RUDN university

5100

V. L. Klyushin

MATHEMATICS.
TEXTBOOK FOR STUDENTS MAJORING IN NON MATHEMATICAL SUBJECTS

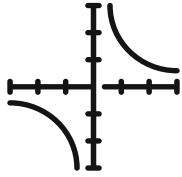
RUDN University, 2019

RUDN university

5100

V. L. Klyushin

MATHEMATICS.
TEXTBOOK FOR STUDENTS MAJORING IN NON MATHEMATICAL SUBJECTS



Moscow
RUDN University
2019

Foreword

This textbook is written by the author based on many years (more than 50 years) of lecturing and conducting higher mathematics classes on the non-mathematical faculties of the RUDN University. Many years of the author's experience in the Faculty of Science are taken into account.

In this textbook, the basics of the higher mathematics are presented. The content of the textbook is necessary for students studying different specialties. In particular, these are economics, medicine, chemistry and engineering, agricultural, humanitarian specialties.

The author tried to give the material strictly but simple to not just share the information about the higher mathematics but to interest students in mathematics, to open their minds and to inculcate the mathematical culture on them.

While writing this textbook the author has used some materials, tricks and finds from the author's textbooks and tutorials published in 2006-2019. However, the amount of such tricks and finds in this textbook is significantly replenished with the new ones which were not mentioned before in the author's publications. But, first of all, it is important to mention the chapters and sections of this textbook which have not been published by the author.

In particular, these are the following materials:

The chapter "The surfaces of the second-order";

The examples of the application of the derivative and the differential in biology and chemical engineering;

Biological applications of the definite integral;

The application of the integral calculus to the study of chemical processes, the process of radioactive decay and the calculation of a mean lifetime of a radioactive atom;

The Chapter “Double integrals”;

The Chapter “Triple integrals”;

The Chapter “Fourier series”, etc.

Experience has shown that for many students starting to study a university course in mathematics, problem-solving is a significant difficulty. That is why this textbook shows how to solve typical examples and problems which illustrate and explain the theoretical material.

A few words should be said about the exposition of the material – it is heterogeneous. At the beginning of the course, as well as when it comes to basic mathematical concepts and theorems of mathematical analysis and analytical geometry, the author adhered to a detailed presentation. It might seem to be too detailed and simple for a strong student. The author thinks that the basic definitions and theorems are the minimum that has to be learnt by all readers without any exceptions. The other parts of the book touching more complicated and deep theoretical questions are presented by the author in a shorter style.

The whole content of this textbook might be given approximately 144 academic hours.

The material of this textbook was tested by the author while giving lectures to students of RUDN University at the Engineering Academy, Agrarian and Technological Institute, Institute of Medicine and the Faculty of Humanities and Social Sciences.

The author hopes that this book, written as a textbook for students, will be useful to all teachers of higher mathematics and all applying its apparatus.

Chapter 1. The basics of set theory

1.1. The definition of a set

In mathematics, a **set** is a well-defined collection of distinct objects in its own right. The definition of a set is known to be a basic mathematical principle, which means it has no strict definition. G. Cantor once said: "A set is gathering together into a whole of definite, distinct objects of our perception or of our thought- which are called elements of the set".

There are some examples of sets: a set of vertices or diagonals of a polygon, a set of all solutions of an equation, a set of all the books that form a library and etc. A set might consist of a finite or infinite amount of objects. Those objects that form a set are called **elements** or **points.** Usually (but not always) a set is denoted using capital letters, while its elements are denoted using lower case. A set can be specified by enumerating its elements or by indicating the characteristic properties of its elements, or, in other words, by properties that every single element has. For instance, $A = \{2,4,7,8\}$ is a set that consists of number 2,4,7,8. Or $A = \{x: x > 0\}$ which is a set of all positive real numbers.

If a is an element of a set A, then it's denoted as: $a \in A$; overwise if a is not an element of a set A, then it's denoted $a \notin A$. The symbol $\in$ is called set membership.

A set that consists of no elements is called **empty** and denoted as $\emptyset$. For instance, a set of real solutions of the equation $x^2 + 1 = 0$ is empty.

A set is called **finite** if it consists of a finite amount of elements. Overwise its called **infinite**.

A set A is called a subset of the set B, if each element of the set A is also an element of the set B (denoted as $A \subset B$). An empty set is a subset of any set by definition. The symbol $\subset$ is called an inclusion.

Two sets are called **equal** if they consist of the same elements. Equality is denoted as $A = B$, which also means that $A \subset B$ and $B \subset A$.

1.2. Basic operations. Countable and uncountable sets

Definition. The **union** (or **addition**) of two sets A and B is the set C of all elements of either A or B; it's denoted by

$C = A \cup B$.

Definition. The **intersection** of sets A and B is the set C of all elements that are members of both A and B; it's denoted by

$C = A \cap B$.

Definition. The **complement of B in A** is the set of all elements that are members of A but not members of B; it's denoted as:

$C = A$.

***Example*1.1.** Sets$A = \{2,3,4,7\}$ and $B = \{1,3,5,8\}$ are given. Find the union and the intersection of sets *A* and *B*.

Solution: $A \cup B = \{1,2,3,4,5,7,8\}$, $A \cap B = \{3\}$, $A = \{2,7\}$.

Properties of $\cup$ and $\cap$:

1. $A \cup B = B \cup A$ and $A \cap B = B \cap A$ (commutativity);
2. $(A \cup B) \cup C = A \cup (B \cup C)$ and $(A \cap B) \cap C = A \cap (B \cap C)$ (associativity);
3. $A \cup A = A$ and $A \cap A = A$ (idempotency);
4. $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ and $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ (distributivity).

An infinite set is called **countable** if all its elements can be enumerated by natural numbers. Overwise it is **uncountable**. It's known that a set

of real numbers is countable and the set of real numbers between 1 and 0 is uncountable.

1.3. Numerical sets and numerical line

The set of all real numbers is denoted as **R**. It's worth mentioning the following subsets of R: **N** is a set of all-natural numbers (in other words positive integer), **Z** is a set of all integers (both positive and negative and zero), **Q** is a subset of all the numbers ration, **I** is a subset of all the numbers irrational. Recall that a number $\frac{m}{n}$,(where m and n - integers, $n \neq 0$.) can be defined as rational. Any rational number is either an integer, or represented by a finite decimal fraction, or by a periodic infinite decimal fraction. Any real number that is not rational is called **irrational.** An irrational number is a non-periodic decimal fraction, numbers like $\sqrt{2}$, $\sqrt{3}$, π, lg7. are irrational. For example, let's prove, that lg7 is an irrational number. Suppose it is a real number: $\lg 7 = \frac{m}{n}$,where *m* and *n* are integers. Then, $10^{\frac{a}{b}} = 7$ or, $10^a = 7^b$ which is impossible, since the left side of this equality is an even number and the right one is odd.

Obviously, $\mathbf{N} \subset \mathbf{Z} \subset \mathbf{Q} \subset \mathbf{R}$, $\mathbf{I} \subset \mathbf{R}$, $\mathbf{Q} \cap \mathbf{I} = \emptyset$, $\mathbf{R} = \mathbf{Q} \cap \mathbf{I}$. . Note that the sets **N**, **Z**, **Q** are countable, and the sets **I** are **R** are uncountable.

Note the **continuity** property of the set **R** of all real numbers:

Let X and Y be two sets of real numbers. Then, if the inequality $x \leq y$, is verified for any numbers $x \in X$ and $y \in Y$, then there exists at least one such number c, that for all x and y the inequalities $x \leq c \leq y$ is verified.

It is easy to see that the set **Q** of all rational numbers is not continuous. For instance, if X is the set of all rational numbers x that are less than π, and Y is the set of all rational numbers y greater than π, there is no rational number c, such that for all x and$y$the inequalities $x \leq c \leq y$ is verified.

A **numerical line** (or **numerical axis**) is a line on which a reference point, a positive direction and a scale are selected, in other words unit of length:

Fig. 1.1. *Number line*

There is a **one-to-one correspondence** between the set **R** of all real numbers and the set of all points of the number line: to each real number there corresponds one definite point of the number line, and vice versa, to each point of the line there corresponds one definite real number. Having established this one-to-one correspondence, we identify the points of the number line and the corresponding real numbers. The concepts of "number *x*" and "point *x*" become indistinguishable. Therefore, often instead of "point *x*", they say "number *x*" and vice versa. We can say, for example: "Take point 5," or, pointing to a point on a number line, say: "Take this number."

Let's note the simplest numerical sets. Let a and b be two numbers, and $a < b$ then:

line segment $[a, b]$ is a set of all the numbers *x*, that satisfy $a \leq x \leq b$;

interval (a, b) is a set of all the numbers *x*, that satisfy $a < x < b$;

half-intervals $(a, b]$and $[a, b)$ are the set of all the numbers that correspond to$a < x \leq b$ and $a \leq x < b$.

In particular intervals and half-intervals can be infinite: $(-\infty, a)$, $(b, +\infty)$, $(-\infty, +\infty)$, $(-\infty, a]$, $[b, +\infty)$ (Obviously, the interval $(-\infty, +\infty)$ is the whole number line.)

The general term for all the above sets is the **gap**. By saying "interval", we mean either a segment, or an interval, or a half-shaft.

The **neighborhood** of a point is any interval containing this point a. The interval $(a-\varepsilon, a+\varepsilon)$ is called the ε-**neighborhood** of a.

1.4. Module of the real number

Definition. **The module** (or **absolute value**) of a real number x is called the number x itself if it is positive, and the number opposite to the number x, if x is negative:

$$|x| = \begin{cases} x, & \text{если } x \geq 0, \\ -x, & \text{если } x < 0. \end{cases}$$

Obviously, $|x| \geq 0$.

In particular the following **properties of modules** are known:

$|x+y| \leq |x| + |y|$; $|x-y| \geq |x| - |y|$; $|xy| = |x| \cdot |y|$;

$\left|\frac{x}{y}\right| = \frac{|x|}{|y|}$.

The modulus of the difference of two numbers $|x-a|$ is the distance between points x and a of the number line, in particular, $|x|$ is the distance from point 0 to point x. The set of points x satisfying the condition $|x-a| < \varepsilon$ is, obviously, the $\mathcal{E}$ -neighborhood of a.

1.5. Mathematical induction

The method of mathematical induction is one of the most important methods of mathematical proof. It is used to prove statements that depend on a positive integer n.

The method of mathematical induction: in order to prove a statement depending on a positive integer n, one must:

1) verify if statement is true at $n = 1$ (or at least at n, where the statement makes sense);

2) verify if statement is true at $n = k$, and then the same for $n = k + 1$.

Then we make a conclusion is the statement true for any n.

Example **1.2.** Prove that $1 + 3 + 5 + \ldots + (2n-1) = n^2$.

Solution. Denote $1 + 3 + 5 + \ldots + (2n-1) = S_n$.

1. Obviously, when $n = 1$ the statement is verified: $1 = 1^2$.

2. We assume that $S_k = k^2$, let's prove that $S_{k+1} = (k+1)^2$. Really,

$S_{k+1} = S_k + [2(k+1) - 1] = k^2 + (2k+1) = (k+1)^2$,

Using the mathematical induction, we make a conclusion that $S_n = n^2$.

***Example* 1.3.** Prove that the compound interest formula $S_n = S_0(1 + \frac{i}{100})^n$, where S_0 is initial capital, *i* – interest rate, *n* is the number of accrual periods is verified.

Solution. 1. When $n = 1$ we have $S_1 = S_0 + S_0 \cdot \frac{i}{100} = S_0(1 + \frac{i}{100})$, i.e. the formula is true.

2. Assume that $S_k = S_0(1 + \frac{i}{100})^k$.

Let's prove that $S_{k+1} = S_0(1 + \frac{i}{100})^{k+1}$:

$$S_{k+1} = S_0\left(1 + \frac{i}{100}\right)^k + S_0\left(1 + \frac{i}{100}\right)^k \cdot \frac{i}{100} =$$

$$= S_0(1 + \frac{i}{100})^k(1 + \frac{i}{100}) = S_0(1 + \frac{i}{100})^{k+1},$$

Q.E.D. The formula is verified.

1.6. Union and Newton's binomial

Union

Let *X* a set consisting of *n* elements: $X = \{x_1, x_2, \ldots, x_n\}$. We will form various subsets out of the elements of *X*, which are called the **union.** Depending on whether the union contains all the elements of the set *X* or part of them, and whether the arrangement of the elements plays a role, three types of compounds are **distinguished:**

- variations;
- permutation;
- combinations.

Definition. Unions containing each *m* elements from the data of *n* elements of the set *X*, which differ from each other either by the elements themselves or by the order of their arrangement, are called **variations** of *n* elements in *m*

For example, when scheduling a specific day in a class where 10 subjects are studied and 5 lessons each day, placement of 5 elements out of 10 is considered.

The number of placements of *n* elements in *m* is denoted by A_n^m. Let us prove that the formula is valid:

$$A_n^m = n(n-1)(n-2)...[n-(m-1)],\ 1 \le m \le n. \qquad (1.1)$$

Let $m = 1$.

We can pick 1 element from n by n ways: $A_n^1 = (n-1+1) = n$

We assume that the formula is valid for $m = k$: $A_n^k = n(n-1)...(n-k+1)$.

Let $m = k+1$.

Considering, that after picking k elements elements left. And we can pick 1 element $n-k$ ways, $n-k = n-(k+1)+1$, we obtain $A_n^{k+1} = n(n-1)...(n-(k+1)+1)$

Q.E.D.

Definition. Compounds, each of which contains *n* elements of the set *X* and which differ only in the order of elements, are called **permutations** of *n* elements**.**

The number of permutations of n elements is denoted by P_n.

Permutations are a special case of variations when $m = n$.According to formula (1.1)

$$P_n = A_n^n = n(n-1)(n-2)\cdots 1, \text{ or } P_n = 1 \cdot 2 \cdots (n-1)n. \qquad (1.2)$$

Multiplication $1 \cdot 2 \cdots (n-1)n$ is «n factorial» and denoted n!. When $n = 0$ we consider $0! = 1$. Formula (12.2) can be represented as:

$P_n = n!$. (1.3)

We can represent formula (1.1) using the symbol $n!$ in the form

$A_n^m = \frac{n!}{(n-m)!}$. (1.1')

Definition. Unions containing each m elements from given n elements of the set X that different from each other by at least one element are called **combinations** of n elements by m.

The arrangement of elements within the combination is not taken into account. The number of combinations of n elements in m is denoted C_n^m. From the definition, it follows that

$A_n^m = C_n^m P_m$.

Thus

$C_n^m = \frac{A_n^m}{P_m} = \frac{n(n-1)(n-2)\cdots[n-(m-1)]}{m!}$, (1.4)

or

$C_n^m = \frac{n!}{m!(n-m)!}$. (1.4')

It follows from the last formula: $C_n^m = C_n^{n-m}$ for all $0 \le m \le n$.

It can be proved that

$C_n^{m+1} + C_n^m = C_{n+1}^{m+1}$. (1.5)

Newton's binomial formula

For any real *n* the formula

$(a+b)^n = C_n^0 a^n + C_n^1 a^{n-1} b + \ldots + C_n^m a^{n-m} b^m + \ldots + C_n^n b^n$ (1.6)

is called **Newton's binomial formula**.

To prove the validity of formula (1.6), we apply the method of mathematical induction.

1. Let $n = 1$.

$(a+b)^1 = C_1^0 a + C_1^1 b = a + b$

(we used $C_1^0 = \frac{1}{0!} = 1$, $C_1^1 = \frac{1}{1!} = 1$).

2. Assuming that formula (1.6) is true for $n = k$, we prove that it is true for$n = k + 1$, i.e. prove that

$$(a+b)^{k+1} = C_{k+1}^0 a^{k+1} + C_{k+1}^1 a^k b + \ldots + C_{k+1}^{m+1} a^{k-m} b^{m+1} + \ldots + C_{k+1}^k \mathrm{ab}^k + C_{k+1}^{k+1} b^{k+1}. \quad (1.7)$$

Next:

$$(a+b)^{k+1} = (a+b)^k (a+b) = (C_k^0 a^k + C_k^1 a^{k-1} b + \ldots + C_k^m a^{k-m} b^m + \ldots + C_k^k b^k)(a+b) =$$

$$C_k^0 a^{k+1} + C_k^1 a^k b + \ldots + C_k^{m+1} a^{k-m} b^{m+1} + \ldots + C_k^k \mathrm{ab}^k + C_k^0 a^k b + \ldots + C_k^m a^{k-m} b^{m+1} + \ldots + C_k^{k-1} \mathrm{ab}^k + C_k^k b^{k+1}$$

We obtain:

$$(a+b)^{k+1} = C_k^0 a^{k+1} + (C_k^0 + C_k^1) a^k b + \ldots + (C_k^m + C_k^{m+1}) a^{k-m} b^{m+1} + \ldots + (C_k^{k-1} + C_k^k) \mathrm{ab}^k + C_k^k b^{k+1}.$$

Considering $C_k^0 = 1 = C_{k+1}^0$, $C_k^0 + C_k^1 = C_{k+1}^1$, $C_k^m + C_k^{m+1} = C_{k+1}^{m+1}$, $C_k^{k-1} + C_k^k = C_{k+1}^k$, $C_k^k = 1 = C_{k+1}^{k+1}$ [see formulas (1.4'), (1.5)], we obtain (1.7). Using the method of mathematical induction we obtain that (1.6) is valid for all *n*.

The coefficients C_n^0, C_n^1, ..., C_n^m, ..., C_n^n in (12.6) are called **binomial coefficients**.

Questions

1. Does any set contain an infinite number of elements
2. Can the following statements be true for sets *A* and *B*: "*A* is a subset of the set *B*" and "*B* is a subset of the set *A*"?
3. In what case does the union of two sets coincide with their intersection?

4. What is the difference between set *A* and set *B*?
5. What is the complement of the set *A* to the set *B*?
6. Is a countable set finite or infinite?
7. What numbers are called rational? Is the set of all rational numbers countable or is it uncountable?
8. What is the one-to-one correspondence between the set of all real numbers and the set of all points of the number line?
9. What general term is used for the name of a numerical set, which is either a segment, or an interval, or a half-interval?
10. Is equality always true$\sqrt{a^2} = a$? If not, what is this root $\sqrt{a^2}$ equal to?
11. What is the geometric meaning of the module of a real number?
12. Is it possible to say that the modulus of the sum of two real numbers is equal to the sum of their modules? Is a similar statement
13. What double inequality is equivalent to inequality $| a | < b$?

ELEMENTS OF ANALYTICAL GEOMETRY

Chapter 2. Lines on the plane

2.1. Basic concepts

Let a coordinate system be given on the plane, and let we have on the plane some line (a straight line or a curve).

Definition The following equation is called the **equation of a line**:

$$F(x, y) = 0. \qquad (2.1)$$

Coordinates of any point belonging to this line satisfy this equation, and the coordinates of any point not belonging to this line do not satisfy this equation.

In short, equation (7.1) is the equation of a line if it satisfies the coordinates of all those and only those points that belong to this line.

***Example* 2.1.** Write the equation of the set of points equidistant from the axis *Ox* и and point $A(0, 2)$.

Solution. It's known that distance between

$M_1(x_1, y_1)$ and $M_2(x_2, y_2)$ is calculated using:

$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$,

and the distance from a point to the *Ox* axis is the ordinate of that point, taken with the corresponding sign.

Let $M(x, y)$ – random point on a line. Then $\mathrm{MM}_0 = \mathrm{MA}$ (fig. 2.1) or

$y = \sqrt{x^2 + (y - 2)^2}$.

(Here, obviously, $y > 0$.) Square both sides of the equation:

$y^2 = x^2 + y^2 - 4y + 4$,

We obtain:

$y = \frac{x^2}{4} + 1.$

This line is a parabola with a vertex at a point (0, 1).

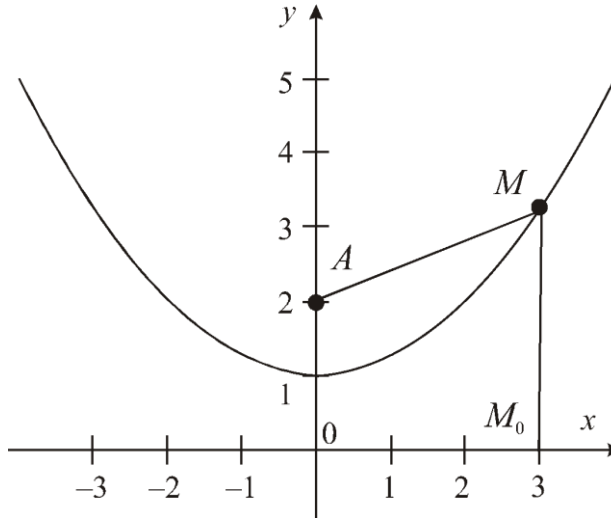


Fig. 2.1. *Parabola* $y = \frac{x^2}{4} + 1$

Note that in many cases, from equation (2.1), we can explicitly express *y* in terms of *x*. Then we get the equation of the line in the form $y = f(x)$.

In those cases when the line is defined by an algebraic equation of the *n*th order (in particular, of the first or second-order), then this line is called a line of the n^{th} order (respectively, of the first or second-order). For example, lines $y = 3x^2$, $x^2 + y^2 - 4 = 0$ are lines of the second order, $2x - 5y + 3 = 0$ are lines of the first order $y = x^3 + 3x + 1$, $x^2y + y^2 - 1 = 0$ lines of the third order.

2.2. General equation of a line of first order. Direct on the plane

The concept of a vector is well known from school mathematics. Recall that on a plane, a vector is defined by its coordinates: $\bar{a} = (a_1, a_2)$ the addition and the multiplication of the vector by a number are defined by coordinates. The scalar product $(\bar{a}, \bar{b})$of vectors $\bar{a} = (a_1, a_2)$ and $\bar{b} = (b_1, b_2)$ is the product of their modules by the cosine of the angle between them, in other words, it is a number $|\bar{a}| \cdot |\bar{b}|\cos\phi$, where ϕ is the angle between the vectors, $(\bar{a}, \bar{b}) = a_1 b_1 + a_2 b_2$. The scalar product of nonzero vectors is equal to zero if and only if the vectors are perpendicular.

The lines of the first order are the lines that are defined by the equation (2.1) that is linear, i.e. an algebraic equation that contains the variables *x* and *y* only to the first degree:

$$\mathrm{Ax} + \mathrm{By} + C = 0. \qquad (2.2)$$

Here $B \neq 0$, than *y* will be:

$y = -\frac{A}{B}x - \frac{C}{B}$,

Or by denoting $k = -\frac{A}{B}$, $b = -\frac{C}{B}$:

$$y = \mathrm{kx} + b. \qquad (2.3)$$

Equation (2.3) is called the **equation of a line with an angular coefficient** *k*. Here $k = \mathrm{tg}\phi$, where ϕ is the angle between the direct and positive direction of the *Ox* axis (Fig. 7.2).

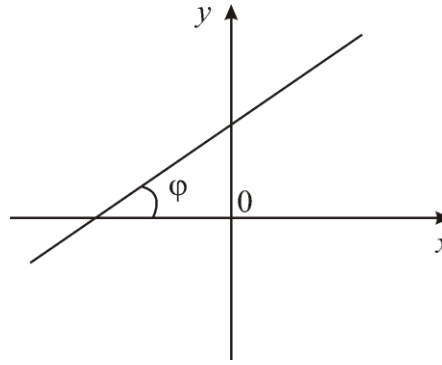


Fig. 2.2. *Line with an angular coefficient* $k = \text{tg}\phi$

If $B = 0$ in equation (2.2), then the straight line is perpendicular to the *Ox* axis, its angular coefficient is not defined, and the equation has the form $x = a$.

It might be handy to know some varieties of the equation of the line.

1. If the *angular coefficient* k and the *point* $M(x_0, y_0)$ through which the line passes are known, then obviously the identity

$$y_0 = \text{kx}_0 + b. \qquad (*)$$

Subtracting this identity from equation (2.3), we obtain the **equation of a line with a given angular coefficient and passing through a given point**:

$$y - y_0 = k(x - x_0). \qquad (2.4)$$

2. If the line *passes through two points* $M_0(x_0, y_0)$ *and* $M_1(x_1, y_1)$, then in addition to the identity (*), the identity also holds:

$$y_1 = \text{kx}_1 + b. \qquad (**)$$

From (*) and (**) we obtain:

$$k = \frac{y_1 - y_0}{x_1 - x_0}$$

and taking into account (7.4) we obtain **the equation of a line passing through given points**:

$$y - y_0 = \frac{y_1 - y_0}{x_1 - x_0}(x - x_0). \qquad (2.5)$$

***Example* 2.2.** Make an equation of a line passing through points $M_0(-2,-1)$ and $M_1(1,5)$.

Solution. Apply those coordinates to (2.5):

$y + 1 = \frac{5+1}{1+2}(x + 2)$, or $y = 2x + 3$.

3. If the point $M_0(x_0, y_0)$ through which the line passes and the vector $\bar{n} = (A, B)$ perpendicular to this line is known, then the equation of the line has the form:

$$A(x - x_0) + B(y - y_0) = 0. \qquad (2.6)$$

Let us prove this. Let $M(x, y)$ be a random point of a given line. Then $\overline{M_0M} = (x - x_0, y - y_0)$. By condition $\overline{M_0M} \perp \bar{n}$, and this is equivalent to $(\overline{M_0M}, \bar{n}) = 0$. Writing this equation in coordinate form, we obtain:

$A(x - x_0) + B(y - y_0) = 0$,

Q.E.D.

Reveal the brackets in the last equation:

$\mathrm{Ax} + \mathrm{By} - \mathrm{Ax_0} - \mathrm{By_0} = 0$.

Denote $-\mathrm{Ax_0} - \mathrm{By_0} = C$, we obtain:

$\mathrm{Ax} + \mathrm{By} + C = 0$. (2.7)

So, the equation of a line is a linear equation.

Let us prove that every first-order equation of the form (2.7) is an **equation of some straight line in the plane**.

Let us assume that we have an equation of the first degree (2.7). Let at least one of the coefficients, *A* or *B*, be nonzero (otherwise it would not be an equation of the first degree). Let, for example, be $A \neq 0$. This equation always has a solution (for example, assuming $y_0 = 1$, we find $x_0 = \frac{-B-C}{A}$). Let (x_0, y_0) be some solution of equation (2.7), i.e.

$$Ax_0 + By_0 + C = 0. \quad (2.72)$$

Subtracting (2.72) from (2.7), we obtain

$A(x - x_0) + B(y - y_0) = 0.$

This equation is equivalent to equation (2.7) since it is obtained from (2.7) using identical transformations. At the same time, it is a direct equation, as proved above. Therefore, equation (2.7) is the equation of the line.

Equation (2.7) is called the **general equation of the line**, and any nonzero vector perpendicular to the line is called its **normal vector**. In particular, $\bar{n}(A, B)$ the vector is the normal vector of the line (7.7).

Consider the cases when the equation is incomplete, i.e. when one of the coefficients is zero.

Consider the cases when the equation Ax + By + C = 0 is **incomplete**, i.e. when one of the coefficients is zero.

C=0; the equation has the form Φx + By = 0 and determines a straight line passing through the origin.

$B = 0$ $(A \neq 0)$; the equation $Ax + C = 0$ has the form and defines a line parallel to the ordinate axis. This equation is reduced to the form $x = a$ where.$a = -\frac{C}{A}$.

$A = 0$ $(B \neq 0)$; the equation has the form $By + C = 0$ and defines a straight line parallel to the abscissa axis.

Now suppose that none of the coefficients in the equation$Ax + By + C = 0$ is equal to zero. Convert it to

$$\frac{x}{-\frac{C}{A}} + \frac{y}{-\frac{C}{B}} = 1.$$

Assuming $a = -\frac{C}{A}$, $b = -\frac{C}{A}$, we obtain

$$\frac{x}{a} + \frac{y}{b} = 1.$$

This equation is called the equation of the line in the "segments". This line intersects the coordinate axes at points $(a, 0)$ and $(0, b)$.

The angle between the lines

I. Let us assume that we have two lines $y = k_1x + b_1$ и $y = k_2x + b_2$, where $k_1 = \text{tg}\phi_1$, $k_2 = \text{tg}\phi_2$. Let ϕ be the angle between the lines (fig. 2.3).

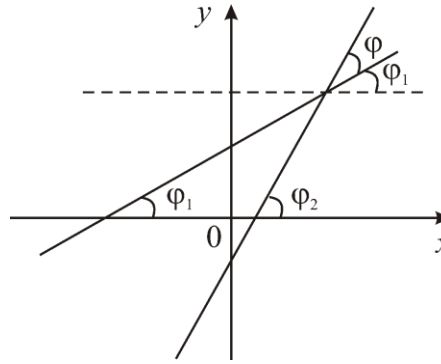


Fig. 2.3. *The angle φ between the lines*

Then $\phi = \phi_2 - \phi_1$, and using the well-known formula:

$\text{tg}\phi = \text{tg}(\phi_2 - \phi_1) = \frac{\text{tg}\phi_2 - \text{tg}\phi_1}{1 + \text{tg}\phi_1 \text{tg}\phi_2}$

or

$$\text{tg}\phi = \frac{k_2 - k_1}{1 + k_1 k_2}. \qquad (2.8)$$

From here, in particular, immediately follows the *parallelism condition:*

$k_1 = k_2$.

It is also easy to obtain *the condition of perpendicularity of the lines*:

$k_1 k_2 = -1$.

***Example* 2.3.** Find the angle between the lines:

$y = 3x + 2$, $y = -2x + 1$.

Solution. Apply $k_1 = 3$, $k_2 = -2$ to (2.8), we obtain

$\text{tg}\phi = \frac{-2-3}{1-6} = 1$.

Thus $\phi = \frac{\pi}{4}$.

Note that formula (2.8) defines one of two angles between intersecting straight lines; the other angle is $\pi - \phi$.

II. Now let two lines l_1 and l_2 be given by general equations:

$l_1 : A_1x + B_1y + C_1 = 0$

$l_2: A_2x + B_2y + C_2 = 0.$

The angle between these lines is equal to the angle between their normal vectors (or complements it to 180 °). Therefore, one of the two angles α between these lines can be calculated by the formula

$$\cos\alpha = \frac{(\bar{n}_1,\bar{n}_2)}{|\bar{n}_1||\bar{n}_2|} = \frac{A_1A_2+B_1B_2}{\sqrt{A_1^2+B_1^2}\cdot\sqrt{A_2^2+B_2^2}}. \quad (2.9)$$

***Example* 2.4.** Find the angle between the lines given by the general equations

$3x - 4y + 7 = 0, 8x - 6y + 15 = 0.$

Solution. Using (2.9):

$\cos\alpha = \frac{3\cdot8+4\cdot6}{\sqrt{3^2+4^2}\sqrt{8^2+6^2}} = \frac{24}{25}$.

Therefore, one of the angles between these lines is $\arccos\frac{24}{25}$.

The parallelism condition for lines l_1 and l_2 is the parallel condition for their normal vectors $\bar{n}_1(A_1, B_1)$ and $\bar{n}_2(A_2, B_2)$:

$$\frac{A_1}{A_2} = \frac{B_1}{B_2}. \quad (2.10)$$

The parallelism condition for lines is the parallel condition for their normal vectors and:

$$A_1A_2 + B_1B_2 = 0. \quad (2.11)$$

Half-plane

Let the line l be given by equation (2.7). Let $\bar{n}(A, B)$ be its normal vector. We divide all points of the plane that do not belong to l into two sets π_1 and π_2 as follows:

$M(x, y) \in \pi_1 \Leftrightarrow \mathrm{Ax} + \mathrm{By} + C > 0$,

$M(x, y) \in \pi_2 \Leftrightarrow \mathrm{Ax} + \mathrm{By} + C < 0$.

The set π_1 is called the *positive half-plane* with respect to the *equation of the line* (2.7), and the set π_2 is called the negative half-plane. Not that that the concept of positive and negative half-planes is defined with respect to the equation of the line, and not to the line itself. Obviously, if we multiply both sides of equation (2.7) by –1, we get the equation of the same straight line, however, in this case, the positive half-plane becomes negative, and the negative becomes positive.

It can be proved (we will not do this here) that the vector $\bar{n}(A, B)$ is directed to that part of the plane that is positive with respect to the equation of the line (2.7):

$\mathrm{Ax} + \mathrm{By} + C = 0$.

Distance from point to line

We derive the formula for the distance d from an arbitrary point $M_0(x_0, y_0)$ to the line (2.7).

The distance from the point M_0 to the line (2.7) is equal to the length of the perpendicular dropped from M_0 to line. We denote by$N(x_1, y_1)$ the base of this perpendicular, i.e. the point of intersection of the perpendicular with the line (2.7). Then according to the formula of the distance between two points

$$d = \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2}. \qquad (2.12)$$

The angular coefficient k of the line (2.7) is obviously equal to

$k = -\frac{A}{B}$.

According to the perpendicularity condition, the angular coefficient of the perpendicular MN is equal to $k' = \frac{B}{A}$, and the equation of this perpendicular (considering that it passes through a point $M_0(x_0, y_0)$) has the form

$$y - y_0 = \frac{B}{A}(x - x_0). \qquad (2.13)$$

It's possible, by solving equations (2.7) and (2.13) together to find the coordinates (x_1, y_1) of the point *N* and substitute them in (2.12). However, despite the simplicity of this solution, we will obtain bulky expressions. Therefore, we will apply another method. We use the fact that the point *N* belongs to M_0N. Thus, the unknown so far coordinates x_1, y_1, points *N* satisfy equation (2.13):

$y_1 - y_0 = \frac{B}{A}(x_1 - x_0)$.

We obtain

$\frac{y_1 - y_0}{B} = \frac{x_1 - x_0}{A}$.

Denote the total value of these fractions by δ:

$\frac{x_1 - x_0}{A} = \frac{y_1 - y_0}{B} = \delta$.

This value of δ is unknown, since x_1 and y_1 are unknown. Let's find it:

$$x_1 - x_0 = A\delta,\ y_1 - y_0 = B\delta. \qquad (*)$$

Apply these differences to the formula (2.12), we obtain

$$d = \sqrt{(A\delta)^2 + (B\delta)^2} = \sqrt{(A^2 + B^2)\delta^2} = |\delta|\sqrt{A^2 + B^2}. \qquad (2.14)$$

Note that we do not know whether the number ™ is positive or negative.

Express x_1 and y_1 from (*):

$x_1 = x_0 + A\delta,\ y_1 = y_0 + B\delta$

and apply these values to (2.7). (Recall that the point $N(x_1, y_1)$ belongs to perpendicular to the line (2.7), and to line to itself.) We obtain

$A(x_0 + A\delta) + B(y_0 + B\delta) + C = 0,$

and

$\delta = -\frac{Ax_0+By_0+C}{A^2+B^2}.$

We apply the found value of ™ to formula (2.14) and make the necessary reductions:

$d = \left|-\frac{Ax_0+By_0+C}{A^2+B^2}\right| \sqrt{A^2+B^2} = \frac{|Ax_0+By_0+C|\sqrt{A^2+B^2}}{A^2+B^2} =$

$= \frac{|Ax_0+By_0+C|}{\sqrt{A^2+B^2}}.$

And finally, we obtain:

$$d = \frac{|Ax_0+By_0+C|}{\sqrt{A^2+B^2}}. \qquad (2.15)$$

So, the *distance to the line defined by the general equation* can be found by substituting the coordinates of the point on the left side of this equation, and then dividing the module of the resulting number by the square root of the sum of the squares of the coefficients of this equation of the line.

***Example* 2.5.** Find distance from point $M_0(2,3)$ to line $4x + 3y + 8 = 0$.

Solution. Use (7.15):

$d = \frac{|4\cdot2+3\cdot3+8|}{\sqrt{4^2+3^2}} = 5.$

***Example* 2.6.** Find the distance between parallel lines l_1 and l_2:

l_1: $4x + 3y - 8 = 0$

l_2: $8x + 6y + 9 = 0.$

Solution. The distance between two parallel lines is obviously equal to the distance from any point on one of these lines to the other line. By assuming $y = 0$ in the equation the first line, we get $x = 2$. Therefore, the point $M_0(2,0)$ belongs to the first line. Find the distance from M_0 to the line l_2:

$d = \frac{|8\cdot2+6\cdot0+9|}{\sqrt{64+36}} = \frac{1}{4}.$

***Example* 2.7.** Obtain the bisector equation of the angle between the lines l_1 and l_2:

l_1: $3x - 4y + 7 = 0$,

l_2: $5x + 12y - 21 = 0$.

