and the target tissue is required, unlike with electric stimulation.

The first documents on magnetic stimulation described the stimulation of the retina by Jacques d'Arsonval (1896) and Silvanus P. Thompson (1910). The retina is known to be very sensitive to stimulation by induced currents, and field strengths as low as 10 mT rms at 20 Hz will cause a stimulation (Nilsson, 1980).

From the pioneering works of d'Arsonval and Thompson it took some time before the magnetic method was applied to neuromuscular stimulation. Bickford and Fremming (1965) used a damped 500 Hz sinusoidal magnetic field and demonstrated muscular stimulation in animals and humans. The first successful magnetic stimulation of superficial nerves was reported by Polson et al. in 1982 (Polson, Barker, and Freeston, 1982).

Transcranial stimulation of the motor cortex is the most interesting application of magnetic stimulation because the magnetic field (unlike the electric current) penetrates through the skull without attenuation. The first transcranial stimulation of the central nervous system was achieved in 1985 (Barker and Freeston, 1985; Barker, Freeston, Jalinous, Merton, and Morton, 1985; Barker, Jalinous, and Freeston, 1985). A more complete history of magnetic stimulation may be found from a review article of Geddes (1991).
9.2.2. STIMULUS PULSE

The experimental stimulator examined by Irwin et al. (1970) had a multicapacitor construction equaling a capacitance of 4760 µF. This was charged to 90-260 V and then discharged with a bank of eight thyristors through the stimulating coil. The result was a magnetic field pulse of 0.1–0.2 T, 5 mm away from the coil. The length of the magnetic field pulse was of the order of 150–300 µs. Today’s commercial magnetic stimulators generate magnetic energies of some 500 J and use typically 3 ... 5 kV to drive the coil. Peak fields are typically 2 T, risetimes of order 100 µs, and peak values of $dB/dt = 5 \times 10^4$ T/s.

The energy required to stimulate tissue is proportional to the square of the corresponding magnetic field. According to Faraday’s induction law, this magnetic field is in turn approximately proportional to the product of the electric field magnitude and the pulse duration (Irwin et al., 1970):

$$W \propto B^2 \propto E^2 t^2$$  \hspace{1cm} (9.2.1)

Thus:

$$E \propto \sqrt{\frac{W}{t}}$$  \hspace{1cm} (9.2.2)

Where $W$ — energy required to stimulate tissue, $B$ — magnetic flux density, $E$ — electric field, $t$ — pulse duration.

The effectiveness of the stimulator with respect to energy transfer is proportional to the square root of the magnetic energy stored in the coil when the current in the coil reaches its maximum value. A simple model of a nerve fiber is to regard each node as a leaky capacitor that has to be charged. Measurements with electrical stimulation indicate that the time constant of this leaky capacitor is of the order of 150–300 µs. Therefore, for effective stimulation the current pulse into the node should be shorter than this (Hess, Mills, and Murray, 1987). For a short pulse in the coil less energy is required, but obviously there is a lower limit too.

9.2.3. ACTIVATION OF EXCITABLE TISSUE BY TIME-VARYING MAGNETIC FIELDS

The actual stimulation of excitable tissue by a time-varying magnetic field results from the flow of induced current across membranes. Without such flow a depolarization is not produced and excitation cannot result. Unfortunately, one cannot examine this question in a general sense but rather must look at specific geometries and structures. To date this has been done only for a single nerve fiber in a uniform conducting medium with a stimulating coil whose plane is parallel to the fiber (Roth and Basser, 1990).

The effect of field risetime on efficiency of stimulation has been quantified (Barker, Freeston, and Garnham, 1990; Barker, Garnham, Freeston, 1991). Stimulators with short risetimes ($< 60$ µs) need only half the stored energy of
those with longer risetimes (> 180 µs). The use of a variable field risetime also enables membrane time constant to be measured and this may contain useful diagnostic information.

What is the essence of the magnetic stimulation of neural tissue?

Magnetic stimulation can be applied to nervous stimulation both centrally and peripherally.

The main effect of magnetic stimulation is that the stimulating current density is not concentrated at the skin, as in electric stimulation, but is more equally distributed within the tissue. This is true especially in transcranial magnetic stimulation of the brain, where the high electric resistivity of the skull does not have any effect on the distribution of the stimulating current. Therefore, magnetic stimulation does not produce painful sensations at the skin, unlike stimulation of the motor cortex with electrodes on the scalp (Mills, Murray, and Hess, 1986; 1988; Rimpiläinen et al., 1990, 1991).

Another benefit of the magnetic stimulation method is that the stimulator does not have direct skin contact. This is a benefit in the sterile operation theater environment.
Part 10. High frequency therapy

10.1. The biophysical principles of the influence of the alternating current on the tissues and organs

Can we state that in each cell there is an ion movement? Does it mean that a change of its velocity results in the electromagnetic field change?

So, the influence of the external electric field on the organs and tissues leads to the movement of the ions both inside and outside the cell, doesn’t it?

Does it mean, that the greater the efficiency of the external alternative electromagnetic field is, the deeper this field permeates inside the organism?

Will it happen with the low reactance, which depends on the frequency of the influence?

Does it mean, that the greater the frequency of the electromagnetic field is, the deeper the radiation permeates inside the organism?

Are there any differences between the influence of the electric field and that of magnetic field on the organism with the equal values of frequency?

The effect of heat is characterized by the influence of an electric and magnetic field.

For the magnetic field, which influences the electrolytic solution in a unit of time, the quantity of heat will be characterized:

\[ Q = \frac{v^2 B^2}{\rho_0}, \]  

where \( v \) — frequency of the magnetic field, \( \rho_0 \) — resistivity, \( B \) — magnetic field.

For the electric field, which influences the electrolytic solution in a unit of time, the quantity of heat can be determined similarly:

\[ Q = \frac{E^2}{\rho_0} \]  

– Are there any tissues which are dielectric?

– The quantity of heat can be characterized:

\[ Q \sim \varepsilon v E^2 \tan \delta \]
10.2. THE METHODS OF THE HIGH FREQUENCY THERAPY

10.2.1. DIATHERMY

The therapeutic effects of heat were known more than 2,000 years ago, where the Romans build hot-spring bathhouses. Electromagnetic methods had begun official start in the medical treatment in 1892.

The term diathermy is derived from the Greek words therma, meaning heat, and dia, meaning through. Diathermy literally means heating through. Diathermy is based on the influences of high-frequency electrical current to heat deep muscular tissues, because electrical impedance is inversely proportional to frequency: the increasing of frequency leads to the decreasing of impedance, i.e. both the increasing of the depth of permeability and heat effect. Purpose to use diathermy in therapeutical operations is to accelerate healing processes of tissues. Diathermy means deep internal heating of tissues and organs where the heat is the result of electrical energy conversion to heat while current is passing through the tissues.

Deep heating of body tissues as a treatment for pain, spasm, stiffness and inflammation is applied in the physiotherapy. In this case diathermy is based on using short-wave current which causes heating of the tissues with higher electrolyte content in response to electromagnetic radiation passed through the body. Heating is caused by the dielectric loss of the tissues when electromagnetic waves pass through.

Diathermy equipment uses two large (e.g. 10 cm diameter) electrodes placed at each side of the body and operating at 27 MHz. Other types direct microwave radiation at the skin to produce a more superficial heating. Both types cause mild internal heating from which the therapeutic effect is derived.

Ultrasound is absorbed by living tissues to a varying degree depending on the frequency employed (the higher is the frequency the greater is the proportion of energy absorbed), and the type of tissue. They are used with frequencies from 1 MHz – till 3 MHz (total power reaches 1–3 W) to heat tissue. It is necessary to recollect the heating which is known to accelerate healing processes and relieve inflammation.

10.2.2. SURGICAL DIATHERMY

Riviere, at the turn of the 20th century, conceived the notion of using very small treatment electrodes in order to concentrate the current density. In 1907 Walter de Keating-Hart and Pozzi introduced the term fulguration (from Latin fulgur, lightning). They asserted that using the spark from the Oudin coil was ideal for treating skin cancer and that the spark could selectively destroy tumor cells by interfering with their source of nutrition.
In 1909 Doyen introduced the term electrocoagulation (from Latin *coagulare*, to curdle) in which tissue was touched directly with both the treatment electrode and an indifferent electrode.

In the case of **surgical diathermy** high-frequency electric current passes through the body between two contact electrodes. **Diathermo-coagulation** is used in surgery by sealing blood vessels with electrically heated probes to stop bleeding, for **cutting and coagulating tissue**.

Surgery diathermy is based on the using of high frequency current (0.5 to 3.0 MHz). In opposite case the using of low-frequency alternating current (e.g. 50 Hz or 60Hz of the main supply) results in **muscle spasm**.

**Area of electrode inversely proportional to current density**!

In the diathermy one electrode is large (about 100 sm²) and attached to the skin whilst the other is a pointed probe held in contact with the tissue. This causes intense heating at the point of contact between the probe and the tissue since the **current density is very high**.

Large electrode causing low density of current refers to little or no heating.

**10.2.3. HYPERTERMIA**

Hyperthermia means the **heating of body tissues** to higher than physiological temperatures of body. Hyperthermia is very effectively used in the oncological field, not only as an alternative to traditional chemotherapy, radiotherapy and immunotherapy, but also in association with them.

The electromagnetic waves used in this type of treatment are in the frequencies interval from microwaves to short and long waves.

By using radiofrequency electromagnetic fields, the target organ is heated until it reaches temperatures of 42 °C or more. It takes approximately 60–90 minutes. The heating to the abovementioned temperatures can be performed repeatedly depending on the protocols adopted, but is usually restricted to three times per week, in order to avoid thermotolerance.

**10.2.4. DARSONVALISATION**

Interest to the use of electromagnetic energy in medicine was inflamed by Luigi Galvani’s experiments (eighteenth century) which illustrated both stimulation and heat effect of electrical current through living organisms.

In early 1890s, **Arsène d'Arsonval** (1851–1940) discovered that at a frequency of 10 kHz or more, electric current produces a **sensation of heat** as it crosses tissues, without causing the painful muscular contraction experienced with **lower** frequencies.

Now darsonvalisation can be defined as an electrotherapeutic method to treat some vascular and skin disorders, diseases of eye, ear and throat and to be used in urology, proctology and dentistry.
The electrical parameters of darsonvalisation method are the following: small current (not resulting in the heat effect), frequency — 500 kHz; voltage — about several kV.

10.2.5. **INDUCTOTHERMY**

The effect of deep heating of tissues and organs can be based on using a spiral or helix of wire to produce an oscillating magnetic field within the body which will induce currents having the same effect. The frequency employed is usually 27 MHz and applied by placing two large insulated metal electrodes, one at each side of the place to be treated. The tissue acts as the dielectric in a capacitor, and heat is produced due to the dielectric loss. This is called *inductothermy* and may be used to apply heating on a single flat surface.

To induce a current into the underlying tissue and organs, 1) strong and 2) 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 $F$. This current is due to the electromotive force (emf) induced by the time-varying flux. The magnitude of emf depends upon the rate of change of the magnetic flux $dF/dt$. The direction of emf is such that the time-varying magnetic field is always opposite to that of $dF/dt$; therefore,

$$EMF = -\frac{dF}{dt} \tag{10.2.1}$$

Where EMF means electromotive force [V], $F$ — magnetic flux [Wb], $t$ — time [s].

Magnetic flux, linking the circuit is given by $F = BA \cos(\vec{B}, \vec{A})$, where $B$ — magnetic field variable.
Flux is defined as: $F = LI$, where $L$ is the inductance of the coil, so:

$$ΕΜF = -\frac{dF}{dt} = -L\frac{dL}{dt}$$  \hspace{1cm} (10.2.2)

where $L$ — inductance of the coil [H], $I$ — current in the coil [A].

Inductance $L$ depends upon the geometry and constitutive property of the medium. The principal factors for the coil system are the shape of the coil, the number of turns on the coil, and the permeability of the core.
Part 11. The sensors

Very often physicians have to know about small changing in living organism processes at both micro and macro levels. For example, What is the change of hemoglobin under some pathological factors? Or, in what way was heart rhythm distorted? Very often features of treatment need long-time observations without hospitalization of a patient. To answer questions like mentioned above in medical practice so-called sensors are used — electrical devices, based on physical effects.

All sensors are classified by two major types: active and passive. 

Active sensors transform all external influences (physical stimulus, such as thermal energy, electromagnetic energy, acoustic energy, pressure, magnetism, or motion) into electromagnetical energy. Physical effects: piezoelectric effect, thermoelectromagnetical phenomenon, etc. are basic of active sensors.

On the other hand some types of sensors never produce electricity. Basic principle of passive sensor action is the changing of its parameters.

- There are several sensors such as the following.

<table>
<thead>
<tr>
<th>Current</th>
<th>Oxygen</th>
</tr>
</thead>
<tbody>
<tr>
<td>Displacement</td>
<td>Pressure</td>
</tr>
<tr>
<td>Inductive</td>
<td>Probe</td>
</tr>
<tr>
<td>Light</td>
<td>Proximity</td>
</tr>
<tr>
<td>Magnetic</td>
<td>Speed</td>
</tr>
<tr>
<td>Optical</td>
<td>Capacitive</td>
</tr>
<tr>
<td>Acoustic</td>
<td></td>
</tr>
</tbody>
</table>

11.1. PIEZOELECTRIC EFFECT

The word piezo is Greek for “push”. The effect known as piezoelectricity was discovered by brothers Pierre and Jacques Curie in 1880.

There are some crystals which generate a voltage (current) when compressed, twisted or distorted and are said to be piezoelectric. And inversely, when an electrical influence (voltage) is applied across certain surfaces of a solid that exhibits the piezoelectric effect, the solid undergoes a mechanical distortion.

This provides a special transducer effect between electrical and mechanical oscillations. Piezoelectric materials consist of crystals with unique symmetry in structure. Under mechanical stress the current will be generated due to the atoms displacement within lattice, or an electrical output from a mechanical input will be produced. In opposite cases, in earphones and ultrasonic radiators a mechanical output from an electrical input is produced.
Barium titanate, lead zirconate and lead titanate are ceramic materials which exhibit piezoelectricity and are used in medicine as pressure sensors. For example, in blood pressure calculating.

11.2. Temperature sensors

The main feature of temperature sensors is dependent resistance of both conductors and semiconductors upon temperature.

Increasing temperature is increasing active resistance of conductors (Resistance temperature detector). Semiconductors with temperature increasing lose resistance. Such semiconductors are called Thermistor (Bulk Semiconductor Sensor).

The mentioned above allows to use these sensors for temperature measurements of different organs.

Temperature Sensors are divided into three categories:

- Head mounted
- Probe style
- All specials and application specific sensors.

The Resistance temperature detector or resistance thermometer acts on the base that the resistance of metals increases with temperature as mentioned above. Examples of RTD are shown schematically on fig. 11.2.2. The resistance-temperature relationship of some materials is illustrated in the following scheme where the $y$-axis is the normalized resistance with respect to resistance at $0 \, ^\circ\text{C} (32 \, ^\circ\text{F})$, $x$-axis is the temperature.
The following metals are used as sensors: platinum, copper, nickel, AL, Au, Ag, e.t.c.

The thermistor uses ceramic semiconducting materials which are inversely proportional to temperature. Examples of thermistors are shown in figure 4.

The shape of a thermistor probe is different, for example a bead, washer, disk or rod as illustrated in the figure above. Typical operation resistances are in the kW range, although the actual resistance may range from several MW to several W.

Mostly of temperature measurement applications use so-called thermocouple sensors, which generate a current proportional to temperature increasing.

\[ \text{thermoemf} = \alpha (T_2 - T_1), \quad (11.2.1) \]

where \( \alpha \) — coefficient of proportionality, \((T_2 - T_1)\) — temperature difference.

Thermocouples are based on the principle that when two different metals are joined the current will be generated that relates to the difference in temperature between the measuring junction and the reference junction.
11.3. Passive sensors

The simplest example is a resistor. Active resistance depends upon resistivity and dimensions:

\[ R = \rho_0 \frac{I}{A} \]  

(11.3.1)

Changing of resistor dimensions means the external influence on it. Like resistor the capacitor and inductor are examples of passive sensors.