Solution. It is known that any point of the bisector is at the same distance from the sides of the corner. Therefore, if $M(x, y)$ is a point that belongs to the bisector of the angle between the straight lines l_1 and l_2, then

$\frac{|3x-4y+7|}{\sqrt{9+16}} = \frac{|5x+12y-21|}{\sqrt{25+144}}$.

It yields

$13|3x - 4y + 7| = 5|5x + 12y - 21|$,

or

$13(3x - 4y + 7) = \pm 5(5x + 12y - 21)$.

We get two bisector equations:

1) $39x - 52y + 91 = 25x + 60y - 105$,

$14x - 112y + 196 = 0$,

$x - 8y + 14 = 0$;

2) $39x - 52y + 91 = -25x - 60y + 105$,

$64x + 8y - 14 = 0$,

$32x + 4y - 7 = 0$.

So, the bisectors of the angles formed by intersecting straight lines l_1 and l_2 are the lines

$x - 8y + 14 = 0$ and $32x + 4y - 7 = 0$.

***Example* 2.8.** Obtain the bisector equation of the internal angle at vertex *B* of triangle *ABC*, where $A(1,1)$, $B(5,-2)$, $C(2,2)$.

Solution. We compose the equations of the parties *AB* and *BC* using (7.5):

AB: $y - 1 = \frac{-2-1}{5-1}(x - 1)$, or $3x + 4y - 7 = 0$;

BC: $y + 2 = \frac{2+2}{2-5}(x - 5)$, or $4x + 3y - 14 = 0$.

We write the equations of both bisectors of the angle between lines *AB* and *BC*:

$\frac{|3x+4y-7|}{\sqrt{9+16}} = \frac{|4x+3y-14|}{\sqrt{16+9}}$,

$|3x + 4y - 7| = |4x + 3y - 14|$,

$3x + 4y - 7 = \pm(4x + 3y - 14)$,

It yields

$l_1: x - y - 7 = 0$,

$l_2: x + y - 3 = 0$.

One of these bisectors is the bisector of the inner corner of the triangle, and the other is the bisector of the outer corner.

Next, we reason as follows:

1. If the point belongs to the bisector of the inner corner of the triangle and is located on the other side from point *B* where the triangle *ABC* is located, then the point $M_0$is in the same half-plane as point *C* with respect to line *AB*, as well as in the same half-plane as point *A* with respect to line *BC*.

2. If the point M_0 lies on the other side of *B*, then it lies in opposite planes both with respect to the line *AB* and with respect to the line *BC* (see Fig. 2.4).

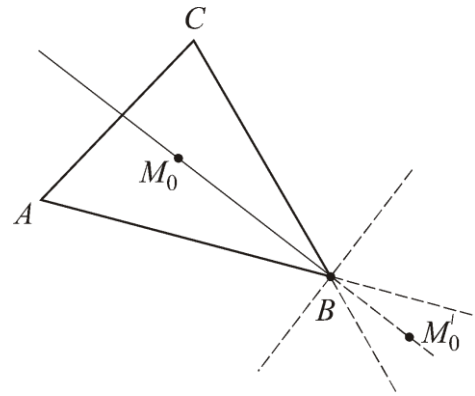


Fig. 2.4. *The bisector of the inner corner of the triangle ABC*

Let's take a random point, for instance let's take $(4, -3)$, on a line l_1. Apply its coordinates in the equation of the line *AB*: $3 \cdot 4 + 4 \cdot (-3) - 7 = -7 < 0$. Apply coordinates of the point *C* to the same line: $3 \cdot 2 + 4 \cdot 2 - 7 = 7 > 0$. Now let's apply coordinates of a point $(4, -3)$ to the line *BC*: $4 \cdot 4 - 3 \cdot 3 - 14 = -7 < 0$. Apply coordinates of a point *A* to: $4 + 3 - 14 = -7 < 0$.

So, the point $(4, -3)$ is in the same half-plane as point *A* in relation to the *BC* line, but with respect to the line AB, it is not in the same half-plane in which point *C*. Therefore, the point $(4, -3)$ belongs to the bisector of the external, not internal angle of the triangle.

So, the line l_1: $x - y - 7 = 0$ is the bisector of the external angle at the vertex *B*. Therefore, the desired bisector is the straight line l_2: $x + y - 3 = 0$.

Questions

1. What is a line equation on a plane? Give examples of line equations.
2. What is the order of an algebraic line?
3. What is the angular coefficient of a straight line in the plane? Is the slope of a straight line parallel to the axis *Oy* defined?
4. What is the normal line vector on a plane? How to determine normal vectors using the general equation of a line?
5. How to determine the acute angle between the lines that are given by the general equations?
6. How to calculate the distance between two parallel lines on a plane that is given by general equations?

Chapter 3. Second order curves

3.1. Circle. Ellipse

Definition. **A circle** is the set of all points of the plane located at the same distance (which is also called **the radius)** from a fixed point called the center of the circle.

Let the radius of the circle be equal to R and the center is a point $C(x_0, y_0)$. We derive the equation of this circle.

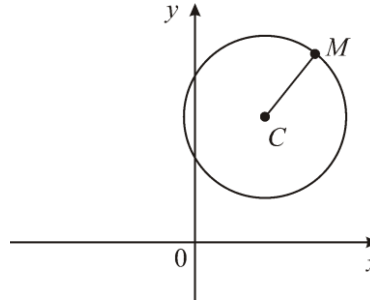


Fig. 3.1. *Circle centered at point C and radius R = CM*

For any point $M(x, y)$ on a circle the equality CM $= R$ is verified, in other words $\sqrt{(x - x_0)^2 + (y - y_0)^2} = R$.

Hence, we obtain the equation of the circle

$(x - x_0)^2 + (y - y_0)^2 = R^2$.

In particular, if the center of the circle coincides with the origin, then the equation of the circle has the form:

$$x^2 + y^2 = R^2. \qquad (3.1)$$

Equation (3.1) is called **the canonical equation of a circle**.

Definition **An ellipse** is a line, for all points of which the sum of the distances to two fixed pointscalled **foci**, is a constant and greater than the distance between the foci.

Let's obtain the ellipse equation. We choose a coordinate system such as that the *Ox* axis passes through the foci F_1 and F_2, and the axis Oy in the middle between the foci (Fig. 3.2).

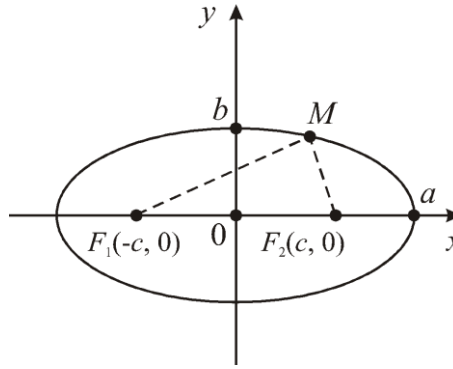


Fig. 3.2. *Ellipse* $\frac{x^2}{a^2}+\frac{y^2}{b^2}=1$

Suppose that the distance between the foci is equal to $2c$, and the sum of the distances of an arbitrary point $M(x, y)$ of the ellipse from the foci is $2a$: (according to the definition, $a > c$). Express the distance from the point $M(x, y)$ to the foci $F_1(-c, 0)$ and $F_2(c, 0)$, accordingly: $r_1 = F_1M = \sqrt{(x + c)^2 + y^2}$, $r_2 = F_2M = \sqrt{(x - c)^2 + y^2}$. Since, $r_1 + r_2 = 2a$then

$$\sqrt{(x + c)^2 + y^2} + \sqrt{(x - c)^2 + y^2} = 2a. \qquad (3.2)$$

This is an ellipse equation. Convert it.

$\sqrt{(x + c)^2 + y^2} = 2a - \sqrt{(x - c)^2 + y^2}$,

$(x + c)^2 + y^2 = 4a^2 - 4a\sqrt{(x - c)^2 + y^2} + (x - c)^2 + y^2$,

$a\sqrt{(x - c)^2 + y^2} = a^2 - \text{cx}$,

$a^2[(x - c)^2 + y^2] = (a^2 - \text{cx})^2$,

$a^2x^2 - 2a^2\text{cx} + a^2c^2 + a^2y^2 = a^4 - 2a^2\text{cx} + c^2x^2$,

$a^2x^2 - c^2x^2 + a^2y^2 = a^4 - a^2c^2$.

Because $a^2 - c^2 > 0$, we can denote $a^2 - c^2 = b^2$.

We obtain:

$b^2x^2 + a^2y^2 = a^2b^2$

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1. \qquad (3.3)$$

It's important to verify that equation (3.3) is the equation of the ellipse. So far, we can only assert that each point $M(x, y)$ satisfying the ellipse equation (3.2) also satisfies equation (3.3). However, equation (3.3) was obtained after double squaring, and we know that when squaring both sides of the equation, an equation can be obtained that is not equivalent to the original. Make sure that this did not happen here. We must prove that every point $M(x, y)$ that satisfy the equation (8.3) is an ellipse point, i.e. that the condition $r_1 + r_2 = 2a$ is fulfilled for her.

So, let $M(x, y)$ be an arbitrary point whose coordinates satisfy equation (3.3). Let's find the distances r_1 and r_2 points M from the foci F_1 and F_2, respectively.

We obtain

$$r_1 = \sqrt{(x + c)^2 + y^2}. \qquad (*)$$

Express y^2 from (3.3):

$y^2 = b^2(1 - \frac{x^2}{a^2})$.

But $b^2 = a^2 - c^2$, therefore,

$y^2 = (a^2 - c^2)(1 - \frac{x^2}{a^2}) = a^2 - c^2 - x^2 + \frac{c^2}{a^2}x^2$.

Apply this value y^2 to (*):

$$r_1 = \sqrt{x^2 + 2cx + c^2 + a^2 - c^2 - x^2 + \frac{c^2}{a^2}x^2} =$$

$$= \sqrt{2cx + a^2 + \frac{c^2}{a^2}x^2} = \sqrt{(a + \frac{c}{a}x)^2}.$$

The value e that is determined by the ratio $e = \frac{c}{a}$ is called the **eccentricity** of the ellipse, in this case $0 < e < 1$. It is defined as the measure of elongation of an ellipse. The greater the eccentricity, the more elongated the ellipse is. The eccentricity is zero if and only if the focal points of the ellipse coincide: $F_1 = F_2$. In this case, the ellipse turns into a circle of radius a.

We obtain $r_1 = \pm\left(a + \frac{c}{a}x\right) = \pm(a + ex)$.

On the left is a positive number r_1. Therefore, on the right you need to choose a sign so that the right side is also positive. It follows from (3.3) that $|x| \leq a$. In addition. $0 < e < 1$, therefore $|ex| < a$. So, regardless of, $x > 0$ or $x < 0$, always $a + ex > 0$, therefore, you need to take the "plus sign "on the right:

$$r_1 = a + ex. \qquad (**)$$

Similarly, way, we obtain

$$r_2 = a - ex. \qquad (***)$$

From (**) and (***) we obtain

$r_1 + r_2 = 2a$,

therefore, the point $M(x, y)$ belongs to an ellipse.

We have proved that equation (3.3) is an ellipse equation. It is called the **canonical equation** of the ellipse.

Here a is the semimajor axis of the ellipse, b is the semimajor axis ($b = \sqrt{a^2 - c^2}$). It follows from equation (3.3) that the axes Ox and Oy are the axes of symmetry of the ellipse, and the point of their intersection, the point *O (0, 0),* is the center of symmetry.

In the particular case when $a = b$, the focal points of the ellipse merge, $c = 0$ and we have a circle of radius a centered at the origin.

It is known that planets move along elliptical trajectories, while the eccentricities of planetary orbits are small. In particular, for instance, the

eccentricity of the orbit of Venus is 0.007. Thus, the planets move almost in circles. Some comets also move in elliptical orbits, but their eccentricities are large, i.e. are close to one. For example, the eccentricity of Halley's comet is 0.9671429. Comets are either approaching the Sun, which is in one of the foci, then moving away from it for many years.

The vertical lines $x = -\frac{a}{e}$ and $x = \frac{a}{e}x$ are called the **directrices** of the ellipse defined by equation (3.3). It is easy to prove that if r is the distance from an arbitrary point $M$of an ellipse to some foci, and δ is the distance from the same point to the directrix corresponding to the same foci, then the ratio $\frac{r}{\delta}$ is a constant, equal to the eccentricity of the ellipse. Let's take the right focus and the right director.

We obtain:

$r = r\_2 = a - ex$ - see (***),

$\delta = \frac{a}{e} - x,$

therefore

$\frac{r}{\delta} = \frac{a-ex}{\frac{a}{e}-x} = \frac{(a-ex)e}{a-ex} = e.$

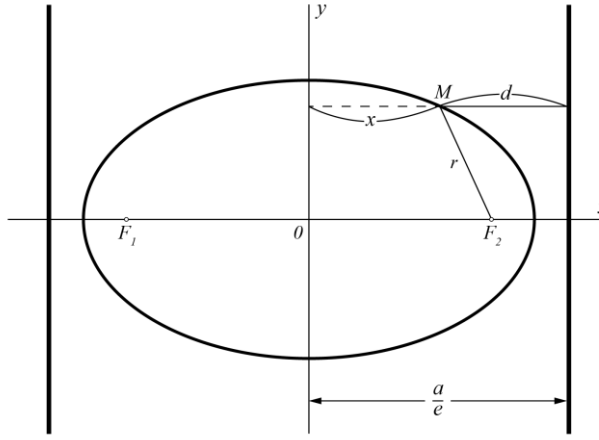


Fig.3.3

3.2. Hyperbola

Definition. **A hyperbola** is a line for all points of which the modulus of the difference in distances to two fixed points, which are called **foci**, is a constant and smaller than the distance between the foci.

Denote, as in the previous case, the distance between the foci F_1 and F_2 as $2c$, choose the coordinate system such that the axis Ox passes through the foci and the axis Oy in the middle between them (Fig. 3.3). We denote the distances from an arbitrary point $M(x, y)$ of the hyperbola to the foci F_1 and F_2, respectively, by r_1 and r_2, we obtain:

$$|r_1 - r_2| = 2a.$$

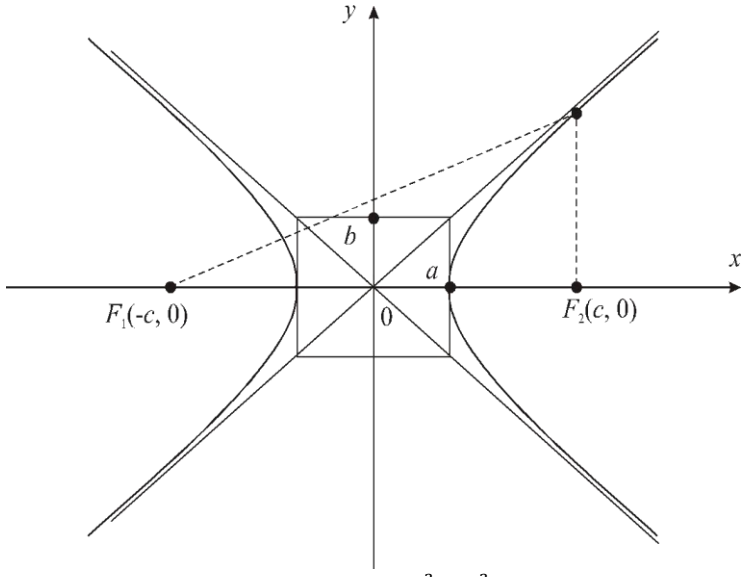


Fig. 3.4. *Hyperbola* $\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$

Assuming $r_1 = F_1M = \sqrt{(x+c)^2 + y^2}$, $r_2 = F_2M = \sqrt{(x-c)^2 + y^2}$, we obtain the hyperbola equation:

$$\sqrt{(x+c)^2 + y^2} - \sqrt{(x-c)^2 + y^2} = \pm 2a.$$

By performing calculations similar to those we conducted with the ellipse equation (see § 8.1), we obtain the **canonical hyperbola equation**:

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1, \qquad (3.4)$$

where $b^2 = c^2 - a^2$.

A hyperbola has two axes of symmetry, the intersection point of which is its center of symmetry.

Let's show that with an increase of x in the branch of the hyperbola, they come close to the lines $y = \pm\frac{b}{a}x$, are called the **asymptotes**[1] **hyperbola**.

For instance, let's take $y = \frac{b}{a}x$ and the hyperbole branch lying in the first quadrant. We obtain the equation of this branch by expressing *y* from equation (3.4):

$y = \frac{b}{a} \cdot \sqrt{x^2 - a^2}$. For each x consider the difference in the ordinates of the specified line and the branch of the hyperbola:

$\frac{b}{a}x - \frac{b}{a} \cdot \sqrt{x^2 - a^2} = \frac{b}{a}(x - \sqrt{x^2 - a^2}) = \frac{b}{a} \cdot \frac{[x^2 - (x^2 - a^2)]}{(x + \sqrt{x^2 - a^2})} = \frac{\text{ab}}{x + \sqrt{x^2 - a^2}}$.

We see that with increasing of *x* this difference becomes arbitrarily small.

The lines $y = \pm\frac{a}{e}x$ are called the **directrix** of the hyperbola given by the canonical equation (3.4). If *r* is the distance from an arbitrary point *M* of the hyperbola to some foci, and δ is the distance from the same point to the directrix corresponding to this focus, then the ratio $\frac{r}{\delta}$ is a constant value equal to the eccentricity of the hyperbola. The proof of this statement does not differ from the above proof of a similar statement for an ellipse.

[1]More on asymptotes in § 11.3.

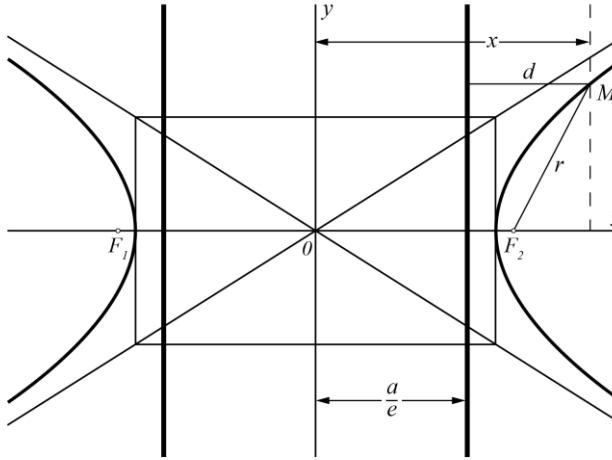


Fig.3.5

3.3. Parabola

***Definition*. A parabola** is a line for all points at which the distance to a fixed point, called the **focus**, is equal to the distance to a fixed line called the **directrix**, which is not passing through the focus.

Let a point F and a line d not passing through this point be given on the plane. We derive the parabola equation with focus F and directrix d. Let the distance from the point F to the line d be equal to p. We choose the coordinate system as follows. Draw the Ox axis through the point F perpendicular to the line d, and the Oy axis in the middle between the point F and the line d (Fig. 3.6).

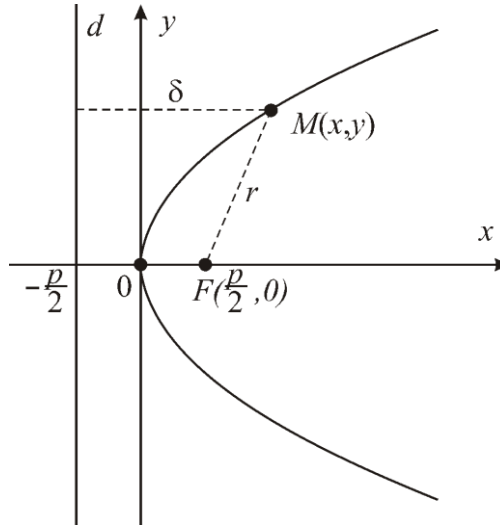


Fig. 3.6. *Parabola* $y^2 = 2px$

Let $M(x, y)$ be an arbitrary point on a parabola. Denote by δ the distance from this point to the directrix and by r the distance to the focus. According to the definition:

$r = \delta$.

Given that $r = \sqrt{(x - \frac{p}{2})^2 + y^2}$, $\delta = \frac{p}{2} + x$, we obtain:

$$\sqrt{(x - \frac{p}{2})^2 + y^2} = \frac{p}{2} + x$$

$$x^2 - px + \frac{p^2}{4} + y^2 = \frac{p^2}{4} + px + x^2$$

$$y^2 = 2px. \qquad (3.5)$$

This is a **canonical equation of a parabola**. A number *p* is called the **parameter** of a parabola.

Example 3.1. A space object is launched from the Earth's surface along a tangent to the Earth's surface and flies along a parabolic path. The

top of the parabola is on the surface of the Earth, the focus is in the center of the globe. Find the speed of flight.

Solution. We choose a coordinate system so that the abscissa axis passes through the focus and the ordinate axis is perpendicular to the abscissa axis, which is tangent to the Earth. The radius of the Earth is 6370000 $m \approx 6400000m$. The equation of the parabola is $y^2 = 2$px, where $\frac{p}{2} = 6400000$. The object is launched in the direction of the ordinate axis, but under the influence of gravity it shifts towards the center of the globe. It is known that in one second a freely falling object flies $4.9m$. We substitute in equation (3.5) $2p = 4 \cdot 640000$, $x = 4{,}9$. We obtain $y = \sqrt{4 \cdot 640000 \cdot 4{,}9} = 11200$. So, in the first second, the object flies 11,200 meters, that is, its speed is 11.2 kilometers per second. This is the second cosmic velocity.

The eccentricity of the parabola is equal to one: $\frac{r}{\delta} = 1$. So, all three considered second-order curves are characterized by the ratio $\frac{r}{\delta}$, where r is the distance from an arbitrary point of the curve to the focus and is the distance from the same point to the corresponding directrix. If this ratio is less than one, then the curve is an ellipse, if it is greater than one, then the hyperbola, if equal to one, then the parabola.

We note that the canonical parabola equation (3.5) differs from the equation familiar from the school course. This is due to the choice of the coordinate system. If we change the coordinate axes, then instead of equation (3.5) we get the usual equation of the form $y = \text{ax}^2$, where a is a constant number. A similar remark applies to the hyperbola equation (3.4). For example, if for the hyperbola $\frac{x^2}{2} - \frac{y^2}{2} = 1$ we take its asymptotes as the axes of the new coordinate system, then in this new coordinate system its equation will have a familiar form $y = \frac{1}{x}$.

3.4. General equation of a second order line

A general equation of a second-order line has the following form:

$a_{11}\,x^2 + 2a_{12}\,xy + a_{22}\,y^2 + 2a_{13}\,x + 2a_{23}\,y + a_{33} = 0;$ (3.6)

Where $a_{11}^2 + a_{12}^2 + a_{22}^2 \neq 0$.

Let us prove that there exist 9 different types of second-order lines. These are:

- **ellipses;**
- **hyperbole;**
- **parabolas;**
- curves degenerating into **a pair of straight lines**.

In this case, ellipses and pairs of lines can be both real and imaginary.

Thus, among the second-order lines, the curves in the usual sense of the word are only an ellipse, a hyperbola, and a parabola. Therefore, they are called the most important second-order curves.

The proof of the statement above is based on the transformation of the general equation (3.6). We give this proof here. But for this we must first study how the coordinates of the points and the equations of the lines change when the coordinate system is changed.

3.5. Coordinate transformation

Before we write a line equation on a plane, we need to select a specific coordinate system. Obviously, the same line will have different equations in different coordinate systems. We know, for example, that the equation of a circle of radius *R*, the center of which has coordinates x_0 and y_0 in the selected system Oxy , looks like this:

$(x - x_0)^2 + (y - y_0)^2 = R^2.$

In another coordinate system $O'x'y'$ where the center O' coincides with the center of this circle, the equation of this circle will have the form

$x'^2 + y'^2 = R^2$.

It will be so if the coordinate systems Oxy and $O'x'y'$ are connected by the relations

$x = x' + x_0$,

$y = y' + y_0$.

Let us consider one more example. Let a line be given in a coordinate system Oxy by the equation

$x - y + 2 = 0$,

and let us choose another coordinate system $O'x'y'$ which is connected to the previous system by the relations

$x = \frac{x'}{\sqrt{2}} - \frac{y'}{\sqrt{2}} - 1$,

$y = \frac{x'}{\sqrt{2}} + \frac{y'}{\sqrt{2}} + 1$.

Then the equation of a considered line in the system $O'x'y'$ will have a quite simple form:

$y' = 0$.

So, we see that a successful choice of a coordinate system allows to simplify the equation of a considered line.

In analytic geometry, the transition from one rectangular coordinate system to another is usually carried out with the rotation and parallel transfer.

The parallel transfer of the coordinate system Oxy **to a point** $M_0(x_0, y_0)$ is defined by the formulas

$$\begin{cases} x = x' + x_0, \\ y = y' + y_0, \end{cases} \tag{3.7}$$

which express old coordinates x, y through new x', y'.

Rotation of the coordinate system by the angle α (counter clockwise) is defined with the formulas

$$\begin{cases} x = x' \cos\alpha - y' \sin\alpha, \\ y = x' \sin\alpha + y' \cos\alpha. \end{cases} \tag{3.8}$$

Let us derive these formulas.

1. We start with the parallel transfer. Let two rectangular coordinate systems be defined on the plane: the "old" system Oxy and the "new"[1] system $O'x'y'$, with the axis $O'x'$ parallel to the axis Ox and the axis $O'y'$ parallel to the axis Oy, and in addition, the directions of the corresponding old and new axes coincide. In other words, the new system $O'x'y'$ is obtained from the old one by *parallel transfer*, or *shift*, at which the origin of coordinates O moves to a point O'.

Let the point O' have coordinates x_0 and $y_0$in the old system. Let us for definiteness consider$x_0 > 0$, $y_0 > 0$. We take an arbitrary point *M* on the plane (Fig. 3.5), and let its coordinates be (x, y) in the old system, and let it be (x', y') in the new system.

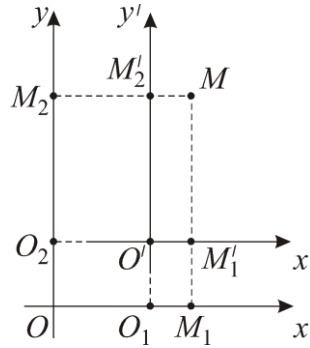


Fig. 3.5. *Parallel transfer of a coordinate system*

Let us consider fig. 3.5 (we took $M(x, y)$ with the coordinates $x > 0$, $y > 0$). So $OO_1 = x_0$, $OO_2 = y_0$, $OM_1 = x$, $OM_2 = y$, $O'M_1' = x'$,

[1] From here quotes in words "old" and "new" are omitted

$O'M_2' = y'$. Then we have $x = OM_1 = OO_1 + O_1M = x_0 + x'$, $y = OM_2 = OO_2 + O_2M = y_0 + y'$. Therefore we obtain (3.7).

The case when the points O' and M have negative coordinates is considered similarly.

Remark. One could reason differently: consider vectors$\overline{OM}$, $\overline{O'M}$ and $\overline{OO'}$. Obviously, $\overline{OM} = \overline{O'M} + \overline{OO'}$. Adding the vectors coordinate-wise, we obtain (3.7).

2. Now we move to a rotation of the coordinate system. Consider the case when the new system $Ox'y'$ is obtained from the old one Oxy by rotating by a certain angle α, counted counterclockwise. Moreover, both systems have a common origin *O*.

Suppose that, as in the previous case, the point M has coordinates (x, y) in the old system, and (x', y') in the new, respectively.

For definiteness, we consider the case when the angle $\alpha = \angle BOC$ is acute. Let M_1 be the projection of the point *M* onto Ox, *B* the projection of the same point *M* onto Ox' (Fig. 3.6). The sides of the angle formed by the straight lines MM_1 and MB are perpendicular to the sides of the angle formed by the axes Ox, Ox' and equal to α. Hence, $\angle AMB = \alpha$.

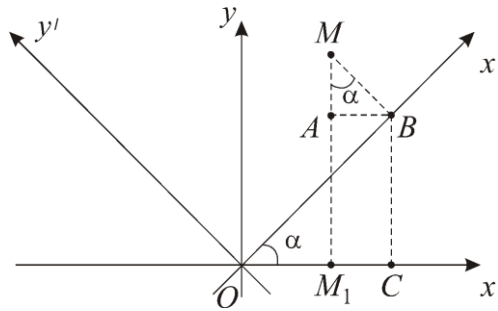


Fig. 3.6. *The rotation of a coordinate system*

According to the notation on fig. 3.6:

$x = OM_1$, $y = M_1M$; $\quad x' = OB$, $y' = BM$.

However

$OM_1 = OC - M_1C = OC - AB.$

From the triangle *OBC* (in which *OC* is a cathetus adjacent to the angle α) we find:

$OC = OB\cos\alpha = x'\cos\alpha.$

From the triangle *AMB* (in which *AB* is a cathetus opposite to the angle α) we find:

$AB = BM\sin\alpha = y'\sin\alpha.$

Thus,

$$x = OM_1 = x'\cos\alpha - y'\sin\alpha. \qquad (*)$$

Analogically we find *y:*

$y = M_1M = M_1A + AM = CB + AM,$

$CB = OB\sin\alpha = x'\sin\alpha,\ AM = BM\cos\alpha = y'\cos\alpha.$

Thus,

$$y = M_1M = x'\sin\alpha + y'\cos\alpha. \qquad (**)$$

From (*) and (**) we obtain (3.8).

Note that for the case when the angle α is not acute, the arguments are similar. Formulas (3.8) are valid for any angle α.

So, we derived formulas (3.7) and (3.8) expressing the old coordinates through the new ones. It might seem that formulas expressing new coordinates through old ones would be more useful. These formulas are easy to obtain. From (3.7) we immediately find:

$$\begin{cases} x' = x - x_0, \\ y' = y - y_0. \end{cases} \qquad (3.7')$$

Further, multiplying in (3.8) the first equality by $\cos\alpha$, and the second by $\sin\alpha$ and adding them, we obtain

$x' = x\cos\alpha + y\sin\alpha.$

Similarly, multiplying the second equality by $cos\,\alpha$ and subtracting from it the first multiplied by $sin\,\alpha$, we get

$y' = -x\,sin\,\alpha + y\,cos\,\alpha.$

Finally we have

$$\begin{cases} x' = x\,cos\,\alpha + y\,sin\,\alpha\,, \\ y' = -x\,sin\,\alpha + y\,cos\,\alpha\,. \end{cases} \qquad (3.8')$$

However, in practice, it is rarely necessary to find new coordinates of points by their old coordinates. Much more often, it is required to obtain the equation in the new system from the equation of the line in the old coordinate system. And for this it is necessary to replace the old coordinates with new ones, i.e. apply formulas (3.7) and (3.8), not (3.7') and (3.8').

Let us now consider the **general case of coordinate transformation**, when it is necessary to move from a rectangular system Oxy to a new rectangular system $O'x'y'$, at which the origin O' does not coincide with the point *O*, and the axes $O'x'$ and $O'y'$ are not parallel to the axes Ox and Oy, respectively. Let the new origin O' have coordinates (x_0, y_0) in the old Oxy system, and the axis $O'x'$ form an angle α with the axis Ox. Then the transition from the Oxy system to the system $O'x'y'$can be carried out in two stages:

1) make a parallel transfer of the Oxy system so that the origin is at a point $O'(x_0, y_0)$;

2) rotate around a point $O'$by the angle α.

It is easy to make sure that in this case the old coordinates will be expressed through the new ones using formulas

$$\begin{cases} x = x'\,cos\,\alpha - y'\,sin\,\alpha + x_0, \\ y = x'\,sin\,\alpha + y'\,cos\,\alpha + y_0. \end{cases}$$

3.6. Transformation of a general equation of a second-order line

Let us the general equation of a second- order line be given

$$a_{11}x^2 + 2a_{12}xy + a_{22}y^2 + 2a_{13}x + 2a_{23}y + a_{33} = 0, \quad (3.6)$$

where $a_{11}^2 + a_{12}^2 + a_{22}^2 \neq 0$.

We denote the left part of the equation (3.6) as $F(x, y)$:

$F(x, y) = a_{11}x^2 + 2a_{12}xy + a_{22}y^2 + 2a_{13}x + 2a_{23}y + a_{33}$.

In this polynomial, second-order terms form a quadratic form:

$$\phi(x, y) = a_{11}x^2 + 2a_{12}xy + a_{22}y^2. \quad (3.9)$$

The first step of the transformation is to transform the quadratic form (3.9) to the canonical form

$a_{11}^{'}x^{'2} + a_{22}^{'}y^{'2}$.

by turning the coordinate system by an angle α.

Note that the transformation (3.8) is a non-degenerate linear transformation.

So we make a transformation (3.8). We obtain

$F(x, y) = a_{11}\left(x^{'2}\cos^2\alpha\, 2x^{'}y^{'}\cos\alpha\sin\alpha + +y^{'2}\sin^2\alpha\right) +$

$+2a_{12}\left(x^{'2}\cos\alpha\sin\alpha - x^{'}y^{'}\sin^2\alpha - y^{'2}\sin\alpha\cos\alpha + x^{'}y^{'}\cos^2\alpha\right) +$

$+a_{22}\left(x^{'2}\,sin^2\,\alpha + 2x^{'}y^{'}cos\,\alpha\,sin\,\alpha + y^{'2}\,cos^2\,\alpha\right) +$

$+2a_1x^{'}cos\,\alpha - 2a_1y^{'}sin\,\alpha + 2a_2x^{'}sin\,\alpha + 2a_2y^{'}cos\,\alpha + a_{33} =$

$a_{11}^{'}x^{'2} + 2a_{12}^{'}x^{'}y^{'} + 2a_{23}^{'}y^{'} + a_{33} = F\left(x^{'}, y^{'}\right)$,

where the new coefficients are as follows:

$a_{11}' = a_{11}\,cos^2\,\alpha + 2a_{12}\,cos\,\alpha\,sin\,\alpha + a_{22}\,sin^2\,\alpha,$

$a_{12}' = -a_{11}\,cos\,\alpha\,sin\,\alpha + a_{12}(cos^2\,\alpha - sin^2\,\alpha) + a_{22}\,cos\,\alpha\,sin\,\alpha,$

$a_{22}' = a_{11}\,sin^2\,\alpha - 2a_{12}\,cos\,\alpha\,sin\,\alpha + a_{22}\,cos^2\,\alpha,$

$a_{13}' = a_{13}\,cos\,\alpha + a_{23}\,sin\,\alpha,$

$$a_{23}' = -a_{13}\,sin\,\alpha + a_{23}\,cos\,\alpha. \quad (3.10)$$

The angle α is determined by the requirement that $a'_{12} = 0$, i.e. so that in the transformed equation there is no term containing the product of unknowns. According to (8.10), the requirement $a'_{12} = 0$ means

$$a_{12}\cos^2\alpha + (a_{22} - a_{11})\cos\alpha\sin\alpha - a_{12}\sin^2\alpha = 0. \qquad (3.11)$$

So, when rotating by the angle α, which satisfies equality (3.11), the quadratic form (3.9) will have a canonical form.

In equation (3.11) it is natural to assume that $a_{12} \neq 0$ (if $a_{12} = 0$, then nothing would have to be transformed because the quadratic form $\phi(x, y)$ would already have the form $a_{11}x^2 + a_{22}y^2$.

Dividing (3.11) by $\cos^2\alpha$ we obtain

$a_{12} + (a_{22} - a_{11})\text{tg}\alpha - a_{12}\text{tg}^2\alpha = 0$

or

$a_{12}\text{tg}^2\alpha - (a_{22} - a_{11}) \cdot \text{tg}\alpha - a_{12} = 0.$

Solving this quadratic (with respect to $\text{tg}\alpha$) equation, we get:

$$\text{tg}\alpha = \frac{a_{22} - a_{11} \pm \sqrt{(a_{22} - a_{11})^{2} + + 4a_{12}^2}}{2a_{12}}. \qquad (3.12)$$

Assuming $(a_{22} - a_{11})^2 + 4a_{12}^2 > 0$, we can always find the necessary angle α from (3.12).