11.3.1. Capacitive sensors

The capacitor consists of two parallel plates separated by the insulator. The capacitance of it depends upon:

- Area of the conductive plates (capacitance increases as the plate size increases);
- Gap (capacitance decreases as the gap increases);
- Relative permittivity (dielectric material will cause the capacitance to increase or decrease depending on the material).

\[ C = \frac{k\varepsilon_0 A}{d} \]  

(11.3.2)

In ordinary capacitance sensing the dielectric material will be constant under mechanical influences. The variables are the plate’s area and the distance between plates.

The electronics is calibrated to output specific voltage changes for corresponding changes in capacitance. A common sensitivity setting is \( \frac{1V}{10^{-4}m} \). It means that for every 100 µm change in the gap of capacitor, the output voltage changes exactly 1.0V. With this calibration, a +5V change in the output means that the target has moved 500 µm closer to the probe.

For example, a new capacitive sensor for displacement measurement allows to calculate dynamical measurements in the range from 0 till 100 Hz. This sensor measures the displacement between two macroscopic opaque surfaces (can calculate displacement with atomic resolution). It consists of a plane capacitor, a high frequency oscillator and a high sensitivity frequency converter.

11.3.2. Inductive displacement sensors

Like capacitors the inductors depend upon dimensions too. They can be applied in some medical measurements of mechanical stress.
Light polarization

12.1. ELECTROMAGNETIC EQUATION

A controversy over the nature of light existed for centuries. In the seventeenth century, Sir Isaac Newton explained many properties of light with a particle model. In the early nineteenth century, Thomas Young performed some interference experiments that could be explained only by assuming that light is a wave. By the end of the nineteenth century, nearly all known properties of light, including many of its interactions with matter, could be explained by assuming that light consists of an electromagnetic wave. By an electromagnetic wave, we mean that light can be produced by accelerating an electric charge. According to Maxwell’s theory a varying electric field sets up a magnetic one which, generally speaking, is also varying. This varying magnetic field sets up an electric field, and so on. Thus, if one use oscillating charges to produce a varying (alternating) electro magnetic field, then in the space surrounding the charges a sequence of mutual transformations of an electric and a magnetic field propagating from point to point will appear. This process will be periodic
in both time and space and, consequently, will be a wave. Thus light can be represented as a transverse electromagnetic wave made up of mutually perpendicular, fluctuating electric and magnetic fields with the same amplitude and frequency.

This sinusoidally varying electric and magnetic field can be written as:

\[ E = E_0 \sin \omega (t - \frac{x}{\nu}), \quad (12.1.1) \]
\[ B = B_0 \sin \omega (t - \frac{x}{\nu}), \quad (12.1.2) \]

where \( \omega = 2\pi\nu \) is the angular frequency, \( t \) — is a time, \( \nu \) — is the velocity, \( x \) — is the coordinate.

If the wave is sinusoidal, then the period \( T \), frequency \( \nu \), and wavelength \( \lambda \), are related by \( \nu = 1/T, \lambda = c/\nu \). The velocity of electromagnetic waves is determined by formula:

\[ \nu = \frac{1}{\sqrt{\varepsilon_0 \mu_0}}, \quad (12.1.3) \]

where \( \varepsilon \) is a permittivity, \( \varepsilon_0 = 8.85 \times 10^{-12} \) F/m, \( \mu \) is a permeability, \( \mu_0 = 1.43 \times 10^{-7} \) Gn/m. In a vacuum (i.e. when \( \varepsilon = \mu = 1 \)), the velocity of electromagnetic waves \( c = \frac{1}{\sqrt{\varepsilon_0 \mu_0}} = 2.98 \times 10^8 \) m/s is maximum. Light travels in a vacuum with a velocity \( c = 3 \times 10^8 \) m s\(^{-1} \) (to an accuracy of 0.07 %). When light travels through matter, its speed is less than this and is given by \( \nu = c/n \), where \( n \) is the index of refraction of the substance. The value of the index of refraction is determined as

\[ n = \frac{c}{\nu} = \sqrt{\varepsilon \mu} \quad (12.1.4) \]

and depends on both the composition of the substance and the wavelength of the light.

12.1.1. THE ELECTROMAGNETIC SPECTRUM OF RADIATION

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 300000 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 (\( \lambda > 10^{-3} \) m) that range from less than a centimeter to tens or even hundreds of meters.

**Infrared** 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 we know as visible light is the portion of the electromagnetic spectrum with wavelengths between 400 to 760 nanometers (760 nm > \( \lambda > 400 \) nm). It is the part of the electromagnetic spectrum that we see, and coincides with the wavelength of greatest intensity of sunlight.

**Ultraviolet** radiation has a range of wavelengths 400 nm > \( \lambda > 80 \) nm. A small dose of ultraviolet radiation is beneficial to humans, but larger doses cause skin cancer and cataracts.

**X-rays** are high energy waves which have great penetrating power and are used extensively in medical applications and in inspecting welds. The wavelength range is 80 nm > \( \lambda > 10^{-5} \) nm.

**Gamma rays** have wavelengths of less than \( \lambda < 10^{-5} \) 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.

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 speed 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

\[
E = h \nu = h \frac{c}{\lambda}.
\]

The proportionality constant \( h \) is called 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. 1 eV = 1.6 \cdot 10^{-19} \text{ J}. It is the energy acquired by an electron that moves through a potential difference of 1 V. We use the number “\( h \) stroke” or “\( h \) bar”: \( h = h/2 \pi = 1.05 \cdot 10^{-34} \text{ J s} = 0.66 \cdot 10^{-15} \text{ eV s} \). In terms of the angular frequency \( \omega = 2\pi \nu \),

\[
E = h \omega.
\]

It should be noticed the shorter the wavelength of the radiation, the higher the energy and the more harmful for biological objects.
12.2. POLARIZATION OF LIGHT

Light can be represented as a transverse electromagnetic wave made up of mutually perpendicular, fluctuating electric and magnetic fields. Fig. 12.2.1 shows the sinusoidally varying electric field in the $xy$ plane, the sinusoidally varying magnetic field in the $xz$ plane and the propagation of the wave with a velocity $v$ in the $x$ direction. The fig. 12.2.1 presents the electric field (denoted by $E$) in the $xy$ plane, the magnetic field (denoted by $B$) in the $xz$ plane and the propagation of the wave in the $x$ direction.

![Fig. 12.2.1. Electromagnetic wave](image)

Conventionally, when considering polarization, the electric field vector is described and the magnetic field is ignored since it is perpendicular to the electric field and proportional to it. Historically, the orientation of a polarized electromagnetic wave has been defined in the optical regime by the orientation of the electric vector, and in the radio regime, by the orientation of the magnetic vector. Vector $E$ is called light vector. At the point of intersection, the electric field is measured and shown as a vector in the observation plane, which is perpendicular to light of propagation. In other words, the length and direction of the vector represents the strength and the direction of the electric field measured at the starting point of the vector. Don't interpret these fig. 12.2.1 as showing waves located in space. The plane in which the light vector oscillates will be called the plane of oscillations.

The shape traced out in an observation plane by the electric vector $E$ as such a plane wave passes over it is a description of the polarization state. If the vector of the electric field (measured at a fixed point of space) oscillates along a straight line then the waves are called plane-polarized or linearly polarized waves. In this case the tip of the vector $E$ traces out a single line in the plane as illustrated in fig. 12.2.2.
At any fixed point in space that is in the line of the propagation of this wave, the electric field vector rotates in a circle while its length remains constant. Such waves are called circularly polarized waves. In this special case the electric vector traces out a circle in the observation plane which is perpendicular to light of propagation, so this special case is called circular polarization. Circular polarization may be referred to as right (R) or left (L), depending on the direction in which the electric field vector rotates. (fig. 12.2.3). Circular polarization is a limiting case of the more general condition of elliptical polarization.

**Elliptical polarization** is the polarization wave such that the tip of the electric field vector describes an ellipse in an observation plane. The magnitude of the electric field vector varies as it rotates (fig. 12.2.4).

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” (natural). Thus unpolarized light is a linear superposition of linearly polarized waves. Light emitted by the sun, by a lamp in the classroom, or by a candle flame is unpolarized light. Such light waves are created by electric charges which vibrate in a variety of directions, thus creating an electromagnetic wave.
which vibrates in a variety of directions. This concept of unpolarized light is rather difficult to visualize. In general, fig. 12.2.5 is to picture unpolarized light as a wave which has a multitude of vibrations.

![Unpolarized light](image)

*Fig. 12.2.5. Unpolarized light*

Partly polarized light can be considered as a mixture of natural (unpolarized) and plane-polarized light. The expression

\[
P = \frac{I_{\text{max}} - I_{\text{min}}}{I_{\text{max}} + I_{\text{min}}} \tag{12.2.1}
\]

is known as the degree of polarization. For plane-polarized light $I_{\text{min}} = 0$, and $P = 1$. For natural light, $I_{\text{min}} = I_{\text{max}}$ and $P = 0$. For partly polarized light, $P < 1$.

![Partly polarized light](image)

*Fig. 12.2.6. Partly polarized light*

It is possible to transform unpolarized light into polarized light. The process of transforming unpolarized light into polarized light is known as polarization. There are a variety of methods of polarizing light.

12.2.1. POLARIZATION BY REFLECTION

When unpolarized light is incident on a boundary between two dielectric surfaces, for example on an air-glass boundary, then the reflected and transmitted components are partially plane polarized. Light with the perpendicular oscillations to the plane of incidence (is said to be s-polarized)
predominate in the reflected ray (in fig. 12.2.7 these oscillations are denoted by points), and oscillations parallel to the plane of incidence (p-polarized light) predominate in the refracted ray (they are depicted in the figure by double-headed arrows). The degree of polarization depends on the angle of incidence.

\[
\tan \theta_B = \frac{n_2}{n_1},
\]

where \( n_1 \) and \( n_2 \) are the refractive indices of two media.

This equation is known as Brewster’s law. Note that, since p-polarized light is refracted (i.e., transmitted), any light reflected from the interface at this angle must be s-polarized.

A simple polarizer can be made by tilting a stack of glass plates at Brewster's angle to the beam (fig. 12.2.8). 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.

\[\text{Fig. 12.2.7. Polarization by reflection}\]

At one particular angle of incidence, however, light with one particular polarization cannot be reflected. This angle of incidence is Brewster’s angle \( \theta_B \):

\[
\tan \theta_B = \frac{n_2}{n_1},
\]

where \( n_1 \) and \( n_2 \) are the refractive indices of two media.

This equation is known as Brewster’s law. Note that, since p-polarized light is refracted (i.e., transmitted), any light reflected from the interface at this angle must be s-polarized.

A simple polarizer can be made by tilting a stack of glass plates at Brewster's angle to the beam (fig. 12.2.8). 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.

\[\text{Fig. 12.2.8. A stack of plates at Brewster's angle to a beam reflects off a fraction of the s-polarized light at each surface, leaving a p-polarized beam.}\]

Full polarization at Brewster’s angle requires many more plates than shown.

12.2.2. Optical anisotropy
Isotropic media is a media 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. However, there are one or two directions such that any light in that direction in the crystal has the same speed, regardless of its state of polarization. This direction is called the optic axis. The crystals have two optic axes are said to be biaxial, which have one optic axes is called uniaxial ones.

It must be borne in mind that an optical axis is not a straight line passing through a point of a crystal, but a definite direction in the crystal. Any straight line parallel to the given direction is an optical axis of the crystal.

A plane passing through an optical axis is called a principal plane of the crystal. Customarily, the principal plane passing through the light ray is used.

Investigation of the ordinary and extraordinary rays shows that they are both completely polarized in mutually perpendicular directions. 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 occur in a plane coinciding with a principal plane.

12.2.3. POLARIZATION IN DOUBLE REFRACTION

Many crystals are anisotropic to light and exhibit properties such as birefringence. 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. 12.2.9).

\[
\sin \beta_o = \frac{\sin \alpha}{n_o} ; \quad \sin \beta_e = \frac{\sin \alpha}{n_e}
\]  

(12.2.3)
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.

12.2.4. 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. 12.2.10).

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.

### 12.2.5. PHENOMENON OF DICHROISM

In some crystals, one of the rays (ordinary or extraordinary) is absorbed to a greater extent than the other as illustrated in fig. 12.2.11.

That is one in which light rays having different polarizations are absorbed by different amounts. This phenomenon is called dichroism. This circumstance has been taken advantage of for manufacturing a polarizing device called a polaroid. A crystal of tourmaline (a mineral of a complex composition) displays very great dichroism in visible rays. 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 not strongly coloured, but is difficult to grow in large crystals. It 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 filter and is described quantitatively by Malus intensity law.

### 12.2.6. POLARIZED LIGHT TRANSITION THROUGH A POLARIZER. MALUS'S LAW

Plane-polarized light can be obtained from originally unpolarized light with the aid of devices called polarizers. These devices freely transmit oscillations
parallel to the plane which we shall call the polarization plane and completely or partly retain the oscillations perpendicular to this plane.

Assume that plane-polarized light of amplitude $E_o$ and intensity $I_o$ falls on a polarizer (fig. 12.2.12).

An oscillation of amplitude $E_o$ occurring in a plane making the angle $\phi$ with the polarizer plane can be resolved into two oscillations having the amplitudes $E_{II} = E_o \cos \phi$ and $E_\perp = E_o \sin \phi$ (fig. 12.2.12). The first, oscillation will pass through the device, the second will be retained. The intensity of the transmitted wave is proportional to $E_{II}^2 = E_o^2 \cos^2 \phi$, i.e. is $I \cos^2 \phi$, where $I$ is the intensity of the oscillation of amplitude $E$. Consequently, an oscillation parallel to the plane of the polarizer carries along a fraction of the intensity equal to $\cos^2 \phi$. In natural light, all the values of $\phi$ are equally probable. Therefore, the fraction of the light transmitted through the polarizer will equal the average value of $\cos^2 \phi$, i.e. one-half. When the polarizer is rotated about the direction of a natural ray, the intensity of the transmitted light remains the same.

If plane-polarized light of amplitude $E_o$ and intensity $I_o$ incidents on a polarizer. The component of the oscillation having the amplitude $E = E_o \cos \phi$, where $\phi$ is the angle between the plane of oscillations of the incident light and the plane of the polarizer, will pass through the device. Hence, the intensity of the transmitted light $I$ is determined by the expression

$$I = I_o \cos^2 \phi,$$

(12.2.4)
where $I_0$ is the initial intensity, and $\varphi$ is the angle between the light's initial plane of polarization and the axis of the polarizer (is the angle between the transmission axes of the polarized beam and the polarizer). This relation is known as Malus's law. It was first formulated by the French physicist Etienne Malus.

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. Since the polarizer transmits only the component parallel to the axis of polarization, one-half of the incident intensity is transmitted.

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 the two axes are orthogonal $\varphi = 90^\circ$, the polarizers are crossed and in theory no light is transmitted (fig. 12.2.13), though again practically speaking no polarizer is perfect and the transmission is not exactly zero. The maximum intensity equal to $\frac{I_0}{2}$ obtained at $\varphi = 0$ (the polarizers axes are parallel).

If a transparent object is placed between the crossed polarizers, any polarization effects present in the sample (such as birefringence) will be shown as increases in transmission. This effect is used in polarimetry to measure the optical activity of a sample.

![Fig. 12.2.13. The crossed polarizers](image)
12.2.7. **Optical Activity**

An optical activity is the rotation of linearly polarized light as it travels through certain materials (fig. 12.2.4). Some substances known as optically active ones have the ability of causing rotation of the plane of polarization of plane-polarized light passing through them. Such substances include crystalline bodies (for example, quartz, cinnabar), pure liquid (turpentine, nicotine), and solutions of optically active substances in inactive solvents (aqueous solutions of sugar, tartaric acid).

![Fig. 12.2.14. 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:

$$\alpha = \alpha_o l,$$  \hspace{1cm} (12.2.5)

The coefficient $\alpha_o$ 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 $c$:

$$\alpha = \alpha_o c l,$$  \hspace{1cm} (12.2.6)

Here $\alpha_o$ 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 (12.2.5) or (12.2.6). When a solution is used, we can determine its concentration $c$ by Eq. (12.2.6) if we know the specific rotational constant $\alpha_o$ 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.
Part 13. Thermal radiation

Bodies can emit electromagnetic waves at the expense of various kinds of energy. The most widespread is thermal radiation, i.e. the emission of electromagnetic waves at the expense of the internal energy of bodies. Thermal radiation occurs at any temperature, but at low temperatures practically only long (infrared) electromagnetic waves are emitted.

13.1. Basic concepts of thermal radiation

We shall characterize the intensity of thermal radiation by the magnitude of the energy flux measured in watts (W) and use the symbol $\Phi$. $\Phi$ can be written in the form:

$$\Phi = \frac{E}{t}, \quad (13.1.1)$$

where $E$ is the radiation energy, $t$ is the time.

The second important characteristic is the radiant emittance. The radiant emittance of the body is the energy flux emitted by unit surface area of a radiating body in all directions. We shall use the symbol $R$ to designate this quantity. The radiant emittance has units of Wm$^{-2}$ and can be written in the form:

$$R = \frac{\Phi}{S}, \quad (13.1.2)$$

where $\Phi$ is the energy flux, $S$ is a surface area.

Radiation consists of 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 $d\lambda$:

$$dR_\lambda = r_\lambda d\lambda \quad (13.1.3)$$

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 Wm$^{-3}$. Like the radiant emittance $R$, the emissivity $r_\lambda$ depends greatly on the temperature of a body. Thus, $r_\lambda$ is a function of the wavelength and temperature.

The radiant emittance $R$ and the emissivity $r_\lambda$ are related by the formula:

$$R_{rT} = \int_{0}^{\infty} dR_{rT} = \int_{0}^{\infty} r_{\lambda T} d\lambda \quad (13.1.4)$$

(to stress that the radiant emittance $R_{rT}$ and the emissivity $r_{\lambda T}$ depend on the temperature, we have provided them with the subscript T).

Assume that the flux of radiant energy $d\Phi_{\lambda T}$ 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 $d\Phi_\lambda$ will be absorbed by the body. The dimensionless quantity

$$\alpha_\lambda = \frac{dU_\lambda}{dI_\lambda}$$  \hspace{1cm} (13.1.5)$$

is called the absorptivity of a body. The absorptivity of a body is a function of the wavelength and temperature.

**13.2. BLACKBODY**

By definition, $\alpha_\lambda T$ cannot be greater than unity. There are three kinds of a body with a different absorptivity: a blackbody, gray bodies and the all other bodies:

1. A blackbody completely absorbs the radiation of all wavelengths falling on it, $\alpha_\lambda T \equiv 1$ at the all wavelengths.
2. Gray bodies have the same absorptivity at the all wavelengths too, but it is smaller of the unity: $\alpha_\lambda T = \alpha_T = \text{const} < 1$.
3. The absorptivity 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 unity 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. 13.2.1. 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. 13.2.1. The model of a blackbody](image-url)

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 than is the absorbing material lining the cavity.
The blackbody radiant emittance $R_b$ and its emissivity $\varepsilon_\lambda$ are related by the formula:

$$R_b = \int_0^\infty \varepsilon_\lambda d\lambda = \int_0^\infty e_\lambda d\lambda \quad (13.2.1)$$

Much work was done on the properties of blackbody or thermal or cavity radiation in the late 1800s and early 1900s. While some properties could be explained by classical physics, others could not. The description of the function $\varepsilon_\lambda(\lambda, T)$ by 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:

$$E = h \nu, \quad (13.2.2)$$

where the constant of proportionality $h$ was subsequently named Planck’s constant. Moreover Planck has obtained formula for the function $\varepsilon_\lambda(\lambda, T)$:

$$\varepsilon_\lambda = \frac{8hc\pi}{\lambda^5} \frac{1}{e^{hc/\lambda kT} - 1} \quad (13.2.3)$$

We will not discuss the history of these developments, but will simply summarize the properties of the blackbody radiation function that are important to us.

Now we will consider the main thermal radiation laws, that have been opened.

### 13.3. THERMAL RADIATION LAWS

#### 13.3.1. 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. 13.3.1).

![Fig. 13.3.1. 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 $\alpha_{\lambda, T}$. Thus, the greater the emissivity $r_{\lambda, T}$ of a body, the greater is its absorptivity $\alpha_{\lambda, T}$. 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: the ratio of the emissivity and the absorptivity does not depend on the nature of a body, it is the same (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 $\alpha_{\lambda, T}$ 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.

### 13.3.2. 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:

$$R_b = \sigma T^4$$

where $\sigma$ is a constant quantity, and $T$ is the absolute temperature.

Relation (13.3.2) between the radiant emittance of a blackbody and its absolute temperature was named the Stefan-Boltzmann law. The constant $\sigma$ is called the Stefan-Boltzmann constant. Its experimental value is $\sigma = 5.67 \times 10^{-8}$ W·m$^{-2}$ K$^{-4}$.

The temperature dependence of the gray body radiant emittance is similar:
where \( \alpha \) is the gray body absorptivity.

### 13.3.3. 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. 13.3.2.

![Fig. 13.3.2. The blackbody radiation function for several temperatures.](image)

The visible spectrum is marked by \( \nu \).

In 1893, 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, that is known as Wien’s displacement law:

\[
\lambda_{\text{max}} = \frac{b}{T},
\]

where the experimental value of the Wien’s constant \( b \) is: \( b = 2900 \ \text{\( \mu \text{m} \cdot \text{K} \))}. 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. 13.3.2). The value of \( \varepsilon_l \) is plotted for several different temperatures in fig. 13.3.2 As the blackbody become hotter, the spectrum shifts toward shorter wavelengths. The visible region of the spectrum is marked on the abscissa in figure 3 even at 1,600 K when the radiating surface appears white, most of the energy is radiated in the infrared.

### 13.4. HEAT TRANSFER MECHANISMS IN COOLING THE HUMAN BODY

Fig. 13.4.1 gives a simplified model of the process by which the human body gives off heat.

![Fig. 13.4.1. A simplified model of the process by which the human body gives off heat.](image)

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:

\[
P = \alpha_{gb} \sigma \cdot S(T_b^4 - T_s^4),
\]

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 m\(^2\)) and \( \alpha_{gb} \) is the absorptivity of the skin (human skin is a near blackbody radiator in the infrared, \( \alpha_{gb} = 0.97 \)), \( \sigma = 5.67 \times 10^{-8} \text{ W m}^{-2}\text{K}^{-4} \) is Stefan-Boltzmann constant.
This model indicates that an unclothed person at rest in a room temperature of 23 Celsius ($T_r$) would be uncomfortably cool. The skin temperature of 34 ºC ($T_b$) is a typical skin temperature taken from physiology texts, compared to the normal core body temperature of 37 ºC. 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.

### 13.5. 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 (-273.15 ºC) 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 $10^8$ W/cm². 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. Infrared radiation has been used for over 40 years to image the body, but the value of the technique is still a matter of debate. According to Wien’s displacement law we can estimate the wavelength \( \lambda_{\text{max}} \) corresponding to the human body maximum emissivity, considering the temperature of skin is \( T = 34 ^\circ \text{C} = 307 \text{ K} \):

\[
\lambda_{\text{max}} = \frac{b}{T} = \frac{2900 \mu \text{mK}}{307 \text{ K}} = 9.5 \mu \text{m}. \quad (13.5.1)
\]

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 \( 5 \mu \text{m} < \lambda \leq 25 \mu \text{m} \), \( \alpha_\lambda T = 0.98 \pm 0.01 \). 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 \( \mu \text{m} \). 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 \( \mu \text{m} \), with a peak at 9 \( \mu \text{m} \) (fig. 13.5.1). The detectors typically respond to wavelengths below 6–12 \( \mu \text{m} \).
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 $\pm 2\%$ or worse accuracy (not as accurate as contact); ability to only measure surface body areas.

Thermography has also been proposed to detect and to diagnose various circulatory problems. Clinical applications of thermography are phlebology — vein trombosis, vascular cancer, ischemy of the limbs. Technical scheme of thermovision is illustrated in fig. 13.5.2, where IR detector is semiconductor InSb, which is cooling by liquid N\(_2\) (-196 °C). Thermography is generally not widely accepted, though it still has its proponents.
Part 14. Spectral analysis

14.1. TYPES OF SPECTRUM

Atomic spectroscopy is the determination of elemental composition of a given 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. 14.1.1). An element's emission spectrum is the relative intensity of electromagnetic radiation of each frequency it emits when it is excited (fig. 14.1.2).

![Fig. 14.1.1. An absorption spectrum](image)

![Fig. 14.1.2. An emission spectrum](image)

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 19th 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.

14.2. BOHR’S THEORY OF THE HYDROGENIC 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 \((10^{-9} \text{ s})\). 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. 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.

Bohr’s theory of the hydrogenic (one-electron) atom is based on the following postulates:

1. The electron revolves in circular orbits around the nucleus which are restricted by the quantization of angular momentum i.e. they revolve in orbits where the angular momentum of electron is an integral multiple of \(\frac{h}{2\pi} \), where \(h\) is Planck's constant.

\[
m \nu r = \frac{nh}{2\pi},
\]

where \(n = 1, 2, 3 \ldots \) — is the principal quantum number, \(m\) — is the mass of the electron, \(\nu\) is the electron velocity, and \(r\) is the orbit radius.

2. The energy of the atom has a definite value in a stationary orbit: \(E_1, E_2, E_3 \ldots E_n\). These orbits are called stationary states. In these orbits of special radius electron does not radiate energy as expected from Maxwell's laws. This is called as Bohr's quantization rule.

3. The electron can jump from one stationary orbit to another. If it jumps from an orbit of higher energy \(E_2\) to an orbit of lower energy \(E_1\), it emits a photon. The energy of the photon is:

\[
h \nu = E_2 - E_1.
\]

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. 14.2.2.

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 of the emitted light can be related to its energy \( \nu = \frac{hc}{\lambda} = \Delta E \). It is usually easier to measure the wavelength of light than to directly measure its energy.

**Fig. 14.2.2.** The various ways of how an 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

### 14.3. 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 \( \text{He}^+ \), \( \text{Li}^{2+} \), \( \text{Be}^{3+} \) etc. Let us consider the case of an ion with the charge of nucleus being \( Ze \) and an electron moving with a constant speed \( \nu \) along a circle of radius \( r \) with the center at the nucleus. The force acting on the electron is that due to Coulomb attraction and is equal to

\[
F = \frac{Ze^2}{4\pi\varepsilon_0 r^2}.
\]  

(14.3.1)

The acceleration of the electron is towards the center and has a magnitude \( \frac{\nu^2}{r} \). If \( m \) is the mass of the electron, from Newton's law \( F = ma \), we obtain

\[
\frac{Ze^2}{4\pi\varepsilon_0 r^2} = \frac{mv^2}{r}.
\]

(14.3.2)

Using Bohr's angular momentum quantization rule (14.2.1) for the value \( n \), the principal quantum number, we obtain both the velocity \( \nu \) and the radius \( r \) as:
\[ v = \frac{Ze^2}{2\varepsilon_0 hn}, \quad r = \frac{\varepsilon_0 h^2 n^2}{m\pi Ze^2} \]  
(14.3.3)

We see that the allowed radii are proportional to \( n^2 \). For each value of \( n = 1, 2, 3, \ldots \), we have an allowed orbit. For \( n = 1 \), we have the first orbit (smallest radius), for \( n = 2 \), we have the second orbit and so on.

The kinetic energy of the electron in the \( n_{th} \) orbit is
\[ E_k = \frac{mv^2}{2} = \frac{mZ^2 e^4}{8\varepsilon_0^2 h^2 n^2}. \]  
(14.3.4)

The potential energy of the atom is
\[ E_p = -e\varphi = -\frac{Ze^2}{4\pi\varepsilon_0 r} = -\frac{mZ^2 e^4}{4\varepsilon_0^2 h^2 n^2}. \]  
(14.3.5)

We have taken the potential energy to be zero when the nucleus and the electron are widely separated. The total energy of the atom is
\[ E = E_p + E_k = -\frac{mZ^2 e^4}{8\varepsilon_0^2 h^2 n^2} < 0. \]  
(14.3.6)

Equations (14.3.4) to (14.3.6) give various parameters of the atom when the electron is in the \( n_{th} \) orbit. The atom is also said to be in the \( n_{th} \) energy state in this case.

From equation (14.3.3) the radius of the smallest circle allowed to the electron is \( (n = 1) \)
\[ r_1 = \frac{\varepsilon_0 h^2}{m\pi Ze^2}. \]  
(14.3.7)

For hydrogen atom \( Z = 1 \) and substituting the values of other constants we get \( r_1 = 0.0529 \) nm. This length 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
\[ r_n = n^2 a_0. \]  
(14.3.8)

From equation (8) the total energy of the atom in the state \( n = 1 \) is
\[ E_1 = -\frac{mZ^2 e^2}{8\varepsilon_0^2 h^2} < 0. \]  
(14.3.9)

For hydrogen atom \( Z = 1 \) and substituting the values of the constants \( E_1 = -13.6 \) eV. Note that the energy is negative and hence a larger magnitude means lower energy. The zero of energy corresponds to the state where the electron and the nucleus are widely separated. This is the energy when the electron revolves in the smallest allowed orbit \( r = a_0 = 0.053 \) nm. We also see from equation (14.3.6) that energy of an electron is proportional to \( \frac{1}{n^2} \).

Thus,
The energy in the state $n = 2$ is $E_2 = E_1/4 = -3.4 \text{ eV}$. In the state $n = 3$ it is $E_1/9 = -1.5 \text{ eV}$ etc. The lowest energy corresponds to the smallest circle. The state of atom with the lowest energy is called is ground state. The higher energy states are called excited states. Thus the energy of a hydrogen atom in the ground state is $-13.6 \text{ eV}$ and in the first excited state is $-3.4 \text{ eV}$. Positive energy states correspond to the ionization atom where the electron is no longer bound. The hydrogen atom energy levels determined by formula (14.3.10) are shown schematically in fig. 14.3.1.

![Energy Levels Diagram](image)

Fig. 14.3.1. Hydrogen energy diagram illustrating Lyman, Balmer and Paschen series formation

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 shows the energy structure of atoms therefore
the studying of this 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. This is revealed most clearly in the spectrum of the simplest atom — hydrogen.

When a hydrogen atom \( (Z = 1) \) passes from the state \( k \) to the state \( n \), a photon is emitted:

\[
h\nu = E_k - E_n.
\]

The frequencies of all the hydrogen atom spectrum lines can be represented by a single formula:

\[
\nu = \frac{E_k - E_n}{h} = \frac{E_0}{h} \left( \frac{1}{n^2} - \frac{1}{k^2} \right), \quad (14.3.11)
\]

where \( n = 1, 2, 3, 4, \ldots \); \( k = n+1, n+2, n+3 \ldots \)

We have arrived at the generalized Balmer formula.

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** is the series of transitions and resulting emission lines of the hydrogen atom as an electron goes from any exited energy levels \( k \geq 2 \) to the ground one \( n = 1 \) (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 (14.3.11) if \( n = 1 \) and \( k = 2, 3, 4, 5, \ldots \)

\[
\nu = \frac{E_0}{h} \left( 1 - \frac{1}{k^2} \right), \quad (14.3.12)
\]

where \( k = 2, 3, 4, 5 \ldots \)

2. **The Balmer series** is characterized by the electron transitions from \( k \geq 3 \) to \( n = 2 \), 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:

\[
\nu = \frac{E_k - E_2}{h} = \frac{E_0}{h} \left( \frac{1}{4} - \frac{1}{k^2} \right), \quad (14.3.13)
\]

where \( k = 3, 4, 5, 6, \ldots \)

3. **The Paschen series** is the emission lines corresponding to an electron transitions from \( k \geq 4 \) to \( n = 3 \). 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:

\[
\nu = \frac{E_0}{h} \left( \frac{1}{9} - \frac{1}{k^2} \right), \quad (14.3.14)
\]

where \( k = 4, 5, 6, 7, \ldots \)
14.4. MOLECULAR SPECTRUM

The internal energy of a molecule $E_{\text{mol}}$ includes the electronic energy $E_{\text{el}}$, the vibrational energy of nuclei $E_{\text{vib}}$, the rotational energy of the whole molecule $E_{\text{rot}}$:

$$E_{\text{mol}} = E_{\text{el}} + E_{\text{vib}} + E_{\text{rot}}. \tag{14.4.1}$$

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. 14.4.1. Thus, 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 correspond to the microwave region of the electromagnetic spectrum. The vibration motion of nuclei is also quantized and described by vibrational quantum number $v$.