Denote for shorter notation $a'_{11} = \lambda_1$, $a'_{22} = \lambda_2$. Now we state the obtained result.

By rotating the coordinate system by the angle α, determined by formula (3.12), one can transform the quadratic form

$\phi(x, y) = a_{11}x^2 + 2a_{12}\text{xy} + a_{22}y^2$

to the canonical form

$\phi'(x', y') = \lambda_1 x'^2 + \lambda_2 y'^2.$

Meanwhile the polynomial $F(x, y)$ is transformed to the form

$$F'(x', y') = \lambda_1 x'^2 + \lambda_2 y'^2 + 2a'_{13}x' + 2a'_{23}y' + a_{33}. \qquad (3.13)$$

Note that both coefficients λ_1 and λ_2 cannot simultaneously vanish: if there were $\lambda_1 = \lambda_2 = 0$, then the quadratic form (3.9) as a result of a linear non-degenerate transformation would turn into an identical zero, which is impossible.

So **two basic cases** are possible:

I. $\lambda_1 \neq 0, \lambda_2 \neq 0$.

II. One of the coefficients λ_1, λ_2 is nonzero, the other one is equal to zero (*parabolic case*).

We consider case I, i.e. first basic case: $\lambda_1 \neq 0, \lambda_2 \neq 0$. When transferring the origin to some point $O'(x_0', y_0')$, i.e. when converting

$x' = x'' + x_0',$

$y' = y'' + y_0',$

the polynomial (3.13) takes the form

$$F''(x'', y'') = \lambda_1 x''^2 + \lambda_2 y''^2 + 2(\lambda_1 x_0' + a_{13}')x'' +$$

$$+2(\lambda_2 y_0' + a_{23}')y'' + a_{33}', \qquad (3.14)$$

where a free term

$a_{33}' = \lambda_1 x_0'^2 + \lambda_2 y_0'^2 + 2a_{13}' x_0' + 2a_{23}' y_0' + a_{33} = F'(x_0', y_0')$.

The second part of the transformation is as follows. We select such coordinates (x_0', y_0') of the new origin so that the coefficients by x'' and y'' in (3.14) turn to zero, i.e. to satisfy the inequalities:

$\lambda_1 x_0' + a_{13}' = 0, \quad \lambda_2 y_0' + a_{23}' = 0.$

Since$\lambda_1 \neq 0, \lambda_2 \neq 0$, then we can find x_0' and y_0':

$$x_0' = -\frac{a_{13}'}{\lambda_1}, \quad y_0' = -\frac{a_{23}'}{\lambda_2}. \qquad (3.15)$$

So, rotating the coordinate axes through an angle α, defined by formula (3.12), and moving the origin of the coordinate system to a point which coordinates x_0', y_0' are determined by equalities (3.15), we transform equation (3.6) to the form

$$\lambda_1 x''^2 + \lambda_2 y''^2 + a_{33}' = 0. \qquad (3.16)$$

Here **two cases** are possible:

λ_1 and $\lambda_2$are of different signs (so called *hyperbolic case*);

λ_1 and $\lambda_2$are of the same sign (so called *elliptic case*).

Consider the first case (*hyperbolic*).

Let $a'_{33} \neq 0$. Obviously, one of the coefficients has the same sign as a'_{33}; let, for example, λ_2 and a'_{33} be of the same sign; then λ_1 and a'_{33} are opposite in sign.

Rewrite the equation (3.16) as

$$\frac{x''^2}{-\frac{a'_{33}}{\lambda_1}-\frac{a'_{33}}{\lambda_2}} = 1.$$

The denominator $-\frac{a'_{33}}{\lambda_1}$ in the first term in a positive number which we denote as a^2; the denominator $-\frac{a'_{33}}{\lambda_2}$ in negative, we denote it as $-b^2$. Then we obtain an equation

$$\frac{x''^2}{a^2} - \frac{y''^2}{b^2} = 1.$$

This is the canonical expression of a hyperbola.

Let now $a'_{33} = 0$. Then (3.16) takes the form

$$\lambda_1 x''^2 + \lambda_2 y''^2 = 0. \qquad (3.17)$$

We assume that $\lambda_1 > 0, \lambda_2 < 0$ (if not, we multiply both parts of (8.17) by –1). Denote $\lambda_1 = a^2, \lambda_2 = -b^2$ and then obtain the equation

$$a^2 x''^2 - b^2 y''^2 = 0.$$

It can be rewritten in the form

$$(ax'' + by'')(ax'' - by'') = 0.$$

It is an equation of a pair of straight lines

$$ax'' + by'' = 0, ax'' - by'' = 0,$$

which intersect in the origin.

Similarly, we consider the elliptic case when both λ_1 and λ_2 are of the same sign and the parabolic case, when one of the coefficients, λ_1 or λ_2, is zero.

In the *elliptic* case we obtain either ellipse $\frac{x''^2}{a^2} + \frac{y''^2}{b^2} = 1$, or imaginary ellipse $\frac{x''^2}{a^2} + \frac{y''^2}{b^2} = -1$, or a pair of imaginary intersecting lines $\frac{x''^2}{a^2} + \frac{y''^2}{b^2} = 0$.

In the *parabolic* case we obtain either parabola $y''^2 = 2px''$, or a pair of parallel lines $y''^2 + a^2 = 0$, or a pair of coinciding straight lines $y''^2 = 0$.

Questions

1) What are the semi-axes of an ellipse?
2) What is the eccentricity of an ellipse? What characterizes the eccentricity of the ellipse and what is the range of its value?
3) How many axes of symmetry does an ellipse have?
4) What curve is called a hyperbola?
5) How many axes of symmetry does a hyperbola have?
6) What are the asymptotes of a hyperbola? How many asymptotes does a hyperbola have?
7) What are the main properties of a hyperbola?
8) What is the parameter of a hyperbola? Is it possible to find a parameter of a parabola knowing the distance from its focus to its vertex?
9) How many axes of symmetry does a parabola have?
10) How many different types of second-order curves exist there?

Chapter 4. Straight lines and planes in the space

4.1. Plane in the space

Let the coordinate system *Oxyz* be given in space and let the plane Π pass through the point $M_0(x_0,y_0,z_0)$ perpendicular to the vector $\overline{N} = (A, B, C)$. These two conditions determine *the only plane* in the space *Oxyz*. A vector $\overline{N}$ is called a **normal plane** vector. We derive the equation of this plane.

Let $M(x, y, z)$ be an arbitrary point on a plane Π. Then a vector $\overline{M_0M} = (x - x_0, y - y_0, z - z_0)$ and a vector $\overline{N} = (A, B, C)$ are mutually perpendicular. Hence, their scalar product is equal to zero: $(\overline{N}, \overline{M_0M}) = 0$. We write this last equation in a scalar form:

$$A \cdot (x - x_0) + B \cdot (y - y_0) + C \cdot (z - z_0) = 0. \qquad (4.1)$$

This is the equation of a plane passing through the point $M_0(x_0, y_0, z_0)$ perpendicular to the given vector $\overline{N} = (A, B, C)$. From (4.1) we obtain

$Ax + By + Cz - Ax_0 - By_0 - Cz_0 = 0.$

Denoting $-Ax_0 - By_0 - Cz_0 = D$ we obtain a **general equation of a plane:**

$$Ax + By + Cz + D = 0. \qquad (4.2)$$

So, the equation of a plane is a linear equation or a first-order equation with three variables.

It is easy to prove that *any first-order equation with three variables is an equation of a plane.*

It is known that a plane is uniquely defined by three points which are not on the same line. Let $M_0(x_0, y_0, z_0)$, $M_1(x_1, y_1, z_1)$ and $M_2(x_2, y_2, z_2)$ be three points not lying on the same line. Then vectors $\overline{M_0M_1} = (x_1 -$

$x_0, y_1 - y_0, z_1 - z_0)$ and $\overline{M_0M_2} = (x_2 - x_0, y_2 - y_0, z_2 - z_0)$ are not parallel to the same line (not collinear). Let $M(x, y, z)$ be an arbitrary point on the plane Π. Then the vector $\overline{M_0M}$ can be decomposed into vectors $\overline{M_0M_1}$ and $\overline{M_0M_2}$. Therefore, these three vectors $\overline{M_0M_1}$, $\overline{M_0M_2}$ and $\overline{M_0M}$are linearly dependent and that why

$$\begin{vmatrix} x - x_0 & y - y_0 & z - z_0 \\ x_1 - x_0 & y_1 - y_0 & z_1 - z_0 \\ x_2 - x_0 & y_2 - y_0 & z_2 - z_0 \end{vmatrix} = 0. \qquad (4.3)$$

This is an equation of a plane passing through three points $M_0(x_0, y_0, z_0)$, $M_1(x_1, y_1, z_1)$ and $M_2(x_2, y_2, z_2)$ which are not on the same straight line.

***Example* 4.1.** Write an equation of a plane which passes through points

$M_0(1, 2, 1)$, $M_1(3, 3, 1)$, $M_2(2, 3, 2)$.

Solution. We substitute the coordinates on these points into the equation (4.3):

$$\begin{vmatrix} x - 1 & y - 2 & z - 1 \\ 3 - 1 & 3 - 2 & 1 - 1 \\ 2 - 1 & 3 - 2 & 2 - 1 \end{vmatrix} = 0, \begin{vmatrix} x - 1 & y - 2 & z - 1 \\ 2 & 1 & 0 \\ 1 & 1 & 1 \end{vmatrix} = 0$$

$$x - 1 - 2 \cdot (y - 2) + z - 1 = 0$$

$$x - 2y + z + 2 = 0.$$

We consider the relative position of two planes. Given two planes:

$A_1x + B_1y + C_1z + D_1 = 0,$

$A_2x + B_2y + C_2z + D_2 = 0.$

Their normal vectors are $\bar{N}_1 = (A_1, B_1, C_1)$ и $\bar{N}_2 = (A_2, B_2, C_2)$.

The angle between these two planes is the angle between $\bar{N}_1$ and $\bar{N}_2$ which is defined by the formula

$$\cos\phi = \frac{A_1A_2 + B_1B_2 + C_1C_2}{\sqrt{A_1^2 + B_1^2 + C_1^2}\sqrt{A_2^2 + B_2^2 + C_2^2}}. \qquad (4.4)$$

The **parallelism** condition for two planes is the condition of proportionality of their normal vectors:

$$\frac{A_1}{A_2} = \frac{B_1}{B_2} = \frac{C_1}{C_2}. \quad (4.5)$$

The **coinciding** planes condition looks as follows:

$$\frac{A_1}{A_2} = \frac{B_1}{B_2} = \frac{C_1}{C_2} = \frac{D_1}{D_2}. \quad (4.6)$$

The **perpendicular** planes condition is the conditioncos$\phi = 0$, i.e.

$$A_1A_2 + B_1B_2 + C_1C_2 = 0. \quad (4.7)$$

4.2. Line in space. Line and a plane in the space

Let the straight line L pass through the point $M_0(x_0, y_0, z_0)$ parallel to the vector $\bar{s} = (l, m, n)$. In this case, the vector $\bar{s}$is called the directing vector of the straight line. Let $M(x, y, z)$ be an arbitrary point of the line L. Obviously, the vectors $\overline{M_0M} = (x - x_0, y - y_0, z - z_0)$ and $\bar{s}$ are proportional. Having written down the condition of their proportionality in coordinate form, we obtain the canonical equation of the straight line L:

$$\frac{x-x_0}{l} = \frac{y-y_0}{m} = \frac{z-z_0}{n}. \quad (4.8)$$

From (4.8) we obtain:

$x - x_0 = lt, \quad y - y_0 = mt, \quad z - z_0 = nt,$

where *t* is a proportionality coefficient. These equations give:

$$x = x_0 + lt, \quad y = y_0 + mt, \quad z = z_0 + nt. \quad (4.9)$$

These are the **parametrical equations of a line** *L*. (Sometimes these are called in a singular form as a *parametric equation of the line.*)

A line in the space also can be defined as an intersection of two planes, i.e. as a set of points, the coordinates of which satisfy the conditions:

$$\begin{cases} A_1x + B_1y + C_1z + D_1 = 0 \\ A_2x + B_2y + C_2z + D_2 = 0. \end{cases} \quad (4.10)$$

The canonical equation (4.8), however, can also be considered as a pair of plane equations, considered together. It is easy to derive the canonical or parametric equation of a line defined in the form (4.10). To do this, it is enough to find some point $M_0(x_0, y_0, z_0)$ that belongs to the line and the direction vector. The coordinates of $M_0$are easy to find since this is any solution to the system (9.10). For example, putting, $z_0 = 0$ from the system (9.10) we find x_0,$y_0$and obtain $M_0(x_0, y_0, 0)$. The coordinates of the direction vector $\bar{s}$ may be numbers:

$$l = \begin{vmatrix} B_1 & C_1 \\ B_2 & C_2 \end{vmatrix},\ m = \begin{vmatrix} C_1 & A_1 \\ C_2 & A_2 \end{vmatrix},\ n = \begin{vmatrix} A_1 & B_1 \\ A_2 & B_2 \end{vmatrix}.$$

Let us consider now the relative position of the line and the plane in space. Let the line L be given:

$\frac{x-x_0}{l} = \frac{y-y_0}{m} = \frac{z-z_0}{n}$

and a plane Π: $Ax + By + Cz + D = 0$.

Obviously, the line L is parallel to the plane Π when the direction vector of the line $\bar{s} = (l, m, n)$ is perpendicular to the normal vector of the plane $\overline{N} = (A, B, C)$, i.e. the **parallel condition** *of the line L and the plane* $\Pi$is the condition:

$$Al + Bm + Cn = 0. \quad (4.11)$$

The condition for the proportionality of these vectors is the ***perpendicularity*** *condition of the line L* and the plane Π:

$$\frac{A}{l} = \frac{B}{m} = \frac{C}{n}. \quad (4.12)$$

The angle between the line and the plane is the angle between the line and its projection onto the plane, and this is the angle additional to the angle between the direction vector $\bar{s}$ of the line L and the normal vector $\overline{N}$ of the plane Π:

$$\sin\phi = \left|\cos\ \widehat{\bar{N},\bar{s}}\right| = \frac{Al+Bm+Cn}{\sqrt{A^2+B^2+C^2}\cdot\sqrt{l^2+m^2+n^2}}. \quad (4.13)$$

The distance from a point to a plane is calculated using a formula similar to the formula for the distance from a point to a line on a plane [see (7.15)]. We show that the distance d from the point $M_0(x_0, y_0, z_0)$ to the plane

$$Ax + By + Cz + D = 0. \quad (4.2)$$

is calculated using the formula

$$d = \frac{|Ax_0+By_0+Cz_0+D|}{\sqrt{A^2+B^2+C^2}}. \quad (4.14)$$

We write the equation of a line passing through a point $M_0(x_0, y_0, z_0)$ perpendicular to the plane (4.2). To do this, we use the parametric equations (4.9):

$x = x_0 + lt, y = y_0 + mt, z = z_0 + nt.$ (4.9)

In order for the line (4.9) to be perpendicular to the plane (4.2), it is necessary that its direction vector$\bar{s} = (l, m, n)$ be parallel to the vector $\bar{N} = (A, B, C)$, i.e. so that the coordinates of the vectors $\bar{s}$ and $\bar{N}$ are proportional. The easiest way, of course, is to take vector $\bar{N}$as the vector $\bar{s}$, i.e. take $l = A$, $m = B$, $n = C$. Then the parametric equations (4.9) will look like this:

$$x = x_0 + At, y = y_0 + Bt, z = z_0 + Ct. \quad (4.9')$$

The straight line (4.9 ') is perpendicular to the plane (4.2) and passes through a point M_0. Consequently, the distance from the point M_0 to the plane (4.2) is the distance between the point M_0 and the point M of the intersection of the line (4.9 ') with the plane (4.2). Let us find the coordinates of this point M. For this, it is necessary to solve equations (4.2) and (4.9 ') together. The easiest way to do this is substitute the expressions for x, y, and z from (4.9 ') into (4.2). We obtain:

$A(x_0 + At) + B(y_0 + Bt) + C(z_0 + Ct) + D = 0,$

$(A^2 + B^2 + C^2)t + (Ax_0 + By_0 + Cz_0 + D) = 0.$

From these we find *t*:

$t = -\frac{Ax_0+By_0+Cz_0+D}{A^2+B^2+C^2}$.

This value of *t* determines the coordinates of the point *M* which is the base of the perpendicular dropped from the point M_0 onto the plane (4.2). Substitute the found *t* in (4.9'):

$$x = x_0 + A(-\frac{\mathrm{Ax_0+By_0+Cz_0+D}}{A^2+B^2+C^2}),$$

$$y = y_0 + B(-\frac{\mathrm{Ax_0+By_0+Cz_0+D}}{A^2+B^2+C^2}), \qquad (4.9'')$$

$$z = z_0 + C(-\frac{\mathrm{Ax_0+By_0+Cz_0+D}}{A^2+B^2+C^2}).$$

The distance *d* from the point M_0 to the plane (4.2) is the length of the perpendicular M_0M or, which is the same, the distance between the points $M_0(x_0, y_0, z_0)$ and $M(x, y, z)$, i.e.

$d = \sqrt{(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2}$.

Since *x*, *y* and *z* are defined in (4.9''), we obtain

$$d = \sqrt{(A^2 + B^2 + C^2)\left(-\frac{Ax_0+By_0+Cz_0+D}{A^2+B^2+C^2}\right)^2} = \text{or}$$

$$= \sqrt{A^2 + B^2 + C^2}\,\frac{|Ax_0+By_0+Cz_0+D|}{A^2+B^2+C^2},$$

$d = \frac{|\mathrm{Ax_0+By_0+Cz_0+D}|}{\sqrt{A^2+B^2+C^2}}$.

That completes the proof.

Example 4.2. *Find the distance between a point* $M_0(1,0,2)$ and a plane $x + 2y - 2z + 9 = 0$.

Solution.

$d = \frac{|1+2\cdot0-2\cdot2+9|}{\sqrt{1^2+2^2+(-2)^2}} = \frac{6}{3} = 2.$

Example 4.3. *Find the distance between a line*

$\frac{x+1}{2} = \frac{y-2}{2} = \frac{z}{1}$

and a plane $4x - 2y - 4z + 9 = 0$.

Solution. The line is parallel to the plane. Indeed, the scalar product of its direction vector and the normal plane vector is zero: $2 \cdot 4 + 2 \cdot (-2) + 1 \cdot (-4) = 0$. Therefore, the distance from a straight line to a plane is equal to the distance from any point M_0 of this straight line to a plane. The most convenient way is to take a point $M_0 = (-1,2,0)$ whose coordinates appear in the equation of a line. We obtain

$$d = \frac{|4 \cdot (-1) - 2 \cdot 2 - 4 \cdot 0 + 9|}{\sqrt{4^2 + (-2)^2 + (-4)^2}} = \frac{1}{6}.$$

***Example* 4.4.** Find the distance between a point $M_0(1,2,3)$ and a line $\frac{x-6}{2} = \frac{y}{-2} = \frac{z-7}{1}$.

Solution. We write the equation of the plane that passes through the given point M_0 and is perpendicular to the given line, and find the coordinates of the point *M* of the intersection of the line and the plane. Obviously, M_0M is a perpendicular dropped from a point M_0 to a given line. Its length is the distance we want to find.

The equation of a plane passing through M_0 perpendicular to a given line is

$$2 \cdot (x - 1) - 2 \cdot (y - 2) + 1 \cdot (z - 3) = 0,$$

or

$$2x - 2y + z - 1 = 0. \qquad (*)$$

Write the equation of this straight line in a parametrical form:

$$x = 6 + 2t,\ y = -2t,\ z = 7 + t. \qquad (**)$$

Find the intersection point of the line (**) and the plane (*). To do this, first we substitute *x*, *y* and *z* from (**) into (*) and find *t*:

$$2 \cdot (6 + 2t) - 2 \cdot (-2t) + 7 + t - 1 = 0,$$

$$9t + 18 = 0,\ t = -2.$$

Now, substituting the value $t = -2$ in (**), we obtain $x = 2$, $y = 4$, $z = 5$. So, the point $M(2,4,5)$ is the base of the perpendicular M_0M. Therefore

$d = M_0M = \sqrt{(2-1)^2 + (4-2)^2 + (5-3)^2} = 3.$

Note that there is another way to solve this and similar problems, based on the concept of a vector product of vectors, which is not considered here.

Questions

1) What is a normal vector to a plane in space?
2) Will the angle between the planes $3x + y - z = 0$ and $x - y + 2z + 5 = 0$be the right angle?
3) Will the planes $3x - 2y + z = 0$ and $6x - 3y + 2z + 12 = 0$ be parallel?
4) Does a point $M_0(1,2,3)$ belong to a plane $2x - 3y + z + 1 = 0$?
5) What is the distance between the origin and a plane $2x - y + 2z + 9 = 0$?
6) What are the coordinates of a point on a line $\begin{cases} x = 2 + t, \\ y = 1 - 2t, \\ z = 3 + 2t, \end{cases}$corresponding to the value $t = -1$?
7) Does the point $M_0(1,\ 3,\ 2)$ belong to a line $\begin{cases} x = 3 - t, \\ y = 1 + t, \\ z = -4 + 3t? \end{cases}$ If it belongs, what is the corresponding value of a parameter *t*?
8) Will the vector $\bar{a} = (2,\ -1,\ 3)$ be parallel to a line $\begin{cases} x = 3 - 4t, \\ y = 1 + 2t, \\ z = 5 - 6t? \end{cases}$
9) What point of a line $\begin{cases} x = 1 + t, \\ y = 2 - 3t, \\ z = -3 + 2t \end{cases}$ corresponds to the parameter value $t = 2$?

10) Do a line $\begin{cases} x = 2 + t, \\ y = 3 + 2t, \\ z = -1 - 2t \end{cases}$ and a plane $x + 2y - 5z + 2 = 0$ intersect? If they intersect, what are the coordinates of the intersection point?

11) Is the line $\frac{x-1}{2} = \frac{y+2}{-1} = \frac{z-3}{-2}$ parallel to a plane $x + 2y + 2z - 7 = 0$?

12) How to determine the coordinates of a direction vector of a line given by a pair of planes $\begin{cases} A_1x + B_1y + C_1x + D_1 = 0, \\ A_2x + B_2y + C_2z + D_2 = 0 \end{cases}$?

13) How to find the distance between the parallel planes?

14) Does a plane $\frac{x-1}{2} = \frac{y+1}{-1} = \frac{z-3}{3}$ go through points $M_1(-1,\ 0,\ 0)$ and $M_2(5,\ -3,\ 9)$?

Chapter 5. Function

5.1. Definition of function

One of the most important definitions in mathematics and its applications is the definition of a function.

***Definition*.** Let us be given two numerical sets *X* and *Y*. Suppose that each element $x \in X$ according to some law *f* is associated with some (unique) element $y \in Y$. Then we say that a **function**[1] $y = f(x)$ is given on the set *X*.

Moreover, *x* is called an **independent variable** (or **argument**), *y* is a **dependent variable**, and the set $D(f) = X$ is called the **domain** of the function. The set $R(f)$ of all values of the function is called the **scope** of the function. Obviously $R(f) \subseteq Y$.

So, the definition of a function consists of three parts:

a domain $D(f) = X$;

a scope $R(f) = f(x)$;

a rule *f* which associates each point $x \in X$ with a unique point $y = f(x) \in R(f)$.

[1] More precisely, **a numerical function**.

A **graph of a function** $y = f(x)$ is a set of points with coordinates $(x, f(x)),\ x \in X$.

If the set *X* is not specifically stated, then the domain of the function is the set of all such values of *x* for which the function $y = f(x)$ makes sense at all (this is the so-called natural domain of definition).

Note that we use different letters to denote a function and its argument, for example:

$$y = y(x),\ y = F(x),\ s = s(t),\ y = \varphi(x).$$

The most common are the following methods for **setting the function**:

1) **analytical –** the relationship between the argument and the function is given in the form of a formula (or formulas). So, the functions $y = 2x + 3$, $y = \frac{1}{x^2}$, $y = x^3 + \frac{2x}{\sqrt{x^2 + 1}}$ are given analytically.

Note that one function can be defined with a set of formulas: different functions (different analytical expressions) describe different parts of the domain. For example:

$$y = \begin{cases} x^2 - 1, & \text{if } x \le 1, \\ 1 - x, & \text{if } x > 1; \end{cases}$$

2) **tabular** – a function is given with a table containing the values of the argument *x* and corresponding values of a function *f(x)*. Examples of such functions are tables of financial statements, a table of logarithms. Databases are also essentially based on the tabular method of specifying, storing and processing information, therefore, they are also based on the tabular form of functional dependence;

3) **graphical** – the function is given graphically if its graph is drawn, i.e. the correspondence between the argument and the function is set by

means of a graph. The advantages of this method include its visibility, the disadvantages include its low accuracy.

There are other less common ways of defining functions, for example, **verbal**, which consists in the fact that the function is described by the rule of its compilation.

Consider an example of a function defined verbally or descriptively. This is the Dirichlet function, usually denoted as $\chi(x)$. It is equal to one of its argument *x* is a rational number and to zero if *x* is an irrational number. The Dirichlet function is defined on the whole number line, and the set of its values consists of two points: 0 and 1. It is impossible to graphically depict it:

$$y = \begin{cases} 1, & \textit{if } x \textit{ is rational}, \\ 0, & \text{if } x \textit{ is irrational}. \end{cases}$$

We move on to consider the **basic properties of functions**.

1) **parity and oddness**. A function $y = f(x)$ defined on an interval symmetric with respect to the coordinate origin is called an **even function** if, for any values *x* from its domain, equality $f(-x) = f(x)$ holds. If $f(-x) = -f(x)$, then the function is called an **odd function**. A function that is not even or odd is called a **general function**.

For example: 1) $y = x^4$ is an even function since $f(-x) = (-x)^4 = x^4 = f(x)$; 2) $y = \sin x$ is an odd function since $f(-x) = \sin(-x) = -\sin x = -f(x)$; 3) $y = x^2 + \sin x$ is a general

function since $f(-x)=(-x)^2+\sin(-x)=x^2-$ $-\sin x$, $f(-x)\neq f(x)$, $f(-x)\neq -f(x)$.

The graph of an even function is symmetric relative to the axis Ox, and the graph of an odd function is relative to the origin.

2) **monotony**. The function $y=f(x)$ is called **increasing** on the interval *X* if for any x_1, $x_2 \in X$ from the inequality $x_2 > x_1$ follows that $f(x_2) > f(x_1)$; a function is called **decreasing** if from $x_2 > x_1$ it follows that $f(x_2) < f(x_1)$.

A function is called **monotonic** on the interval *X* if it either grows on the entire interval or decreases on it.

Note that we gave a definition of a monotonic function in the strict sense. In general, monotonic functions include **non-decreasing** functions, i.e. those for which from $x_2 > x_1$ it follows that $f(x_2) \geq f(x_1)$ and **non-increasing** functions, i.e. those for which from $x_2 > x_1$ it follows that $f(x_2) \leq f(x_1)$.

3) **boundedness**. A function $y = f(x)$ is called **bounded** in a given domain if there exists a number $M > 0$ such that $|f(x)| \leq M$ for all *x* from this domain.

For example, the function $\frac{1}{x^2+1}$ is bounded on the whole number line since $\left|\frac{1}{x^2+1}\right| \leq 1$ for any $x \in \mathbf{R}$.

4) **periodicity**. A function $y = f(x)$ is called **periodic** if there exists a number $T \neq 0$ such that $f(x+T) = f(x)$ for all x from the domain of the function.

In this case, T is called the **period** of the function. Obviously, if T is the period of a function $y = f(x)$, then the periods of this function are also $2T$, $3T$ etc. Therefore, usually, the period of the function is called the smallest positive period (if it exists). For example, function $y = sin\, x$ has a period $T = 2\pi$, and function $y = \text{tg}\ 3x$ has a period $T = \frac{\pi}{3}$. It should be noted that not every periodic function has the shortest period. In particular, the Dirichlet function is periodic and any real number is its period, but it does not have the smallest period.

5.2. Basic elementary functions

We list the basic elementary functions and briefly recall their basic properties known from the school course in mathematics.

1. **A power function**, $y = x^a$, here a is any real number.

Consider this function for different values of a:

1) a is a natural number. The domain of the function is the entire number line. The function is *odd* if a is odd and even if a is even.

If a is odd, then the function increases on $(-\infty; +\infty)$; if a is even, then it decreases on $(-\infty, 0)$ and increases on $(0. +\infty)$.

2) a is a negative integer. In this case, the function is defined for all values of x except $x=0$.

A function is odd if a is an odd number and even if a is an even number. If a is odd, then the function decreases on $(-\infty, 0)$ and $(0, +\infty)$.

3) $a = \frac{1}{n}$, $n \neq 0$. If n is an even number, then the function is defined on $[0, +\infty)$; if n is an odd number, then on $(-\infty, +\infty)$. The function increases throughout the definition area.

Fig. 5.1–5.4 show graphs of the power function for various values of a.

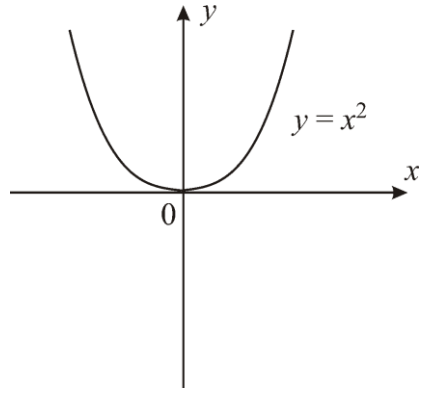
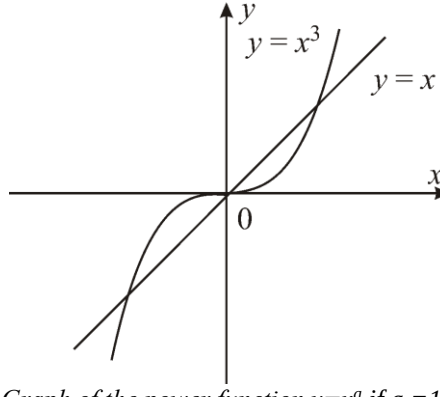
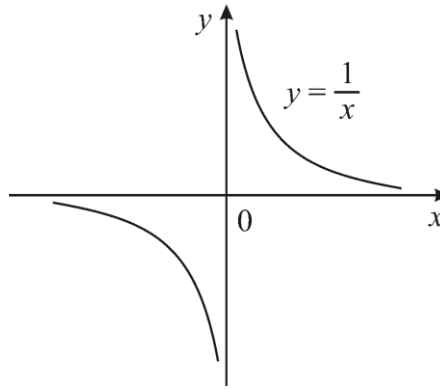


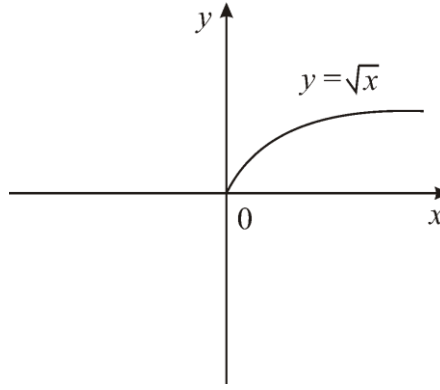
Fig. 5.1. *Graph of the power function $y=x^a$ if a = 2*



Puc. 5.2 Graph of the power function $y=x^a$ if a =1 and a = 3



Puc. 5.3. *Graph of the power function $y=x^a$ if $a = –1$*



Puc. 5.4. *Graph of the power function $y=x^a$ if $a=½$*

The power function is non-periodic for any *a.*

2. **Exponential function** $y = a^x$, $a > 0$, $a \neq 1$. This function is defined on the whole number line. It is a general function; increases on

$(-\infty, +\infty)$ at $a > 1$ (Fig. 5.5, *a*), decreases on $(-\infty, +\infty)$ at $0 < a < 1$ (Fig. 5.5, *b*). Non-periodic.

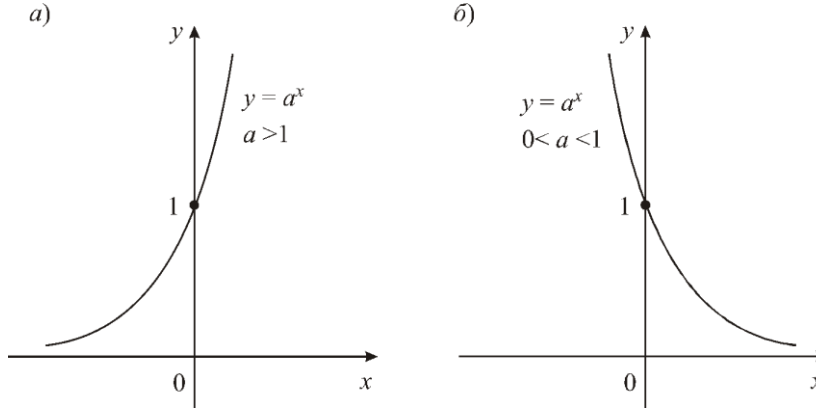


Fig. 5.5. *Graph of the exponential function $y=a^x$ at $a>1$ (a) and $0<a<1$ (b)*

3. **Logarithmic function** $y = \log_a x$, $a > 0$, $a \neq 1$. Logarithmic function defined on $(0, +\infty)$, it is a function of a general form; increases on $(0, +\infty)$ at $a > 1$; decreases on $(0, +\infty)$ at $0 < a < 1$. Non-periodic.

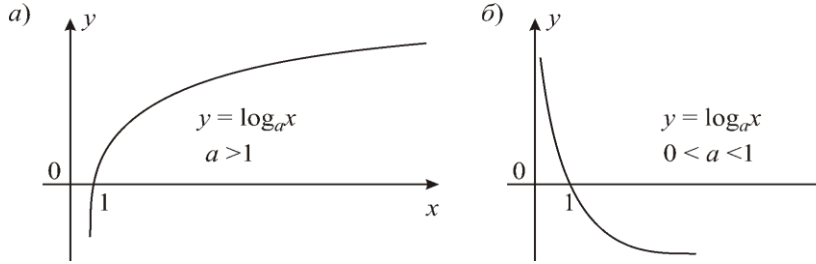


Fig. 5.6. *Graph of the logarithmic function $y=\log_a x$ at $a>1$ (a) and $0<a<1$ (b)*

Recall that the exponential function $y = a^x$ and the logarithmic function $y = \log_a x$ are mutually inverse functions.

4. **Trigonometric functions:**

1) $y = \sin x$ (Fig. 5.7). Odd periodic function with the period $T = 2\pi$, defined on $(-\infty, +\infty)$.

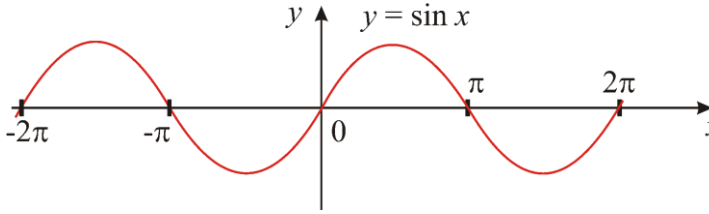


Fig. 5.7. *Graph of the function y=sin x*

2) $y = \cos x$ (Fig. 5.8). Even periodic function with the period $T = 2\pi$, defined on $(-\infty, +\infty)$.

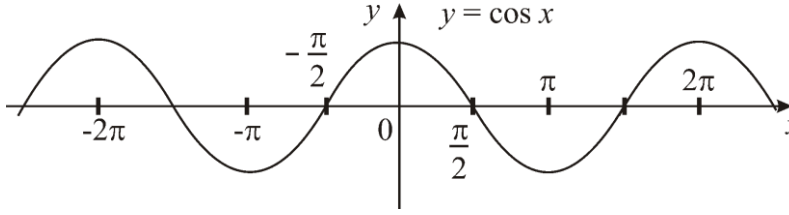


Fig. 5.8. *Graph of the function y=cos x*

3) $y = \operatorname{tg} x$ (Fig. 5.9). The domain is $\left(-\frac{\pi}{2} + \pi n, \quad \frac{\pi}{2} + \pi n\right)$, $n \in \mathbf{Z}$. The function is odd, increases on $\left(-\frac{\pi}{2} + \pi n, \quad \frac{\pi}{2} + \pi n\right)$, $n \in \mathbf{Z}$. It is periodic with the period $T = \pi$.