![Molecular energy diagram illustrating an absorption and emission band appearance](image)
Vibration and rotational energy levels are very closely spaced while the energy spacing of electronic levels are much larger:

\[ E_{el} \gg E_{vib} \gg E_{rot}. \]  
(14.4.2)

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 level \( E_1 (n = 1) \), if this is the ground state, the molecule will remain in this state unless got excited. When an radiation of frequency \( \nu \) is incident on the species, there is a finite probability that the molecule will absorb the incident energy and jump from its ground electronic state to one of the various vibrational-rotational states in excited electronic level \( E_2 (n = 2) \) or \( E_3 (n = 3) \). In this case the absorption lines are caused by a transition between closed located quantized energy states. They are composed of more than one wavelength of light and this spectral line are broadening. 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 molecule in \( \tau \sim 10^{-11}-10^{-12} \) seconds 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. Energy is emitted, the wavelength of which refers to the discrete lines of the emission spectrum. Note however that the emission extends over a range of frequencies, thus spectral lines are broadening. Wavelength 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 (fig. 14.4.2). That is why emitted light always has a longer wavelength than the absorbed light due to the internal conversion.

![Stokes shift](image)

*Fig. 14.4.2. Stokes shift. Emitted photon has less frequency \( \nu \) than the absorbed photon, this frequency difference determines the Stokes shift*
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.

14.5. THE SPECTRAL DEVICES

The devices which measure the interaction between light and materials as
a function of wavelength are spectrometer and spectrograph. A spectrometer is
used in spectroscopy for producing spectral lines and measuring their
wavelengths and intensities. Spectrometer is a term that is applied to instruments
that operate over a very wide range of wavelengths, from gamma rays and
X-rays into the far infrared. If the region of interest is restricted to near
the visible spectrum, the study is called spectrophotometry.

Early spectroscopes were simply a prism with graduations marking
wavelengths of light. Modern spectroscopes, such as monochromators, generally
use a diffraction grating, a movable slit, and some kind of photodetector, all
automated and controlled by a computer.

When a material is heated to incandescence it emits light that is
characteristic of the atomic makeup of the material. Particular light frequencies
give rise to sharply defined bands on the scale which can be thought of as
fingerprints.

In the original spectroscope design in the early 19th century, 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. This image was then viewed through a tube with a scale that was
transposed upon the spectral image, enabling its direct measurement

(Fig. 14.5.1).

*Fig. 14.5.1. 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

With the development of photographic film, the more accurate spectrograph was created. It was based on the same principle as the spectroscope, but it had 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. Arrays of photosensors are also used in place of film in spectrographic systems. Such spectral analysis, or spectroscopy, has become an important scientific tool for analyzing the composition of material in physical and analytical chemistry for the identification of substances through the spectrum emitted from or absorbed by them.
Part 15. 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 ($\lambda = \frac{c}{\nu}$).

There are different manners of atom and molecule exiting and the following kinds of luminescence are known:

1. **Photoluminescence** is a process in which a substance absorbs photons (electromagnetic radiation) and then emits photons. Quantum mechanically, 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 such as a phosphor, 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 (luminescence) due to excitation in the result of a chemical reaction.

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 ~ 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:

- **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:
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.

After the radiation causing luminescence has stopped the intensity of the luminescence decays exponentially with time and is described by the formula:

$$I_{\text{lum}} = I_0 e^{-t/\tau},$$

where $\tau$ is the delay time during that the intensity of the luminescence decreases in $e = 2.7$ times.

There are two principal varieties of luminescence: fluorescence and phosphorescence, distinguished by the delay time. If the luminescence decays in very short time ($\sim 10^{-8}–10^{-7}$ s) it is known as fluorescence (in this case $\tau < 10^{-7}$ s). If the luminescence continues after the radiation causing it has stopped, then it 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.

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.
Part 16. 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.

16.1. 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 the fig. 16.1.1, 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.

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. 16.1.2). One photon interacting with an excited atom results in two photons being emitted.

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, 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, the material is in a state of “Population Inversion” (fig. 16.3).

Fig. 16.1.3. A state of “Population Inversion” and a process of stimulated emission

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.

16.2. 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. 16.2.1.

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. 16.1.3).

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 3 level and 4 level systems, while the simple 2 level system is not used in practice, as it is difficult to achieve population inversion in a 2 level system.

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.
Fig. 16.2.1. Laser components and laser operation

When laser is pumped, photons are spontaneously emitted as excited electrons return to lower energy levels.

Photons reflect off mirrors and start to stimulate other electrons to emit their photons.

A chain reaction of photons begins and the laser begins to "lase."

Full operation.
**Resonator:** A Fabry-Perot cavity that has a pair of mirrors, one at each end of the laser is used as a resonator in most lasers. 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 $(R - 1) \%$ of the beam. This $(R - 1) \%$ is the output beam which one see. 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.

16.3. **Characteristics of Laser Light**

1. **Coherent:** different parts of the laser beam are related to each other in phase. 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 from one set of atomic 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.

16.4. **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. 16.4.1 gives a diagram of the energy levels of the chromium ion $\text{Cr}^{3+}$ (level 3 is a band formed by a collection of closely arranged levels).

The excitation of the ions as a result of pumping is depicted by arrow W13. The lifetime of level 3 is very small (~$10^{-8}$ s). Most of the ions, however, will pass to metastable level 2. The lifetime of level 2 is larger (~$10^{-3}$ s). When the pumping power is adequate, the number of chromium ions at level 2 becomes greater than their number at level 1. Consequently, levels 1 and 2 become inverted and the population inversion occurs.
Arrow T21 depicts a spontaneous transition from the metastable level to the ground one. The emitted photon may produce stimulated emission of additional photons which, in turn, will produce stimulated emission, etc. A cascade of photons is formed as a result.

Fig. 16.4.1. A diagram of the energy levels of the chromium ion Cr$^{3+}$ in the ruby laser

### 16.5. 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. **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.

2. **Gas lasers** (helium and helium-neon, HeNe, are the most common gas lasers) have a primary output of a visible red light. CO$_2$ lasers emit energy in the far-infrared, 10.6 micrometers, and are used for cutting hard materials.

3. **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.

4. **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.

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.

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. 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.

16.6. 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 optical 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. Medical lasers are not always used for therapy. Some of them rather assist the diagnosis e. g. via methods of laser microscopy or spectroscopy.
Part 17. Physics of eye vision

Optic — this is science concerned with the genesis and propagation of light, the changes that it undergoes and produces, and other phenomena closely associated with it. There are two major branches of optics, physical and geometrical. Physical optics deals primarily with the nature and properties of light itself. Geometrical optics has to do with the principles that govern the image-forming properties of lenses, mirrors, and other devices that make use of light.

17.1. LENS OPTIC

The main part of optical instruments, a lens is normally a glass body bounded on two sides by spherical surfaces. Each surface can be convex (bulging outwards from the lens), concave (depressed into the lens), or planar (flat). The line joining the centres of the spheres making up the lens surfaces is called the axis of the lens. Typically the lens axis passes through the physical centre of the lens.

Lenses are classified by the curvature of the two optical surfaces. A lens is biconvex (or just convex) if both surfaces are convex. A lens with two concave surfaces is biconcave (or just concave). If one of the surfaces is flat, the lens is plano-convex or plano-concave depending on the curvature of the other surface. A lens with one convex and one concave side is convex-concave or meniscus. It is this type of lens that is most commonly used in corrective lenses.
Fig. 17.1.1. a — convex lens; b — concave lens

If the lens is biconvex or plano-convex, a parallel beam of light passing through the lens will be focused to a spot on the axis, at a certain distance behind the lens (known as the **focal length**). In this case, the lens is called a **positive** or **converging** lens.

If the lens is biconcave or plano-concave, a parallel beam of light passing through the lens is diverged; the lens is thus called a **negative** or **diverging** lens. The beam after passing through the lens appears to be emanating from a particular point on the axis in front of the lens; the distance from this point to the lens is also known as the focal length, although it is negative with respect to the focal length of a converging lens.

A **thin lens** is a lens with a thickness (distance along the optical axis between the two surfaces of the lens) that is negligible compared to the focal length of the lens. The thin lens approximation ignores optical effects due to the thickness of lenses and simplifies ray tracing calculations.

The focal length of a lens in air can be calculated from the equation:

\[
\frac{1}{f} = (n - 1) \left( \frac{1}{R_1} - \frac{1}{R_2} \right)
\]

where \(f\) is the focal length of the lens; \(n\) is the refractive index of the lens material; \(R_1\) is the radius of curvature of the lens surface closest to the light source; \(R_2\) is the radius of curvature of the lens surface farthest from the light source, and \(d\) is the thickness of the lens, which is small compared to \(R_1\) and \(R_2\).

The focal length \(F\) is positive for converging lenses, and negative for diverging lenses.

The value

\[
D = \frac{1}{F}
\]

is known as the **optical power** of the lens. **Optical power** is the degree to which a lens converges or diverges light. It is equal to the reciprocal of the focal length of the device. The **dioptre** is the most common unit of measurement of optical power:

\[1 \text{ dioptre} = 1\text{m}^{-1}\]

The shorter the focal length, the stronger the refraction in the lens and the larger the value of the optical power. For converging lenses the optical power is positive, while for diverging lenses, it is negative.

A converging lens in air will focus a parallel beam travelling along the lens axis to a **focal point** at a distance \(f\) from the lens. Conversely, a point source of light placed at the focal point will be converted into a parallel beam by the lens. These two cases are examples of image formation in lenses. Fig. 17.1.2 shows the rays with which the construction can easily be carried out. The first ray passes through an optical center without changing its direction. The second ray is incident on the lens parallel to the optical axis. After the refraction, it will pass through the rear focus. The third ray passes through the front focus. After the refraction, it will emerge parallel to the optical axis.
These rays are constructed without any difficulty. Any other ray emerging from point $A$ would be more difficult to construct. But this is not necessary since any other refracted ray will pass through the same point $A_1$.

If an object is placed at a distance $d$ along the axis in front of a positive lens of focal length $F$, a screen placed at a distance $f$ behind the lens will have a sharp image of the object projected onto it, as long as $d > F$. The image in this case is known as a real image.

Note that if $d < F$, $f$ becomes negative, the image is apparently positioned on the same side of the lens as the object. Although this kind of image, known as a virtual image, cannot be projected on a screen, an observer looking through the lens will see the image in its apparent calculated position.

The magnification of the lens is given by:

$$ M = \frac{H}{h} = \frac{f}{d} = \frac{F}{d - F} \tag{17.1.3} $$

where $H$ is a size of an image, $h$ is a size of an object. $M$ is the magnification factor; if $|M| > 1$, the image is larger than the object.

Notice that, if $M$ is negative, as it is for real images, the image is upside-down with respect to the object. For virtual images, $M$ is positive and the image is upright.

The formulas above may also be used for negative (diverging) lens by using a negative focal length ($f$), but for these lenses only virtual images can be formed.

### 17.2. Aberrations

Lenses do not form perfect images. There are several different types of aberration which can affect image quality.

**Spherical aberration** causes beams parallel to, but distant from the lens axis to be focused in a slightly different place than beams close to the axis. This
manifests itself as a blurring of the image. Spherical aberration can be minimised by careful choice of the curvature of the surfaces for a particular application.

![Fig. 17.2.1. Spherical aberration](image)

**Chromatic aberration** is caused by the dispersion of the lens material — the variation of its refractive index $n$ with the wavelength of light. It follows that different wavelengths of light will be focused to different positions. Chromatic aberration of a lens is seen as fringes of colour around the image. It can be minimised by using an achromatic doublet (or *achromat*) in which two materials with differing dispersion are bonded together to form a single lens. This reduces the amount of chromatic aberration over a certain range of wavelengths, though it does not produce perfect correction.

![Fig. 17.2.2. Chromatic aberration](image)

**Astigmatism** is an aberration when an optical system has different foci for rays that propagate in two perpendicular planes. If an optical system with astigmatism is used to form an image of a cross, the vertical and horizontal lines will be in sharp focus at two different distances.
A magnifying glass (also called a loupe) is a convex lens which is used to produce a magnified virtual image of an object behind the lens. This lens have a focal lengths from 10 to 100 mm.

A magnifying glass is placed in front of the eye as close to it as possible. The distance between the lens and the object must be shorter than the focal length of the lens $d < F$.

We can determine the magnification of the magnifying glass by formula:

$$\Gamma = \frac{D}{F}$$

(17.3.1),

where $D$ is the distance of normal vision (usually equal to 25 sm), $F$ is a focal length.

It may seem that a magnifying glass can be used for obtaining very large magnification when its focal length is short. For instance, the magnification of a magnifying glass with a focal length of 0.25 mm is 1000. However, it is practically impossible to employ loupes with very short focal lengths, and hence with very small diameters. So magnifying glasses with a magnification exceeding 40 are not used.

17.4. EYE STRUCTURE

The human eye is an approximate sphere. A simplified cross section of the eye is shown in fig. 17.4.1. The principal components through which the light passes are the curved, thin, transparent cornea, the aqueous, the lens, the vitreous, and the retina.
The wall of the eye consists of three layers (tunics) that are discontinuous in the place where the optic nerve enters the eye. The outermost tunic is a fibrous layer of dense connective tissue that includes the cornea and the sclera. The cornea comprises the front curved surface of the eye, contains collagen fibers and no blood vessels, and is transparent to visible light. The cornea serves as a coarse focusing element to project light onto the retina. The sclera, is an opaque and resilient sheath to which the eye muscles are attached.

The second layer of the wall is a vascular tunic termed the choroids. It contains capillaries that supply blood to the tunics. Pigments in the choroid reduce internal light reflection that otherwise would blur the images. The ciliary body contains the muscles that support and focus the lens. It also contains capillaries that secrete fluid into the anterior segment of the eyeball. The iris is the colored part of the eye that has a central aperture termed the pupil. The diameter of the aperture can be altered through the action of muscles in the iris to control the amount of light entering the eye.

The lens of the eye provides the fine focusing of incident light onto the retina. It is a convex lens that can change in thickness by action of the ciliary muscles. The retina contains photoreceptors that translate light energy into electrical signals that serve as nerve impulses to the brain. It is more as an outpost of the brain rather than as simply another component of the eye.

There are no photoreceptors where the optic nerve enters the eye, creating a blind spot. Near the blind spot is the macula lutea, an area of about 3 mm² over which the retina is especially thin. Within the macula lutea is the fovea centralis, a slight depression about 0,4 mm in diameter. The fovea is on
the optical axis of the eye and is the area where the visual cones are concentrated to yield the greatest visual acuity.

17.5. Formation of Image by Eye

The lens system of the eye is composed of four refractive interfaces: (1) the interface between air and the anterior surface of the cornea, (2) the interface between the posterior surface of the cornea and the aqueous humor, (3) the interface between the aqueous humor and the anterior surface of the lens of the eye, and (4) the interface between the posterior surface of the lens and the vitreous body. The refractive index of air is 1; the cornea — 1.38; the aqueous humor — 1.33; the crystalline lens (on average) — 1.40; and the vitreous humor — 1.34.

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.

Fig. 17.5.1. Graphical representation of the eye looking at a tree. Point C is the optical center of the lens

About two thirds of the 60 diopters of refractive power of the eye is provided by the anterior surface of the cornea (not by the eye lens). The principal reason for this is that the refractive index of the cornea is markedly different from that of air, while the refractive index of the eye lens is not greatly different from the indices of the aqueous humor and vitreous. The total refractive power of the lens of the eye, as it normally lies in the eye surrounded by fluid on each side, is only 20 diopters, about one third the total refractive power of the eye.

The image on the retina is upside down relative to the original object, and it is also reversed right to left.

225
17.6. ACCOMMODATION

Accommodation is the process by which the eye increases optical power to maintain a clear image (focus) on an object as it draws near the eye.