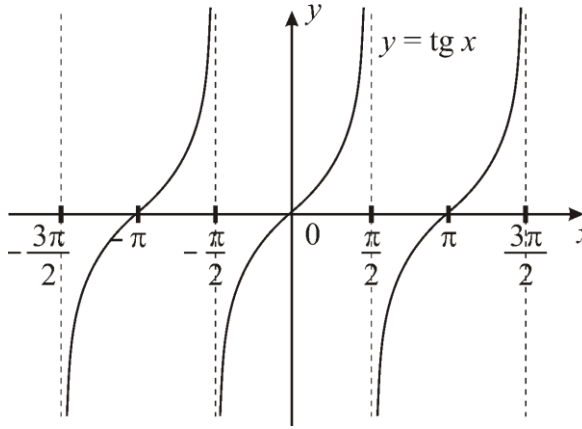


Fig. 5.9. *The graph of the function y=tg x*

4) $y = \operatorname{ctg} x$ (Fig. 5.10). The domain: $(\pi n, \pi + \pi n)$, $n \in \mathbf{Z}$. An odd periodic function, decreases on $(\pi n, \pi + \pi n)$, $n \in \mathbf{Z}$; the period is $T = \pi$.

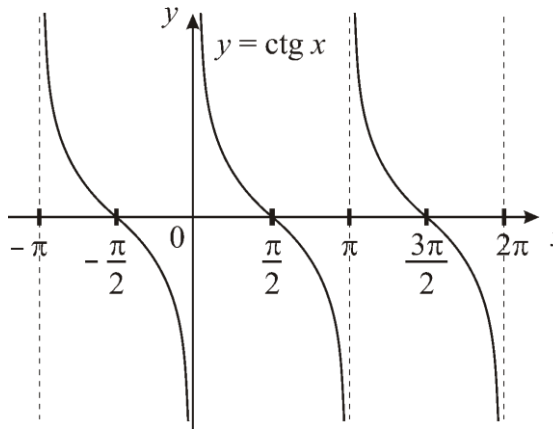


Fig. 5.10. *The graph of the function y=ctg x*

5. **Inverse trigonometric function**:

1) $y = \arcsin x$; 2) $y = \arccos x$;

3) $y = \operatorname{arctg} x$; 4) $y = \operatorname{arcctg} x$.

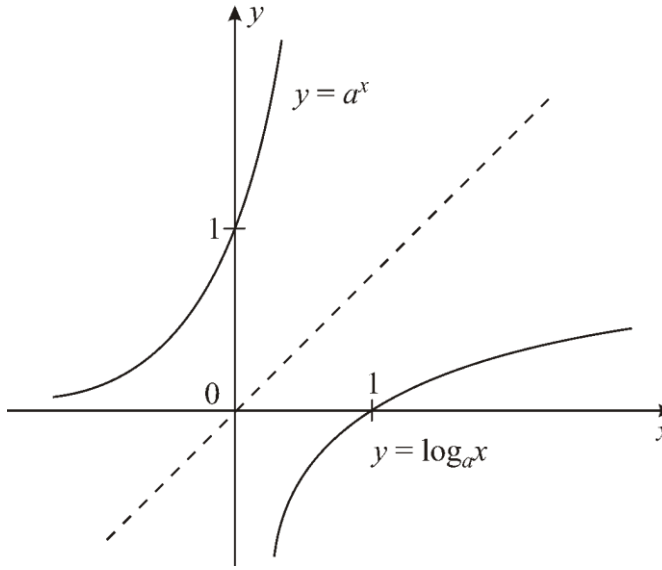
The functions $\arcsin x$ and $\arccos x$ are defined on $[-1, 1]$, functions $\operatorname{arctg} x$ and $\operatorname{arcctg} x$ are defined on the whole number line.

5.3. Elementary functions

Let function $y = f(x)$ be defined on interval *X*, its range of variation is *Y*, and let different values of *x* correspond to different values of *y*. Then for each value $y \in Y$ there is a single number $x \in X$ at which *f(x)=y*. Then the resulting function $x = \varphi(y)$ defined on *X* with the range of variation of *Y* is called the **inverse function**.

Since an independent variable is usually denoted by x and a function by y, the function inverse to the function $y = f(x)$ is also denoted as $y = f^{-1}(x)$.

Mutually inverse functions are, in particular, $y = a^x$ and $y = \log_a x$ (Fig. 10.11), $y = \sin x$ and $y = \arcsin x$, etc. The graphs of mutually inverse functions are symmetric with respect to the line $y = x$.



Puc. 5.11. *Graphs of mutually inverse functions* $y = a^x$ *and* $y = \log_a x$

It is well known that arithmetic operations can be performed on functions: addition, subtraction, multiplication, division.

Consider another action on functions called *taking a function from a function* or *constricting a complex function*. Let a function $y = f(u)$ be defined on a set *U* and its range of variation is *Y*, its argument *u* be a function of *x*: $u = \varphi(x)$, defined on a set *X* with a range of *U*. Then function $y = f(\varphi(x))$ defined on *X* is called a **composite function** or a **function of a function (a superposition of functions)**

For example, two functions $y = \lg u$ and $u = 1 - x^2$ define a composite function $y = \lg(1 - x^2)$ with domain $(-1, 1)$.

Note that the operation of taking a function from a function can be performed any number of times. For example, a function $y = \sqrt{\lg \sin x^2}$ is obtained as a result of the following operations:

$y = \sqrt{u}$, $u = \lg v$, $v = \sin w$, $w = x^2$.

Definition. An **elementary function** is called a function that is obtained from the basic elementary functions and constants using a finite number of operations of addition, subtraction, multiplication, division, and taking a function of a function.

Elementary functions are divided into *algebraic* and *transcendental*.

Algebraic functions include:

a) polynomials:

$$y = a_0 x^n + a_1 x^{n-1} + \ldots + a_n;$$

b) fractional rational functions:

$$y = \frac{a_0 x^n + a_1 x^{n-1} + \ldots + a_n}{b_0 x^m + b_1 x^{m-1} + \ldots + b_m},$$

i.e. functions defined as the ratio of two polynomials;

c) irrational functions, i.e. functions obtained by a finite number of superpositions and arithmetic operations on power functions with integer and fractional exponents and which are not rational.

The examples of the irrational functions are $y = x^2 + \sqrt[3]{x}$, $y = \frac{\sqrt{x^2 + 1}}{x^3 + \sqrt{x}}$, etc.

Generally, function $y = f(x)$ is called algebraic if it satisfies the equations of form $P_0 y^n + P_1 y^{n-1} + ... + P_n = 0$,

where P_0, P_1, …, P_n are polynomials depending on *x*.

A function that is not algebraic is called **transcendental**.

Transcendental functions include exponential, logarithmic and trigonometric functions.

5.4. Application of functions in the economics

Functions are widely used in economic theory. Here we give the most commonly used functions of a single argument.

The **utility function** (see Fig. 5.12) is a subjective numerical assessment by a given individual of the utility *u* of the quantity *x* of a certain product for him. In a broad sense, the utility function is the dependence of the utility (effect) of a certain action on its level (intensity).

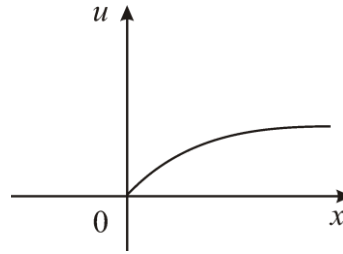


Fig. 5.12. *Graph of the utility function*

The **output function** (one-factor production function; see Fig. 5.13) is the dependence of the volume y of the output on the volume x of the processed resource. The output function is a particular type of **production function** that expresses the dependence of the result of production activity on the factors that caused it.

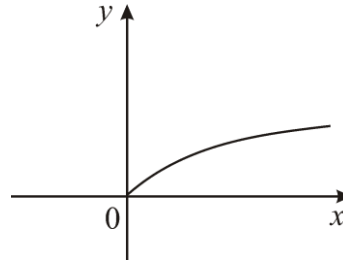


Fig. 5.13. *Graph of the output function*

The **cost function** is the dependence of production costs on the volume of products. The cost function is also a particular type of production function.

The **supply and demand function** is the dependence of the volume of demand D and supply S on the price of goods p.

Consider some product. Let $D(p)$ be the quantity (number of units) of a product that a buyer wants to buy at a given price p per unit. The function

$D = D(p)$ is called the *demand function* for the product. This function is decreasing. Usually, it has the form:

$$D = kp^a + c, \qquad (5.1)$$

where $a < 0$ (Fig. 5.14).

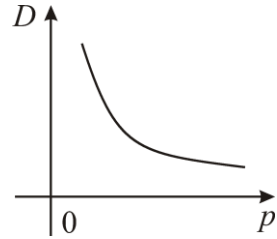


Fig. 5.14. *Graph of the demand function*

On the other hand, let $S(p)$ be the number of units of goods offered by the sellers in the market at the price *p*. Obviously, supply increases with rising prices. Therefore, *the function of the proposal* $S = S(p)$ is an increasing function. It usually has the form:

$$S = p^b + d, \qquad (5.2)$$

where $b \geq 1$ (Fig. 5.15).

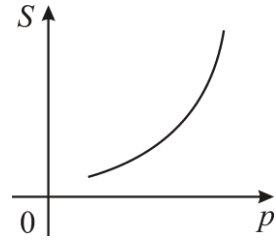


Fig. 5.15. *Graph of the proposal function*

For the economics of interest is the condition when demand is *equal* to the supply:

$$D(p) = S(p). \tag{5.3}$$

The price $p = p_0$ at which equality (5.3) holds is called **equilibrium**. The intersection point of the curves *D* and *S* (graphs of functions $D = D(p)$ and $S = S(p)$) is called the **equilibrium point**.

With an increase in the well-being of the population, the constant *c* in formula (5.1) increases, curve *D* rises, the equilibrium point shifts to the right (Fig. 5.16).

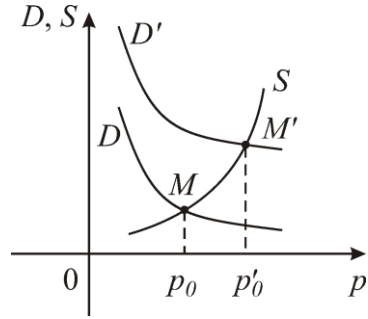


Fig. 5.16. *The position of the equilibrium point depending on the welfare of the population*

Questions

1) What is the natural domain of a function?
2) What are the ways to define functions?
3) What property does the graph of an odd function have?
4) What is the general term used to refer to increasing and decreasing functions?
5) How many different periods does a periodic function have?

6) Let function $y = f(x)$ have the smallest period *T*. Is function $y = f(3x)$ periodic and if so, what is its smallest period?
7) How to get the graph of the inverse function from the graph of the function itself?
8) Which function is the inverse of function $y = x^3$?
9) Does function $y = 2x + 3$ belong to the basic elementary functions?
10) How can one get elementary functions from basic elementary functions?
11) Will the sum of elementary functions be an elementary function? And the square root of the elementary function?
12) What functions of one variable are most often used in economics?

Chapter 6. Limits

6.1. Sequence. Limit of a sequence

If according to some law, each positive integer n is assigned one specific real number x_n, then they say that a **numerical sequence** is given:

$$x_1,\ x_2,\ \ldots,\ x_n,\ \ldots . \tag{6.1}$$

In other words, a numerical sequence is a function of a natural argument: $x_n = f(n)$, i.e. function defined on the set of natural numbers.

Sequence (6.1) is written briefly in the form $\{x_n\}$. Numbers x_1, x_2, …, x_n, … are called the **members** of the sequence and the n^{th} member x_n is called the **general member** of the sequence.

A sequence is considered given if its general member is specified or a method for obtaining any of its elements is specified. For example, a formula $x_n = \dfrac{2n+1}{5n+2}$ defines a sequence

$$\frac{3}{7},\ \frac{5}{12},\ \frac{7}{17},\ \frac{9}{22},\ \ldots,\ \frac{2n+1}{5n+2},\ \ldots . \tag{6.2}$$

The sequence may be *monotonic* or *nonmonotonic*, *limited* or *unlimited*. (There is no need to define these concepts since in § 5.1 definitions of a monotone and bounded function have already been given.)

In particular, sequence (6.2) is monotonically decreasing. Indeed, consider the difference $x_n - x_{n+1}$:

$$x_n - x_{n+1} = \frac{2n+1}{5n+2} - \frac{2\cdot(n+1)+1}{5\cdot(n+1)+2} = \frac{2n+1}{5n+2} - \frac{2n+3}{5n+7} =$$

$$= \frac{(2n+1)\cdot(5n+7)-(2n+3)\cdot(5n+2)}{(5n+2)\cdot(5n+7)} = \frac{10n^2+19n+7-10n^2-19n-6}{(5n+2)\cdot(5n+7)} =$$

$$= \frac{1}{(5n+2)\cdot(5n+7)}.$$

So, $x_n - x_{n+1} > 0$, i.e. $x_{n+1} < x_n$ for any *n*.

Consider the same sequence (6.2). If, for example, $n = 100$, then $x_n = \frac{201}{502}$; if $n = 100\,000$, then $x_n = \frac{200\,001}{500\,002}$. We see that with increasing *n*, the members of the sequence x_n are less and less different from $\frac{2}{5}$ and this difference can become arbitrarily small.

In particular, if $n = 100\,000$

$$x_n - \frac{2}{5} = \frac{200\,001}{500\,002} - \frac{2}{5} = 0{,}0000003999... < 0{,}000001.$$

Definition The number *a* is called the **limit of the sequence** $\{x_n\}$ if for any (arbitrarily small) number $\varepsilon > 0$ there exists number *N* such that for all $n > N$

$$\lim_{n\to\infty} x_n = a\,. \qquad (6.4)$$

If a sequence $\{x_n\}$ has limit a, then it is called **convergent** (to the number a). In this case, we write:

$$\lim_{n\to\infty} x_n = a\,. \qquad (6.4)$$

Sometimes instead of (6.4), it can be written: $x_n \to a$ as $n \to \infty$.

If we return to sequence (6.2), then we see that for $\varepsilon = 0{,}000001$ inequality (6.3) holds as $n > 100\,000$.

Let us find out the **geometric meaning of the limit of the numerical sequence**. Inequality (6.3) is equivalent to the double inequality:

$-\varepsilon < x_n - a < \varepsilon$, or $a - \varepsilon < x_n < a + \varepsilon$.

This means that for all $n > N$ all members of the sequence $\{x_n\}$ are in ε-neighborhood of the point a (Fig. 6.1).

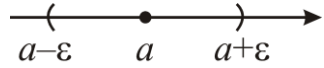


Fig. 6.1. *ε-neighborhood of a*

Therefore, there can be only a finite number of members of this sequence outside this neighborhood.

6.2. Limit of a function

Limit of a function at infinity

Considering the limit of sequence $x_n = f(n)$, we were dealing with a function whose argument *n* increasing, assumed only natural values. Now consider the function $y = f(x)$. Its argument *x* in the process of change can take any (not only natural and not only integer) values.

Definition Number *b* is called the **limit of the function** $y = f(x)$ **as** *x* **tends to infinity** if for any (arbitrarily small) $\varepsilon > 0$ there is a number $M > 0$ such that for all *x* satisfying the condition $|x| > M$ the inequality $|f(x) - b| < \varepsilon$ holds.

In this case, we write:

$$\lim_{x \to \infty} f(x) = b. \qquad (6.5)$$

Sometimes instead of (6.5), it can be written: $f(x) \to b$ as $x \to \infty$.

The meaning of the definition of the limit of a function at infinity is basically the same as for the limit of a sequence:

$\lim_{n \to \infty} x_n = a$ means that the members of the sequence differ arbitrarily little from *a* if *n* is large enough;

$\lim_{x \to \infty} f(x) = b$ means that the values of the function differ arbitrarily little from *b* if *x* is large enough in absolute value.

Remark. If in the definition stated above we replace the condition $|x| > M$ with the condition $x > M$, then we obtain the definition of the limit of the function as $x \to +\infty$. If we replace it with a condition $x < -M$, then we obtain the definition of the limit as $x \to -\infty$.

Limit of a function at a point

Let function $y = f(x)$ be defined in some neighborhood of the point *a*, except, perhaps, the point *a* itself.

Definition . N umber *b* is called the **limit of the function** $y = f(x)$ **as x tends to** *a* if for any (arbitrarily small) number $\varepsilon > 0$ there is a number $\delta > 0$ such that for all $x \neq x_0$ satisfying the condition $|x - a| < \delta$ inequality

$|f(x) - b| < \varepsilon$ holds.

In this case, we write:

$$\lim_{x \to a} f(x) = b . \qquad (6.6)$$

The geometric meaning of this definition is as follows: for any ε-neighborhood of point *b* (on the *Oy* axis) there exists a δ-neighborhood of point *a* (on the *Ox* axis) such that as soon as *x* falls into this δ-neighborhood of the point *a*, the corresponding value of function $f(x)$ belongs to an ε-neighborhoods of *b*: $x \in (a - \delta, a + \delta) \Rightarrow f(x) \in (b - \varepsilon, b + \varepsilon)$.

6.3. Infinitely small quantities. Infinitely big quantities.

Definition. A function $\alpha = \alpha(x)$ is called an **infinitely small quantity** (or simply **infinitesimal**) as $x \to x_0$ (as $x \to \infty$) if its limit is zero: $\lim\limits_{x \to x_0} \alpha(x) = 0$ ($\lim\limits_{x \to \infty} \alpha(x) = 0$).

An infinitesimal sequence is defined similarly.

Let us now consider the **relationship of a variable with its limit**.

Theorem 6.1. N umber *b* is the limit of the function $f(x)$ as $x \to x_0$ (as $x \to \infty$) if and only if

$$f(x) = b + \alpha(x), \tag{6.7}$$

where $\alpha(x)$ is infinitesimal as $x \to x_0$ (as $x \to \infty$).

Proof. 1. *Necessity*. Let $\lim\limits_{x \to x_0} f(x) = b$. Denote $\alpha(x) = f(x) - b$. Let $\varepsilon > 0$. Then there exists such $\delta > 0$ that for all $x \neq x_0$ satisfying the condition $|x - x_0| < \delta$ inequality $|f(x) - b| < \varepsilon$ holds, i.e. $|\alpha(x)| < \varepsilon$ and this means that $\alpha(x)$ is infinitely small.

2. *Sufficiency*. Let $f(x) = b + \alpha(x)$ where $\alpha(x)$ is infinitesimal. Then the difference $f(x) - b$ is infinitesimal, i.e. for every $\varepsilon > 0$ there exists

such $\delta > 0$ that for all $x \neq x_0$ satisfying the condition $|x - x_0| < \delta$ the inequality $|f(x) - b| < \varepsilon$ holds and this means that $\lim\limits_{x \to x_0} f(x) = b$.

We proved the theorem for the case $x \to x_0$. The proof is similar if $x \to \infty$.

Definition Function $y = f(x)$ is called an **infinitely large quantity** (or simply **infinitely large**) as $x \to x_0$ if for every $M > 0$ there exists such $\delta > 0$ that for all *x* not equal to x_0 and satisfying the condition $|x - x_0| < \delta$ inequality $|f(x)| > M$ holds.

In this case, we also say that $f(x)$ has an infinite limit as $x \to x_0$ or that $f(x)$ tends to infinity as $x \to x_0$. We write: $\lim\limits_{x \to x_0} f(x) = \infty$, or $f(x) \to \infty$ as $x \to x_0$.

Infinitely large is determined similarly as $x \to \infty$.

There is an obvious connection between the concepts of infinitesimal and infinitely large: if $\alpha(x)$ is infinitely small as $x \to x_0$ ($x \to \infty$), then $f(x) = \dfrac{1}{\alpha(x)}$ is infinitely large as $x \to x_0$ ($x \to \infty$); if $f(x)$ infinitely large, then $\alpha(x) = \dfrac{1}{f(x)}$ is, infinitely small.

Properties of infinitesimal:

1. *The algebraic sum of a finite number of infinitesimals is infinitely small.*

2. *The product of an infinitesimal quantity by a limited quantity is infinitely small.*

The corollary of this statement is the following statements:

1) *the product of an infinitesimal by a constant is infinitely small*;

2) *the product of two infinitesimal is infinitesimal*.

Let us prove property 2 as an example. Let $\alpha = \alpha(x)$ be infinitesimal as $x \to x_0$ and let *y* be a bounded quantity, i.e. $|y| < M$. Let $\varepsilon > 0$. Then for $\varepsilon' = \dfrac{\varepsilon}{M}$ there exists such $\delta > 0$ that for all $x \neq x_0$ satisfying the condition $|x - x_0| < \delta$ the following inequality holds:

$$|\alpha| < \varepsilon' = \frac{\varepsilon}{M}.$$

Then

$$|\alpha y| < \frac{\varepsilon}{M} \cdot M = \varepsilon, \text{ i.e. } |\alpha y| < \varepsilon.$$

And this means that $|\alpha y|$ is infinitesimal.

Infinitesimal can be compared. In particular, if $\lim\limits_{x \to x_0} \dfrac{\alpha(x)}{\beta(x)} = 0$, then we say that $\alpha(x)$ is **infinitesimal of a higher order** than $\beta(x)$ and write $\alpha(x) = o(\beta(x))$.

6.4. Basic theorems about limits

Uniqueness of a limit

Theorem 6.2. If a function has a limit, then this limit is unique.

Proof. Assume the opposite: $\lim\limits_{x\to a} f(x)=b_1$, $\lim\limits_{x\to a} f(x)=b_2$ and $b_1 \neq b_2$. Then, according to Theorem 6.1:

$f(x)=b_1+\alpha_1(x)$ and $f(x)=b_2+\alpha_2(x)$

here $\alpha_1(x)$ and $\alpha_2(x)$ are infinitesimal. Subtracting these equalities term by term, we obtain

$0=b_1-b_2+[\alpha_1(x)-\alpha_2(x)]$

This gives

$\alpha_1(x)-\alpha_2(x)=b_1-b_2=c=\text{const}\neq 0$.

This equality is impossible since the difference $\alpha_1(x)-\alpha_2(x)$ is infinitesimal. Therefore, the assumption of the existence of two different limits is false.

The limit of the sum, product, quotient

We will consider the limits of functions $u=u(x)$ and $v=v(x)$ as $x\to a$ or as $x\to\infty$. A short notation $\lim u$ will mean either $\lim\limits_{x\to a} u(x)$ or $\lim\limits_{x\to\infty} u(x)$. $\lim v$ is similar.

1. *The limit of the algebraic sum is equal to the algebraic sum of the limits*:

$\lim(u+v)=\lim u+\lim v$.

2. *The limit of the product is equal to the product of the limits:*

$\lim(uv)=\lim u\cdot\lim v$.

3. *The limit of the quotient is equal to the quotient of the limits* (provided that the limit of the divisor is nonzero):

$$\lim\frac{u}{v}=\frac{\lim u}{\lim v}.$$

We state these statements more clearly as $x\to a$ and prove it. (The proof in case of $x\to\infty$ is similar.)

1. If $u=u(x)$ and $v=v(x)$ have limits $\lim\limits_{x\to a}u(x)=b_1$, $\lim\limits_{x\to a}v(x)=b_2$ as $x\to a$, then its sum $u(x)+v(x)$ has a limit $\lim\limits_{x\to a}[u(x)+v(x)]=b_1+b_2$.

Proof. Since $\lim\limits_{x\to a}u(x)=b_1$, $\lim\limits_{x\to a}v(x)=b_2$, then by Theorem 14.1 the functions $u(x)$ and $v(x)$ can be written in the form $u(x)=b_1+\alpha_1(x)$, $v(x)=b_2+\alpha_2(x)$, where $\alpha_1(x)$ and $\alpha_2(x)$ are infinitesimals, $\lim\limits_{x\to a}\alpha_1(x)=0$, $\lim\limits_{x\to a}\alpha_2(x)=0$. Hence,

$$u(x)+v(x)=[b_1+\alpha_1(x)]+[b_2+\alpha_2(x)]=(b_1+b_2)+[\alpha_1(x)+\alpha_2(x)].$$

Here b_1+b_2 is a constant, $\alpha_1(x)+\alpha_2(x)$ is infinitesimal. Applying Theorem 6.1 again, we obtain

$\lim\limits_{x\to a}\left[u(x)+v(x)\right]=b_1+b_2$, i.e.

$\lim\limits_{x\to a}\left[u(x)+v(x)\right]=\lim\limits_{x\to a}u(x)+\lim\limits_{x\to a}v(x).$

2. If the functions $u=u(x)$ and $v=v(x)$ have limits $\lim\limits_{x\to a}u(x)=b_1$, $\lim\limits_{x\to a}v(x)=b_2$ as $x\to a$, then the product $u(x)v(x)$ has a limit and $\lim\limits_{x\to a}\left[u(x)v(x)\right]=b_1b_2$.

Proof. Since $\lim\limits_{x\to a}u(x)=b_1$, $\lim\limits_{x\to a}v(x)=b_2$, according to Theorem 14.2 $u(x)=b_1+\alpha_1(x)$, $v(x)=b_2+\alpha_2(x)$ where $\alpha_1(x)$ and $\alpha_2(x)$ are infinitesimal as $x\to a$. We have:

$$u(x)v(x)=\left[b_1+\alpha_1(x)\right]\left[b_2+\alpha_2(x)\right]=b_1b_2+b_1\alpha_2(x)+b_2\alpha_1(x)+\alpha_1(x)\alpha_2(x).$$

Denote $b_1\alpha_2(x)+b_2\alpha_1(x)+\alpha_1(x)\alpha_2(x)=\alpha(x)$. As defined above, $b_1\alpha_2(x)$, $b_2\alpha_1(x)$ and $\alpha_1(x)\alpha_2(x)$ are infinitesimals, therefore, its sum $\alpha(x)$ is also infinitesimal. So,

$$u(x)v(x)=b_1b_2+\alpha(x),$$

where $\alpha(x)$ is infinitesimal. That means that $u(x)v(x)$ has a limit which is equal to b_1b_2. That completes the proof.

Corollary. A constant multiplier can be taken out of the limit sign:

$\lim\limits_{x\to a}cu(x)=c\lim\limits_{x\to a}u(x)$.

Indeed, if $\lim\limits_{x\to a} u(x) = b$, $c = \text{const}$, then $\lim\limits_{x\to a} c = c$. Then

$$\lim_{x\to a}\left[cu(x)\right] = \lim_{x\to a} c \cdot \lim_{x\to a} u(x) = c \cdot \lim_{x\to a} u(x)$$

3. If $u = u(x)$ and $v = v(x)$ have limits $\lim\limits_{x\to a} u(x) = b_1$, $\lim\limits_{x\to a} v(x) = b_2$ as $x \to a$ and $b_2 \neq 0$, then the function $\dfrac{u(x)}{v(x)}$ also has the limit as $x \to a$ and $\lim\limits_{x\to a} \dfrac{u(x)}{v(x)} = \dfrac{b_1}{b_2}$.

Proof. Let $\lim\limits_{x\to a} u(x) = b_1$, $\lim\limits_{x\to a} v(x) = b_2 \neq 0$. By Theorem 6.1 $u(x) = b_1 + \alpha_1(x)$, $v(x) = b_2 + \alpha_2(x)$, where $\alpha_1(x)$ and $\alpha_2(x)$ are infinitesimals. We do simple identity transformations:

$$\frac{u(x)}{v(x)} = \frac{b_1 + \alpha_1(x)}{b_2 + \alpha_2(x)} = \frac{b_1}{b_2} + \left(\frac{b_1 + \alpha_1(x)}{b_2 + \alpha_2(x)} - \frac{b_1}{b_2}\right) = \frac{b_1}{b_2} + \frac{b_2\alpha_1(x) - b_1\alpha_2(x)}{b_2\left[b_2 + \alpha_2(x)\right]}$$

.

So,

$$\frac{u(x)}{v(x)} = \frac{b_1}{b_2} + \frac{b_2\alpha_1(x) - b_1\alpha_2(x)}{b_2\left(b_2 + \alpha_2(x)\right)}.$$

Here $\dfrac{b_1}{b_2}$ is a constant, and a fraction $\dfrac{b_2\alpha_1(x) - b_1\alpha_2(x)}{b_2\left[b_2 + \alpha_2(x)\right]}$ is infinitesimal. It follows from the properties of infinitesimals: $b_2\alpha_1(x) - b_1\alpha_2(x)$ is infinitesimal, and $\dfrac{1}{b_2\left[b_2 + \alpha_2(x)\right]}$ is bounded.

The limit passage in inequalities

We will assume that the inequalities discussed below are fulfilled in some neighborhood of point *a* (excluding, perhaps, this point) or for sufficiently large *x*.

4. If function $u = u(x)$ is non-negative: $u \geq 0$, then $\lim u \geq 0$.

5. If inequality $u \geq v$ holds for functions $u = u(x)$ and $v = v(x)$, then $\lim u \geq \lim v$.

6. If the inequality $u \leq v \leq w$ holds for functions $u = u(x)$, $v = v(x)$, $w = w(x)$ and $\lim u = \lim w = b$, then $\lim v = b$.

For example, we state the last statement in more detail and prove it.

Theorem 6.3. If conditions $u(x) \leq v(x) \leq w(x)$ are satisfied in some neighborhood of *a* and functions $u = u(x)$ and $w = w(x)$ have the same limit as $x \to a$: $\lim\limits_{x \to a} u(x) = \lim\limits_{x \to a} w(x) = b$, then the function $v(x)$ has the same limit: $\lim\limits_{x \to a} v(x) = b$.

Proof. Let an arbitrary $\varepsilon > 0$ be given. Then since $\lim\limits_{x \to a} u(x) = b$, then there exists $\delta_1 > 0$ such that for all $x \neq a$ satisfying the condition $|x - a| < \delta_1$ the following inequality holds:

$|u(x)-b<\varepsilon|$. (*)

Since $\lim_{x\to a} w(x) = b$, then there exists such $\delta_2 > 0$ that for all $x \neq a$ satisfying the condition $|x-a| < \delta_2$ the inequality holds:

$$|w(x)-b| < \varepsilon. \qquad (**)$$

If ™ is the smallest of δ_1 and δ_2, then for all $x \neq a$ satisfying the condition $|x-a| < \delta$ both inequalities (*) and (**) hold simultaneously, i.e. at the same time

$b-\varepsilon < u(x) < b+\varepsilon$, $b-\varepsilon < w(x) < b+\varepsilon$.

From the last inequalities we get:

$b-\varepsilon < u(x) \le v(x) \le w(x) < b+\varepsilon$,

therefore $b-\varepsilon < v(x) < b+\varepsilon$, i.e.

$|v(x)-b| < \varepsilon$.

And that means that $\lim_{x\to a} v(x) = b$. (The proof is similar for $x \to \infty$.)

One-side limits

If $f(x)$ tends to the limit *b* as *x* tends to *a* and $x < a$, then *b* is called the **limit of function** $f(x)$ **as** *x* **approaches** *a* **from the left** or **left-side limit**. In this case, we write $\lim_{x\to a-0} f(x) = b$ (or $\lim_{x\to a-} f(x) = b$).

Similarly, the limit of $f(x)$ as $x \to a$, $x > a$ is called the **right-side limit** and is written in the form $\lim\limits_{x \to a+0} f(x) = b$ (or $\lim\limits_{x \to a+} f(x) = b$).

It is easy to prove that a function $f(x)$ has a limit as $x \to a$ if and only if there are simultaneously left-side and right-side limits and they are equal to each other:

$$\lim_{x \to a-} f(x) = \lim_{x \to a+} f(x) = b.$$

In this case, the limit in the usual sense is also equal to *b*:

$$\lim_{x \to a} f(x) = b.$$

A sufficient criterion of the existence of a limit

Theorem 6.4. A monotonic bounded sequence has a finite limit.

In particular, if a sequence $\{x_n\}$ increases and is bounded above (i.e., there exists an *M* such that $x_n < M$ for all *n*), then it has a limit. Similarly, decreasing and bounded below sequence also has a limit. Moreover, the increase and decrease can be understood in a broad sense (i.e. $x_{n+1} \geq x_n$ and $x_{n+1} \leq x_n$ respectively for all *n*).

The validity of this theorem seems almost obvious, but we do not give a strict proof of it since it is based on information from the theory of real numbers that are not considered in this book.

Let's look at some examples. Moreover, we take into account that the quantities $\frac{1}{n}$, $\frac{1}{n^2}$ (and generally $\frac{1}{n^\alpha}$ for> 0) are infinitesimals.

Example 6.1. Find $\lim\limits_{n\to\infty}\dfrac{2n^2-n+3}{5n^2+3n-4}$.

Solution. The numerator and denominator of the fraction tend to infinity as $x\to\infty$. This is the so-called "$\frac{\infty}{\infty}$" uncertainty. Therefore, it is impossible to apply the limit theorem. We first convert this expression by dividing the numerator and denominator by n^2. Then we apply the statement about the limit of the quotient and about the limit of the sum.

$$\lim_{n\to\infty}\frac{2n^2-n+3}{5n^2+3n-4}=\lim_{n\to\infty}\frac{2-\frac{1}{n}+\frac{3}{n^2}}{5+\frac{3}{n}-\frac{4}{n^2}}=\frac{2-0+0}{5+0-0}=\frac{2}{5}.$$

Example 6.2. $\lim\limits_{n\to\infty}\dfrac{\sqrt{x^2+1}+2n}{4n+3}$.

Solution.

$$\lim_{n\to\infty}\frac{\sqrt{n^2+1}+2n}{4n+3}=\lim_{n\to\infty}\frac{\sqrt{1+\frac{1}{n^2}}+2}{4+\frac{3}{n}}=\frac{3}{4}$$

Here we divide the numerator and denominator by *n* and note that $\dfrac{\sqrt{n^2+1}}{n}=\sqrt{\dfrac{n^2+1}{n^2}}=\sqrt{1+\dfrac{1}{n^2}}$.

Example 6.3. Find $\lim\limits_{x\to+\infty}(\sqrt{x^2+4x}-\sqrt{x^2+1})$.

Solution. Here is the uncertainty of the form "$\infty-\infty$". We multiply and divide this difference of radicals by their sum (i.e., by the conjugate expression). We get:

$$\lim_{x\to+\infty}(\sqrt{x^2+4x}-\sqrt{x^2+1}=$$

$$\lim_{x\to+\infty}\frac{(\sqrt{x^2+4x}-\sqrt{x^2+1})\cdot(\sqrt{x^2+4x}+\sqrt{x^2+1}}{\sqrt{x^2+4x}+\sqrt{x^2+1}}=$$

$$=\lim_{x\to+\infty}\frac{4x-1}{\sqrt{x^2+4x}+\sqrt{x^2+1}}=\lim_{x\to+\infty}\frac{4-\frac{1}{x}}{\sqrt{1+\frac{4}{x}}+\sqrt{1+\frac{1}{x^2}}}=2$$

6.5. Two remarkable limits

1. Let us prove that there exists the limit of function $\frac{\sin x}{x}$ as $x\to 0$, and this limit is equal to 1:

$$\lim_{x\to 0}\frac{\sin x}{x}=1. \qquad (6.8)$$

This limit is usually called the **first remarkable limit**.

Proof.

Consider a circle of radius *R*: $OA=OM=R$ (Fig 6.2 shows its section). Let x be a radial measure of acute angle AOM: $0<x<\frac{\pi}{2}$.

Then $MB=OM\cdot\sin x=R\sin x$, $NA=OA\cdot\operatorname{tg}x=R\operatorname{tg}x$.

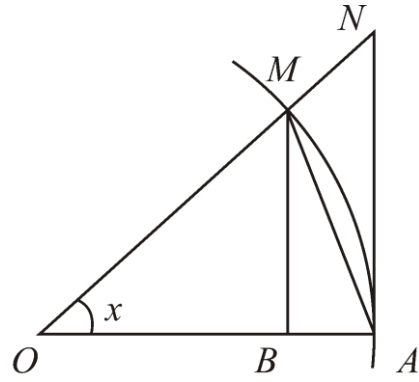


Fig. 6.2

Let S_1 be the area of triangle OMA: $S_1 = \frac{1}{2} OA \cdot MB = \frac{1}{2} R^2 \sin x$.