Light rays from distant objects are nearly parallel and do not need as much refraction to bring them to a focus. Rays from objects closer than 6 m from the observer are brought to a focus behind the retina, and consequently the objects appear blurred. Therefore these light rays require more refraction for focusing.

The object distance \(d\), image distance \(f\) and focal length \(F\) are related by the thin-lens equation:

\[
\frac{1}{d} + \frac{1}{f} = \frac{1}{F}
\]

(17.6.1)

The problem of bringing diverging rays from close objects to a focus on the retina can be solved by increasing the distance between the lens and the retina or by increasing the curvature or refractive power of the lens. In mammals, the problem is solved by increasing the curvature of the lens.

When the object is infinitely far away the image distance is equal to the focal length of the lens, \(f = F\). As the object is brought closer to the eye \(f\) can not change, but the lens changes to decrease the focal length.
Fig. 17.6.2. Deformation and movement of lens. Accommodation for near point (solid line) and accommodation for far point (dotted line)

At rest, the lens is held under tension by the lens ligaments. Because the lens capsule has considerable elasticity, the lens is pulled into a flattened shape. When the look is directed at a near object, the ciliary muscles contract. This decreases the distance between the edges of the ciliary body and relaxes the lens ligaments. When their pull is removed for near vision, the natural elasticity of the lens causes it to become more spherical. This more spherical shape provides additional bending of the light rays, which is important to focus near objects on the retina. In young individuals, the change in shape may add as many as 12 diopters to the refractive power of the eye. When these fibers contract, they pull the whole ciliary body forward and inward.

17.7. Visual acuity

The optical centre of the human eye lies in the lens near its rear surface. The distance $f$ from the optical centre to the retina is known as the depth of the eye and normally amount to 17 mm. The image on the retina of the eye is always real, diminished and inverted. The angle $\varphi$ under which object $AB$ is seen from the optical centre O of the eye is called the angle of view.

The retina has a complicated structure and consists of separated light-sensitive elements. Therefore, two points of an object lying so close to each other that their images on the retina get into the adjacent elements are perceived by the eye as a single point. The minimum angle of view 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. The eye poorly recognizes the details of an object seen at an angle less than 1'. It accords to the image dimension on retina equal to 5 µm.
The angle $1'$ is an angle at which a segment having a length of 1 sm 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 $1^\circ$.

![Figure 17.7.1. Angle of view](image)

**Visual acuity** is the ability to identify black symbols on a white background at a standardized distance as the size of the symbols is varied. It quantitative measures as

$$
\gamma = \frac{1'}{\varphi_{\text{min}}}
$$

(17.7.1)

For example, if $\varphi_{\text{min}} = 2'$, then the visual acuity for this patient: $\gamma = \frac{1}{2} = 0.5$.

By bringing an object close to the eye, we increase the angle of view, and hence make it possible to resolve finer details. However, objects can not 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 $d_0 = 25$ sm. At this distance the eye recognized details well enough without being tired. This is the **distance of normal vision**.

The retina has a finite spatial resolution, so the image of a point still appears sharp, even if it is slightly out of focus. The retina is behind the plane in which the image is in focus. In dim light, the pupil of the eye is fully open and light from a point object is spread out over the larger circle on the retina. In brighter light the pupil is smaller, and light from the same point object is confined to the smaller circle. As long as this circle is smaller than the spatial resolution, the image is sharp. This is why we can see better in brighter light. The maximum photopic (bright-light) resolution of the eye is limited by four effects: diffraction of the light passing through the circular aperture of the pupil, spacing of the receptors, chromatic and spherical aberrations.

The eye **resolution limit** for the distance of normal vision is equal to:

$$
AB = 1' \cdot d_0 = 2.9 \cdot 10^{-4} \text{ rad} \cdot 250\text{mm} = 73\mu\text{m}.
$$

(17.7.2)
17.8. **Errors of Refraction**

The eye is considered to be normal, or *emmetropic*, if parallel light rays from distant objects are in sharp focus on the retina when the ciliary muscle is completely relaxed. 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.

*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. To overcome this abnormality, the ciliary muscle must contract to increase the strength of the lens. By using the mechanism of accommodation, a farsighted person is capable of focusing distant objects on the retina. If the person has used only a small amount of strength in the ciliary muscle to accommodate for the distant objects, he or she still has much accommodative power left, and objects closer and closer to the eye can also be focused sharply until the ciliary muscle has contracted to its limit. In old age, when the lens becomes *presbyopic*, a farsighted person is often unable to accommodate the lens sufficiently to focus even distant objects.

![Diagram of common defects of the optical system of the eye](image)

*Fig. 17.8.1. Common defects of the optical system of the eye*

In *myopia*, or nearsightedness, when the ciliary muscle is completely relaxed, the light rays coming from distant objects are focused in front of the retina. 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. No mechanism exists by
which the eye can decrease the strength of its lens to less than that which exists when the ciliary muscle is completely relaxed. A myopic person has no mechanism by which to focus distant objects sharply on the retina. However, as an object moves nearer to the person’s eye, it finally gets close enough that its image can be focused. Then, when the object comes 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.

It is known that light rays passing through a concave lens diverge. If the refractive surfaces of the eye have too much refractive power, as in myopia, this excessive refractive power can be neutralized by placing in front of the eye a concave spherical lens, which will diverge rays.

Conversely, in a person who has hyperopia — the abnormal vision can be corrected by adding refractive power using a convex lens in front of the eye.

Astigmatism is a condition in which the curvature of the cornea is not uniform. When the curvature in one meridian is different from that in others, light rays in that meridian are refracted to a different focus, so that part of the retinal image is blurred. A similar defect may be produced if the lens is pushed out of alignment or the curvature of the lens is not uniform, but these conditions are rare. Astigmatism can usually be corrected with cylindrical lenses placed in such a way that they equalize the refraction in all meridians.

![Fig. 17.8.2. Astigmatism](image)

17.9. ANATOMY AND FUNCTION OF THE RETINA

After light passes through the lens system of the eye 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. 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. **Rods** are distributed over the entire retina, except in the blind spot and the fovea centralis. The retina contains about 125 million rods. Active elements in the rods (and in the cones as well) are replenished throughout an individual’s lifetime. Rods have a low but variable threshold to light, and they respond to very low intensities of incident light. Vision under low illumination levels (e.g., night vision) is attributable almost entirely to the rods.

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 can not 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 can not discriminate among these photoreceptors to identify the origin of an action potential transmitted along the ganglion.

The retina contains about 7 million **cones**, packed tightly in the fovea and diminishing rapidly across the macula lutea. Cones are maximally sensitive to light of about 550 nm, in the yellowgreen 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. Cones are responsible for color vision through mechanisms that are imperfectly understood at present.
The fovea is a small area in the center of the retina, it is especially capable of acute and detailed vision. The central fovea, only 0.3 millimeter in diameter, is composed almost entirely of cones; these cones have a special structure that aids their detection of detail in the visual image.

Let us consider essential components of a photoreceptor (either a rod or a cone). As shown in fig. 17.9.2, the outer segment of the cone is conical in shape. Major functional segments of 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.

There are large number of disk in the outer segments of the rods and cones. Each of the discs is actually an infolded shelf of cell membrane. There are as many as 1000 discs in each rod or cone.
17.10. GENERATION OF RECEPTOR POTENTIAL

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 the cis-form of retinal into an trans-form. 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.

Because the three-dimensional orientation of the reactive sites of the trans-retinal, it begins to divide into opsin and a 11-cis-retinal.

This process excites electrical changes in the rods, and the rods then transmit the visual image 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 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.

**Fig. 17.10.1.** Cis- and trans-forms of retinal

### 17.11. Light and Dark Adaptation

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 (JND)*. 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:

\[
dS = k \frac{dI}{I}
\]  
(17.11.1)

where \( I \) is the intensity of stimulus, \( dS \) is an increment of perception (termed a *limen*), and \( k \) is a coefficient. The integral form of this expression is known as the *Weber–Fechner law*:

\[
S = k \log \frac{I}{I_0}
\]  
(17.11.2)

The Weber–Fechner law is similar to the expression for the intensity of sound in decibels.

If a person has been in bright light for hours, large portions of the photochemicals in both the rods and the cones will have been reduced to retinal and opsins. Because of this, the sensitivity of the eye to light is correspondingly reduced. This is called *light adaptation*. 
Conversely, if a person remains in darkness for a long time, the retinal and opsins in the rods and cones are converted back into the light-sensitive pigments. This is called dark adaptation.

In addition to adaptation caused by changes in concentrations of rhodopsin or color photochemicals, the eye has other mechanisms for light and dark adaptation. One of these is a change in pupillary size. This can cause adaptation because of changes in the amount of light allowed through the pupillary opening.

Between the limits of maximal dark adaptation and maximal light adaptation, the eye can change its sensitivity to light as much as 500,000 to 1 million times, the sensitivity automatically adjusting to changes in illumination. Because registration of images by the retina requires detection of both dark and light spots in the image, the sensitivity of the retina always adjusts so that the receptors respond to the lighter areas but not to the darker areas.

Conversely, when a person first enters darkness, the sensitivity of the retina is usually so slight that even the light spots in the image can not excite the retina. After dark adaptation, the light spots begin to register.

17.12. COLOR VISION

Different cones are sensitive to different colors of light. Let 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 when only red, green, and blue monochromatic lights 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 (fig. 17.12.1).

For example, an orange monochromatic light with a wavelength of 580 nanometers 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.
Fig. 17.12.1

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.
Part 18. X-Rays

X-rays (or Röntgen rays) are a form of electromagnetic radiation with a wavelength in the range of 80 to $10^{-5}$ nanometers. They are longer than Gamma rays but shorter than UV 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. We divide X-rays into brems-strahlung and characteristic X-rays according to the mechanism of their formation.

18.1. BREMS-STRAHLUNG X-RAYS

According to laws of electrodynamics when a charged particle undergoes acceleration or deceleration, it emits photons. The radiation is called deceleration radiation, braking radiation, or brems-strahlung radiation. The frequency rises with the increase in a module of the charge acceleration. If a charged particle has a large kinetic energy and experiences an abrupt deceleration, the electromagnetic radiation appears in an X-ray photon. In medicine X-rays are produced in a highly evacuated glass bulb, called an X-ray tube (fig. 18.1.1). It contains two electrodes — an anode made of platinum, tungsten, or another heavy metal of high melting point, and a cathode. When a high voltage is applied between the electrodes, streams of electrons (cathode rays) are accelerated from the cathode to the anode and produce X rays as they strike the anode. The focusing electrode directs the electron beam towards the anode. Some part of the electron kinetic energy turns into the energy of X-rays. The other part of this energy passes into heat so the temperature of the anode rises. Therefore anode is kept cool by means of air cooling arrangement.

![Fig. 18.1.1. X-ray tube](image-url)
Let us describe quantitatively the formation of bremsstrahlung X-rays. The electron obtains its kinetic energy in the electric field between the cathode and the anode:

$$\frac{mv^2}{2} = eU,$$  \hspace{1cm} (18.1.1)

where \( m \) is the mass of an electron, \( v \) — its velocity, \( e \) — its charge, \( U \) — is the electrical potential difference between the cathode and the anode.

The energy of the electron turns into the form of a quantum \( h\nu \) of the bremsstrahlung radiation. A part of the energy goes into heat \( Q \):

$$eU = h\nu + Q,$$  \hspace{1cm} (18.1.2)

where \( h\nu \) is the photon of energy (\( h \) — Planck’s constant, \( \nu \) — frequency), \( Q \) — heat.

A relation between summands in right part (18.1.2) is random. Therefore we can observe different frequencies in this radiation. The spectrum of the bremsstrahlung X-rays is a continuous distribution of frequencies up to some maximum value. This shows that the X-ray beam emitted by the target is made of a number of continuous wavelengths of different intensity and that the intensity rises sharply at certain definite wavelengths. The continuous spectrum has a definite short wavelength \( \lambda_{\min} \) below which there is no radiation. The minimum wavelength \( \lambda_{\min} \) corresponds to maximum frequency \( \nu_{\max} \). If we know the magnitude of the voltage \( U \) we can calculate the numerical value of \( \lambda_{\min} \). Let’s assume that \( Q = 0 \) (all energy turns into the radiation). Then:

$$h\nu_{\max} = \frac{hc}{\lambda_{\min}} = eU \Rightarrow \lambda_{\min} = \frac{hc}{eU}.$$  \hspace{1cm} (18.1.3)

where \( c \) is velocity of light.

Putting the values of constants, we get

$$\lambda_{\min} (\text{nm}) = \frac{1.23}{U} (\text{kV}).$$  \hspace{1cm} (18.1.4)
The spectrum of the bremsstrahlung X-rays we can see on fig. 18.1.2. \( \Phi_{\lambda} \) is spectral concentration of a radiant flux, where the radiant flux \( \Phi \) is determined by the number of light quanta falling on the surface in time unite.

As we can see on fig. 18.1.2 for each voltage, the intensity curve starts at a particular minimum wavelength, rises rapidly to a maximum and drops gradually but indefinitely towards the longer wavelengths. According to (18.1.3) and (18.1.4) the wavelength at which the intensity is maximum depends on the accelerating voltage, being shorter for higher voltages. The radiation becomes harder.

We can calculate a radiant flux \( \Phi \) of X-rays per unit of area by formula:

\[
\Phi = \kappa I U^2 Z, \quad (18.1.5)
\]

where \( I \) and \( U \) are current and voltage in the X-rays tube, \( Z \) — atomic number, \( \kappa = 10^{-9} \) — constant.

As we can see from fig. 18.1.3 the temperature of the cathode rises as the current in the X-rays tube increases when \( U \) is constant. In this case the radiant flux rises but the radiation hardness remains unchanged.

18.2. Characteristic X-rays

Let us consider high magnitudes of the voltage. In this case a number of spectral lines superimposes on a continuous background. When the voltage applied to the X-ray tube is altered, the positions of the spectral lines remain unchanged. If, instead, a different material is used as a target in an X-ray tube, an entirely different set of spectral lines is obtained. Thus the spectral lines
which are characteristic of the material of the target are said to form a spectrum called characteristic X-ray spectrum. Let’s think of its appearance.

When an electron from cathode having large velocity strikes the target, it may go deep into the atoms of the target. The electron may then strike an electron of any shell and eject it due to collision. The hole is usually filled when an electron from a higher energy level drops down on the unoccupied level. As it does so, the atom emits a characteristic X-ray — a photon with energy equal to the difference in energies of the two levels:

\[ h \nu_{21} = E_2 - E_1. \]  

(18.2.1)

This leaves a new hole, which is then filled with an electron from a still higher level with the emission of another X-ray. This phenomenon is called an Auger effect.

The spectral lines in characteristic X-ray spectrum are found to occur in several groups. The group having the lowest wavelength is the most intense and is called K group or K-series (fig. 18.2.1).

The frequency \( \nu \) of the characteristic X-rays rises as the atomic number \( Z \) increases. This relation is known as a Moseley law: \( \sqrt{\nu} = A(Z - B) \), where \( A \) and \( B \) are constant.

18.3. INTERACTION BETWEEN X-RAY AND MATTER

Let us consider an interaction between quanta of X-ray and atoms and molecules of the matter. Obviously, the result of this interaction depends on the energy of the quantum. We can examine several different cases.
1) The quantum energy is smaller than the energy of the atomic ionization \((h \nu < A_i)\). Such interaction is called a coherent scattering. It is a process in which the photon is scattered on the entire atom (fig. 18.3.1, a). That is, the internal energy of the atom does not change. We can say that the energy of the incident photon equals the energy of the scattered photon:

\[
h \nu_0 = h \nu. \tag{18.3.1}
\]

Only soft X-ray experiences a coherent scattering. This is not an ionizing interaction.

2) The quantum energy is slightly greater than the energy of the atomic ionization \((h \nu \geq A_i)\). This phenomenon is known as a photoelectric effect or photoeffect (fig. 18.3.1, c). In this case, the photon is absorbed by the atom and a single electron is ejected. The initial photon energy \(h \nu_0\) is equal to the final energy. The recoil kinetic energy of the atom is very small because its mass is large, so the final energy is the kinetic energy of the electron plus the excitation energy of the atom:

\[
h \nu_0 = \frac{mv^2}{2} + A \tag{18.3.2}
\]

3) The quantum energy is much greater than the energy of the atomic ionization \((h \nu \gg A_i)\). A photon interacts with an electron, but in contrast to the photoelectric effect, only a part of the photon energy is transferred to

![Fig. 18.3.1. Interaction between X-ray and matter](image)
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. 18.3.1, b). The electron is still emitted from its shell. In addition the electron obtains a kinetic energy $E_k$:

$$h\nu = A_i + h\nu' + E_k. \quad (18.3.3)$$

If the electron is ejected from interior shells then the characteristic X-ray appears (fig. 18.3.1, d).

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.

High-energy photons experience more Compton scattering than low energy photons. Unfortunately, Compton scattering is the major source of background noise in X-ray images. In addition, Compton scattering is the major source of tissue damage due to X-rays. For these reasons, this phenomenon is very undesirable.

### 18.4. ATTENUATION OF X-RAYS

When the X-rays pass through the material, its intensity falls due to its absorption and scattering by the matter (fig. 18.4.1). The character of attenuation depends on the energy of X-rays, nature (i.e. wavelength) and thickness of matter. Let $I_o$ be an initial intensity of X-rays incident normally on a material and $I$ be an intensity of the X-rays after traveling a distance $x$ in the material. We can describe the attenuation of X-rays by the exponential law:

$$I = I_o e^{-\mu x} \quad (18.4.1)$$

where $\mu = \mu_{\text{absor}} + \mu_{\text{disp}}$ — is the constant and is known as linear attenuation coefficient of material.

![Fig. 18.4.1. Attenuation of X-rays](image)

We can also use a mass attenuation factor: $\mu_m = \mu/\rho$, where $\rho$ is a density.
The beam of X-rays encloses quanta with different energy. They have different penetrating power. Therefore the coefficient $\mu$ in (18.4.1) is constant only for monoenergetic X-ray photons. We can use the effective attenuation coefficient $\mu_{\text{eff}}$ for beams of photons with different energies.

Let’s estimate the penetrating power of X-rays. We can use a half-value layer, which is the thickness required to attenuate the beam intensity by 50% (fig. 18.4.2). We can relate the half-value layer to the linear attenuation coefficient analytically. If in (18.4.1) $x = d_{1/2}$, then $I = I_o/2$:

\[
I_o/2 = I_o e^{-\mu d_{1/2}}
\]

\[
e^{\mu d_{1/2}} = 2
\]

\[
ln e^{\mu d_{1/2}} = ln 2
\]

\[
\mu d_{1/2} = ln 2 = 0.69
\]

Thus:

\[
d_{1/2} = ln 2/\mu = 0.69/\mu.
\]  

(18.4.2)

![Fig. 18.4.2. Definition of the half-value layer](image)

For example the half-value layer for X-ray is equal 10 mm of water or 1 mm of aluminium when the voltage is 60 kV.

The half-value layer is a function of the energy of X-ray beam. Therefore a spectral composition of X-ray changes 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”.

18.5. PHYSICAL PRINCIPLES OF THE X-RAY DIAGNOSTICS

The absorption coefficient of X-rays depends on the matter composition and the wavelength:
\[ \mu_n = k \lambda^3 Z^3, \]  
\[ (18.5.1) \]

where \( k \) is a coefficient of proportionality; \( Z \) — an atomic number of the material; \( \lambda \) — a wavelength.

We can see that the absorption coefficient is not constant. It increases with the increasing of the atomic number and depends on the photon energy. This is a basis of the medical X-ray diagnostics. The purpose of this diagnostic is to measure features of the internal anatomy of a patient through differences in the attenuation of rays passing through different parts of the body.

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 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 and other soft tissues) in applications such as breast imaging (mammography) 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 (~11 %) and carbon (~57 %) and a lower concentration of nitrogen (~1 %), oxygen (30 %), and high-Z trace elements (<1 %) Hence, the effective atomic number of fat \( (Z_{\text{eff}} = 5.9 \text{ to } 6.3) \) is less than that of soft tissues \( (Z_{\text{eff}} = 7.4) \) or bones \( (Z_{\text{eff}} = 11.6 \text{ to } 13.8) \). 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. They are fluoroscopy, mammography, and xeroradiography.

X-rays can be captured on film, or viewed directly on a fluorescent screen. X-rays diagnostic creates a significant radiation dose especially a fluoroscopic procedure because the patient is being continuously exposed to X-ray radiation.

There are different methods of dose reducing. For example fluorescent screens, by absorbing more of the photons, reduce the dose for a patient.

Radiographs provide only an integrated value of the attenuation coefficient. The radiographic image is often difficult to interpret because of its integration along one axis. For example, it may be difficult to visualize the kidneys because of the overlying intestines.

Several types of computed tomography (tomos means slice) have been developed in the last few decades. They include transmission computed
tomography (CT), single photon emission computed tomography (SPECT), and positron emission tomography (PET).

In modern machines all of the electrical connections are made through slip rings. This allows continuous rotation of the gantry and scanning in a spiral as the patient moves through the machine. Interpolation in the direction of the axis of rotation (the $z$ axis) is used to perform the reconstruction for a particular value of $z$. This is called spiral CT or helical CT. Spiral CT provides $\mu(x, y, z)$, and the 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%. The resolution limit of a tomography is no less than 2 mm. A scanning time for receiving one section can reach several seconds.
Part 19. Radioactivity

Radioactive decay is the process in which an unstable atomic nucleus (called radionuclides) loses energy by emitting radiation in the form of particles or electromagnetic waves. This decay, or the loss of energy, results in an atom of one type, called the parent nuclide transforming to an atom of a different type, called the daughter nuclide. This is a random process on the atomic level, in which it is impossible to predict when a particular atom will decay. On average the decay rate for a large number of similar atoms is predictable.

19.1. General 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. A proton has an electric charge of one positive fundamental unit \( e = 1.6 \cdot 10^{-19} \text{ C} \) and a mass which is about of one atomic mass unit (1 amu = \( 1.6726 \cdot 10^{-27} \text{ kg} \)). The mass of a neutron is equal to 1,001 amu.

Let’s point basic properties of the nuclear force. The nuclear force is related to a strong interaction. At short distances, the nuclear force is stronger than the Coulomb force; it can overcome the Coulomb repulsion of protons inside the nucleus. On the other hand its range is limited to distances about a few \( x10^{-15} \) meters. The nuclear force is nearly 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.

The sizes of nuclei grow steadily through the periodic table. The nuclear radius \( R \) and the atomic mass number are related by

\[
R = 1.5 \cdot 10^{-15} \cdot \sqrt[3]{A} \text{ (m)}
\]  

(19.1.1)

The number of protons in the nucleus, \( Z \), is called the atomic number. It determines what chemical element the atom is. The number of neutrons in the nucleus is marked by \( N \). The atomic mass number of the nucleus:

\[
A = Z + N
\]  

(19.1.2)

The given element can have many different isotopes, which differ from each other by the number of neutrons contained in the nuclei. In a neutral atom, the number of electrons orbiting the nucleus equals the number of protons in the nucleus. The nuclear charge is equal to:

\[
q = Ze
\]  

(19.1.3)

The nuclear symbol is \( ^A_Z\text{X} \). 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).
**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. The atomic nuclei in \(N/Z > 1.6\) undergo a radioactive disintegration. The atomic nucleus, with more than 83 protons, irrespective of the number of neutrons, is always unstable and undergoes radioactive disintegration. Thus, bismuth \((Z = 83)\) is the heaviest element in which at least one isotope \((^{209}_{83} \text{Bi})\) is stable.

### 19.2. Modes of Radioactive Decay

The process of spontaneous nuclear transformation generally occurs because of instability in the neutron-proton ratio or because the atom is an excited state following the previous transformation. This process is termed radionuclide decay or **radioactivity**. Radioactivity simply refers to the property of unstable atoms to transform themselves and move to a more stable configuration. The mode of decay depends on the nuclear structure, the neutron-proton ratio and the character of a nuclei force. All modes of decay obey laws of conservation of energy and charge.

**Alpha decay** is a type of radioactive decay in which an atomic nucleus emits an alpha particle (two protons and two neutrons bound together into a particle identical to a helium nucleus) and transforms into an atom with a mass number by time on 4 less and atomic number by time on 2 less.

Decay proceeds according to the following scheme:

\[
\frac{A}{Z} X \rightarrow \frac{A-4}{Z-2} Y + \frac{4}{2} \alpha \quad \text{(19.2.1)}
\]

For example:

\[
^{239}_{94} \text{Pu} \rightarrow ^{235}_{92} \text{U} + \frac{4}{2} \alpha.
\]

It is known that the total rest energy of the produced nucleus and alpha particle are always a little smaller than the rest energy of plutonium nucleus. The difference between these energies determines the kinetic energy of the alpha-particle and uranium nucleus. A daughter nucleus can be in an excited state. Its transition into a nonexcited state is accompanied by the emission of \(\gamma\)-quantum.

According to it:

\[
^{239}_{94} \text{Pu} \rightarrow ^{235}_{92} \text{U} + \frac{4}{2} \beta + \gamma.
\]

The unit of energy commonly used in atomic and nuclear physics is equal to the energy gained by an electron (a charged particle carrying unit electronic charge when the electrical potential at the electron increases by one volt). The electron volt equals \(1.6 \times 10^{-19}\) J. Therefore:

\[
1 \text{ MeV} = 1.6 \times 10^{-19} \text{ C} \cdot 10^6 \text{ V} = 1.6 \times 10^{-13}\text{ J}.
\]
Alpha particle has a strongly defined value of the energy for every particular nuclear decay. *Consequently alpha-decay has a linear energy spectrum.* The energy spectrum for $^{239}_{94}$Pu decay (the maximal energy of $\alpha$-particle is equal 5.15 MeV, of $\gamma$-photon is equal to 0.38 MeV) is shown on fig. 19.2.1.

![Energy Spectrum for $^{239}_{94}$Pu Decay](image)

*Fig. 19.2.1. The energy spectrum for $^{239}_{94}$Pu decay*

It often happens that the nuclei appeared as a result of radioactive transformation are also radioactive and decay. The new decay product may again be radioactive until a stable nucleus is formed.

**Beta-decay** is accompanied by the interconversion between neutrons and protons inside a nucleus. There are three varieties of *beta-decay*: $\beta^-$-decay, $\beta^+$-decay, and electron capture.

1. In *negative beta decay* an unstable nucleus ejects from itself an energetic electron (of a relatively negligible mass) and an antineutrino (with no rest mass), and a neutron in the nucleus is converted into a proton that remains in the product nucleus:

$$n \rightarrow p + \beta^- + \bar{\nu}$$  \hspace{1cm} (19.2.2)

Thus, negative beta decay results in a daughter nucleus, the proton number (the atomic number) of which is one more than its parent but the mass number (the total number of neutrons and protons) of which is the same:

$$A \ X \rightarrow A+1 \ Y + \beta^- + \bar{\nu}$$  \hspace{1cm} (19.2.3)

Beta-decay (a transformation of $^8_3$Li into $^8_4$Be) is shown in fig. 19.2.2.
Fig. 19.2.2. Beta-decay (a transformation of $^7\text{Li}$ into $^8\text{Be}$).

The energy lost by the nucleus is carried away by the ejected electron and the antineutrino, so that beta particles from a radioactive material have the energy ranging from zero to a distinct maximum, characteristic of the unstable parent. We can see a distribution of electron energy on fig. 19.2.3 for bismuth $^{211}_{83}\text{Bi}$.

Fig. 19.2.3. The energy spectrum for β-decay

Energy spectrum is continuous for all types of β-decay. There is a distinctive maximum energy $E_{\text{max}}$ for every specified radionuclide.

Neutrino is an elementary particle that travels close to the speed of light, having no electric charge, little or no mass. It is able to pass through ordinary matter almost undisturbed and is thus extremely difficult to detect. The difference between neutrino and antineutrino consists of the opposite direction of spins.

2. In positive beta-decay a proton in the parent nucleus transforms into a neutron that remains in the daughter nucleus and ejects a positron, which is a positive particle like an ordinary electron in mass but of opposite charge, along with a neutrino, which has no mass:
Thus, positive beta decay produces a daughter nucleus, the atomic number of which is one less than its parent and the mass number of which is the same:

$$\left(19.2.4\right)$$

$$\overset{\Lambda}{A} X \rightarrow \overset{\Lambda}{A} Y + \overset{0}{A} B + \overset{0}{A} H.$$

For example:

$$\overset{30}{15} P \rightarrow \overset{30}{14} Si + \overset{0}{1} B + \overset{0}{1} H.$$

3. In electron capture, an electron orbiting around the nucleus combines with a nuclear proton to produce a neutron, which remains in the nucleus, and a neutrino, which is ejected. Most commonly the electron is captured from the innermost, or K, shell of electrons around the atom; for this reason, the process is often called K-capture (fig. 19.2.4).

As in positron emission, the nuclear positive charge and hence the atomic number decreases by one unit, and the mass number remains the same:

$$\left(19.2.5\right)$$

$$\overset{4}{7} P + \overset{0}{4} B \rightarrow \overset{4}{n} n + \overset{0}{0} H.$$

An example is the transformation of beryllium into lithium:

$$\overset{4}{7} Be + \overset{0}{4} B \rightarrow \overset{3}{3} Li + \overset{0}{0} \nu.$$

### 19.3. SIMPLE NUCLEAR REACTIONS. METHODS OF RADIONUCLIDES PRODUCTION

A nuclear reaction is a process in which two nuclei or nuclear particles collide to produce products different from the initial particles.

The first nuclear reaction (1919) was carried by Ernest Rutherford, who bombarded nitrogen with alpha particles and identified the ejected lighter particles as hydrogen nuclei or protons and the product nuclei as a rare oxygen isotope:
Thus the first method of the nuclear reaction realization is a charged particles bombarding. 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 protons.

Radionuclides can be obtained in atomic reactors by means of neutron irradiation of stable elements (neutron activation). A stable nucleus absorbs a neutron and changes into a radionuclide of this element:

\[ ^{\lambda}ZX + ^{0}n \rightarrow ^{\lambda+1}ZX. \]  

(19.3.1)

It is possible to obtain radioactive cobalt this way:

\[ ^{59}\text{Co} + ^{0}n \rightarrow ^{60}\text{Co}. \]

Radioactive cobalt \(^{60}\text{Co}\) is subjected to electron decay:

\[ ^{60}\text{Co} \rightarrow ^{60}\text{Ni} + ^{0}\text{B} + \gamma + ^{0}\text{H}. \]

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.

So another method of the nuclear reaction realization is neutron activation. Radionuclides can also be obtained out of the fission product of uranium (for example radioactive iodine).

### 19.4. RADIOACTIVE DECAY LAW

Individual radioactive nuclei transform independently of one another. We may therefore consider that the number of nuclei \(dN\) decay during the small time interval \(dt\) is proportional both to the number of available nuclei \(N\) and to the time interval \(dt\):

\[ dN = -\lambda N \, dt. \]  

Here \(\lambda\) is a constant characteristic of the given radioactive substance and known as the decay constant. The minus has been taken to allow us to consider \(dN\) as an increment of the number of undecayed nuclei. Let introduce an initial condition: \(N = N_0\) at \(t = 0\).

Integration of the equation (19.4.1) leads to the expression:

\[ N = N_0 \, e^{\lambda t}. \]  

(19.4.2)

Equation (19.4.2) expresses the law of radioactive decay: the number of undecayed nuclei decreases exponentially with time.