Denote as S_2 the area of sector OMA. Then $S_2 = \frac{1}{2} R^2 x$. Denote as S_3 the area of sector AON. Then $S_3 = \frac{1}{2} OA \cdot NA = \frac{1}{2} R^2 \operatorname{tg} x$. Obviously,

$S_1 < S_2 < S_3$,

hence,

$\frac{1}{2} R^2 \sin x < \frac{1}{2} R^2 x < \frac{1}{2} R^2 \operatorname{tg} x$.

This gives

$\sin x < x < \operatorname{tg} x$.

Dividing the last inequality by $\sin x$, we obtain:

$1 < \frac{x}{\sin x} < \frac{1}{\cos x}$, or $\cos x < \frac{\sin x}{x} < 1$.

So, the function $\frac{\sin x}{x}$ is located between the functions $u(x) = \cos x$ and $w(x) = 1$ which have the same limit 1 as $x \to 0$ [1]. According to Theorem 6.3 we obtain the equality:

$$\lim_{x \to 0} \frac{\sin x}{x} = 1.$$

The proof is similar for $x < 0$.

***Example* 6.4.** Find $\lim_{x \to 0} \frac{\sin ax}{ax}$.

Solution. Make a substitution $\alpha = ax$. Obviously, $x \to 0$ equals to $\alpha \to 0$. We obtain

$$\lim_{x \to 0} \frac{\sin ax}{ax} = \lim_{\alpha \to 0} \frac{\sin \alpha}{\alpha} = 1.$$

Example 6.5. Find $\lim_{x \to 0} \frac{\sin ax}{bx}$.

Solution. $\lim_{x \to 0} \frac{\sin ax}{bx} = \lim_{x \to 0} \frac{\sin ax}{ax} \cdot \frac{a}{b} = \frac{a}{b} \cdot \lim_{x \to 0} \frac{\sin ax}{ax} = \frac{a}{b} \cdot 1 = \frac{a}{b}$.

***Example* 6.6.** Find $\lim_{x \to 0} \frac{\operatorname{tg} x}{x}$.

---

[1] tends to 1 as since .

Solution. $\lim\limits_{x\to 0}\dfrac{\operatorname{tg} x}{x}=\lim\limits_{x\to 0}\dfrac{\sin x}{x}\dfrac{1}{\cos x}=\lim\limits_{x\to 0}\dfrac{\sin x}{x}\lim\limits_{x\to 0}\dfrac{1}{\cos x}=1\cdot 1=1$.

Note that:

$$\lim_{x\to 0}\frac{\operatorname{tg} x}{x}=1.$$

***Example* 6.7.** Find $\lim\limits_{x\to 0}\dfrac{1-\cos x}{x^2}$.

Solution.

$$\lim_{x\to 0}\frac{1-\cos x}{x^2}=\lim_{x\to 0}\frac{(1-\cos x)(1+\cos x)}{x^2(1+\cos x)}=\lim_{x\to 0}\frac{1-\cos^2 x}{x^2(1+\cos x)}=$$

$$=\lim_{x\to 0}\frac{\sin^2 x}{x^2(1+\cos x)}=\lim_{x\to 0}\left(\frac{\sin x}{x}\right)^2\frac{1}{1+\cos x}=1\cdot\frac{1}{2}=\frac{1}{2}.$$

(We can calculate this example using the formula $1-\cos x=2\sin^2\dfrac{x}{2}$.)

2. Consider the sequence $\{a_n\}$ with the general term:

$$a_n=\left(1+\frac{1}{n}\right)^n.$$

We prove that this sequence increases monotonously and is bounded. Hence, it has a limit by Theorem 6.4.

Using the binomial expression, we obtain

$$\left(1+\frac{1}{n}\right)^n = 1+n\frac{1}{n}+\frac{n(n-1)}{1\cdot 2}\left(\frac{1}{n}\right)^2+\frac{n(n-1)(n-2)}{1\cdot 2\cdot 3}\left(\frac{1}{n}\right)^3+\ldots+$$

$$+\frac{n(n-1)(n-2)\cdots[n-(n-1)]}{1\cdot 2\cdot 3\cdots n}\left(\frac{1}{n}\right)^n.$$

Transform the expression

$$\left(1+\frac{1}{n}\right)^n = 1+1+\frac{1}{1\cdot 2}\frac{n(n-1)}{n^2}+\frac{1}{1\cdot 2\cdot 3}\frac{n(n-1)(n-2)}{n^3}+\ldots+$$

$$+\frac{1}{1\cdot 2\cdot 3\cdots n}\frac{n(n-1)(n-2)\cdots[n-(n-1)]}{n^n}=1+1+\frac{1}{1\cdot 2}\left(1-\frac{1}{n}\right)+$$

$$+\frac{1}{1\cdot 2\cdot 3}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)+\ldots+\frac{1}{1\cdot 2\cdot 3\cdots n}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{n-1}{n}\right).$$

(*)

The last equality gives $\left(1+\frac{1}{n}\right)^n \geq 2$ and it shows that the considered sequence increases as *n* increases.

Indeed, each term of the last sum (starting from the third one) increases as the index increases from *n* to $n+1$:

$$\frac{1}{1\cdot 2}\left(1-\frac{1}{n}\right)<\frac{1}{1\cdot 2}\left(1-\frac{1}{n+1}\right)$$ etc.

adding another term (positive).

So, the sequence $\{a_n\}$ increases monotonously.

Let us prove now that $\{a_n\}$ is bounded. Obviously, $1-\frac{1}{n}<1$, $\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)<1$ etc. Therefore from (*) we obtain the inequality:

$$\left(1+\frac{1}{n}\right)^n<1+1+\frac{1}{1\cdot 2}+\frac{1}{1\cdot 2\cdot 3}+\ldots+\frac{1}{1\cdot 2\cdot 3\cdots n}.$$

Noting that

$$\frac{1}{1\cdot 2\cdot 3}<\frac{1}{2^2},\ \frac{1}{1\cdot 2\cdot 3\cdot 4}<\frac{1}{2^3},\ \ldots,\ \frac{1}{1\cdot 2\cdot 3\cdots n}<\frac{1}{2^{n-1}},$$

we obtain:

$$\left(1+\frac{1}{n}\right)^n<1+1+\frac{1}{2}+\frac{1}{2^2}+\ldots+\frac{1}{2^{n-1}}.$$

The terms of the right part (starting from the second term) form the geometric progression:

$$1+\frac{1}{2}+\frac{1}{2^2}+\ldots+\frac{1}{2^{n-1}}=\frac{1-\left(\frac{1}{2}\right)^n}{1-\frac{1}{2}}=2-\left(\frac{1}{2}\right)^{n-1}<2.$$

Hence,

$$\left(1+\frac{1}{n}\right)^n<3.$$

So,

$$2\le\left(1+\frac{1}{n}\right)^n<3.$$

We have proved that the sequence $\{a_n\}$, where $a_n = \left(1+\frac{1}{n}\right)^n$ is increasing and bounded. It has a limit by Theorem 6.4.

Definition The limit $\lim\limits_{n\to\infty}\left(1+\frac{1}{n}\right)^n$ is called the **number e**:

$$\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n = \mathrm{e}. \tag{6.9}$$

The limit (6.9) is called the **second remarkable limit**.

The number e is an irrational number. Moreover, it is a transcendental number, i.e. is not the root of any algebraic equation with integer coefficients.

It is known that

$$\mathrm{e} = 2{,}7182818284...$$

In most cases, in practice it is believed.

Consider an example. Let the initial contribution to the bank be S_0 monetary units. The bank pays annually *p*%. Then after the end of the year, the deposit amount will be $S_0 + \frac{p}{100}\cdot S_0 = S_0\cdot\left(1+\frac{p}{100}\right)$, i.e. multiplied by $\left(1+\frac{p}{100}\right)$. In two years it will again be multiplied by $\left(1+\frac{p}{100}\right)$ and will be $S_0\cdot\left(1+\frac{p}{100}\right)^2$, etc. Thus, at *p*% per annum after *n* years, the deposit amount will be equal to:

$$S_n = S_0 \cdot \left(1 + \frac{p}{100}\right)^n . \qquad (6.10)$$

This is a **compound interest formula** (see Example 1.3).

It should be noted that interest on a deposit may not be accrued once a year but, for example, quarterly, monthly, every day. Formula (6.10) allows you to calculate the amount of the deposit S_n after *n* periods at an interest rate of *p*% per period (regardless of how long these periods are).

Imagine that a bank located in Moscow having finished a working day transfers (taking into account the time difference) a certain amount S_0 to a bank located in Vladivostok for 12 hours from 20 hours of the current day to 8 hours of the next day Moscow time. Vladivostok Bank returns money to the beginning of the work of the Moscow Bank, paying 1% for the use of this short-term loan. Then, the next day, the Moscow bank repeats this operation but with the received amount of 101% of S_0 etc. (Such an agreement between banks is hardly possible in practice but the rate of 1% per day in the early 1990s was real.)

After 300 days, the Moscow bank will receive the amount:

$$S_{300} = S_0\left(1 + \frac{1}{100}\right)^{300} = S_0\left[\left(1 + \frac{1}{100}\right)^{100}\right]^3 \approx S_0 \mathrm{e}^3 \approx S_0 2{,}7^3 \approx 19{,}68 S_0 ,$$

i.e. the initial amount will increase almost 20 times over the year.

In general, let amount S_0 be placed in the bank for *t* years at *p*% per annum and interest accrued *n* times a year. Then the interest rate for the $\frac{1}{n}$

part of the year will be $\frac{p}{n}$ %, and the deposit for *t* years (with *nt* charges) will be:

$$S_n = S_0 \cdot \left(1+\frac{p}{100n}\right)^{nt},$$

or denoting $\frac{p}{100} = r$:

$$S_n = S_0 \cdot \left(1+\frac{r}{n}\right)^{nt}.$$

Convert this last expression:

$$S_n = S_0 \cdot \left[\left(1+\frac{r}{n}\right)^{\frac{n}{r}}\right]^{rt}.$$

We introduce the notation: $\frac{n}{r} = m$. Let $n \to \infty$, then $m \to \infty$. We obtain

$$S = \lim_{n\to\infty} S_n = \lim_{m\to\infty} S_0 \left[\left(1+\frac{1}{m}\right)^{m}\right]^{rt} = S_0 \mathrm{e}^{rt}.$$

This formula

$$S = S_0 \mathrm{e}^{rt} \text{ or } S = S_0 \mathrm{e}^{\frac{pt}{100}}$$

is called the **continuous interest formula**.

Now consider the function:

$$f(x)=\left(1+\frac{1}{x}\right)^x.$$

Here *x* changes continuously, taking any (and not only natural) values.

Theorem 6.8. The limit of function $f(x)=\left(1+\frac{1}{x}\right)^x$ as $x\to\infty$ exists and is equal to e.

Proof. Let $x\to+\infty$. For each value *x* there exists a natural number *n* such that

$$n\le x\le n+1.$$

From these inequalities we obtain:

$$\frac{1}{n}\ge\frac{1}{x}\ge\frac{1}{n+1},$$

$$1+\frac{1}{n}\ge1+\frac{1}{x}\ge1+\frac{1}{n+1},$$

$$\left(1+\frac{1}{n}\right)^{n+1}\ge\left(1+\frac{1}{x}\right)^x\ge\left(1+\frac{1}{n+1}\right)^n.$$

Obviously, if $x\to\infty$, then $n\to\infty$. We find the limits of the variables between which the function $\left(1+\frac{1}{x}\right)^x$ is enclosed:

$$\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^{n+1}=\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n\cdot\left(1+\frac{1}{n}\right)=\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n\cdot\lim_{n\to\infty}\left(1+\frac{1}{n}\right)=\mathrm{e}\cdot1=\mathrm{e}$$

$$\lim_{n\to\infty}\left(1+\frac{1}{n+1}\right)^{n}=\lim_{n\to\infty}\frac{\left(1+\frac{1}{n+1}\right)^{n+1}}{1+\frac{1}{n+1}}=\ldots=\frac{\mathrm{e}}{1}=\mathrm{e}.$$

So, both variables between which the function $\left(1+\frac{1}{x}\right)^{x}$ is enclosed have the same limit e. Therefore, by Theorem 6.3

$$\lim_{x\to\infty}\left(1+\frac{1}{x}\right)^{x}=\mathrm{e}.$$

For the case of $x\to-\infty$ making the substitution $x=-y$, it is easy to prove that also

$$\lim_{x\to-\infty}\left(1+\frac{1}{x}\right)^{x}=\mathrm{e}.$$

So,

$$\lim_{x\to\infty}\left(1+\frac{1}{x}\right)^{x}=\mathrm{e}. \qquad (6.11)$$

We make a replacement $\frac{1}{x}=\alpha$ in (6.11). Then $x\to\pm\infty$ is equivalent to $\alpha\to 0$. We obtain

$$\lim_{\alpha\to 0}(1+\alpha)^{\frac{1}{\alpha}}=\mathrm{e}. \qquad (6.12)$$

We have obtained three formulas for the number e: (6.9), (6.11) and (6.12).

Number e is one of the fundamental quantities in mathematics.

Exponential function with base e (Fig. 6.3)

$$y = \mathrm{e}^x$$

(and generally the function of the form $y = \mathrm{e}^{ax}$) plays an important role in mathematics and its applications. It is used in statistics, physics, chemistry, in the study of demographic processes, etc. This function (and its graph) is called the **exponent**.

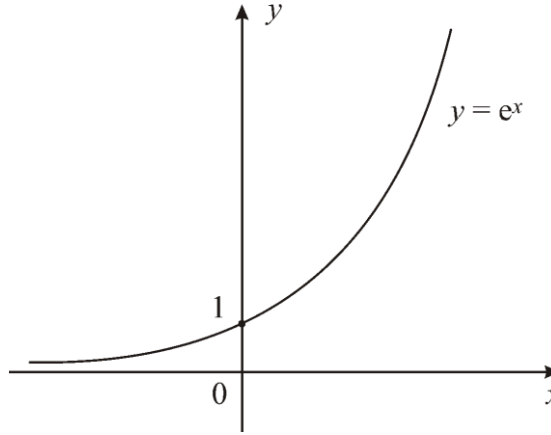


Fig. 6.3. *Exponent*

The logarithm with base e is called the **natural logarithm** and is denoted by the symbol $\ln$: $\log_e x = \ln x$.

Let's look at some examples.

Example 6.9. Find $\lim\limits_{x\to\infty}\left(1+\dfrac{4}{x}\right)^x$.

Solution. We apply the substitution of variable by setting $\frac{4}{x} = \alpha.$ Then $x = \frac{4}{\alpha}$. We get $\lim\limits_{\alpha \to 0} (1+\alpha)^{\frac{4}{\alpha}} = \lim\limits_{\alpha \to 0} \left[(1+\alpha)^{\frac{1}{\alpha}} \right]^4 = e^4.$

Example 6.10. Find $\lim\limits_{x \to \infty} \left(1 + \frac{5}{x}\right)^{2x}$.

Solution.

$$\lim_{x \to \infty} \left(1 + \frac{5}{x}\right)^{2x} = \lim_{x \to \infty} \left[\left(1 + \frac{5}{x}\right)^{\frac{x}{5}} \right]^{\frac{5}{x} \cdot 2x} = \lim_{x \leftarrow \infty} \left[\left(1 + \frac{5}{x}\right)^{\frac{x}{5}} \right]^{10} = e^{10}.$$

Example 6.11. Find $\lim\limits_{x \to \infty} \left(1 - \frac{1}{x}\right)^{x}$.

Solution. $\lim\limits_{x \to \infty} \left[\left(1 - \frac{1}{x}\right)^{-x} \right]^{-1} = \lim\limits_{\alpha \to 0} \left[(1+\alpha)^{\frac{1}{\alpha}} \right]^{-1} = e^{-1} = \frac{1}{e.}$

(Here we applied the substitution $-\frac{1}{x} = \alpha$).

Example 6.12. Find $\lim\limits_{x \to \infty} \left(\frac{2x+3}{2x+1}\right)^{x+1}$.

Solution. Divide the numerator and denominator by 2*x*:

$$\lim_{x \to \infty} \left(\frac{1 + \frac{3}{2x}}{1 + \frac{1}{2x}} \right)^{x+1} = \lim_{x \to \infty} \frac{\left(1 + \frac{3}{2x}\right)^{x} \cdot \left(1 + \frac{3}{2x}\right)}{\left(1 + \frac{1}{2x}\right)^{x} \cdot \left(1 + \frac{1}{2x}\right)} = \frac{e^{\frac{3}{2}}}{e^{\frac{1}{2}}} = e.$$

Questions

1) What is a general term of a sequence?
2) Can the same numbers correspond to different numbers in a numerical sequence?
3) Let all members of the monotonic sequence be multiplied by –1. Will the resulting sequence be monotonic?
4) Let number 4 be the limit of the number sequence. Is it possible to say that outside the interval (3, 5) there is only a finite number of members of the sequence?
5) Does the sequence of 1, 0, 1, 0, 1, 0, ... have a limit?
6) Let the number 5 be the limit of the number sequence. Can this sequence have negative terms?
7) Can the number –1 be the limit of a numerical sequence, all members of which are positive?
8) Can a sequence have two different limits?
9) Let the inequality $f(x) \geq 5{,}001$ hold for all *x*. Could it be $\lim\limits_{x \to 1} f(x) = 5$?
10) What is number e?
11) What function is called the exponent? Which curve is called the exponent?

Chapter 7. Continuity of a function

7.1. Main definitions

The concept of continuity of function is one of the basic concepts of mathematical analysis.

Definition 1. A function $f(x)$ is called **continuous** at a point x_0 if the following three conditions are satisfied: 1) $f(x)$ is defined in a certain neighborhood of the point x_0; 2) there is a finite limit of $f(x)$ as $x \to x_0$; 3) this limit is equal to the value of the function at the point x_0, i.e.

$$\lim_{x \to x_0} f(x) = f(x_0). \qquad (7.1)$$

Note that equality (7.1) and the continuity condition of $f(x)$ at a point x_0 can be written in the form:

$$\lim_{x \to x_0} f(x) = f\left(\lim_{x \to x_0} x\right).$$

From a geometric point of view, a continuous function is a function whose graph is a continuous curve.

There are several equivalent definitions of continuity.

Denote the difference $x - x_0$ as Δx. We say that when passing from value x_0 to value *x*, the argument receives an *increment* $\Delta x = x - x_0$. In this case, the function $y = f(x)$ receives a corresponding increment

$\Delta y = f(x_0 + \Delta x) - f(x_0)$. In view of the above, equality (7.1) is equivalent to equality

$$\lim_{\Delta x \to 0} \Delta y = 0 .$$

Definition 2. A function $y = f(x)$ is called **continuous** at a point x_0 if it is defined at this point and some neighborhood of it and

$$\lim_{\Delta x \to 0} \Delta y = 0 . \qquad (7.2)$$

(This definition is easy to remember in the following form: *a function is continuous if an infinitesimal increment of the argument corresponds to an infinitely small increment of the function*.)

Example 7.1. We show that the function $y = \sin x$ is continuous at an arbitrary point *x*. Give the argument an increment Δx. Then the function will get the increment

$$\Delta y = \sin(x + \Delta x) - \sin x = 2\cos\frac{(x + \Delta x) + x}{2}\sin\frac{(x + \Delta x) - x}{2} =$$

$$= 2\cos\left(x + \frac{\Delta x}{2}\right)\cdot\sin\left(\frac{\Delta x}{2}\right).$$

If $\Delta x \to 0$, then $\sin\frac{\Delta x}{2} \to 0$ (since $\left|\sin\frac{\Delta x}{2}\right| < \left|\frac{\Delta x}{2}\right|$); while $\cos\left(x + \frac{\Delta x}{2}\right)$ is limited. Therefore

$$\lim_{\Delta x \to 0} \Delta y = \lim_{\Delta x \to 0} 2\cos\left(x + \frac{\Delta x}{2}\right)\sin\frac{\Delta x}{2} = 0$$

Therefore, the function $y = \sin x$ is continuous.

In a similar way, one can prove that *any basic elementary function is continuous at every point at which it is defined.*

The following ***statements*** are true:

1. If the functions $f(x)$ and $g(x)$ are continuous at a point x_0, then their sum $\varphi(x) = f(x) + g(x)$ is also continuous at this point.

2. The product of two continuous functions is a continuous function.

3. The quotient of two continuous functions is a continuous function if the denominator at the point in question does not vanish (that is, if both $f(x)$ and $g(x)$ are continuous at x_0 and $g(x_0) \neq 0$, then $\varphi(x) = \dfrac{f(x)}{g(x)}$ is continuous at x_0).

4. If $u = \varphi(x)$ is continuous at $x = x_0$ and $f(u)$ is continuous at a point $u_0 = \varphi(x_0)$, then the complex function $y = f(\varphi(x))$ is continuous at a point x_0.

The proofs of these statements are simple and based on the properties of the limits.

Let us *prove*, for example, statement 2. Let $f(x)$ and $g(x)$ be continuous at the point x_0: $\lim\limits_{x \to x_0} f(x) = f(x_0)$, $\lim\limits_{x \to x_0} g(x) = g(x_0)$, and let $\varphi(x) = f(x) g(x)$. Since the limit of the product is equal to the product of the limits, then $\lim\limits_{x \to x_0} \varphi(x) = \lim\limits_{x \to x_0} f(x) \lim\limits_{x \to x_0} g(x) =$

$= f(x_0)\,g(x_0) = \varphi(x_0)$. So, $\lim\limits_{x \to x_0} \varphi(x) = \varphi(x_0)$ i.e. $\varphi(x)$ is continuous at a point x_0.

Theorem 7.1. Every elementary function is continuous at every point at which it is defined.

The proof follows from statements 1–4 formulated above[1].

If the function $f(x)$ is not continuous at a point x_0, then the point x_0 is called the **discontinuity point** of the function $f(x)$. There are **removable discontinuities** when there are finite limits $\lim\limits_{x \to x_0 -} f(x)$ and $\lim\limits_{x \to x_0 +} f(x)$ and **jump discontinuities** when at least one of these one-sided limits is infinite or does not exist. Among the points of discontinuity of the first kind, it should also be noted the **essential discontinuities**, when the limit of the function $f(x)$ as $x \to x_0$ exists, but either it is not equal to $f(x_0)$ or the function is not defined at $x = x_0$.

Example 7.2.

1. $f(x) = \operatorname{arctg} \dfrac{1}{x}$. Here $x_0 = 0$ is the removable discontinuity since $\lim\limits_{x \to 0-} \operatorname{arctg} \dfrac{1}{x} = -\dfrac{\pi}{2}$, $\lim\limits_{x \to 0+} \operatorname{arctg} \dfrac{1}{x} = \dfrac{\pi}{2}$.

[1] Given that the basic elementary functions are continuous.

2. $f(x)=\frac{1}{x}$. Here $x_0=0$ is the and jump discontinuity since $\lim\limits_{x\to 0+}\frac{1}{x}=+\infty$, $\lim\limits_{x\to 0-}\frac{1}{x}=-\infty$.

3. $f(x)=\begin{cases}\frac{\sin x}{x}, & x\neq 0\\ 0, & x=0.\end{cases}$ Here $x_0=0$ is the essential discontinuity as the limit $\lim\limits_{x\to 0}\frac{\sin x}{x}=1$ exists. This discontinuity can be eliminated by changing the value of the function at a point $x=0$ by setting $f(0)=1$.

If a function $y=f(x)$ is continuous at every point of a certain interval, then it is said this function to be **continuous on this interval**.

7.2. Properties of continuous functions on a segment

Theorem 7.2 **(the first Weierstrass theorem).** If the function *f(x)* is continuous on a segment [*a,b*], then it is bounded on this segment.

Theorem 7.3 **(the second Weierstrass theorem).** If the function *f(x)* is continuous on the segment [*a,b*], then on this segment it reaches the smallest value *m* and the largest value *M* (i.e., there exists a point c_1 on this segment at which $f(c_1)=m$ and a point c_2 at which $f(c_2)=M$).

Theorem 7.4 **(the first Bolzano-Cauchy theorem).** Let the function *f(x)* be continuous on a segment [*a,b*] and at its ends takes values of different signs[1*]. Then inside of [*a,b*] there exists a point *c* such that *f(c)*=0.

[1*] For instance, , .

Theorem 7.5 **(the second Bolzano-Cauchy theorem).** Let the function *f(x)* be continuous on [*a*,*b*] and let *m* be the smallest and *M* the largest values of *f(x)* on [*a*,*b*]. Then, for any *C* satisfying the condition $m < C < M$, there exists a point *c* from [*a*,*b*] such that $f(c) = C$.

It is not easy to prove these theorems, and we will not do this. However, all of them are special cases of the following **statement** (which intuitively seems obvious): *if a function f(x) is continuous on a segment [a,b], then the area of its change is a segment.*

7.3. Economic interpretation of continuity

Most of the functions used in the economics are continuous. Such, in particular, the previously mentioned functions of supply and demand, the utility function, the output function (see § 5.4). Among the functions used in the economics, there are discontinuous functions.

1. The tax rate (Fig. 7.1) is a function expressing the dependence of the tax rate *N* as a percentage of the annual income *q*. This function is discontinuous at the ends of the gaps, and these discontinuities are of the first kind.

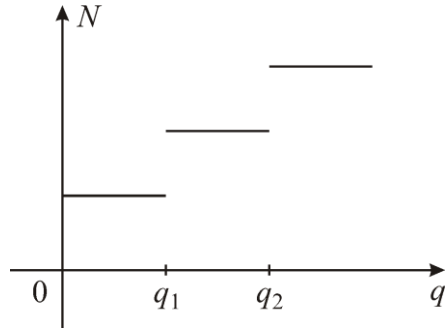


Fig. 7.1. *Tax rate*

However, the value of income tax *P* itself is a continuous function of annual income *q* (Fig. 7.2):

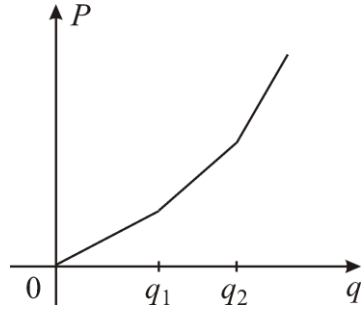


Fig. 7.2. *Income tax*

From the continuity of the function $P = P(q)$ it follows, in particular, that if the income of taxpayers does not differ significantly, then the difference in their income tax is also small.

2. As you know, there are two main categories of market relations: *supply* and *demand*. Both that and another depends on many factors, among which the main thing is the *price* of the goods. Let us denote the price of the goods *p* (price), the volume of *demand* as *D* (demand), the value of supply as *S* (supply). By their meaning, the functions $D = D(p)$ and $S = S(p)$ continuously depend on *p*. This means that with small price fluctuations, supply and demand change insignificantly. However, sometimes demand changes spasmodically. This usually happens for reasons of a psychological nature, in particular, when "breaking through" the round price. It happens that when the price of a certain product rises, demand decreases slightly for some time, but as soon as the price exceeds a certain amount (for example, 100 monetary units), demand drops sharply. In this case, the function $d = d(p)$ has a discontinuity at the indicated value of *p*.

In the analysis of functions[1] $D = D(p)$ and $S = S(p)$ (if we consider them on such a price segment, where they have no discontinuities), we can use the properties of continuous functions (see § 15.2). Consider the difference $D(p) - S(p)$.

For small *p*, obviously $D(p) - S(p) > 0$ (demand exceeds supply), and for large *p*, on the contrary, $D(p) - S(p) < 0$. Applying the first Bolzano-Cauchy theorem to the difference $D(p) - S(p)$, we conclude that there exists such a price p_0 for which $D(p_0) - S(p_0) = 0$, i.e. $D(p_0) = S(p_0)$. This price is called the **equilibrium** price (we mentioned it in § 13.4).

7.4. Comparison of the infinitesimals

Let simultaneously consider several infinitesimal quantities α, β, γ, …, which are functions of the same argument *x*, and these quantities are infinitesimals when *x* tends to some finite limit a or to infinity.

In many cases, it is of interest to compare these infinitely small with each other in the nature of their tendency to zero. It is a question of the comparative "speed" of their tendency to zero: which of the infinitely small tends to zero "faster" and which is "slower".

To compare two infinitesimals, we usually study their *ratio*. Moreover, considering the fraction $\frac{\alpha}{\beta}$ (or $\frac{\beta}{\alpha}$), it is assumed that the variable standing

[1] We consider these functions on a segment where the functions don't have discontinuities.

in the denominator does not vanish, at least for values of *x* sufficiently close to *a* (or for sufficiently large in absolute value as $x \to \infty$).

I. If the relation $\frac{\beta}{\alpha}$ has a finite limit other than zero, then the infinitely small α and β are called **infinitesimal of the same order**.

In this case, obviously, the relation $\frac{\alpha}{\beta}$ has a finite limit.

Example 7.3. Infinitesimals $\alpha = 3x$ and $\beta = \sin 2x$ are infinitesimals of the same order as $x \to 0$ since (see Example 6.2)

$$\lim_{x\to 0} \frac{\beta}{\alpha} = \lim_{x\to 0} \frac{\sin 2x}{3x} = \frac{2}{3}.$$

Example 7.4. Infinitesimals $\alpha = x$ and $\beta = \sqrt{1+x} - 1$ are also infinitesimals of the same order, since

$$\lim_{x\to 0} \frac{\beta}{\alpha} = \lim_{x\to 0} \frac{\sqrt{1+x}-1}{x} = \lim_{x\to 0} \frac{x}{\sqrt{x+1}+1} = \lim_{x\to 0} \frac{1}{\sqrt{1+x}+1} = \frac{1}{2}.$$

II. If the ratio $\frac{\beta}{\alpha}$ itself turns out to be infinitesimal, i.e. $\lim_{x\to a} \frac{\beta(x)}{\alpha(x)} = 0$ ($\lim_{x\to a} \frac{\alpha(x)}{\beta(x)} = \infty$), then they say that an infinitesimal β is an **infinitesimal of a higher order** with respect to an infinitesimal α (and an infinitesimal α is an **infinitesimal of a lower order** with respect to an infinitesimal β).

Example 7.5. Infinitesimal $\beta = 1 - \cos 2x$ is infinitesimal of a higher order with respect to $\alpha = x$. Indeed,

$$\lim_{x\to 0}\frac{1-\cos 2x}{x}=\lim_{x\to 0}\frac{2\sin^2 x}{x}=0.$$

Note that if β is an infinitesimal of a higher order with respect to an infinitesimal α, then this circumstance is written as follows:

$$\beta = o(\alpha).$$

In particular, Example 7.5 shows that

$$1-\cos 2x = o(x).$$

III. An infinitesimal β is called an **infinitesimal of k[th] order** with respect to an infinitesimal α if β and a^k are infinitesimal of the same order, i.e. if there is a finite limit of the ratio $\frac{\beta}{\alpha^k}$ other than zero.

Example 7.6. If $\alpha = x$ and $\beta = 1-\cos x$, then infinitesimal β is an infinitesimal of the second order with respect to infinitesimal α as $x\to 0$. Indeed,

$$\lim_{x\to 0}\frac{\beta}{\alpha^2}=\lim_{x\to 0}\frac{1-\cos x}{x^2}=\lim_{x\to 0}\frac{2\sin^2\frac{x}{2}}{x^2}=\frac{1}{2}.$$

Example 7.7. If $\alpha = x$ and $\beta=\sqrt{1+x^3}-1$, then infinitesimal β is an infinitesimal of the third order with respect to infinitesimal α as $x\to 0$. Make sure of this:

$$\lim_{x\to 0}\frac{\beta}{\alpha^3}=\lim_{x\to 0}\frac{\sqrt{1+x^3}-1}{x^3}=\lim_{x\to 0}\frac{x^3}{x^3\left(\sqrt{1+x^3}+1\right)}=\frac{1}{2}.$$

IV. Infinitesimal α and β are called **equivalent** if the limit of their ratio is unity:

$$\lim \frac{\beta}{\alpha} = 1.$$

If α and β are equivalent, then we write $\alpha \sim \beta$.

Theorem 7.6. Infinitesimals α and β are equivalent, if and only if their difference γ=β-α is infinitely small of higher order with respect to α and β.

Proof. 1. *Necessity*. Let $\gamma = o(\alpha)$, $\gamma = o(\beta)$. Then $\frac{\gamma}{\alpha} = \frac{\beta}{\alpha} - 1$, therefore $\lim \frac{\beta}{\alpha} = 1$ (since $\lim \frac{\gamma}{\alpha} = 1$).

2. *Sufficiency*. Let $\alpha \sim \beta$, i.e. $\lim \frac{\beta}{\alpha} = 1$. Then from the equality $\frac{\gamma}{\alpha} = \frac{\beta}{\alpha} - 1$ we obtain $\lim \frac{\gamma}{\alpha} = 1 - 1 = 0$.

Note that the limit of the ratio of infinitesimal may not exist at all. In this case, they say that the infinitesimals are incomparable. Consider a traditional example.

Example 7.8. Infinitesimals $\alpha = x$ and $\beta = x \sin \frac{1}{x}$ (as $x \to 0$) are incomparable. Indeed, the relation of these infinitesimals $\frac{\beta}{\alpha} = \sin \frac{1}{x}$ has no limit as $x \to 0$.

Questions

1) Which of the basic elementary functions are continuous?
2) How are discontinuities of a function classified?
3) Is the tax rate a continuous function of the amount of income?
4) Is income tax a continuous function of annual income?
5) Let $D = D(p)$ be a function expressing the dependence of demand *d* on price *p*. What kind of discontinuity does the function $D = D(p)$ have in case of a spasmodic change in demand?
6) What is the comparison of infinitesimal based on?
7) Are equivalent infinitesimals of the same order?
8) Let α and β be two infinitesimals of different orders. Which of them tends to zero faster - the one of a higher order, or the one of a lower order?

Are any two infinitesimals comparable with each other?

DIFFERENTIAL CALCULUS

Chapter 8. Derivative functions. Differential

8.1. Derivative

Let the function $y = f(x)$ be defined on some interval *X*. We give the argument $x_0 \in X$ an arbitrary increment Δx such that a point $x_0 + \Delta x$ is also in *X*. Then the function $f(x)$ will receive the corresponding increment $\Delta y = f(x_0 + \Delta x) - f(x_0)$.

Definition. **The derivative of the function** $y = f(x)$ **at a point** is called a limit of the ratio of the function increment to the argument increment at the point x_0 as $\Delta x \to 0$:

$$\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}$$

(if this limit exists).

The derivative is denoted by $f'(x_0)$, or $y'(x_0)$, or y', or $\frac{dy}{dx}$. (Economists also use the notation $Mf(x)$ for the derivative $f'(x)$ and the term «marginal value of the function f at the point x»).

By definition:

$$f'(x_0) = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}. \quad (8.1)$$

If the function $f(x)$ has a derivative at each point of the set *X*, then the derivative $f'(x)$ is also a function of the argument of *x*, defined on *X*.

Geometric meaning of the derivative

To clarify the geometric meaning of the derivative, it is necessary to formulate a definition of a tangent to the graph of the function at a given point.

Definition. **The tangent** to the graph of the function $y = f(x)$ at the point M_0 is the limit position of the secant M_0M as the point *M* tends to the point M_0 along the curve $y = f(x)$.

Let the point $M_0(x_0, y_0)$ be fixed on the curve $y = f(x)$, where $y_0 = f(x_0)$ (fig. 8.1). We give the argument the increment Δx, i.e. move from $x = x_0$ to $x_0 + \Delta x$.

We get $M(x_0 + \Delta x, y_0 + \Delta y)$ on the curve. From the triangle M_0MA we have:

$$tg\ \varphi = \frac{MA}{M_0A} = \frac{\Delta y}{\Delta x} = \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}.$$

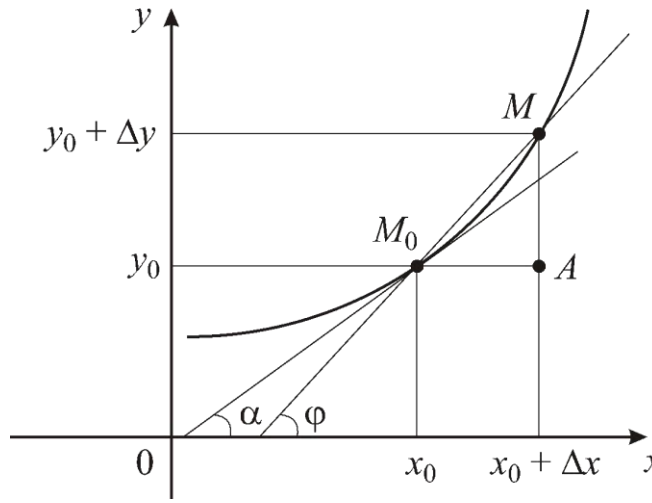


Fig. 8.1. *The geometric meaning of the derivative*

Let $\Delta x \to 0$. Then the point *M* will move along the curve and coincide with the point M_0 in the limit. Here

$$tg\ \alpha = \lim_{\Delta x \to 0} \operatorname{tg} \varphi = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x}.$$

If the derivative $f'(x)$ at the point x_0 exists, then, according to the definition of the derivative, we obtain:

$$\operatorname{tg} \alpha = f'(x_0).$$

So, the derivative $f'(x_0)$ is equal to *the angular coefficient*[1] *of the tangent* to the graph of the function $y = f(x)$ at the point $M_0(x_0, f(x_0))$.

Physical meaning of the derivative

Suppose that the function $s = f(t)$ describes the law of motion of a point at a straight line as the dependence of the distance *s* on time *t*. By the time t_0 the distance is $s_0 = f(t_0)$, and by the time $t_0 + \Delta t$ distance $s = f(t_0 + \Delta t)$. Then, over a period of time Δt the distance $\Delta s = s - s_0$ is passed and the average speed over time Δt is the ratio $\frac{\Delta s}{\Delta t}$. The limit of this ratio as $\Delta t \to 0$

$$\lim_{\Delta t \to 0} \frac{\Delta s}{\Delta t} = f'(t),$$

defines *the instantaneous speed* of a point at the time t_0 as a derivative of the distance with respect to time.

8.2. Application of a derivative in economy

1. Labor productivity. Let the function $q = q(t)$ be a value of *q* products produced over time *t* and let it be required to find labor

[1] The angular coefficient is the tangent of the angle of inclination to the positive direction of the axis *Ox*.

productivity at the moment t_0. Consider the time period from t_0 to $t_0 + \Delta t$. During this period, the volume of production $\Delta q = q(t_0 + \Delta t) - q(t_0)$. *The average productivity for this period is* $\frac{\Delta q}{\Delta t}$. Then, labor productivity at the moment t_0 can be defined as the marginal value of average productivity as $\Delta t \to 0$:

$$q'(t) = \lim_{\Delta t \to 0} \frac{\Delta q}{\Delta t}.$$

As we can see, mathematically the problem at the moment t_0 does not differ from the problem of finding the instantaneous speed of movement (see § 8.1).

Another formulation of the problem is also possible. Let the quantity of products *q* be dependent only on the applied labor *x* (for a company it's just the number of employees): $q = q(x)$.

The average productivity is used to evaluate production efficiency. It denoted by $\frac{q}{x}$.

However, the question arises: how will the volume of production change when the number of personnel changes? The answer to this question can be obtained by introducing the concept of **marginal productivity**. The marginal productivity is a derivative of products *q* by the amount of labor *x*:

$$q' = \frac{dq}{dx}.$$

The marginal productivity with this formulation of the problem is approximately equal to *the change in the volume of products with the change in the number of personnel* **per unit**.

If the number of employees *a* is large, then the increment $\Delta a = 1$ can be considered small enough to take advantage of the approximate equality $q'(a) \approx \frac{\Delta q}{\Delta a} = \frac{q(a+1) - q(a)}{1} = q(a+1) - q(a)$, which gives $q(a+1) = q(a) + q'(a)$. In this case, $q'(a)$ is an additional product produced by new employees per unit of time.

Let *v* be the product price and *p* is employee salary per unit of time. If $vq'(a) > p$, then we need to hire another employee as he brings the company more than it pays him. This rule is called **the "golden" rule of Economics**.

2. Production cost. Consider the dependence of cost *C* manufactured products on its volume *q*: $C = C(q)$.

Marginal cost is the value

$$MC \approx \lim_{\Delta q \to 0} \frac{\Delta C}{\Delta q} = C'(q).$$

Along with the cost price in microeconomics, an important role is played by another marginal indicator - *elasticity*. We consider it later - in the study of so-called logarithmic derivative (see § 20.2).

8.3. Differentiability of a function. Communication between differentiability and continuity

Definition. A function $y = f(x)$ is called a **differentiable function at the point** x_0, if its increment at this point can be represented as

$$\Delta y = A\Delta x + \alpha\Delta x, \quad (8.2)$$

here *A* is an arbitrary number (independent of Δx) and $\alpha = \alpha(\Delta x)$ is the infinitesimal as $\Delta x \to 0$.

***Theorem* 8.1.** The function $y = f(x)$ is differentiable at the point x_0 if and only if it has a finite derivative at that point.

Proof. 1. *Necessity.* Let the function $f(x)$ be differentiable at the point x_0, i.e. its increment can be represented as (8.2). Dividing this equality by $\Delta x \neq 0$, we obtain:

$$\frac{\Delta y}{\Delta x} = A + \alpha.$$

Passing to the limit as $\Delta x \to 0$ ($\alpha \to 0$ as $\Delta x \to 0$), we obtain

$$\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = f'(x_0) = \lim_{\Delta x \to 0}(A + \alpha) = A,$$

i.e. the derivative of the function $f(x)$ exists at the point x_0 and equals to *A*.

2. *Sufficience*. Now let the function $f(x)$ have the derivative at the point x_0, i.e. there is a limit

$$\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = A .$$

Then, in accordance with Theorem 6.1:

$$\frac{\Delta y}{\Delta x} = A + \alpha ,$$

here α is the infinitesimal as $\Delta x \to 0$. Hence

$$\Delta y = A\Delta x + \alpha \Delta x .$$

Therefore, $f(x)$ is differentiable.

Theorem 8.1 allows us to call a function of one argument **differentiable** if it has a derivative. The operation of finding the derivative is called **differentiation**.

Continuity of a differentiable function

***Theorem* 8.2.** If a function is differentiable at the point x_0 then the function is continuous at this point.

Proof. Since $f(x)$ is differentiable at the point x_0, then its increment at this point has the form (8.2). Passing to the limit in this equality as $\Delta x \to 0$, we obtain:

$$\lim_{\Delta x \to 0} \Delta y = \lim_{\Delta x \to 0} \left(A\Delta x + \alpha \Delta x \right) = 0 ,$$

i.e. $\lim_{\Delta x \to 0} \Delta y = 0$, which means that the function is continuous.

8.4. Calculating the derivative

The scheme of calculating the derivative of the function $f(x)$:

1. Give x an increment Δx and find the corresponding value of the function $f(x+\Delta x)$.
2. Find the increment of the function $\Delta y = f(x+\Delta x) - f(x)$.
3. Compose the ratio $\frac{\Delta y}{\Delta x}$.
4. Calculate the limit of this ratio as $\Delta x \to 0$:

$$y' = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x}.$$

Example 8.1. If $y = c = \text{const}$, then $y' = 0$. Indeed, $\Delta y = 0$ for any Δx, so that $y' = 0$. So, if $c = \text{const}$, then

$$c' = 0.$$

Example 8.2. Find the derivative of $y = x^2$.

Solution. 1. Give x an increment Δx and find $f(x+\Delta x) = (x+\Delta x)^2$.

2. Find the increment of the function:

$$\Delta y = (x+\Delta x)^2 - x^2 = x^2 + 2x\Delta x + \Delta x^2 - x^2 = 2x\Delta x + \Delta x^2.$$

3. Compose the ratio $\frac{\Delta y}{\Delta x} = 2x + \Delta x$.

4. Calculate the limit:

$$y' = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} (2x + \Delta x) = 2x$$.

So, $(x^2)' = 2x$.

Example 8.3. Find the derivative of $y = \sin x$.

1. $f(x + \Delta x) = \sin(x + \Delta x)$.

2. $$\Delta y = \sin(x + \Delta x) - \sin x = 2\cos\frac{(x + \Delta x) + x}{2}\sin\frac{(x + \Delta x) - x}{2} = 2\sin\frac{\Delta x}{2}\cos\left(x + \frac{\Delta x}{2}\right)$$.

3. $$\frac{\Delta y}{\Delta x} = \frac{2\sin\frac{\Delta x}{2}\cos\left(x + \frac{\Delta x}{2}\right)}{\Delta x} = \frac{2\sin\frac{\Delta x}{2}}{\Delta x}\cos\left(x + \frac{\Delta x}{2}\right) = \frac{\sin\frac{\Delta x}{2}}{\frac{\Delta x}{2}} \cdot \cos\left(x + \frac{\Delta x}{2}\right)$$.

4. $$y' = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{\sin\frac{\Delta x}{2}}{\frac{\Delta x}{2}} \lim_{\Delta x \to 0} \cos\left(x + \frac{\Delta x}{2}\right) = 1 \cdot \cos x = \cos x$$

(taking into account the first remarkable limit (see § 14.5) and the continuity of the function $\cos x$).

Then, $(\sin x)' = \cos x$.

Rules of differentiation

Assume that $u = u(x)$ and $v = v(x)$ are differentiable functions, $c = \text{const}$.

I. $(cu)' = cu'$. **III.** $(uv)' = u'v + uv'$.

II. $(u \pm v)' = u' \pm v'$. **IV.** $\left(\frac{u}{v}\right)' = \frac{u'v - uv'}{v^2}$.

Let us formulate these rules in more detail and prove them.

I. If $y = cu(x)$, $c = \text{const}$, then $y' = cu'(x)$,

i.e. *the constant multiplier can be taken out of the sign of the derivative*.

Proof. 1. Give *x* an increment Δx. Then

$y + \Delta y = cu(x + \Delta x)$.

2. Find the increment of the function:

$\Delta y = cu(x + \Delta x) - cu(x) = c\left[u(x + \Delta x) - u(x)\right]$.

3. Compose the ratio $\frac{\Delta y}{\Delta x}$:

$\frac{\Delta y}{\Delta x} = c\frac{u(x + \Delta x) - u(x)}{\Delta x}$.

4. Calculate the limit of this ratio as $\Delta x \to 0$, i.e. find y':

$y' = \lim\limits_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = c \lim\limits_{\Delta x \to 0} \frac{u(x + \Delta x) - u(x)}{\Delta x} = cu'(x)$.

(We took advantage of the fact that the constant multiplier can be carried beyond the limit sign (see § 6.4).)

II. If $y' = u(x) \pm v(x)$, то $y' = u'(x) \pm v'(x)$,

i.e. *the derivative of the algebraic sum of differentiable functions is equal to the algebraic sum of the derivatives of these functions*.

Proof. 1. Give *x* an increment Δx. Then functions $u = u(x)$, $v = v(x)$ take corresponding values $u + \Delta u = u(x + \Delta x)$, $v + \Delta v = v(x + \Delta x)$, then $y + \Delta y = (u + \Delta u) \pm (v + \Delta v)$.

2. Find the increment of the function *y*:

$$\Delta y = (u + \Delta u) \pm (v + \Delta v) - (u \pm v) = \Delta u \pm \Delta v.$$

3. Compose the ratio $\frac{\Delta y}{\Delta x}$:

$$\frac{\Delta y}{\Delta x} = \frac{\Delta u}{\Delta x} \pm \frac{\Delta v}{\Delta x}.$$

4. Calculate the limit of this ratio as $\Delta x \to 0$

$$y' = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{\Delta u}{\Delta x} \pm \lim_{\Delta x \to 0} \frac{\Delta v}{\Delta x} = u' \pm v'.$$

So, $(u \pm v)' = u' \pm v'$.

(We took advantage of the fact that the limit of the algebraic sum is equal to the algebraic sum of the limits (see § 6.4).)

This statement can be extended to any number of terms. In particular, $(u + v + w)' = u' + v' + w'$.

III. If $u = u(x)$, $v = v(x)$, then $(uv)' = u'v + uv'$,

i.e. *the derivative of product of two differentiable functions is equal to the product of the first derivative of these functions and the second function plus the product of the first function and the derivative of the second function*.

Proof. 1. Give x an increment Δx. Then functions u and v take corresponding values $u+\Delta u$, $v+\Delta v$, and their composition $y=uv$ takes value $(u+\Delta u)(v+\Delta v)$.

2. Find the increment of the function y:

$$\Delta y=(u+\Delta u)(v+\Delta v)-uv=uv+\Delta uv+u\Delta v+\Delta u\Delta v-uv=$$

$$\Delta uv+u\Delta v+\Delta u\Delta v.$$

3. Compose the ratio $\dfrac{\Delta y}{\Delta x}$:

$$\frac{\Delta y}{\Delta x}=\frac{\Delta u}{\Delta x}v+u\frac{\Delta v}{\Delta x}+\Delta u\frac{\Delta v}{\Delta x}.$$

4. Calculate the limit of this ratio as $\Delta x\to 0$:

$$y'=(uv)'=\lim_{\Delta x\to 0}\frac{\Delta y}{\Delta x}=\lim_{\Delta x\to 0}\frac{\Delta u}{\Delta x}v+u\lim_{\Delta x\to 0}\frac{\Delta v}{\Delta x}+\lim_{\Delta x\to 0}\Delta u\lim_{\Delta x\to 0}\frac{\Delta v}{\Delta x}=$$

$$=u'v+uv'+0\cdot v'=u'v+uv'.$$

(Here $\lim\limits_{\Delta x\to 0}\Delta u=0$, since u is continuous function.)

Then, $(uv)'=u'v+uv'$.

According to this statement it is easy to obtain the rule of differentiation of the product of three and, in general, any finite number of functions.

Let $y=uvw$ be the product of three functions. Let us present this composition in the form of $u(vw)$:

$$y'=u'(vw)+u(vw)'=u'vw+u(v'w+vw')=u'vw+uv'w+uvw'.$$

It is easy to understand that for the case of n terms, the same method can be used to obtain a similar formula for the derivative of the product:

$$(u_1u_2u_3\cdots u_n)' = u_1'u_2u_3\cdots u_n + u_1u_2'u_3\cdots u_n + u_1u_2u_3'\cdots u_n + \ldots + u_1u_2u_3\cdots u_n'$$

.

IV. If $y = \frac{u}{v}$, then $y' = \frac{u'v - uv'}{v^2}$,

i.e. *the derivative of a fraction (the ratio of two functions) is equal to a fraction whose denominator is the square of the denominator of the given fraction, the numerator is the difference between the product of the denominator on the derivative of the numerator and the product of the numerator on the derivative of the denominator*.

Proof. 1. Give x an increment Δx. Then functions u and v take corresponding values $u + \Delta u$, $v + \Delta v$ and their ratio $y = \frac{u}{v}$ takes value $y + \Delta y = \frac{u + \Delta u}{v + \Delta v}$.

2. $\Delta y = \frac{u + \Delta u}{v + \Delta v} - \frac{u}{v} = \frac{v\Delta u - u\Delta v}{v(v + \Delta v)}$.

3. $\frac{\Delta y}{\Delta x} = \frac{\frac{v\Delta u - u\Delta v}{\Delta x}}{v(v + \Delta v)} = \frac{\frac{\Delta u}{\Delta x}v - u\frac{\Delta v}{\Delta x}}{v(v + \Delta v)}$.

4. $y' = \lim\limits_{\Delta x \to 0}\frac{\Delta y}{\Delta x} = \lim\limits_{\Delta x \to 0}\frac{\frac{\Delta u}{\Delta x} - u\lim\limits_{\Delta x \to 0}\frac{\Delta v}{\Delta x}}{v\lim\limits_{\Delta x \to 0}(v + \Delta v)}$.

Hence, since $\lim\limits_{\Delta x \to 0}\Delta v = 0$ (since the function v is continuous), we obtain

$$y' = \frac{u'v - uv'}{v^2}, \text{ or } \left(\frac{u}{v}\right)' = \frac{u'v - uv'}{v^2}.$$

Derivative of a complex function

Now we formulate and prove the rule **V**.

V. Let the function $u = \varphi(x)$ have the derivative $u'_x = \varphi'(x_0)$ at the point x_0, and let the function $y = f(u)$ have the derivative $y'_u = f'(u_0)$ at the point $u_0 = \varphi(x_0)$. Then the complex function $y = f(\varphi(x))$ has a derivative at the point x_0 and the following formula holds:

$$y'_x = y'_u u'_x.$$

Proof. Give *x* an increment Δx. Let Δu be the corresponding increment of $u = \varphi(x)$, here Δy is the increment of $y = f(u)$ caused by the increment Δu. Replacing *x* with *u* we rewrite (8.2) in the form:

$$\Delta y = y'_u \Delta u + \alpha \Delta u$$

(here α depends on Δu; $\alpha \to 0$ as $\Delta u \to 0$). Dividing this equality by Δx, we obtain:

$$\frac{\Delta y}{\Delta x} = y'_u \frac{\Delta u}{\Delta x} + \alpha \frac{\Delta u}{\Delta x}.$$

If $\Delta x \to 0$, then $\Delta u \to 0$ (since *u* is continuous), therefore $\alpha \to 0$. Hence, there is a limit

$$\lim_{\Delta x\to 0}\frac{\Delta y}{\Delta x}=y'_u\lim_{\Delta x\to 0}\frac{\Delta u}{\Delta x}=y'_u u'_x,$$

i.e.

$$y'_x=y'_u u'_x.$$

Example 8.4. Find the derivative of the function $y=\sin^2 x$.

Solution. $y=u^2$, here $u=\sin x$. Using the rule **V** and taking into account examples 8.2 and 8.3 we obtain:

$$y'=2u\cdot u'_x=2\sin x(\sin x)'=2\sin x\cos x=\sin 2x.$$

Derivative of the inverse function

Let $y=f(x)$ be differentiable and strictly monotone function on some interval *X* and the function $x=\varphi(y)$ is the inverse function. We can show that $\varphi(y)$ is the continuous function on the corresponding interval *Y*.

Theorem 8.3. Let a function $f(x)$ be strictly monotone and continuous on *X* and have a finite and non-zero derivative $f'(x_0)$ at the point x_0. Then there also exists a derivative of the inverse function $x=\varphi(y)$ at the corresponding point $y_0=f(x_0)$ and

$$\varphi'(y_0)=\frac{1}{f'(x_0)}. \qquad (8.3)$$

Proof. Give $y = y_0$ an arbitrary increment Δy. Then the function $x = \varphi(y)$ takes the corresponding increment Δx. Note, if $\Delta y \neq 0$ then $\Delta x \neq 0$ due to the uniqueness of the function $y = f(x)$. We have

$$\frac{\Delta x}{\Delta y} = \frac{1}{\frac{\Delta y}{\Delta x}}.$$

Now let $\Delta y \to 0$. Then $\Delta x \to 0$ since $\varphi(y)$ is a continuous function. But the denominator of the right side of the written equality tends to the limit $f'(x_0) \neq 0$. Therefore, there is a limit for the left side of the equality. This limit is equal to $\frac{1}{f'(x_0)}$ and it is a derivative $\varphi'(y)$. So,

$$x'_y = \frac{1}{y'_x}. \tag{8.4}$$

The last equality can be rewritten in the following form:

$$y'_x = \frac{1}{x'_y}. \tag{8.5}$$

That completes the proof

8.5. Derivatives of the basic elementary functions

Derivative of logarithmic function

Let us first derive the formula for the derivative of $y = \ln x$.

$$\Delta y = \ln\left(x + \Delta x\right) - \ln x = \ln \frac{x + \Delta x}{x} = \ln\left(1 + \frac{\Delta x}{x}\right)$$

$$\frac{\Delta y}{\Delta x} = \frac{1}{\Delta x} \ln\left(1 + \frac{\Delta x}{x}\right)$$

$$y' = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{1}{\Delta x} \ln\left(1 + \frac{\Delta x}{x}\right).$$

Denote $\frac{\Delta x}{x} = t$; hence $\Delta x = tx$. Obviously, $\Delta x \to 0$ if and only if $t \to 0$. We obtain

$$y' = \lim_{\Delta t \to 0} \frac{1}{tx} \ln\left(1 + t\right) = \frac{1}{x} \lim_{\Delta t \to 0} \ln\left(1 + t\right)^{1/t}.$$

Hence, taking into account the second remarkable limit (see § 6.5) and the continuity of the logarithmic function, we obtain:

$$y' = \frac{1}{x} \ln \lim_{\Delta t \to 0} \left(1 + t\right)^{1/t} = \frac{1}{x} \ln \mathrm{e} = \frac{1}{x},$$

i.e.

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

Now let $y = \log_a x$. Obviously $\log_a x = \frac{\ln x}{\ln a}$. We obtain

$$y' = (\log_a x)' = \left(\frac{\ln x}{\ln a}\right)' = \frac{1}{\ln a}(\ln x)' = \frac{1}{x \ln a},$$

i.e.

$$(\log_a x)' = \frac{1}{x \ln a}.$$

Derivative of exponential function

Let $y = a^x$. Take a logarithm of this function

$$\ln y = x \ln a. \qquad (*)$$

According to the rule of differentiation of a complex function

$$(\ln y)' = \frac{1}{y} y' = \frac{y'}{y}.$$

Differentiating the equality (*), we obtain

$$\frac{y'}{y} = \ln a, \quad y' = y \ln a,$$

or

$$y' = a^x \ln a,$$

i.e.

$$(a^x)' = a^x \ln a.$$

As $a = \mathrm{e}$, obviously,

$$(\mathrm{e}^x) = \mathrm{e}^x.$$

The derivative of the exponential function

Let $y = x^n$, where *n* is any real number. Take a logarithm of this function:

$\ln y = n \ln x$.

Differentiating both parts of equality, we get

$\frac{y'}{y} = n\frac{1}{x}$, $y' = \frac{ny}{x} = \frac{nx^n}{x} = nx^{n-1}$,

i.e.

$(x^n)' = nx^{n-1}$.

Derivatives of trigonometric functions

Now we derive formulas for trigonometric functions:

1. $\boldsymbol{y = \sin x}$. We have already found a derivative of this function (example 8.3):

$(\sin x)' = \cos x$

2. $\boldsymbol{y = \cos x}$.

$$\Delta y = \cos(x + \Delta x) - \cos x = -2\sin\frac{x + \Delta x + x}{2}\sin\frac{x + \Delta x - x}{2} =$$

$$= -2\sin\frac{\Delta x}{2}\sin\left(x + \frac{\Delta x}{2}\right),$$

$$\frac{\Delta y}{\Delta x} = -\frac{\sin\frac{\Delta x}{2}}{\frac{\Delta x}{2}}\sin\left(x + \frac{\Delta x}{2}\right)$$

By virtue of continuity of $\sin x$ and passing to the limit as $\Delta x \to 0$ we obtain:

$y' = -\sin x$, i.e. $(\cos x)' = -\sin x$.

3. $\boldsymbol{y = \operatorname{tg} x}$.

$$y' = \left(\frac{\sin x}{\cos x}\right)' = \frac{(\sin x)'\cos x - \sin x(\cos x)'}{\cos^2 x} = \frac{\cos^2 x + \sin^2 x}{\cos^2 x} = \frac{1}{\cos^2 x},$$

i.e.

$$(\operatorname{tg} x)' = \frac{1}{\cos^2 x}.$$

4. $y = \mathbf{ctg}\, x$. Similarly, we get

$$(\operatorname{ctg} x)' = -\frac{1}{\sin^2 x}.$$

Derivatives of inverse trigonometric functions

Finally, we derive formulas for derivatives of inverse trigonometric functions:

1. $y = \mathbf{arcsin} x$. This function is the inverse function for $x = \sin y$. By the inverse function derivative Theorem (see § 8.4)

$$y'_x = \frac{1}{x'_y} = \frac{1}{\cos y} = \frac{1}{\sqrt{1-\sin^2 y}}$$

(the root is taken with a plus sign, since $\cos y > 0$ as $-\frac{\pi}{2} < y < \frac{\pi}{2}$). Since $\sin y = x$, then finally we obtain

$$(\arcsin x)' = \frac{1}{\sqrt{1-x^2}}.$$

2. $y = \mathbf{arccos} x$:

$$(\arccos x)' = -\frac{1}{\sqrt{1-x^2}}.$$

The calculation is similar to the previous one.

3. $y = \mathbf{arctg}\ x$. This function is the inverse function for $x = \operatorname{tg} y$. Since $x'_y = \dfrac{1}{\cos^2 y}$, then

$$y'_x = \frac{1}{x'_y} = \cos^2 y = \frac{\cos^2 y}{\cos^2 y + \sin^2 y} = \frac{1}{1 + \operatorname{tg}^2 y} = \frac{1}{1 + x^2},$$

i.e.

$$(\operatorname{arctg} x)' = \frac{1}{1 + x^2}.$$

4. $y = \mathbf{arcctg}\ x$.

$$(\operatorname{arcctg} x)' = -\frac{1}{1 + x^2}.$$

The calculation is similar to the previous one.

We have derived formulas for derivatives of all basic elementary functions. Let us now tabulate them and recall once again the rules of differentiation.

Table of derivatives

1. $c' = 0$.

2. $(x^n)' = nx^{n-1}$ (n – is any real number).

3. $(\log_a x)' = \dfrac{1}{x \ln a}$; $(\ln x)' = \dfrac{1}{x}$.

4. $(a^x)' = a^x \ln a$; $(\mathrm{e}^x)' = \mathrm{e}^x$.

5. $(\sin x)' = \cos x$.

6. $(\cos x)' = -\sin x$.

9. $(\arcsin x)' = \dfrac{1}{\sqrt{1 - x^2}}$.

10. $(\arccos x)' = -\dfrac{1}{\sqrt{1 - x^2}}$.

7. $(\operatorname{tg} x)' = \dfrac{1}{\cos^2 x}$.

8. $(\operatorname{ctg} x)' = -\dfrac{1}{\sin^2 x}$.

11. $(\operatorname{arctg} x)' = \dfrac{1}{1+x^2}$.

12. $(\operatorname{arcctg} x)' = -\dfrac{1}{1+x^2}$.

Rules of differentiation

I. $(cu)' = cu'$.

II. $(u \pm v)' = u' \pm v'$.

III. $(uv)' = u'v + uv'$.

IV. $\left(\dfrac{u}{v}\right)' = \dfrac{u'v - uv'}{v^2}$.

V. If $y = f(u),\ u = \varphi(x)$, then $y'_x = y'_u u'_x$.

Formulas 1-12 and rules I-V form the basis for practical differentiation.

8.6. Differential

The function $y = f(x)$ is called *differentiable at the point* x_0, if its increment Δy can be presented in a form (8.2):

$$\Delta y = A\Delta x + \alpha\Delta x,$$

where $\alpha \to 0$ as $\Delta x \to 0$.

The quantity $A\Delta x$ is the main term of the decomposition Δy as $A \neq 0$.

Definition. **The differential** *dy* **of the function** $y = f(x)$ at the point x_0 is called a main linear part of the increment of the function with respect to Δx at that point:

$$dy = A\Delta x.$$

For $A=0$ the differential is also determined by the formula (*), i.e. in this case, $dy=0$.

It follows from Theorem 8.1 that $A=f'(x_0)$, then

$$dy=f'(x_0)\Delta x. \tag{8.6}$$

The differential dx of the independent variable *x* is the increment Δx of this variable, and we can write equality (8.6) in the form:

$$dy=f'(x_0)\,dx, \tag{8.6′}$$

which gives $f'(x_0)=\dfrac{dy}{dx}$. Now we see that $\dfrac{dy}{dx}$ is not just a symbolic designation of the derivative but the ratio of the differential of a function dy to the differential of its argument dx. Due to (8.6), formula (8.2) can be rewritten in the form

$$\Delta y=f'(x_0)\Delta x+o(\Delta x)$$

or

$$\Delta f(x_0)=f'(x_0)\Delta x+o(\Delta x). \tag{8.2′}$$

Geometrical meaning of the differential

Let the point *M* on the graph of the function $y=f(x)$ correspond to the value of the argument $x=x_0$, let the point *N* correspond to the value of the argument $x=x_0+\Delta x$ (fig. 8.2). Then $MA=\Delta x$, $AN=\Delta y$. Draw a tangent to the curve $y=f(x)$ at the point *M*. Let α be the angle between this tangent and the axis *Ox*. We know that $\operatorname{tg}\alpha=f'(x_0)$.

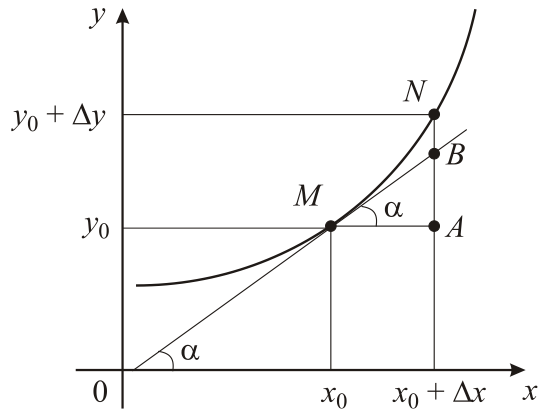


Fig. 8.2. *The geometric meaning of the differential*

Consider a right triangle MAB. Obviously, $MA = \Delta x$, $AB = MA \cdot \operatorname{tg}\alpha = \Delta x \operatorname{tg}\alpha = f'(x_0)\Delta x = dy$.

Hence, while Δy is the increment of the ordinate of the curve, dy is the corresponding *increment of the ordinate of the tangent*.

Application of differential in approximate calculations.

Approximate differential calculations are based on the approximate replacement of the function increment by its differential. Since the differential is the main part of the increment of the function, then

$\Delta y \approx dy$,

or

$\Delta f(x_0) = f(x_0 + \Delta x) - f(x_0) \approx f'(x_0)\Delta x$.

Hence

$f(x_0 + \Delta x) \approx f(x_0) + f'(x_0)\Delta x$.

Example 8.5. Calculate approximately $\sqrt[3]{8{,}24}$.

Solution. Let $f(x)=\sqrt[3]{x}$, $x_0=8$, $f(x_0)=2$, $\Delta x=0{,}24$. Due to $\Delta y \approx dy = f'(x_0)\Delta x$:

$$f'(x)=(x^{1/3})'=\frac{1}{3}x^{-2/3}=\frac{1}{3\sqrt[3]{x^2}}.$$

$$f'(x_0)=\frac{1}{3\sqrt[3]{8^2}}=\frac{1}{12},\quad \Delta y \approx dy=\frac{1}{12}\cdot 0{,}24=0{,}02.$$ Hence

$$f(x_0+\Delta x)=\sqrt[3]{8{,}24}\approx 2+0{,}02=2{,}02.$$

The problem of finding the differential of a function is obviously reduced to finding derivative and multiplying it by the differential of the argument. Therefore, the majority of theorems and formulas related to the derivatives holds for the differentials. In particular:

I. $d(cu)=cdu$ ($c=\text{const}$).

II. $d(u\pm v)=du\pm dv$.

III. $d(uv)=vdu+udv$.

IV. $d\left(\frac{u}{v}\right)=\frac{vdu-udv}{v^2}$.

Let us deduce, for example, the differential of a fraction:

$$\left(\frac{u}{v}\right)'=\frac{u'v-uv'}{v^2},\ d\left(\frac{u}{v}\right)=\left(\frac{u}{v}\right)'dx=\frac{u'vdx-uv'dx}{v^2},$$

Since $u'dx=du,\ \ v'dx=dv$, then

$$d\left(\frac{u}{v}\right)=\frac{vdu-udv}{v^2}.$$

Let us find the expression for the differential of a complex function. Let $y=f(u)$, $u=\varphi(x)$ or $y=f(\varphi(x))$. If $y=f(u)$ and $u=\varphi(x)$ are differentiable functions of u and x respectively, then $y'=f'(u)u'$.

The differential of the function

$$dy=f'(x)dx=f'(u)u'dx=f'(u)du,$$

since $u'dx = du$. Thus

$dy = f'(x)dx$ and $dy = f'(u)du$,

i.e. the differential does not depend on the function argument - an independent variable or a function of another argument. This property of the differential is called **the invariance of the differential form**.

From the invariance of the differential form it follows that we can apply formulas I -IV when u, v are functions of an independent variable and when they are complex functions.

Higher order derivatives and differentials

If a function $f(x)$ is defined on *X* and has a derivative $f'(x)$ at all points of *X*, then this derivative itself is a function of the argument *x*: $f'(x) = g(x)$. Derivative of the first derivative function $f(x)$, i.e. $(f'(x))'$, is called the second derivative, or second-order derivative, and denoted by $f''(x)$, or y''. So,

$f''(x) = \left(f'(x)\right)'$, or $y'' = \left(y'\right)'$.

The third order derivative is defined similarly:

$f'''(x) = (f''(x))'$, or $y''' = (y'')'$ and so on.

N-th derivative is denoted by $f^{(n)}(x)$ (or $y^{(n)}$) and it is defined in accordance with the described scheme:

$f^{(n)}(x) = (f^{(n-1)}(x))'$,

here $f^{(n-1)}(x)$ is the derivative of order $(n-1)$.

***Examples* 8.6:**

1) $y = e^{kx}$, $y' = ke^{kx}$, $y'' = k^2 e^{kx}$, …, $y^{(n)} = k^n e^{kx}$;

2) $y = \sin x$, $y' = \cos x$, $y'' = -\sin x$, $y''' = -\cos x$, $y^{(4)} = \sin x$

It is easy to show that $(\sin x)^{(n)} = \sin\left(x + \frac{\pi}{2}n\right)$.

A differential of the second order (or second differential) of the function $y = f(x)$ is called the differential of the differential of this function, i.e. $d(dy)$, and denoted by d^2y:

$$d^2y = d(dy). \tag{8.7}$$

Obviously, $d^2y = d(dy) = d(f'(x)dx) = (f'(x)dx)'dx = f''(x)dx^2$, where $dx^2 = (dx)^2$. We consider dx to be a constant since $dx = \Delta x$ is independent of *x*. Hence,

$$d^2y = f''(x)dx^2. \tag{8.8}$$

The third differential is defined similarly $d^3y = d(d^2y)$; finally, **n-th differential** is the differential from the differential of the order $(n-1)$:

$$d^ny = d(d^{n-1}y). \tag{8.9}$$

We find the expression for d^ny in the same way as it was done above for d^2y:

$$d^ny = f^{(n)}(x)dx^n. \tag{8.10}$$

Hence

$$f^{(n)}(x) = \frac{d^ny}{dx^n}.$$

It should be noted that second and higher order differentials *do not have the form invariance property*, in contrast to the first order differential. Let us show it.

Let $y = f(u)$, $u = \varphi(x)$, or $y = f(\varphi(x))$, then

$d^2 y = d^2 f(\varphi(x)) = d(df(\varphi(x)))$.

By condition, $df(\varphi(x)) = f'(u)du$, $u = \varphi(x)$. Hence

$$d(f'(u)du) = d(f'(u))du + f'(u)d(du). \qquad (**)$$

Let $f'(u) = g(u)$, then

$d(f'(u)) = dg(u) = g'(u)du = (f'(u))'du = f''(u)du$.

Moreover, $d(du) = d^2 u = d^2 \varphi(x)$.

Thus, from (**) we obtain

$$d^2 f(u) = f''(u)(du)^2 + f'(u)d^2 u, \quad u = \varphi(x). \qquad (8.11)$$

Obviously, the second differential of a complex function $f(\varphi(x))$ exists if functions $f(u)$ and $\varphi(x)$ have finite derivatives up to the second order.

It follows from the formula (8.11) that the second differential of a complex function does not have form invariance:

if $y = f(x)$, *x* is an *independent* argument, then

$dy = f''(x)dx^2$;

if $y = f(u)$, *u* is *a dependent* argument, $u = \varphi(x)$, then

$dy = f''(u)\,du^2 + f'(u)\,d^2 u$.