The time during which a half of the initial number of nuclei decays is called the half-life \(T\). It is determined by the condition

\[ \frac{1}{2} N_0 = N_0 \, e^{\lambda T} \Rightarrow 2 = e^{\lambda T}. \]
Finally:

\[ T = \frac{\ln 2}{\lambda} \approx \frac{0.69}{\lambda} \]  \hspace{1cm} (19.4.3)

We can determine the half-life time from the following figure (fig. 19.4.1.):

If the decay quantity is the number of discrete elements of a set, it is possible to compute the average length of time during which an element remains in the set. This is called the **mean lifetime**, and it can be shown that it relates to the decay rate:

\[ \tau = \frac{1}{\lambda} = \frac{T}{0.69}. \]  \hspace{1cm} (19.4.4)

![Graph showing the determination of the half-life time](image)

*Fig. 19.4.1. The determine the half-life time*

When a radiopharmaceutical is given to a patient for either diagnosis or therapy, the nuclei appear in different organs in varying amounts. Suppose that in addition to physical decay with decay constant \( \lambda \), the pharmaceuticals move to another organ while it is still radioactive. Such a process can be complicated, perhaps involving storage in the gut or bladder. In other cases, the disappearance from a particular organ may be close to exponential with **biological disappearance constant** \( \lambda_j \). Let us assume for now that all the radioactive nuclei could disappear biologically. If some are bound in different chemical forms, it might be impossible. If \( N \) is the number of radioactive nuclei in the organ (not the total number originally administered), then the rate of change of \( N \) is:

\[ dN = -(\lambda + \lambda_j) N \, dt. \]

The decay constant \( \lambda + \lambda_j \) is larger than the physical decay constant. **The effective half-life** \( T_{\text{eff}} \) may be found from equation:

\[ \frac{1}{T_{\text{eff}}} = \frac{1}{T} + \frac{1}{T_j}. \]

Therefore:
\[ T_{\text{eff}} = \frac{T_j \cdot T}{T_j + T} \]  

(19.4.5)

The time of their disappearance from an organism for short-lived radionuclides is generally equal to their half-life. For example for \(^{131}\text{I}\) \(T = 8\) days, \(T_j = 138\) days, e. g. \(T << T_j\). Consequently the organism «is cleared» from a radioactive iodine by means of radionuclide decay.

On the contrary when \(T >> T_j\), the time of an organism washout is determined only by the process of the excretion. For example for \(^{137}\text{Cs}\) \(T = 30\) years, \(T_j = 70\) days.

### 19.5. Activity of a Radioactive Substance

The activity of a radioactive preparation is defined as a number of disintegrations occurring in it unite time:

\[ \dot{A} = -\frac{dN}{dt}. \]  

(19.5.1)

Let’s put (19.4.2) in (19.5.1) and differentiate with respect to time:

\[ \dot{A} = -\frac{d}{dt} \left( N_0 \cdot e^{-\lambda t} \right) = -(-\lambda)N_0 \cdot e^{-\lambda t} = \lambda N_0 \cdot e^{-\lambda t}. \]

Then:

\[ \dot{A} = \lambda N \quad \text{or} \quad \dot{A} = \frac{N}{T} \cdot 0.69. \]  

(19.5.2)

Therefore the activity \(A(t)\) is the number of radioactive transformations per second. The SI unit of activity is the becquerel (Bq):

\[ 1 \text{ Bq} = 1 \text{ transformations} \cdot \text{s}^{-1}. \]

The earlier unit of activity, which is still used occasionally, is the Curie (Ci):

\[ 1 \text{ Ci} = 3.7 \cdot 10^{10} \text{ Bq} \]

One curie is defined to be equal to the disintegration rate of 1 gm of \(^{226}\text{Ra}\), or \(3.7 \cdot 10^{10}\) disintegrations per second (d/s).

Activity decreases with time exponentially:

\[ A = A_0 \cdot e^{-\lambda t} \]  

(19.5.3)
Let’s relate the activity to a mass of radionuclide. The number of undecayed nuclei:

\[ N = \frac{m}{m_N} \]

where \( m \) is nuclei mass, \( m_N \) is one nucleus mass.

\[ m_N = \frac{M}{N_A}, \]

where \( M \) is an atomic mass, \( N_A \) is Avogadro constant. According to the formula (19.5.2):

\[ A = \frac{0.69 \cdot N}{T \cdot TM} = \frac{0.69 \cdot 6.02 \cdot 10^{23} \cdot m}{4.17 \cdot 10^{23} \cdot m} \cdot \frac{m}{TM}. \] (19.5.4)

The **unit-mass activity** (measured in Bq/kg or Ci/kg) is determined as

\[ A_m = \frac{A}{m}. \] (19.5.5)

Similarly the **specific volume activity** (unit is Bq/m³, Bq/l, Ci/l) is:

\[ A_v = \frac{A}{V}. \] (19.5.6)

**The specific surface activity** (Bq/m², Ci/m²) is equal to:

\[ A_s = \frac{A}{S}. \] (19.5.7)

### 19.6. **Interaction of the Ionizing Radiation with the Matter**

Ionizing radiation produces ions during the interaction with atoms in the matter. The toxic effect of ionizing radiation is related to the ionization. It is believed that ionization of tissues, composed mainly of water, generates \( \text{H}_2\text{O}^+ \)
and $\text{H}_2\text{O}^-$ ions, which in turn form H and OH radicals. As radicals are chemically very reactive, biological damage, such as attacks on DNA and proteins, occurs.

There are several types of ionizing radiation: alpha particles, beta particles, neutrons, and positrons are examples of particulate ionizing radiation. Gamma rays and X rays are electromagnetic ionizing radiation.

Three important quantities associated with the passage of charged particles through matter are specific ionization, linear energy transfer and mean linear range of the particle in the absorber.

**Linear specific ionization** is the total number of ion pairs $dn$ produced per length unit $dl$ of the path of the incident radiation (fig. 19.6.1): $i = dn/dl$.

Specific ionization increases with decreasing energy of the charged particle because of the increased probability of interaction at low energies. Therefore, the charged particle shows a sharp increase in ionization towards the end of the travel. This phenomenon is predominant for heavy charged particles, whereas it is negligible for electrons.

**Linear energy transfer (LET)** is the amount of energy deposited per length unit of the path by the radiation: $\text{LET} = dE/dl$.

The LET is expressed in units of keV/μm and it is very useful in concepts of radiation protection. Electromagnetic radiation and $\beta$-particles interact with matter, losing only a little energy per interaction and therefore they have low LETs. In contrast, heavy particles ($\alpha$-particles, neutrons and protons) lose energy very rapidly, producing many ionizations in a short distance, and thus they have high LETs.

**Mean linear range** of a charged particle is an absorber is the straight-line distance traversed by the particle. The range of a particle depends on the mass, charge and kinetic energy of the particle and also on the density of the absorber. The heavier and more highly charged particles have shorter ranges than lighter and lower charged particles. The range of charged particle increases with the energy of the particle.

Let us consider features of the interaction with matter for different particles.

1) **Alpha particles** are easily absorbed by materials because of their charge and large mass, and can travel only a few centimeters in air. 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 (most often because radioactive material has been inhaled or ingested), it is the most destructive form of ionizing radiation because of this high mass and strong absorption. It is the most strongly ionization that with doses large enough can cause any or all of the symptoms of radiation poisoning. It is estimated that chromosome damage from alpha particles is about 100 times greater than that caused by the same amount of other
radiation. Alpha emitter polonium-210 is suspected of playing a role in lung and bladder cancer related to tobacco smoking. The relationship between linear specific ionization and length of the path is shown in fig. 19.6.1.

\[ i = \frac{n}{l} \]

\[ (\text{ion pairs/sm}) \]

\[ l \text{(sm)} \]

\[ 2 \times 10^4 \]

\[ 8 \times 10^4 \]

\[ 2 \times 4 \times 6 \times 8 \]

\[ \text{i (ion pairs/sm)} \]

\[ \text{l (sm)} \]

**Fig. 19.6.1.** The relationship between linear specific ionization and length of the path

2) **Beta-particles** have an electrical charge and mass less than alpha-particle charge. The speed of individual beta-particles depends on the amount of the energy they have, and varies greatly. Beta particles are much more penetrating than alpha particles. 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. Therefore beta sources do present an external radiation hazard. They are, however, relatively easy to deal with by shielding. If the beta emitter is taken into the body it presents a far greater hazard, though generally less so than the alpha emitters.

Electron deceleration within water causes ionization and a bremsstrahlung X-ray as well. Positrons are able to perform an annihilation reaction with electrons and produce gamma-quants:

\[ _0^1 \text{e} + _0^1 \text{e} \rightarrow 2 \gamma \]

3) **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. Gamma rays consist of high energy photons with energies above about 100 keV. Due to their high energy content, gamma rays can cause serious damage when absorbed by living cells. 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). Gamma radiation frequently accompanies alpha and beta emissions. Gamma rays are much more penetrating than charged particles. For example, a beta particle emitted by I\(^{131}\) (radioactive iodine) has a range in tissue of a few millimeters whereas the accompanying gamma ray can pass right through the body.

Gamma rays are better stopped or shielded by dense materials, such as lead or depleted uranium.
4) **Neutrons.** Since free neutrons are electrically neutral, they pass unhindered 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 to a large extent on their energies. Neutrons will have a low probability of interaction because of the small size of the nucleus in relation to the atom as a whole, and could thus travel considerable distances in matter.

### 19.7. PRINCIPLES OF RADIONUCLIDE METHODS OF DIAGNOSTICS

Radionuclide diagnostics is based on the radionuclides incorporation in biological tissue. Incorporated radionuclides are the $\gamma$-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 $\gamma$ 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 $\beta^-$ or $\beta^+$ 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: **scintigraphy** and **radiography** (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).

**Radiography (quantitative scintigraphy)** is based 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. 19.7.1). 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). We can perform this measurement for each kidney and make a comparative assessment their work.
Gamma scintigraphy is a radiographic image technique for visualizing the distribution of an injected radionuclide within the given organ as a means of studying 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 colors or shades of grey. We can see a typical gamma-topogram on fig. 19.7.2.

19.8. Basic Physics 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 tumours 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 tumour, 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 tumours. The intensity of X-ray decrease sharply as depth increases (fig. 19.8.1, dotted line).

**Gamma-rays** have deep penetration and cause a minimum of surface-tissue irradiation. It allows to destroy deeply located tumours. Also it decreases damage to the skin and healthy tissues. Gamma radiation from $^{60}$Co has been usually used in cancer therapy.

**Electron beams** with energy 25 MeV produce a maximum ionization at depth of 1–3 sm. They are used for irradiation of not deeply situated tumours.

**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 an 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).
Part 20. Radiation dosimetry

Ionizing radiations are generally characterized by their ability to excite and ionize atoms of matter with which they interact. The important types of ionizing radiation are:

1. \( \gamma \)-rays: Electromagnetic radiation emitted from a nucleus or in annihilation reactions between matter and antimatter. The quantum energy of any electromagnetic photon is \( E = h \nu \).

2. X-rays: Electromagnetic radiation emitted by charged particles (usually electron) in changing atomic energy levels (characteristic X-rays) or when slowing down in Coulomb force field (bremsstrahlung X-rays). X-ray and \( \gamma \)-ray photons of the given quantum energy have identical properties, different only in mode origin.

3. Fast electrons: If positive in charge, they are called positrons. If they are emitted from a nucleus they are usually referred to as \( \beta \)-rays (positive or negative).

4. Heavy charged particles: usually obtained by Coulomb force field in cyclotron. Alpha-particles are also emitted by some radioactive nuclear force. Types include: proton (the hydrogen nucleus), deuteron (the deuterium nucleus, consisting of a proton and neutron bound together by nuclear forces), \( \alpha \)-particles (the helium nucleus), other heavy charged particles consisting of the heavier atoms nuclei.

5. Neutrons: Neutral particles obtained from nuclear reactions.

The primary goal of radiation dosimetry is a quantitative estimation of the energy absorption in tissue and estimation of the biological effects. Let us consider main radiation dose quantities and units.

20.1. EXPOSURE DOSE

Exposure dose at a given point is a measurement of radiation in relation to its ability to produce ionization:

\[
X = \frac{Q}{m}.
\]  

(20.1.1)

where \( Q \) is total positive ions charge, that has been produced in air mass \( m \).

The exposure dose is used only for air and only for X-rays or \( \gamma \)-rays.

The unit of measurement of the exposure dose is 1 C/kg. We can also use the off-system unit — roentgen. It is an X-ray dose or an \( \gamma \)-rays dose which action produces \( 2.08 \times 10^9 \) ion pairs in 1 sm\(^3\) (0.001293 g) of air under favorable condition.

Let’s establish a connection between these units. If the exposure dose is equal to 1R then a charge occurs: \( Q = n \cdot q = 2.08 \cdot 10^{-9} \cdot 1.6 \cdot 10^{-19} = 3.33 \cdot 10^{-10} \) C.
(n — the number of ion pairs, q — one an ion charge). This charge appears in air mass: \( m = \rho V = 1.29 \times 10^6 \) kg (\( \rho = 1.29 \) kg/m\(^3\) — air density, \( V = 10^6 \) m\(^3\) — volume).

Therefore: \( 1 \text{ R} = \frac{3.33 \times 10^{-10}}{1.29 \times 10^{-6}} = 2.58 \times 10^{-4} \text{ C/kg or } 1 \text{ C/kg} = 3876 \text{ R.} \)

**Exposure dose rate** is defined as a relation of the exposure dose according to time:

\[
\dot{X} = \frac{X}{t} \quad (20.1.2)
\]

The unit of measurement of the exposure dose rate is 1A/kg or roentgen/sec.

### 20.2. Absorbed Dose

Chemical and biological changes in the tissue exposed to ionizing radiation depend upon the energy absorbed in the tissue from the radiation. **Absorbed dose** \( D \) delivered to a small mass \( m \) in kilograms is

\[
D = \frac{E}{m}. \quad (20.2.1)
\]

where \( E \) is the energy absorbed in the medium mass \( m \) from any type of ionizing radiation.

**Kerma** is an acronym for kinetic energy released per mass unit. Kerma is the sum of the initial kinetic energies of all the charged particles emanated by uncharged ionizing radiation in a volume element, divided by the mass of matter in the volume element. Under the conditions of charged-particle equilibrium with negligible energy loss by bremsstrahlung, kerma and the absorbed dose are identical.

To describe the energy absorbed in a medium from any type of ionizing radiation, the quantity of radiation should be described in SI units of gray. One **Gray** (Gy) represents the absorption of one joule of energy per kilogram of absorbing material:

\[
1 \text{ Gy} = 1 \text{ J/kg}
\]

The traditional unit of the absorbed dose is **rad** (radiation absorbed dose): 1 Gy = 100 rad.

Note that the absorbed dose is not a good indicator of the likely biological effect. For example 1 Gy of alpha-radiation would be much more biologically damaging than 1 Gy of gamma-radiation.

The **absorbed dose rate** \( \dot{D} \) is described similarly to the exposure dose rate:

\[
\dot{D} = \frac{D}{t} \quad (20.2.2)
\]
and it is measured in Gy/s, rad/s and subunits.

If we know the exposure dose then we can estimate the absorbed dose. Under the external irradiation the absorbed dose is proportional to the exposure dose:

\[ D = f \times X, \]

where \( f \) — is a coefficient which depends on the irradiated material structure and photons energy and on the units.

Let us evaluate this coefficient for the air. If the exposure dose is equal to 1R for 1 kg of air then \( 2,08 \cdot 10^9 \) ion pairs are generated in 1 sm\(^3\) (1,29\( \cdot 10^{-6} \) kg) of air. It is necessary to spend the energy equal to \( 34 \text{ eV} = 34 \cdot 1,6 \cdot 10^{-19} \text{ J} \) for one pair ions production. Then the absorbed dose is equal to:

\[ D = \frac{E}{m} = \frac{34 \cdot 1,6 \cdot 10^{-19} \cdot 2,08 \cdot 10^9}{1,29 \cdot 10^{-6}} = 88 \cdot 10^{-4} \text{ Gy} = 0,88 \text{ rad}. \]

Therefore in exposure dose of \( X = 1\text{R} \) the absorbed dose will be 0,88 rad/R. Consequently coefficient \( f = 0,88\text{rad/R} \) for this case. For bone tissue this coefficient depends on photon energy and we may consider it takes on the value from 1 to 4,5 rad/R, decreasing with the quantum energy increase.

### 20.3. Equivalent Dose

The same absorbed dose delivered by different types of radiation may result in different degrees of biological damage to body tissues. When radiation is absorbed by biological material, the energy is deposited along the tracks of charged particles in a pattern that is characteristic of the radiation type involved. After the exposure to X-rays or gamma-rays, the ionization density would be quite low. After the exposure to neutrons, protons, or alpha-particles, the ionization along the tracks would occur much more frequently, producing a much denser pattern of ionizations. These differences in density of ionizations are the major reason that neutrons, protons, and alpha particles produce more biological effects per unit of the absorbed radiation dose than do more sparsely ionizing radiations such as X-rays, gamma-rays, or electrons.

The relative biological effectiveness for the given test radiation, is calculated in the same way as the dose of a reference radiation, usually X-rays, required to produce the same biological effect. It was seen with a test dose of another radiation. Thus, for the same biological endpoint:

\[ RBE = \frac{\text{Absorbed dose from reference radiation}}{\text{Absorbed dose from test radiation}} \]  

(20.3.1)

We can use the half-lethal dose test for animals to determine a biological effect. It is the dose which causes death of half of the irradiated animals during 30 days.
If it took 200 mGy of X-rays but only 20 mGy of neutrons to produce the same biological effect, the RBE would be \(\frac{200}{20} = 10\) using X-rays as the reference radiation. If we normalize the results to a single reference radiation, the value of the RBE provides a quantitative index of the effectiveness per unit of the absorbed dose of any radiation. This index is called the **radiation quality coefficient**.

This factor is selected for the type and energy of the radiation incident on the body, or in the case of sources within the body, emitted by the source. The value of \(k\) is 1 for X-rays, gamma-rays and beta-particles, but it is higher for protons, neutrons (\(k = 10\) for energy 0.1 to 10 MeV), alpha particles (\(k = 20\) for energy less than 10 MeV) etc.

The **equivalent dose** allows to relate different relative biological effects of ionizing radiation. The equivalent dose is more biologically significant than the absorbed dose.

The equivalent dose is calculated by multiplying the absorbed dose to the organ or tissue with the radiation quality coefficient \(k\):

\[
H = kD,
\]

where \(D\) is the absorbed dose to tissue.

If the absorbed dose unit is Gray, the measure of the equivalent dose is given in **sieverts**. One **sievert** is generally defined as the amount of radiation roughly equivalent in biologic effectiveness to one gray (or 100 rads) of gamma radiation: \(1\text{Sv} = 1\text{J/kg}\). 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 of the equivalent dose is also used: **Röntgen equivalent man** (REM or rem), 100 rem = 1 sievert.

**Equivalent dose rate** \((H)\) — is known as a relation of the exposure dose according to time:

\[
\dot{H} = \frac{H}{t}.
\]

The units of the equivalent dose rate are Sv/s, mSv/hr, rem/s etc.

### 20.4. EFFECTIVE EQUIVALENT DOSE

Let us estimate the damage of the certain equivalent dose has been delivered to some target organs. We assume that the probability of developing cancer in a target organ depends on the dose delivered to that organ but not on the dose to any other part of the body. We also assume that the probabilities are small, so that if several organs have received a the radiation dose, the probability of developing cancer is the sum of the probabilities for each organ.
The **effective dose** is the sum of the products of the equivalent to various organs or tissues \((H_i)\) and the weighting factors \((w_i)\) applicable to each of the body organs or tissues that are irradiated:

\[
H_{\text{eff}} = \sum w_i H_i. \tag{20.4.1}
\]

The tissue **weighting factor** or **coefficient of radiation risk** \(w_i\) is the radiation detriment of the damage to the whole body during constant irradiation. By definition, the sum of \(w_i\) over all organs equals: \(\sum w_i = 1\).

**Example.** If human lungs are exposed to the equivalent dose 1 Sv the probability of developing radiation-induced cancer is \(P = 2 \times 10^{-3}\). If the whole body were to receive the equivalent dose of 1 Sv, the probability of radiation induced cancer is \(P_0 = 1.65 \times 10^{-2}\). The weighting factor for lungs is:

\[
w = \frac{P}{P_0} = \frac{2 \times 10^{-3}}{1.65 \times 10^{-2}} \approx 0.12.
\]

The table shows \(w_i\) permissible to each target organ:

<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>

**Example.** Lungs were to receive the equivalent dose of 5 mSv and thyroid glands – 10 mSv. The effective equivalent dose is equal to:

\[
H_{\text{eff}} = 0.12 \times 5 + 0.03 \times 10 = 0.36 \text{ mSv}.
\]

**20.5. COLLECTIVE EFFECTIVE DOSE**

The equivalent dose characterizes the consequences of radiation exposure for a particular organ, while the effective equivalent dose — for the whole organism. To estimate radiation effects on a large group of people the collective effective dose is used.

The **collective effective dose** is defined as the sum of the individual effective doses \(H_{i\text{eff}}\) received by a people population from exposure to a specified source of radiation in a given time period:

\[
S = \sum H_{i\text{eff}} N_i, \tag{20.5.1}
\]
where \( N_i \) is the total number of individuals in the given group.

It is used to predict the magnitude of stochastic effects on the population. The collective dose is usually measured in units of **person-sieverts** or man-sieverts.

### 20.6. Ionizing-Radiation Detector

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, the track of an 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* 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.

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 (see fig. 20.6.1). 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. 20.6.2) 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.
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.

Let us consider the detection features of different particles (fig. 20.6.3). 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 $\gamma$-rays is the simplest due to their long path in matter. They may be registered even from a deep-seated object layer.

\[ \gamma \text{-rays energy (MeV)} \begin{array}{c|c}
0 & 0 \\
0.5 & 10 \\
1 & 10^4 \\
1.31 \text{ MeV} & \\
0.48 \text{ MeV} & \\
end{array} \gamma \text{-rays intensity (relative units)} \]

**Fig. 20.6.4.** The gamma-spectrum

To estimate of radiation pollution of the environment and organism it is necessary to know the elemental radiation composition. For gamma-rays this is an energy gamma-spectrum –i.e. dependence of its intensity on $\gamma$-quanta energy. Every radioactive transformation has a certain energy spectrum. The gamma-spectrum which occur under $^{47}\text{Ca}$ transformation to $^{47}\text{Sc}$ is shown in fig. 20.6.4. The device for gamma-spectrum measurement is called a *gamma-ray spectrometer*.

### 20.7. 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 individual's exposure to a dangerous environment, particularly when the hazard is cumulative over long intervals of time. 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 (see fig. 20.7.1). 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:

$$q_1 - q_2 = \Delta q = C (U_1 - U_2).$$

where $C$ is electrocapasity.

Therefore the exposure dose is equal to:

$$X = \frac{\Delta q}{m} = \frac{C(U_1 - U_2)}{c \cdot V} = k (U_1 - U_2) = k \Delta U,$$  

where $V$ is a camera volume, $m$ — air mass, $\rho$ — 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.

Exposure rate measure is based on the current determination in the detector:

$$\dot{X} = \frac{dX}{dt} = \frac{dQ}{dt \cdot m} = \frac{I}{m}.$$  

Exposure 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 or mass activity. 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 activity is calculated by the formula:

$$A_v = \frac{N_2}{t_2} - \frac{N_1}{t_1}P,$$

(20.7.4)

where coefficient $P$ take into account a specimen volume and radiometer sensitivity to different radiation types.

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.

20.8. 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 or she lives, the food that is consumed, the radon levels in the house, and so on. Let us consider it in detail.

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 (4.5·10^9 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, 238U and 235U, and one of thorium 232T, 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, 40K and 87Rb make the greatest contribution to the background dose so the dose accrue 0.65 mSv per year.

Internal background is the dose imposed by the isotopes contained in our bodies. A small percentage of the potassium in the human body is 40K. This radioactive nuclide emits both locally absorbed beta radiation and more penetrating gamma radiation. Similarly, 14C, which comprises a small percentage of the carbon atoms found in organic molecules throughout our bodies, contributes to the total dose of 1.35 mSv/yr from internal background.

Radon, as part of the 238U 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 226Ra 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 238U. 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 (218Po and 214Po), 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 mSv per year.

Various human activities add to the annual radiation background. At first it is diagnostic radiology (0.39 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 3H, a low-energy beta emitter, and smoke detectors that contain 241Am, an alpha emitter. Luminous dials on watches, clocks, and instruments contained 3H and 147Pm, 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.

The table consists of certain examples of the equivalent dose rates:

<table>
<thead>
<tr>
<th>Radiation source</th>
<th>Average equivalent dose rate for all body (µSv/h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flight by plan (height is 8 km)</td>
<td>1.35</td>
</tr>
<tr>
<td>Color television set at a distance:</td>
<td></td>
</tr>
<tr>
<td>5 sm away from the screen</td>
<td>100</td>
</tr>
<tr>
<td>250 sm away from the screen</td>
<td>2.5⋅10^{-3}</td>
</tr>
<tr>
<td>District of thermal power station (radius of 20 km)</td>
<td>0.6⋅10^{-3}</td>
</tr>
</tbody>
</table>
Content

Part 1. Mathematical description of medico-biological processes and medical data processing ............................................. 3
1.1. The derivative and differential of a function ................................................................. 3
  1.1.1. The derivative of a function ........................................................................ 3
  1.1.2. The multiple derivatives ........................................................................... 6
  1.1.3. The searching of the function extremums.................................................. 8
  1.1.4. Partial derivatives ....................................................................................... 8
  1.1.5. Differential of function .............................................................................. 10
  1.1.6. Partial differentials. Total differential of function ..................................... 11
1.2. Antiderivative function. Indefinite and definite integral .......................................... 11
1.3. Differential equations .............................................................................................. 19
1.4. Probability theory ................................................................................................. 25
  1.4.1. The main laws of probability theory.......................................................... 26
  1.4.2. Additive Rule for probabilities of disjoint events ...................................... 27
  1.4.3. The multiplication rule for probabilities of independent events ............. 27
  1.4.4. The multiplication rule for probabilities of the dependent events .......... 27
  1.4.5. Practical tasks ............................................................................................. 32
1.5. Random variables ................................................................................................. 34
  1.5.1. The low of distribution of the discrete random variables ......................... 34
  1.5.2. The distribution law of the continuous random variables.
  Density of probability distribution ........................................................................ 35
  1.5.3. Numerical characteristics of random variable distributions ................... 36
  1.5.4. The normal law of random variable distribution ....................................... 38
  1.5.5. The normal distribution’s properties ......................................................... 40
  1.5.6. Problems to practical employment ............................................................ 41
1.6. The elements of math statistics ............................................................................. 41

Part 2. Mechanic oscillations and waves. Sound and hearing ......................................... 48
2.1. Harmonic Oscillations .......................................................................................... 48
2.2. Damped Harmonic Oscillation ............................................................................. 53
2.3. The Driven Harmonic Oscillations ....................................................................... 54
2.4. Characteristics of a wave. Sound .......................................................................... 55
2.5. The Weber–Fechner law ....................................................................................... 61
2.6. Ultrasound, its characteristics and applying in medicine ..................................... 65
2.7. The applying of ultrasound in medicine. The Doppler’s effect ....................... 67

Part 3. Biophysical principles of biorheology and hemodynamics ................................. 73
3.1. Bernoulli’s equation ............................................................................................ 73
3.2. The laminar (streamlined) and turbulence flow ................................................... 75
3.3. Hagen–Poiseuille law ......................................................................................... 80
3.4. Surface Tension. Capillarity ............................................................................. 83
3.5. Surfactant ........................................................................................................... 87
<table>
<thead>
<tr>
<th>Part 11. The Sensors</th>
<th>166</th>
</tr>
</thead>
<tbody>
<tr>
<td>11.1. Piezoelectric effect</td>
<td>166</td>
</tr>
<tr>
<td>11.2. Temperature sensors</td>
<td>167</td>
</tr>
<tr>
<td>11.3. Passive sensors</td>
<td>168</td>
</tr>
<tr>
<td>11.3.1. Capacitive sensors</td>
<td>168</td>
</tr>
<tr>
<td>11.3.2. Inductive Displacement Sensors</td>
<td>169</td>
</tr>
<tr>
<td>Part 12. Electromagnetic waves. Light polarization</td>
<td>170</td>
</tr>
<tr>
<td>12.1. Electromagnetic equation</td>
<td>170</td>
</tr>
<tr>
<td>12.1.1. The electromagnetic spectrum of radiation</td>
<td>171</td>
</tr>
<tr>
<td>12.2. Polarization of light</td>
<td>172</td>
</tr>
<tr>
<td>12.2.1. Polarization by reflection</td>
<td>175</td>
</tr>
<tr>
<td>12.2.2. Optical anisotropy</td>
<td>176</td>
</tr>
<tr>
<td>12.2.3. Polarization in double refraction</td>
<td>177</td>
</tr>
<tr>
<td>12.2.4. The Nicol prism</td>
<td>177</td>
</tr>
<tr>
<td>12.2.5. Phenomenon of dichroism</td>
<td>178</td>
</tr>
<tr>
<td>12.2.6. Polarized light transition through a polarizer. Malus's law</td>
<td>179</td>
</tr>
<tr>
<td>12.2.7. Optical activity</td>
<td>181</td>
</tr>
<tr>
<td>Part 13. Thermal Radiation</td>
<td>183</td>
</tr>
<tr>
<td>13.1. Basic concepts of thermal radiation</td>
<td>183</td>
</tr>
<tr>
<td>13.2. Blackbody</td>
<td>184</td>
</tr>
<tr>
<td>13.3. Thermal radiation laws</td>
<td>185</td>
</tr>
<tr>
<td>13.3.1. Kirchhoff law</td>
<td>185</td>
</tr>
<tr>
<td>13.3.2. Stefan–Boltzmann law</td>
<td>186</td>
</tr>
<tr>
<td>13.3.3. Wien’s Displacement law</td>
<td>187</td>
</tr>
<tr>
<td>13.4. Heat transfer mechanisms in cooling the human body</td>
<td>188</td>
</tr>
<tr>
<td>13.5. Infrared Radiation from the human body</td>
<td>189</td>
</tr>
<tr>
<td>Part 14. Spectral analysis</td>
<td>192</td>
</tr>
<tr>
<td>14.1. Types of spectrum</td>
<td>192</td>
</tr>
<tr>
<td>14.2. Bohr’s theory of the hydrogenic atom</td>
<td>192</td>
</tr>
<tr>
<td>14.3. Energy States of a Hydrogen Atom</td>
<td>194</td>
</tr>
<tr>
<td>14.4. Molecular spectrum</td>
<td>198</td>
</tr>
<tr>
<td>14.5. The spectral devices</td>
<td>200</td>
</tr>
<tr>
<td>Part 15. Luminescence</td>
<td>202</td>
</tr>
<tr>
<td>Part 16. Laser</td>
<td>204</td>
</tr>
<tr>
<td>16.1. Processes of absorption and emission in atomic system</td>
<td>204</td>
</tr>
<tr>
<td>16.2. Construction of a laser</td>
<td>207</td>
</tr>
<tr>
<td>16.3. Characteristics of laser light</td>
<td>209</td>
</tr>
<tr>
<td>16.4. The Ruby Laser</td>
<td>209</td>
</tr>
<tr>
<td>16.5. Types of lasers</td>
<td>210</td>
</tr>
<tr>
<td>16.6. Medical Applications</td>
<td>211</td>
</tr>
</tbody>
</table>
Part 17. Physics of eye vision

17.1. Lens optic
17.2. Aberrations
17.3. Magnifying glass
17.4. Eye structure
17.5. Formation of image by eye
17.6. Accommodation
17.7. Visual acuity
17.8. Errors of refraction
17.9. Anatomy and function of the retina
17.10. Generation of receptor potential
17.11. Light and dark adaptation
17.12. Color vision

Part 18. X-Rays

18.1. Brems-strahlung X-ray
18.2. Characteristic X-rays
18.3. Interaction between X-ray and matter
18.4. Attenuation of X-rays
18.5. Physical principles of the X-ray diagnostics

Part 19. Radioactivity

19.1. General characteristics of nucleus
19.2. Modes of radioactive decay
19.3. Simple nuclear reactions. Methods of radionuclides production
19.4. Radioactive decay law
19.5. Activity of a radioactive substance
19.6. Interaction of the ionizing radiation with the matter
19.7. Principles of radionuclide methods of diagnostics
19.8. Basic physics of the radiation therapy

Part 20. Radiation dosimetry

20.1. Exposure dose
20.2. Absorbed dose
20.3. Equivalent dose
20.4. Effective equivalent dose
20.5. Collective effective dose
20.6. Ionizing-radiation detector
20.7. Radiation monitoring instruments
20.8. Background radiation