Questions

1) What is the geometric meaning of the derivative?
2) Let $f'(3)=\sqrt{3}$ be a derivative of the function $y=f(x)$. What is the angle between *Ox* axis and the tangent to the graph at the point $x=3$?
3) What is the marginal productivity? How is this concept related to the concept of derivative?
4) What is the "Golden" rule of Economics?
5) What is the marginal cost of production?
6) How the concept of differentiability of a function $y=f(x)$ is defined at the point x_0 ?
7) Why the function $f(x)$ is called differentiable at the point x_0 ?
8) Suppose a function have a derivative at the point $x=2$. Is this function continuous at that point?
9) Let the function $y=f(x)$ be continuous at the point $x=5$. Is it possible to say that this function has a derivative at that point?
10) Is the line $y=4x-4$ tangent to the parabola $y=x^2$? And the line $y=-4x-4$?
11) Are the statements equivalent: «the function $y=f(x)$ is differentiable at the point x_0» and «the function $y=f(x)$ has a finite derivative at the point x_0»?

12) What is the algorithm for finding the derivative of an arbitrary function?
13) Is any of the basic elementary functions differentiable at each point at which it is defined?
14) What rule defines the differentiation of a complex function? Give examples other than those listed in the book.
15) What is the geometric meaning of the differential?
16) Can the differential of a function $f(x)$ be greater than the increment of that function?
17) The derivative y' is often denoted as $\frac{dy}{dx}$. What is the meaning of this notation?
18) What is the basis of the differential application in approximate calculations?
19) What is the invariance of the form of the differential of a complex function?

Chapter 9. Properties of differentiable functions

9.1. Basic theorems of the differential calculus

Theorem 9.1 **(Fermat theorem).** Let the function $y = f(x)$ be defined on $(a,\ b)$ and have the largest (smallest) value at some point $x_0 \in (a,\ b)$. Then if there exists a finite derivative $f'(x_0)$ at this point, this derivative is equal to zero, i.e. $f'(x_0) = 0$.

Proof. Let us prove the theorem for the case when the function has the greatest value at the point x_0 (for the smallest value, the proof is similar). In this case, for every $x \in (a,\ b)$ inequality$f(x) \le f(x_0)$ holds. It means that $\Delta y = f(x_0 + \Delta x) - f(x_0) \le 0$ for any point $x = x_0 + \Delta x \in (a,\ b)$. If $\Delta x > 0$, then $\frac{\Delta y}{\Delta x} \le 0$, therefore,

$$\lim_{\Delta x \to 0+} \frac{\Delta y}{\Delta x} \le 0; \qquad (*)$$

if $\Delta x < 0$, then $\frac{\Delta y}{\Delta x} \ge 0$, therefore,

$$\lim_{\Delta x \to 0-} \frac{\Delta y}{\Delta x} \ge 0. \qquad (**)$$

By definition of the derivative

$$f'(x_0) = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x},$$

moreover, this limit does not depend on whether Δx tends to zero, being positive or negative. But limits (*) and (**) coincide only when they are zero:

$$\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0+} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0-} \frac{\Delta y}{\Delta x} = 0;$$

which gives $f'(x_0) = 0$. That completes the proof.

The geometric meaning of Fermat theorem is obvious (fig. 9.1): a differentiable function takes the largest (smallest) value at the point x_0, then the tangent to the graph of this function is parallel to the axis *Ox* at the point $M(x_0,f(x_0))$.

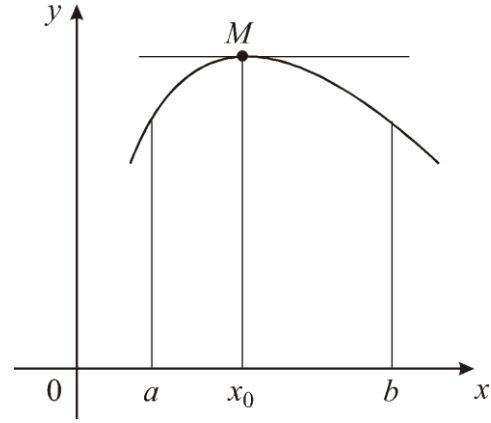


Fig. 9.1. *Geometric meaning of Fermat theorem*

Theorem 9.2 **(Rolle theorem).** Let a function $f(x)$ satisfy the following three conditions:

1) continuous on $[a, b]$;

2) differentiable on (a, b);

3) takes equal values at the ends of the segment: $f(a) = f(b)$.

Then there is at least one point $\xi \in (a,\ b)$ at which the derivative is equal to zero inside the segment:

$f'(\xi) = 0$.

Proof. Since $f(x)$ is continuous on $[a,\ b]$, then, by virtue of the second Weierstrass theorem (see § 7.2), it reaches its largest value *M* and its lowest value *m* on $[a,\ b]$.

There are two possible cases.:

1. $M = m$. Then $f(x) = M = m = \text{const}$; so $f'(x) = 0$ at all points, so that we can take any point ξ on (a, b).

2. $M \neq m$. Both of these values cannot be reached at the ends of the segment (since $f(a) = f(b)$). Therefore, at least one of these values is reached at some internal point $\xi \in (a, b)$ and by virtue of Fermat theorem $f'(\xi) = 0$. That completes the proof.

The geometric meaning of Roll theorem is as follows: if the extreme ordinates of the curve $y = f(x)$ are equal, then there is a point on the curve where the tangent is parallel to the axis *Ox* (fig. 9.2).

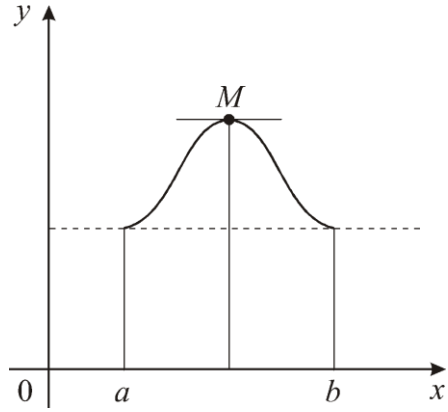


Fig. 9.2. *Geometric meaning of Roll theorem*

It should be noted that all the conditions of Rolle theorem are essential, and if at least one of them fails, the conclusion of the theorem may turn out to be incorrect.

Theorem 9.3 **(Lagrange theorem).** Let a function $f(x)$ be continuous on $[a, b]$and differentiable on (a, b). Then there exists a point $\xi \in (a, b)$, such that:

$\frac{f(b)-f(a)}{b-a} = f'(\xi)$.

Proof. Consider a function:

$F(x) = f(x) - f(a) - \frac{f(b)-f(a)}{b-a}(x-a)$.

The function *F(x)* satisfies all the conditions of Roll theorem: it is continuous on [*a,b*] (since *f(x)* is continuous), differentiable on (a, b):

$F'(x) = f'(x) - \frac{f(b)-f(a)}{b-a}$,

and also takes the same values at the ends of the segment [*a,b*]:

$F(b) = F(a) = 0$.

According to Rolle theorem, there exists a point $\xi \in (a, b)$, such that $F'(\xi) = 0$, i.e.

$$f'(\xi) - \frac{f(b) - f(a)}{b - a} = 0.$$

Hence

$$\frac{f(b) - f(a)}{b - a} = f'(\xi).$$

That completes the proof.

Note that the Lagrange theorem implies the equality:

$$f(b) - f(a) = f'(\xi) \cdot (b - a), \qquad (9.1)$$

called **the Lagrange formula**.

Rolle's theorem is a special case of the Lagrange theorem.

The geometric meaning of the Lagrange theorem is seen on Fig. 9.3. The chord passing through the points $M_1(a,\ f(a))$ and $M_2(b,\ f(b))$ has the angular coefficient which is equal to:

$$\operatorname{tg}\alpha = \frac{M_2N}{M_1N} = \frac{f(b) - f(a)}{b - a}.$$

The Lagrange theorem states that there exists a point $M(\xi, f(\xi))$ on (a, b), where the tangent to the graph of the function is parallel to the chord M_1M_2: its angular coefficient $f'(\xi)$ is equal to the angular coefficient of the chord M_1M_2.

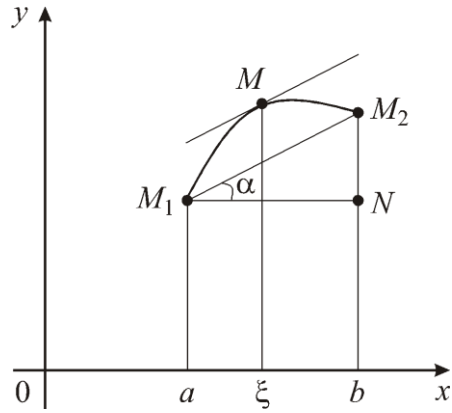


Fig. 9.3. *Geometric meaning of the Lagrange theorem*

In the previous chapter, we talked about approximate calculations based on replacing the increment of a function with a differential. Let us find out what is the accuracy of this replacement.

Evaluation of the accuracy of the equality Δy≈dy.

Let a function $f(x)$ have continuous derivatives $f'(x)$ and $f''(x)$ on $[a,b]$. Let x_0 and $x_0 + \Delta x$ be points on $[a,b.]$ According to the Lagrange formula

$$\Delta f(x_0) = f(x_0 + \Delta x) - f(x_0) = f'(\tilde{x})\Delta x,$$

where $\tilde{x}$ lies between x_0 and $x_0 + \Delta x.$

On the other hand,

$df(x_0) = f'(x_0)\Delta x.$

Repeated application of the Lagrange formula gives

$f'(\tilde{x}) - f'(x_0) = f''(\hat{x})(\tilde{x} - x_0),$

here $\hat{x}$ lies between x_0 and $\tilde{x}$. Therefore,

$\Delta f(x_0) - df(x_0) = f''(\hat{x})(\tilde{x} - x_0)\Delta x.$

Denote by M the highest value $|f'(x)|$. Since $|\tilde{x} - x_0| < |\Delta x|,$ then, replacing x_0 for $x,$ we obtain the estimate

$|\Delta f(x) - df(x)| < M(\Delta x)^2.$

Theorem 9.4 **(Cauchy theorem).** Let functions $f(x)$ and $g(x)$ be continuous on $[a, b]$ and differentiable on (a, b), where $g'(x) \neq 0$. Then there exists a point $\xi \in (a, b)$, such that

$$\frac{f(b)-f(a)}{g(b)-g(a)} = \frac{f'(\xi)}{g'(\xi)}. \qquad (9.2)$$

Proof. First of all, we verify that the denominator in the left-hand side of equality (9.2) is not equal to zero. Indeed, if $g(b) - g(a) = 0$, i.e. $g(b) = g(a)$, then by Rolle theorem $g'(\xi) = 0$ at some point $\xi \in (a, b)$ and this contradicts the condition of the theorem being proved.

Consider a function:

$F(x) = f(x) - f(a) - \frac{f(b)-f(a)}{g(b)-g(a)} \cdot [g(x) - g(a)].$

It is easy to verify that this function satisfies all the conditions of Rolle theorem on $[a, b]$: it is continuous on $[a, b]$ (due to the continuity of $f(x)$ and $g(x)$), differentiable on (a, b), so, its derivative has the form:

$$F'(x) = f'(x) - \frac{f(b) - f(a)}{g(b) - g(a)} \cdot g'(x)$$

and $F(a) = F(b) = 0$. Therefore, there exists a point $\xi \in (a, b)$ such that $F'(\xi) = 0$, i.e.

$$f'(\xi) - \frac{f(b) - f(a)}{g(b) - g(a)} \cdot g'(\xi) = 0.$$

From here (taking into account that $g'(\xi) \neq 0$) we obtain the formula:

$$\frac{f(b) - f(a)}{g(b) - g(a)} = \frac{f'(\xi)}{g'(\xi)}.$$

That completes the proof.

Formula (9.2) is called the **Cauchy formula**.

The Lagrange theorem is a special case of the Cauchy theorem as $g(x) = x$.

9.2. L'Hospital's rule

Indeterminate form $\frac{0}{0}$

Assume the following ratio $\frac{f(x)}{g(x)}$ is an indeterminate form $\frac{0}{0}$, if $\lim\limits_{x\to a} f(x) = \lim\limits_{x\to a} g(x) = 0$.

***Theorem* 9.5.** Let functions $f(x)$ and $g(x)$ satisfy Cauchy's theorem on interval $[a,\ b]$, and let $f(a) = g(a) = 0$. If there exists the limit $\lim\limits_{x\to a} \frac{f'(x)}{g'(x)}$, then the limit $\lim\limits_{x\to a} \frac{f(x)}{g(x)}$ exists too. Moreover

$$\lim_{x\to a} \frac{f(x)}{g(x)} = \lim_{x\to a} \frac{f'(x)}{g'(x)}. \qquad (9.3)$$

Proof. Let $[a,\ x] \subset [a,\ b]$. Apply Cauchy's theorem to functions $f(x)$ and $g(x)$ on$[a,\ x]$:

$\frac{f(x)-f(a)}{g(x)-g(a)} = \frac{f'(\xi)}{g'(\xi)}$,

where ξ ξ is a point between *a* and *x*. As we stated $f(a) = g(a) = 0$, so

$\frac{f(x)}{g(x)} = \frac{f'(\xi)}{g'(\xi)}$.

Let $x \to a$. Then $\xi \to a$ (because $a < \xi < x$). If the limit $\lim\limits_{x\to a} \frac{f'(x)}{g'(x)} =$ *K* exists, and $\lim\limits_{x\to a} \frac{f'(\xi)}{g'(\xi)}$exits too and it's equal to K, thus

$$\lim_{x\to a} \frac{f(x)}{g(x)} = \lim_{x\to a} \frac{f'(\xi)}{g'(\xi)} = \lim_{\xi\to a} \frac{f'(\xi)}{g'(\xi)} = \lim_{x\to a} \frac{f'(x)}{g'(x)} = K .$$

Or,

$$\lim_{x\to a} \frac{f(x)}{g(x)} = \lim_{x\to a} \frac{f'(x)}{g'(x)},$$

Q.E.D.

This theorem is known as **L'Hospital's rule**

***Remark* 1.** Theorem 9.5 remains valid even w hen functions $f(x)$ and $g(x)$ are not defined when $x=a$, but $\lim\limits_{x\to a} f(x)=\lim\limits_{x\to a} g(x)=0$. In this case, it suffices to redefine the functions at the point *a* using $f(a)=\lim\limits_{x\to a} f(x)=0$, $g(a)=\lim\limits_{x\to a} g(x)=0$, so they become continuous at this point and satisfy the theorem.

***Remark* 2.** L'Hospital's rule might be reapplied if both $f'(x)$ and $g(x)$ suffice theorem, as do the source functions $f(x)$ and $g(x)$.

***Example* 9.1.** Calculate limits:

а) $\lim\limits_{x\to 0}\dfrac{\sin 3x}{4x}$; б) $\lim\limits_{x\to 0}\dfrac{e^{3x}-1}{x}$; в) $\lim\limits_{x\to 0}\dfrac{e^{x}+e^{-x}-2}{1-\cos x}$.

Solution:

а) $\lim\limits_{x\to 0}\dfrac{\sin 3x}{4x}=\lim\limits_{x\to 0}\dfrac{(\sin 3x)'}{(4x)'}=\lim\limits_{x\to 0}\dfrac{3\cos 3x}{4}=\dfrac{3}{4}$;

б) $\lim\limits_{x\to 0}\dfrac{e^{3x}-1}{x}=\lim\limits_{x\to 0}\dfrac{3e^{3x}}{1}=3$;

в) $\lim\limits_{x\to 0}\dfrac{e^{x}+e^{-x}-2}{1-\cos x}=\lim\limits_{x\to 0}\dfrac{e^{x}-e^{-x}}{\sin x}=\lim\limits_{x\to 0}\dfrac{e^{x}+e^{-x}}{\cos x}=\dfrac{2}{1}=2$.

***Remark* 3.** L'Hospital's rule could be applied even if $\lim\limits_{x\to\infty} f(x)=0$ and $\lim\limits_{x\to\infty} g(x)=0$.

To proof it, let $x=\dfrac{1}{z}$. Then $z\to 0$ while $x\to\infty$, therefore,

$\lim\limits_{z\to 0} f\left(\frac{1}{z}\right)=0$, $\lim\limits_{z\to 0} g\left(\frac{1}{z}\right)=0$.

Now we can apply theorem 9.5 to functions of variable z $f\left(\frac{1}{z}\right)$ and $g\left(\frac{1}{z}\right)$. So, we obtain

$$\lim_{x\to\infty}\frac{f(x)}{g(x)}=\lim_{z\to 0}\frac{f\left(\frac{1}{z}\right)}{g\left(\frac{1}{z}\right)}=\lim_{z\to 0}\frac{f'\left(\frac{1}{z}\right)\cdot\left(-\frac{1}{z^2}\right)}{g'\left(\frac{1}{z}\right)\cdot\left(-\frac{1}{z^2}\right)}=\lim_{z\to 0}\frac{f'\left(\frac{1}{z}\right)}{g'\left(\frac{1}{z}\right)}=\lim_{x\to\infty}\frac{f'(x)}{g'(x)},$$

or

$$\lim_{x\to\infty}\frac{f(x)}{g(x)}=\lim_{x\to\infty}\frac{f'(x)}{g'(x)}. \qquad (9.4)$$

Indeterminate form $\frac{\infty}{\infty}$

The L'Hospital's rule can be applied even when functions $f(x)$ and $g(x)$ *tend to infinity* while $x\to a$.

Let $\lim\limits_{x\to a} f(x)=\infty$, $\lim\limits_{x\to a} g(x)=\infty$ and let's assume that ratio $\frac{f'(x)}{g'(x)}$ has a limit $\lim\limits_{x\to a}\frac{f(x)}{g(x)}=K$. Then ratio $\frac{f(x)}{g(x)}$ also has a limit while $x\to a$ and equation (17.3) is verified:

$$\lim_{x\to a}\frac{f(x)}{g(x)}=\lim_{x\to a}\frac{f'(x)}{g'(x)}.$$

(We accept this statement without proof.)

Note that functions $f(x)$ and $g(x)$, tend to infinity with $x\to\infty$, L'Hospital's rule is verified:

$$\lim_{x\to\infty}\frac{f(x)}{g(x)}=\lim_{x\to\infty}\frac{f'(x)}{g'(x)}.$$

9.3. Taylor series

Let us assume that function $y=f(x)$ has derivative in point $x=x_0$ and has derivatives up to $(n+1)$ order including.

The n-order **Taylor polynomial** for function $y=f(x)$ can be defined by the following equation:

$$P_n(x)=f(x_0)+\frac{f'(x_0)}{1!}\cdot(x-x_0)+\frac{f''}{2!}\cdot(x-x_0)^2+$$
$$+\ldots+\frac{f^{(n)}(x_0)}{n!}\cdot(x-x_0)^n. \qquad (9.5)$$

This polynomial and its derivatives in the point $x=x_0$ have the same values as the function $f(x)$ and its derivatives respectively:

$$f(x_0)=P_n(x_0),\ f'(x_0)=P_n'(x_0),\ f''(x_0)=P_n''(x_0),\ \ldots,$$
$$f^{(n)}(x_0)=P_n^{(n)}(x_0). \qquad (9.6)$$

(it is easy to obtain these equalities). So, we can consider polynomial (9.5) an approximation of a function $f(x)$. The order of the approximation is measured by difference $R_n(x)=f(x)-P_n(x)$. We obtain

$$f(x)=P_n(x)+R_n(x)$$

or:

$$f(x)=f(x_0)+\frac{f'(x_0)}{1!}\cdot(x-x_0)+\frac{f''(x_0)}{2!}\cdot(x-x_0)^2+\ldots+$$
$$+\frac{f^{(n)}(x_0)}{n!}\cdot(x-x_0)^n+R_n(x). \qquad (9.7)$$

The equation (9.7) is called a **Taylor formula**, and $R_n(x)$ is a **remainder term**.

Next, we are going to find the difference between function $f(x)$ and polynomial $P_n(x)$ with different values of variable *x*, in other words we will estimate the value $R_n(x)$.

Rewrite remainder term as

$$R_n(x) = \frac{Q(x)}{(n+1)!}(x - x_0)^{n+1}, \tag{9.8}$$

where $Q(x)$ is the function which we will find.

Considering (9.8) formula (9.7) becomes

$$f(x) = f(x_0) + \frac{f'(x_0)}{1!}(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \ldots +$$
$$+ \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n + \frac{Q(x)}{(n+1)!}(x - x_0)^{n+1}.. \tag{9.9}$$

Fix *x* variable. Let us assume that $x > x_0$. Then function $Q(x)$ will have a fixed valued. Denote it as *Q*.

Denote the variable with values ranging from x_0 to *x* as *t* and consider a new function on interval $[x_0, x]$

$$F(t) = f(x) - f(t) - \frac{f'(t)}{1!}(x - t) - \frac{f''(t)}{2!}(x - t)^2 - \ldots -$$
$$- \frac{f^{(n)}(t)}{n!}(x - t)^n - \frac{Q}{(n+1)\,!}(x - t)^{n+1}, \tag{9.10}$$

where *Q* has a numerical values which can be defined by (9.9) with fixed *x*.

Let's find derivative $F'(t)$:

$$F'(t) = -f'(t) - \frac{f''(t)}{1!}(x - t) + f'(t) - \frac{f'''(t)}{2!}(x - t)^2 + \frac{2f''(t)}{2!}(x - t) -$$
$$\ldots - \frac{f^{(n+1)}(t)}{n!}(x - t)^n + \frac{nf^{(n)}(t)}{n!}(x - t)^{n-1} + \frac{(n+1)\cdot Q}{(n+1)\,!}(x - t)^n.$$

Here, the corresponding terms with mutually opposite signs are mutually annihilated. And we obtain

$$F'(t) = -\frac{f^{(n+1)}(t)}{n!}(x-t)^n + \frac{Q}{n!}(x-t)^n. \qquad (9.11)$$

Function $F(t)$ has a derivative (9.11) on $[x_0, x]$. Moreover, it follows from (9.10) that $F(x) = F(x_0) = 0$. Therefore we can apply Rolle's theorem to function $F(t)$ on $[x_0, x]$, thus there exists $\xi \in (x_0, x)$, such as $F'(\xi) = 0$. Thence with respect to (9.11) we obtain:

$$-\frac{f^{(n+1)}(\xi)}{n!}(x-\xi)^n + \frac{Q}{n!}(x-\xi)^n = 0,$$

therefore,

$$Q = f^{(n+1)}(\xi).$$

Appling this to (9.8), we obtain:

$$R_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-x_0)^{n+1}. \qquad (9.12)$$

The expression (9.12) is called **remainder term in Lagrange form**.

Appling $R_n(x)$ to (9.7), we obtain:

$$f(x) = f(x_0) + \frac{f'(x_0)}{1!}(x-x_0) + \frac{f''(x_0)}{2!}(x-x_0)^2 + \ldots +$$
$$+ \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n + \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-x_0)^{n+1}. \qquad (9.13)$$

The formula (9.13) is called a **Taylor formula with a remainder term in Lagrange form**.

The formula (9.13) is used when we need to substitute $f(x)$ with polynomial $P_n(x)$ (when $x \neq x_0$) and find the value of the error which occurs during this substitution. However, in so me cases it is necessary

to know the behavior of the remainder term when *x* is tending to x_0, rather than certain values of *x*. To do so we need to rewrite the remainder term in different form

Let's proof that, when $x \to x_0$ the remainder term $R_n(x)$ is infinitesimal with order higher then $(x-x_0)^n$:

$$R_n(x) = o\left((x-x_0)^n\right). \qquad (9.14)$$

This is the **remainder term in the form of Peano**.

Let us assume, that in some neighborhood of a point x_0 exist derivatives of functions $f(x)$ up to order *n* and $f^{(n)}(x)$ is continuous at x_0.

In formula (9.13) substitute *n* to $n-1$:

$$f(x) = f(x_0) + \frac{f'(x_0)}{1\rightleftarrows!}(x-x_0) + \frac{f''(x_0)}{2\rightleftarrows!}(x-x_0)^2 + \ldots +$$
$$+ \frac{f^{(n-1)}(x_0)}{(n-1)!}(x-x_0)^{n-1} + \frac{f^{(n)}(\xi)}{n!}(x-x_0)^n, \qquad (9.15)$$

where ξ is in between x_0 and *x*. We represent the last term in the form

$$\frac{f^{(n)}(\xi)}{n!} = \frac{f^{(n)}(x_0)}{n!} + \alpha(x). \qquad (9.16)$$

If $x \to x_0$, then $\xi \to x_0$ (because ξ is in between x_0 and *x*). Then $f^{(n)}(\xi) \to f^{(n)}(x_0)$, because $f^{(n)}(x)$ is continuous. $\alpha(x) \to 0$ in because of (17.16). It means that $\alpha(x)(x-x_0)^n = o\left((x-x_0)^n\right)$.

Thus

$$\frac{f^{(n)}(\xi)}{n!}(x-x_0)^n = \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n + o\left((x-x_0)^n\right).$$

It follows from (17.15) that

$$f(x) = f(x_0) + \frac{f'(x_0)}{1!}(x-x_0) + \frac{f''(x_0)}{2!}(x-x_0)^2 + \ldots +$$
$$+ \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n + o((x-x_0)^n) . \qquad (9.17)$$

Where

$R_n(x) = o\left((x-x_0)^n\right)$,

Q.E.D.

If we assume that $x_0 = 0$ in formula (9.13), then we obtain the **Maclaurin formula** (which is a special case of the Taylor formula)

$$f(x) = f(0) + \frac{f'(0)}{1!}x + \frac{f''(0)}{2!}x^2 + \ldots + \frac{f^{(n)}}{n!}x^n + \frac{f^{(n+1)}(\xi)}{(n+1)!}x^{n+1}, \qquad (9.18)$$

where ξ is a point with values between 0 and *x*.

The Maclaurin formula with the remainder term in the form of Peano can be defined by the following equation

$$f(x) = f(0) + \frac{f'(0)}{1!}x + \frac{f''}{2!}x^2 + \ldots + \frac{f^{(n)}}{n!}x^n + o(x^n). \qquad (9.19)$$

Maclaurin expansion of some elementary functions

The simplest elementary functions are polynomials. The Taylor and Maclaurin formulas make it possible to represent the function $f(x)$ as a *polynomial*, the coefficients of which can be easily calculated. These expansions are used for the approximate calculation of functions.

In particular, the following approximate equalities hold (when $x \to 0$):

$e^x \approx 1 + \frac{x}{1!} + \frac{x^2}{2!} + \ldots + \frac{x^n}{n!}$;

$ln\ (1+x) \approx x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \ldots + (-1)^{n+1}\frac{x^n}{n}$;

$sin\, x \approx x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots + \frac{(-1)^k x^{2k+1}}{(2k+1)\,!}$;

$cos\, x \approx 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \ldots + \frac{(-1)^k x^{2k}}{(2k)\,!}$;

$(1+x)^\alpha \approx 1 + \frac{\alpha}{1!}x + \frac{\alpha(\alpha-1)}{2!}x^2 + \ldots + \frac{\alpha(\alpha-1)\cdots(\alpha-n+1)}{n!}x^n$.

In those approximations the error is infinitesimal of a higher order than x^n (in case of sinus function $n = 2k+1$, and for cosines $n = 2k$).

Let's take a closer look at the expansion of the exponent and sinus.

1. $\boldsymbol{f(x) = e^x}$. Obviously, $f'(x) = e^x$, $f''(x) = e^x$, …, $f^{(n)}(x) = e^x$; $f(0) = 1$, $f'(0) = 1$, …, $f^{(n)}(0) = 1$. Apply those equations to (9.19), we obtain

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \ldots + \frac{x^n}{n!} + o(x^n).$$

2. $\boldsymbol{f(x) = \sin x}$. Differentially differentiating and Appling $x = 0$, we obtain: $f(0) = 0$, $f'(x) = \cos x$, $f'(0) = 1$, $f''(x) = -\sin x$, $f''(0) = 0$, $f'''(x) = -cos\, x$, $f'''(0) = -1$, …, $f^{(n)}(x) = sin\left(x + n\frac{\pi}{2}\right)$, $f^{(n)}(0) = \sin n\frac{\pi}{2}$. Apply it to (9.19), we obtain

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots + \frac{(-1)^k x^{2k+1}}{(2k+1)!} + o(x^{2k+1}).$$

Consider another form of Taylor's formula. In formula (9.17), we transfer $f(x_0)$ to the left hand side of the equality and denote as $x - x_0 = \Delta x$. Then the difference $f(x) - f(x_0) = f(x_0 + \Delta x) - f(x) = \Delta f(x_0)$. We obtain

$$\Delta f(x_0) = f'(x_0)\Delta x + \frac{f''(x_0)}{2!}\Delta x^2 + \ldots + \frac{f^{(n)}(x_0)}{n!}\Delta x^n + o(\Delta x^n). \quad (9.17')$$

This formula is a generalization of the formula (8.2′):

$$\Delta f(x_0) = f'(x_0)\Delta x + o(\Delta x),$$

which, obviously, is obtained from (17.17 '), if put $n = 1$. Similarly, from (9.13) we obtain:

$$\Delta f(x_0) = f'(x_0)\Delta x + \frac{f''(x_0)}{2!}\Delta x^2 + \ldots + \frac{f^{(n)}(x_0)}{n!}\Delta x^n + \frac{f^{(n+1)}(\xi)}{(n+1)!}\Delta x^{n+1}. \quad (9.12')$$

If we replace the increment of the independent variable Δx with dx in the formulas (9.13 ') and (9.17') (because $dx = \Delta x$) and consider that

$f'(x_0)dx = df(x_0)$, $f''(x_0)dx^2 = d^2f(x_0)$, …, $f^{(n)}(x_0)dx = d^nf(x_0)$,

$f^{(n+1)}(\xi)dx^{n+1} = d^{n+1}(\xi)$,

then after Applying those equations to (9.13′) and (9.17′), we obtain

$$\Delta f(x_0) = df(x_0) + \frac{1}{2!}d^2f(x_0) + \cdots + \frac{1}{n!}d^nf(x_0) + \frac{1}{(n+1)!}d^{n+1}f(\xi), \quad (9.13'')$$

$$\Delta f(x_0) = df(x_0) + \frac{1}{2!}d^2f(x_0) + \ldots + \frac{1}{n!}d^nf(x_0) + o(\Delta x^n). \quad (9.17'')$$

Thus (when $\Delta x \to 0$) with formulas (9.13″) and (9.17″) it is possible to extract from the infinitesimal increment $\Delta f(x_0)$ not only its main term (the first differential) but also members of higher orders of smallness. They are successive differentials of the second, third, etc. orders with coefficients respectively

$$\frac{1}{2!}, \frac{1}{3!}, \ldots, \frac{1}{n!}.$$

Each of the formulas (9.13 ") and (9.17") is called a **Taylor formula in differential form**. In studying the multivariable functions, we will use just such a representation of the Taylor formula.

Questions

1. What is the geometric meaning of the Lagrange theorem?
2. Is the Lagrange theorem a special case of the Cauchy's theorem?
3. Let $\lim\limits_{x\to 2} f(x)=1$, $\lim\limits_{x\to 2} g(x)=0$. Can the L'Hospital's rule be applied to find the limit $\lim\limits_{x\to 2}\dfrac{f(x)}{g(x)}$?
4. Let $\lim\limits_{x\to 1} f(x)=+\infty$, $\lim\limits_{x\to 1} g(x)=-\infty$. Can the L'Hospital's rule be applied to find the limit $\lim\limits_{x\to 1}\dfrac{f(x)}{g(x)}$?
5. What is the Taylor polynomial? What are its properties?
6. How is the Taylor polynomial of a function $f(x)$ related to the Taylor formula for this function?
7. What is the Maclaurin formula?
8. What is the mathematical equation of the Taylor formula in differential form for function $f(x)$?

Chapter 10. Curve sketching with the use of the first derivative

10.1. Monotonic test

***Theorem* 10.1.** Let $f(x)$ continuous on interval *X* and has a finite derivative inside it. In order for function $f(x)$ to be monotonically increasing (decreasing) on *X*, the condition $f'(x)>0$ ($f'(x)<0$) is sufficient on *X*.

Proof (monotonically increasing) Let $f'(x)>0$; x_1, $x_2 \in X$, $x_2 > x_1$. Apply Lagrange's theorem to $f(x)$ on $[x_1, x_2]$:

$$f(x_2)-f(x_1)=f'(c)(x_2-x_1),$$

where $x_1 < c < x_2$. $f(x_2)>f(x_1)$ because $f'(c)>0$, therefore, $f(x)$ is an increasing function.

Note that the proved condition is not necessary. For example, the theorem remains verified if the derivative vanishes at a finite number of interior points of the interval *X*.

10.2. Extremum

Definition. The point x_0 is called the **local maximum point** of the function $f(x)$, if in some neighborhood of the point x_0 the inequality $f(x_0) > f(x)$ is verified.

The point x_0 is called **local minimum point** of the function $f(x)$, if in some neighborhood of the point x_0 the inequality $f(x_0) < f(x)$ is verified.

If x_0 is the point of local maximum (minimum), then the value of the function $f(x_0)$ is called the **local maximum (minimum)**.

The general term for a local maximum and a local minimum is a **local extremum.**

The ***necessary*** *condition for the extremum* of a differentiable function follows from Fermat's theorem proved in § 9.1: in order for the differentiable function $f(x)$ to have a local extremum at the point x_0, it is necessary that the equality $f'(x_0) = 0$ *is verified* at this point.

Since x_0 is an extremum, then there is an interval containing a point x_0, where the value $f(x_0)$ is largest or smallest. Then by Fermat's theorem we obtain that $f'(x_0) = 0$.

Note that the condition $f'(x_0) = 0$ is not a sufficient condition for the extremum. For instance, function $y = x^3$ increases on the whole number line and has no extremum, but its derivative is equal to zero at the point $x_0 = 0$: $f'(x_0) = 3x_0^2 = 0$.

In addition, the function may have an extremum at some point, but not be differentiable at this point.

Points at which the derivative of the function is equal to zero or does not exist are called critical (or stationary). Obviously, if there is an extremum at any point, then this point is critical.

Points at which the derivative of the function is equal to zero or does not exist are called **critical** (or **stationary**). Obviously, if there is an extremum[1] at any point, then this point is critical.

10.3. The first sufficient condition of extremum

***Theorem* 10.2.** Let $f(x)$ function be continuous on any interval containing a critical point x_0, and differentiable at all points of this interval, except, perhaps, the point itself x_0. If, when passing through a point x_0, the derivative changes sign from plus to minus, then the point x_0 has a local maximum, and if from minus to plus, then the minimum.

Proof. For definiteness, let the derivative change sign from plus to minus: $f'(x) > 0$ when $x < x_0$, $f'(x) < 0$ when $x > x_0$ (for all *x*, on the considered interval). We apply the Lagrange's theorem to $f(x)$ on $[x, x_0]$:

$$f(x_0) - f(x) = f'(c) \cdot (x_0 - x), \quad c \in (x, x_0).$$

Because $f'(c) > 0$ and $x_0 - x > 0$, then $f(x) < f(x_0)$.

[1] Often we simply say extremum (maximum, minimum), referring to the local extremum (local maximum, a local minimum).

We apply the Lagrange's theorem on interval $[x_0, x]$, where $x > x_0$, we obtain:

$$f(x) - f(x_0) = f'(c)(x - x_0), \; c \in (x_0, x).$$

Since point *c* is now on the right of x_0, then $f'(c) < 0$. Moreover, $x - x_0 > 0$. Thus $f(x) - f(x_0) < 0$. We obtain $f(x) < f(x_0)$. So, for all *x* on the considered interval, the following equation is verified:

$$f(x_0) > f(x).$$

Thus, there is a local maximum at the point x_0.

The case of a local minimum is similar.

Based on Theorems 10.1 and 10.2, the following scheme is used to **find the extremum of the function using the first derivative**.

1. Calculate the derivative $y' = f'(x)$.
2. Find critical points.
3. Determine the sign of the derivative to the left and right of each critical point and conclude that there are local extrema of the function.
4. Find function values at local extremum points.

***Example* 10.1.** Find the extremum $f(x) = 3x^4 - 4x^3 - 12x^2 + 10$.

Solution. Calculate the derivative:

$$f'(x) = 12x^3 - 12x^2 - 24x = 12x \cdot (x^2 - x - 2) = 12x \cdot (x+1) \cdot (x-2)$$

.

By solving equation $f'(x) = 0$, or $x(x+1)(x-2) = 0$, we find critical points: $x_1 = -1$, $x_2 = 0$, $x_3 = 2$. After determining the sign of the derivative (fig. 10.1), we obtain: $x = -1$, $x = 2$ are local minimum points, $f(-1) = 5$, $f(2) = -22$ are minimum function values; $x = 0$ is a point of

local maximum, $f(0)=10$ is a maximum value of the function at this point..

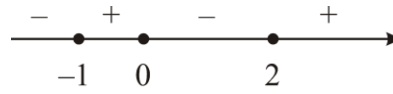


Fig. 10.1

10.4. Largest and smallest values of the function on the interval

Many economic problems are formulated as problems of finding the largest (smallest) value of a function on a certain set. Let us consider the simplest case when it is required to find the largest (smallest) value on an interval $[a, b]$. According to the second Weierstrass's theorem, if a function is continuous on an interval $[a, b]$, then it takes on it the largest and smallest values. Note that the largest or smallest value of the function can be achieved both at the points of the local extremum and the ends of the segment.

The following scheme is used to **find the largest and smallest values of a function on a segment.**

1. Calculate the derivative $f'(x)$.
2. Find critical points.
3. Find the values of the function at critical points and at the ends of the segment and choose the largest and smallest from them.

Note that in this case there is no need to find an extremum at critical points.

***Example* 10.2.** Find the largest and smallest values of the function $f(x)=x^2\mathrm{e}^x$ on $[-3, 1]$.

Solution. 1. $f'(x) = 2x\mathrm{e}^x + x^2\mathrm{e}^x = x(x+2)\mathrm{e}^x$.

2. $f'(x) = 0$: $x(x+2)\mathrm{e}^x = 0$. critical points: $x_1 = 0$, $x_2 = -2$.

3. $f(-3) = 9\mathrm{e}^{-3}$, $f(-2) = 4\mathrm{e}^{-2}$, $f(0) = 0$, $f(1) = \mathrm{e}$.

$f_{\text{наиб}} = f(1) = \mathrm{e}$, $f_{\text{наим}} = f(0) = 0$.

So, the greatest value is achieved at the right end of the segment, and the smallest - at one of the critical points.

***Example* 10.3.** At point *A* is a deposit of raw materials. The distance from point *A* to the nearest point *B* on the railway is 200 km. The railway passes through city *C*, where the processing plant for the mentioned raw materials is located. The distance from *B* to *C* is 1000 km. To deliver raw materials to the plant, the *AD* highway is being built, connecting the field with the railway. The cost of transportation on the highway is double that of the railway. At what distance should point *D* be from *A* so that the total cost of transporting raw materials from field *A* to city *C* along the *ADC* route is minimal?

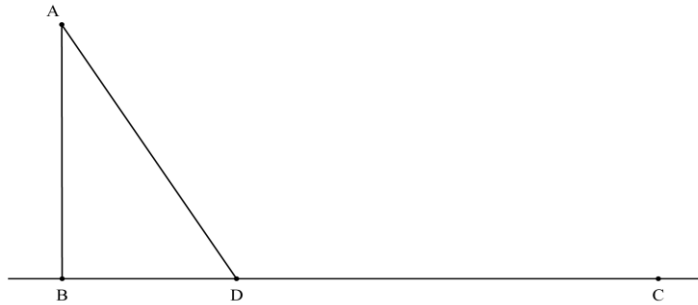


Fig. 10.2

Solution. Denote: $BD = x$. Then $DC = 1000 - x$.

Let a monetary unit cost the transportation of one ton of cargo by rail. Then transportation of one ton on the highway costs. By the Pythagorean's

theorem, we calculate the length of the highway *AD*: $AD = \sqrt{x^2 + 200^2}$. The cost of transporting one ton on the *ADC* route is

$$f(x) = 2a\sqrt{x^2 + 200^2} + a(1000 - x).$$

Obviously, we need to find the smallest value on the interval $[0, 1000]$. In this case, the *ABC* route corresponds to the value $x = 0$, while the *AC* route corresponds to the value $x = 1000$.

We calculate the derivative:

$$f'(x) = \frac{2ax}{\sqrt{x^2 + 200^2}} - a.$$

Find the critical points, equating the derivative to zero:

$$\frac{2ax}{\sqrt{x^2 + 200^2}} - a = 0,$$

$$\frac{2ax - a\sqrt{x^2 + 200^2}}{\sqrt{x^2 + 200^2}} = 0,$$

$$2x = \sqrt{x^2 + 200^2},$$

$$3x^2 = 200^2.$$

We need only positive value *x*:

$$x = \frac{200}{\sqrt{3}}.$$

It means that $BD \approx 115{,}4$ km.

Let's make sure that the value at the point $x = \frac{200}{\sqrt{3}}$ is the smallest. To do so we calculate the values $f(x)$ at the considered point, at points $x = 0$, $x = 1000$ and compare them: $f\left(\frac{200}{\sqrt{3}}\right) \approx 1346a$,

$f(0) = 1400a$,

$f(1000) = 2a\sqrt{1040000} > 2000a$.

So, the smallest value is reached at the critical point $x = \frac{200}{\sqrt{3}}$.

Questions

1) Let us assume that function $y = f(x)$ is increasing on $[0, +\infty)$. Is it inequality $f'(x) > 0$ verified for all $x \in [0, +\infty)$?
2) What is a local extremum?
3) What point is called the critical (stationary) point of a given function?
4) Can a function have two local minimums?
5) Does the function $y = 4 - x^2$ has a local minimum?
6) Let the function $f(x)$ be continuous on X, $x_0 \in X$, $f'(x) < 0$ when $x < x_0$ and $f'(x) > 0$ when $x > x_0$, and at the point $x = x_0$ derivative $f'(x)$ does not exist. Is there an extremum at the point x_0, and if so, which one is it?
7) How many extreme points does the function have $y = \sin x$ on $[0, 2\pi]$?
8) Let the derivative of the function $y = f(x)$ be 1 on $(-1, 3)$. Will the function increase in this interval?
9) The function $y = f(x)$ is differentiable on (a, b) and $f'(x) = 0$ at six points of this interval. Can $f(x)$ have (a, b) four minimums?

10) If a function $y = f(x)$ has a maximum at a point x_0 then will the function $y = (f(x))^2$ have a maximum at this point?
11) Does the function$y = 3x - 4$ have extremum?
12) Can a function $y = f(x)$ at some point $x \in (a, b)$ have a value less than any of the minima of this function on (a, b)?
13) Can the smallest function value $y = f(x)$, $x \in [a, b]$ be at the point $x = b$?
14) Let a function $y = f(x)$ have a local maximum and a local minimum on $[a, b]$. Can its greatest value not coincide with a local maximum, and the smallest - with a local minimum?
15) Is it possible to find the largest and smallest values of a function on an interval without finding a local extremum, but knowing only its values at critical points?

Chapter 11. Curve sketching with the use of the second derivative. Full curve sketching and plotting

11.1. Second sufficient condition of extremum

***Theorem* 11.1.** Let $f(x)$ and derivatives $f'(x)$ and $f''(x)$ exist and are continuous in some neighborhood of the point x_0 and $f'(x)=0$. Then:

1) if $f''(x_0)<0$, then x_0 is a local maximum point

2) if $f''(x_0)>0$, then x_0 is a local minimum point.

Proof. Let $f'(x_0)=0$, $f''(x_0)<0$. Because $f''(x)$ is continuous, then $f''(x)<0$ not only at the point x_0, but in its neighborhood. But $f''(x)$ is the first derivative of the first derivative. Thus, it follows from $f''(x)<0$ that the first derivative is decreasing in that neighborhood. But at the point x_0 derivative equals zero, it means that, $f'(x)$ is positive on

the left of x_0 and negative on the right. According to theorem 18.2 the local maximum exists at the point x_0.

The case $f''(x_0) > 0$ is similarly.

Note that if both derivatives are zero $f'(x_0) = 0$ and $f''(x_0) = 0$ at the point x_0, then this theorem does not answer the minimum and maximum question. In this case, one can either apply the first sufficient condition for the extremum, or involve higher derivatives.

11.2. Convexity and concavity of the function graph. Inflection point

Consider a curve $y = f(x)$ on the plane, that is a graph of a function $f(x)$.

A curve $y = f(x)$ on (a, b) has an **upward convexity,** if all points of the curve lie *below* its tangent in this interval.

A curve $y = f(x)$ (b, c) has a **downward convexity**, if all points of the curve lie *above* its tangent in this interval.

In fig. 11.1 the convexity of the curve on (a, b) is directed upward, and on (b, c) is downward.

If the convexity of the curve is directed upwards, then the curve is called **convex**; if the bulge is directed downward is called **concave**.

The curve shown in fig. 11.1 is convex on (a, b) and concave on (b, c).

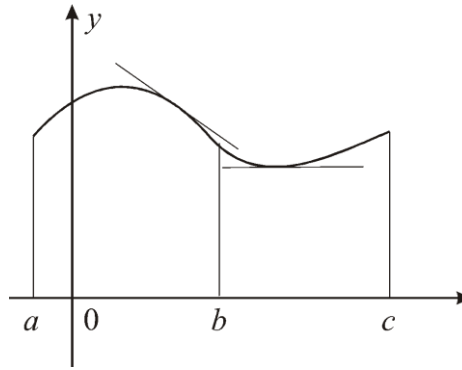


Fig. 11.1. *Graph convexity direction*

***Theorem* 11.2.** If function $y = f(x)$ has a second derivative on the interval (a, b) and $f''(x) > 0$ on (a, b), then the graph of this function has a convexity directed downward on (a, b); if $f''(x) < 0$ on (a, b), then the graph has a bulge upward on (a, b).

Proof. Here we consider the case where $f''(x) < 0$. Let x_0 a random point on (a, b). Ehe equation of the tangent to the graph of the function $y = f(x)$ passing through the point $M_0(x_0, f(x_0))$:

$$Y = f(x_0) + f'(x_0)(x - x_0). \qquad (11.1)$$

Where *Y* is a current ordinate of tangent.

We represent the function $f(x)$ in a neighborhood of a point x_0 using the Taylor formula with $n = 1$:

$$y = f(x) = f(x_0) + \frac{f'(x_0)}{1!}(x - x_0) + \frac{f''(\xi)}{2!}(x - x_0)^2,$$
$$\xi \in (a,\ b). \qquad (11.2)$$

$$R_n = R_1 = \frac{f''(\xi)}{2!}(x - x_0)^2$$

(where .)

Subtract (11.1) from (11.2):

$$y - Y = \frac{f''(\xi)}{2!}(x - x_0)^2. \qquad (11.3)$$

We know that $f''(x) < 0$ on (a, b), therefore $y < Y$ for all $x \in (a, b)$. And this means that the curve $y = f(x)$ is below the tangent. q.e.d.

Definition. The point separating the convex part of the curve from the concave is called an **inflection point**.

The necessary condition for the inflection at point x_0 for a graph of a function $f(x)$ which has at this point a continuous second derivative is that

$$f''(x_0) = 0.$$

Assume the contrary, that $f''(x) > 0$, we obtain that in the neighborhood of the point x_0 the curve has a convexity directed downward, and the point x_0 cannot be an inflection point.

The sufficient condition for the inflection is to change the sign of the second derivative of the function $y = f(x)$ when passing through a point x_0. In other words, if the second derivative has different signs to the left and to the right of x_0, then the graph of the function has an inflection at $x = x_0$.

In this case, the directions of the convexity of the graph to the left and to the right of x_0 are different, and this means the presence of an inflection at the point x_0.

To sum it up the following scheme is used to **find the inflection points:**

1) Find the points where $f''(x)$ is zero or does not exist (points like this also known as *critical points*);

2) Find the sign of the second derivative to the left and right of each such point.

***Example* 11.1.** Find the inflection points and convex directions of the function graph $f(x) = (1-x)\,\mathrm{e}^x$.

Solution. Let's find the first and second derivatives:

$$f'(x) = -\mathrm{e}^x + (1-x)\,\mathrm{e}^x = -x\mathrm{e}^x, \quad f''(x) = -\mathrm{e}^x - x\mathrm{e}^x = -(x+1)\,\mathrm{e}^x.$$

Equating the second derivative to zero, we find the critical point $x = -1$. Obviously, the second derivative is positive to the left of this point, and negative to the right. Thus, when $x < -1$ the convexity of the graph is directed down, and when $x > -1$ it is directed up. Point $x = -1$ is the inflection point.

11.3. Asymptotes

Definition. A straight line is called the **asymptote** of the graph of the function $y = f(x)$ if the distance from the point *M* lying on the graph to this straight line tends to zero when the point *M* is unboundedly from the origin.

There are three types of asymptotes: *vertical*, *horizontal* and *oblique*.

The line $x = a$ is called **vertical asymptote** for $y = f(x)$, if at least one of the limit values $\lim\limits_{x \to a+} f(x)$ or $\lim\limits_{x \to a-} f(x)$ equals $+\infty$ or $-\infty$.

The line $y = b$ is called **horizontal asymptote** for $y = f(x)$ when $x \to \infty$ ($x \to -\infty$), if $\lim\limits_{x \to +\infty} f(x) = b$ ($\lim\limits_{x \to -\infty} f(x) = b$).

The line $y = kx + b$ is called **oblique asymptote** for $y = f(x)$ when $x \to +\infty$ ($x \to -\infty$), if function $f(x)$ can be represented as $f(x) = kx + b + \alpha(x)$, where $\alpha(x) \to 0$ when $x \to +\infty$ ($x \to -\infty$).

The presence of an oblique asymptote is due to the existence of two limits:

$$\lim_{x \to \infty} \frac{f(x)}{x} = k, \quad \lim_{x \to \infty} [f(x) - kx] = b.$$

(the cases $x \to +\infty$ and $x \to -\infty$ should be considered separately).

Examples of vertical and horizontal asymptotes are well known from school mathematics course. In particular, the graph of the function $y = \frac{1}{x}$ has a vertical asymptote $x = 0$ and a horizontal asymptote $y = 0$; the graph of the function $y = \operatorname{tg} x$ has infinitely many vertical asymptotes: $x = \pm\frac{\pi}{2}, x = \pm\frac{3\pi}{2}, \ldots$.

***Example* 11.2.** Find the oblique asymptote for $y = \frac{x^3}{x^2 - x + 1}$.

Solution. The oblique asymptote's equation is $y = kx + b$, find *k* and *b*:

$$k = \lim_{x \to \infty} \frac{f(x)}{x} = \lim_{x \to \infty} \frac{x^3}{x \cdot (x^2 - x + 1)} = \lim_{x \to \infty} \frac{x^2}{x^2 - x + 1} = 1,$$

$$b=\lim_{x\to\infty}\left[f(x)-kx\right]=\lim_{x\to\infty}\left(\frac{x^3}{x^2-x+1}-x\right)=\lim_{x\to\infty}\frac{x^2-x}{x^2-x+1}=1.$$

Asymptote's equation: $y=x+1$.

***Example* 11.3.** The model of consumer demand uses, in particular, the *Tornquist functions*, which model the relationship between the value of income and the value of consumer demand for: a) essential goods; b) essential goods; c) luxury goods:

a) $y=\dfrac{b_1(x-a_1)}{x-c_1}\quad(x>a_1)$;

b) $y=\dfrac{b_2(x-a_2)}{x-c_2}\quad(x>a_2)$;

c) $y=\dfrac{b_3x\cdot(x-a_3)}{x-c_3}\quad(x>a_3)$.

The graphs of the first two of these functions have horizontal asymptotes $y=b_1$ and $y=b_2$:

$$\lim_{x\to+\infty}\frac{b_1(x-a_1)}{x-c_1}=b_1;\quad\lim_{x\to+\infty}\frac{b_2(x-a_2)}{x-c_2}=b_2,$$

while the last graph has oblique asymptote:

$$k=\lim_{x\to+\infty}\frac{f(x)}{x}=\lim_{x\to+\infty}\frac{b_3x\cdot(x-a_3)}{x\cdot(x-c_3)}=\lim_{x\to\infty}\frac{x^2}{x^2-x+1}=b_3,$$

$$b=\lim_{x\to+\infty}\left[\frac{b_3x\cdot(x-a_3)}{x-c_3}-b_3x\right]=$$

$$=\lim_{x\to+\infty}\frac{b_3x^2-b_3a_3x-b_3x^2+b_3c_3x}{x-c_3}=b_3c_3-b_3a_3.$$

Oblique asymptote's equation: $y=b_3x+b_3(c_3-a_3)$.

Here we traditionally denote the argument by *x*, and the function by *y*. Note that other notations are usually used for these functions:

a) $x = \frac{\alpha I}{I + \beta}$; b) $x = \frac{\alpha \cdot (I - \gamma)}{I + \beta}$; c) $x = \frac{\alpha I \cdot (I - \gamma)}{I + \beta}$.)

11.4. Curve sketching and function plotting scheme

Let's give **curve sketching and function plotting scheme** below.

1. Find the domain of the function.
2. Find the break points of the function.
3. Find the intervals of functions increasing and decreasing.
4. Find the minimums and maximums
5. Find the direction of convexity of the function graph, inflection point.
6. Find the asymptotes.

In addition, we might consider the parity (or oddness) of the function, its periodicity, the points of intersection of the graph with the coordinate axes

Based on the curve sketching the graph is plotted. It might be handy to outline the elements of the graph in parallel with the curve sketching.

***Example* 11.4.** Perform curve sketching on $y = \frac{x^3}{2(x-1)^2}$ and plot the graph.

Solution.

1. Domain of the function: $(-\infty, 1) \cup (1, +\infty)$, i.e. $x \neq 1$.
2. $x = 1$ – second degree break points because

$$\lim_{x \to 1-} \frac{x^3}{2 \cdot (x-1)^2} = \lim_{x \to 1+} \frac{x^3}{2 \cdot (x-1)^2} = +\infty .$$

3. Calculate the derivative

$f'(x) = \frac{3x^2(x-1)^2 - 2x^3(x-1)}{2\,(x-1)^4} = \frac{3x^2(x-1) - 2x^3}{2\,(x-1)^3} = \frac{x^3 - 3x^2}{2\,(x-1)^3} = \frac{x^2(x-3)}{2\,(x-1)^3}.$

Find the areas of increasing and decreasing functions:

$x \in (-\infty,\ 1) \Rightarrow f'(x) > 0 \Rightarrow$ function increasing;

$x \in (1,\ 3) \Rightarrow f'(x) < 0 \Rightarrow$ function decreasing;

$x \in (3,\ \infty) \Rightarrow f'(x) > 0 \Rightarrow$ function increasing.

4. Equating the derivative to zero, we find the critical point $x = 3$. The derivative changes sign from minus to plus at the point $x = 3$ ($f'(x) < 0$ when $1 < x < 3$; $f'(x) > 0$ when $x > 3$). Thus, there is a minimum $f_{\min} = f(3) = \frac{27}{8}$ at the point $x = 3$.

5. Calculate the second derivative:

$$f''(x) = \frac{(3x^2 - 6x)(x-1)^3 - 3(x^3 - 3x^2)(x-1)^2}{2(x-1)^6} =$$

$$= \frac{(3x^2 - 6x)(x-1) - 3(x^3 - 3x^2)}{2(x-1)^4} = \frac{3x}{(x-1)^4}.$$

Define the direction of the convexity and the inflection point:

$x < 0 \Rightarrow f''(x) < 0 \Rightarrow$ upward convexity;

$x > 0 \Rightarrow f''(x) > 0 \Rightarrow$ downward convexity;

$x = 0 \Rightarrow f''(x) = 0 \Rightarrow (0, 0)$ is inflection point.

6. Find the asymptotes. Obviously, $x = 1$ is the vertical asymptote. Find the oblique asymptote:

$k = \lim\limits_{x \to \pm\infty} \frac{f(x)}{x} = \lim\limits_{x \to \pm\infty} \frac{x^3}{2x\,(x-1)^2} = \lim\limits_{x \to \pm\infty} \frac{x^2}{2\,(x-1)^2} = \frac{1}{2},$

$b = \lim\limits_{x \to \pm\infty} \left[\frac{x^3}{2\,(x-1)^2} - \frac{1}{2}x\right] = \lim\limits_{x \to \pm\infty} \frac{2x^2 - x}{2\,(x-1)^2} = 1.$

So, $y = \frac{1}{2}x + 1$ is oblique asymptote.

The graph of the considered function is shown in Fig. 11.2.

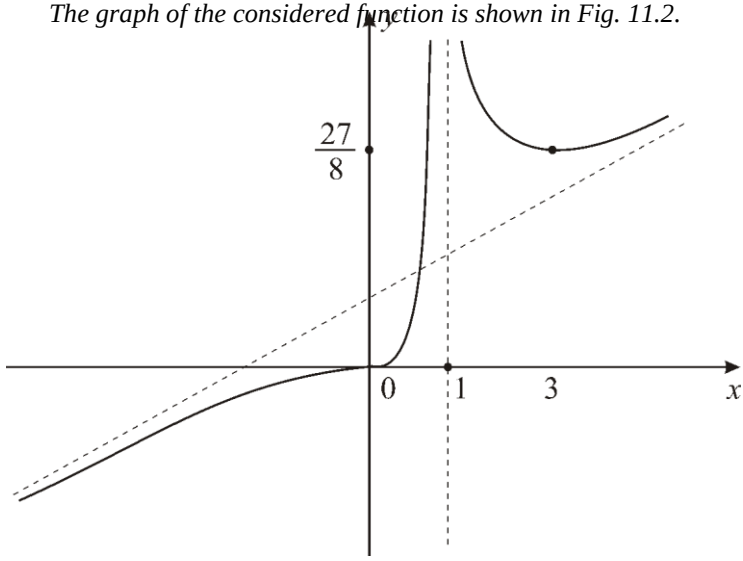


Fig. 11.2. *function* $y = \frac{x^3}{2\,(x-1)^2}$

***Example* 11.5.** In probability theory and statistics, a differential function of the normal distribution plays a very important role:

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}.$$

Let's perform curve sketching by the methods of differential calculus using the scheme above and plot its graph. Note that this graph is called the normal curve (Gaussian curve)

Solution.

1. Domain of the function is the *Ox* axis.
2. Function is continuous on the *Ox* axis.
3. Calculate the first derivative:

$$f'(x)=\frac{1}{\sigma\sqrt{2\pi}}\mathrm{e}^{-\frac{(x-a)^2}{2\sigma^2}}\left(-\frac{(x-a)^2}{2\sigma^2}\right)'=-\frac{x-a}{\sigma^3\sqrt{2\pi}}\mathrm{e}^{-\frac{(x-a)^2}{2\sigma^2}}.$$

Obviously, $f'(x)>0$ while $x<a$, $f'(x)<0$ when $x>a$. Therefore, in the interval $(-\infty, a)$ the function increases, and in the interval $(a, +\infty)$, it decreases

4. Equating the derivative to zero, we find the critical point $x=a$. At the point $x=a$, the derivative changes sign from plus to minus, therefore, it has a maximum

$$f_{\max}=f(a)=\frac{1}{\sigma\sqrt{2\pi}}.$$

5. Calculate the second derivative:

$$f''(x)=-\frac{1}{\sigma^3\sqrt{2\pi}}\mathrm{e}^{-\frac{(x-a)^2}{2\sigma^2}}-\frac{x-a}{\sigma^3\sqrt{2\pi}}\mathrm{e}^{-\frac{(x-a)^2}{2\sigma^2}}\left(-\frac{(x-a)^2}{2\sigma^2}\right)=$$

$$=-\frac{1}{\sigma^3\sqrt{2\pi}}\mathrm{e}^{-\frac{(x-a)^2}{2\sigma^2}}\left(1-\frac{(x-a)^2}{\sigma^2}\right).$$

Ehe second derivative is zero when

$$1-\frac{(x-a)^2}{\sigma^2}=0,$$

i.e. when $x=a+\sigma$ and $x=a-\sigma$.

Next

$x\in(-\infty, a-\sigma)\Rightarrow f''(x)>0\Rightarrow$ downward convexity;

$x\in(a-\sigma, a+\sigma)\Rightarrow f''(x)<0\Rightarrow$ downward convexity;

$x\in(a+\sigma, +\infty)\Rightarrow f''(x)>0\Rightarrow$ downward convexity.

When passing through points $x = a + \sigma$, $x = a - \sigma$, the second derivative changes sign. The value of the function at both of these points is the same:

$$f(a+\sigma) = f(a-\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}} = \frac{1}{\sigma\sqrt{2\pi e}}.$$

Thus, inflection point is

$\left(a-\sigma, \frac{1}{\sigma\sqrt{2\pi e}}\right)$ и $\left(a+\sigma, \frac{1}{\sigma\sqrt{2\pi e}}\right)$.

6. There are obviously no vertical asymptotes. The limit of the function when $x \to \pm\infty$ equals to zero:

$$\lim_{x\to+\infty} f(x) = \lim_{x\to-\infty} f(x) = 0.$$

Therefore, the *O* axis is the horizontal asymptote of the graph (obviously, $\lim\limits_{x\to+\infty} \frac{f(x)}{x} = 0$ there are no inclined asymptotes).

While plotting, we additionally take into account that for all values of the argument $f(x) > 0$, i.e. the curve is located above the *Ox* axis, as well as the fact that the curve is symmetric with respect to the straight line $x = a$ (Fig. 11.3), since the difference $x - a$ in the analytical expression of the function is squared.

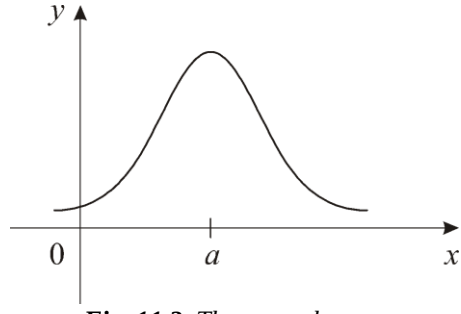


Fig. 11.3. *The normal curve*

11.5. FINDING MAXIMUM AND MINIMUM USING HIGH ORDER DERIVATIVES

If at some point x_0 both the first and second derivatives equal zero: $f'(x_0)=0$, $f''(x_0)=0$, then at this critical point there can be either a maximum or a minimum, or there is neither one nor the other. In this case, higher derivatives can be used.

Let the function $f(x)$ have derivatives up to the *n* order $f^{(n)}(x)$ in a neighborhood of a point $x=x_0$ and be continuous. Let all derivatives up to the $(n-1)$ order inclusively at this point equal zero:

$$f'(x_0)=f''(x_0)=\ldots=f^{(n-1)}(x_0)=0, \qquad (*)$$

and $f^{(n)}(x_0)\neq 0$. Represent the difference $f(x)-f(x_0)$ in powers of the difference $x-x_0$ using Taylor formula with the remainder term in the form of Peano:

$$f(x)-f(x_0)=\frac{f'(x_0)}{1!}(x-x_0)+\frac{f''(x_0)}{2!}(x-x_0)^2+\cdots+$$

$$+\frac{f^{(n-1)}(x_0)}{(n-1)!}(x-x_0)^{n-1}++\frac{f^{(n)}(x_0)}{n!}(x-x_0)^n+o((x-x_0)^n).$$

Here $o\left((x-x_0)^n\right)$ is $\alpha(x)(x-x_0)^n$, where $\alpha(x)\to 0$ where $x\to x_0$. Moreover, according to (*) first $(n-1)$ terms on the right side of the last equality vanish. Therefore

$$f(x)-f(x_0)=\frac{f^{(n)}(x_0)}{n!}(x-x_0)^n+\alpha(x)\ (x-x_0)^n.$$

Let $\alpha(x)=\frac{\beta(x)}{n!}$. Obviously, $\beta(x)$ infinitesimal when $x\to x_0$: $\lim\limits_{x\to x_0}\beta(x)=0$.

We obtain

$$f(x) - f(x_0) = \frac{f^{(n)}(x_0)+\beta(x)}{n\,!}(x - x_0)^n. \qquad (**)$$

Because $\beta(x) \to 0$ when $x \to x_0$, then for values of *x*, sufficiently close to x_0, the sign of the sum $f^{(n)}(x_0)+\beta(x)$, in the numerator coincides with the sign $f^{(n)}(x_0)$ for both $x < x_0$ and $x > x_0$. $x > x_0$. Let's consider two cases.

1) *n* is an odd number, $n = 2k + 1$. Then, when passing from *x* values smaller than x_0, to *x* values larger than x_0, the expression $(x - x_0)^{2k+1}$ will change its sign to the opposite:

$(x - x_0)^{2k+1} < 0$ when $x < x_0$,

$(x - x_0)^{2k+1} > 0$ when $x > x_0$.

In this case, the sign of the first factor in (**), coinciding with the sign of $f^{(n)}(x_0)$, will not change. Thus, the sign of the difference $f(x) - f(x_0)$ will change. Therefore, at a point x_0, the function $f(x)$ cannot have an extremum, since near this point it takes values both less than $f(x_0)$ and greater than $f(x_0)$;

2) *n* is an even number, $n = 2k$. In this case, the difference $f(x) - f(x_0)$ does not change sign when passing from *x* smaller than x_0 to values greater than x_0, since, obviously $(x - x_0)^{2k} > 0$, for all values of *x*. Obviously, near x_0 both left and right, the sign of the difference $f(x) - f(x_0)$ coincides with the sign of $f^{(n)}(x_0)$. Therefore, if $f^{(n)}(x_0) > 0$, then $f(x) > f(x_0)$ in some neighborhood of the point x_0,

therefore, the function $f(x)$ has a minimum at the point x_0; if $f^{(n)}(x_0)<0$, then the function has a maximum.

We introduced the next **rule**: if when $x=x_0$

$$f'(x_0)=f''(x_0)=\ldots=f^{(n-1)}(x_0)=0, \qquad (*)$$

And $f^{(n)}(x_0)\neq 0$ when n is odd, then $f(x)$ has neither maximum or minimum at the point $x=x_0$.

If the first derivative that is not equal to zero at the point x_0 is a derivative of even order, then the function has an extremum at the point x_0; maximum if $f^{(n)}(x_0)<0$, and minimum if $f^{(n)}(x_0)>0$.

***Example* 11.6.** Find maximum and minimum of the following function $f(x)=x^4+8x^3+24x^2+24x$.

Solution. Find the critical points:

$$f'(x)=4x^3+24x^2+48x+24=4\left(x^3+6x^2+12x+8\right).$$

From $4\left(x^3+6x^2+12x+8\right)=0$ we obtain critical point $x=-2$.

Consider the values of the derivatives at the point $x=-2$:

$f''(x)=12x^2+48x+48$, $f''(-2)=0$;

$f'''(x)=24x+48$, $f'''(-2)=0$

$f^{(4)}(x)=24>0$.

Thus, $f(x)$ has minimum at $x=-2$.

***Example* 11.7.** Find the extremum of the following function $f(x)=e^x-e^{-x}-2\sin x$.

Solution. Let's calculate the derivative:

$f'(x) = e^{x} - e^{-x} - 2\sin x$.

Obviously, the point $x = 0$ is critical: $f'(0) = 0$. Next:

$f''(x) = e^{x} - e^{-x} + 2\sin x$, $f''(0) = 0$;

$f'''(x) = e^{x} + e^{-x} + 2\cos x$, $f'''(0) = 4$.

Here, the first derivative that does not vanish at a critical point has a third, i.e. it is odd order. Therefore, at this critical point there is no extremum.

Questions

1. Let x_0 – critical point of $y = f(x)$ and let $f''(x_0) = 0$. Is there an extremum at a point?
2. Let the graph of the function $y = f(x)$ have a convexity directed upwards. Where is the convexity of the curve $y = \lambda f(x)$ directed: a) when $\lambda > 0$, б) when $\lambda < 0$?
3. Let the graph of the function $y = f(x)$ have three inflection points x_1, x_2 и x_3 ($x_1 < x_2 < x_3$) on (a, b) and let$y = f(x)$ is convex curve on (a, x_1). Is this curve convex or concave on (x_3, b)?
4. Let $f''(x_0) = 0$. Is point x_0 the inflection?
5. Let $y = f(x)$ have a horizontal asymptote for $x \to +\infty$. What is the limit $\lim\limits_{x\to+\infty} \frac{f(x)}{x}$?
6. Can a graph of a function $y = f(x)$ have two different oblique asymptotes?
7. How to find an extremum of the function $y = f(x)$ at a point x_0, if $f'(x_0) = 0$ and $f''(x_0) = 0$?

Chapter 12. Derivative applications in economic theory

12.1. Profit Maximization

Consider the economic interpretation of Fermat's theorem.

Let $S = S(x)$ be a function of costs, $D = D(x)$ be a function of income, $P = P(x)$ be a function of profit. Then $P(x) = D(x) - S(x)$. The optimal level of production is the level such that the profit $P(x)$ is maximum, i.e. the value of output x_0 at which the profit function $P(x)$ has a maximum. By virtue of Fermat's theorem, at this point $x = x_0$ the derivative is equal to zero $P'(x_0) = 0$:. But $P'(x) = D'(x) - S'(x)$ therefore

$$D'(x_0) = S'(x_0). \qquad (*)$$

The derivative $S'(x)$ expresses marginal costs MS , and the derivative $D'(x)$ expresses marginal revenue MD . Thus, equality (*) obtained using Fermat's theorem takes the form:

$MS(x_0) = MD(x_0)$.

The last equality is an expression of one of the basic laws of microeconomics*: maximum profit is achieved when the marginal cost and marginal revenue are equal.*

12.2. Elasticity

Now let's consider the logarithmic derivative and its applications. Let the function $y = f(x)$ be positive and differentiable at the point *x*. As it was noted, in particular, in deriving the derivative formula for an exponential function (see § 8.5), the derivative of the function $\ln y = \ln f(x)$ has the form

$$[\ln f(x)]' = \frac{1}{f(x)} f'(x) = \frac{f'(x)}{f(x)}, \text{ or } (\ln y)' = \frac{y'}{y}.$$

This expression is called the **logarithmic derivative** of the function $f(x)$. The logarithmic derivative is also called the **rate of change** T_y of the function *y*:

$$T_y = (ln\, y)' = \frac{y'}{y}. \qquad (12.1)$$

Let $S = S(t)$ be the value of the contribution at a time t. Let us find out whether it is possible to approximately determine the nominal annual rate of bank *i* interest by function $S(t)$. If interest is accrued once per period of time Δt, then interest for the specified period will amount to $Si\Delta t$ (here Δt is the share of the year). Then the increment of the deposit and the interest on the deposit are one and the same $\Delta S = Si\Delta t.$, then

$$i = \frac{\Delta S}{S\Delta t}$$

If $S(t)$ is a differentiable function, then we can replace the increment ΔS with a differential $dS = S'\Delta t.$ We obtain

$$i \approx \frac{S'\Delta t}{S\Delta t} = \frac{S'}{S} = (\ln S)'.$$

And this means that the bank interest rate i coincides with the logarithmic derivative of the contribution.

In many problems, the concept of *elasticity* of a function is used.

Definition. The **elasticity** $E_x(y)$ of a function $y = f(x)$ is the limit of the ratio of the relative increment of the function *y* to the relative increment of the argument *x* for $\Delta x \to 0$:

$$E_x(y) = \lim_{\Delta x \to 0} \left(\frac{\Delta y}{y} : \frac{\Delta x}{x} \right) = \frac{x}{y} \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \frac{x}{y} \cdot y' = x \cdot \frac{y'}{y}. \quad (12.2)$$

The elasticity of the function approximately expresses the percentage change in the function $y = f(x)$ when the argument *x* changes by 1%.

From the formula (12.2) it follows that the *elasticity of the function is equal to the product of the independent variable* **x** by the rate of change of the function T_y:

$$E(y) = E_x(y) = xT_y. \quad (12.3)$$

Note the **elasticity properties**:

$$E(uv) = E(u) + E(v), \quad (12.4)$$

$$E\left(\frac{u}{v}\right) = E(u) - E(v), \quad (12.5)$$

obviously following from the corresponding properties of logarithms.

The elasticity of the function is used in the analysis of supply and demand. Let $D = D(p)$ be a function of demand on the price of goods *p*.

The **elasticity of demand** relative to price is determined by the ratio:

$$E = \frac{\text{Процентное изменение спроса}}{\text{Процентное изменение цены}}. \quad (11.6)$$

Percent change in demand is $\frac{\Delta D}{D}\cdot 100$, and percent change in price is $\frac{\Delta p}{p}\cdot 100$. Therefore

$$E=\left(\frac{\Delta D}{D}\cdot 100\right):\left(\frac{\Delta p}{p}\cdot 100\right),$$

or

$$E=\frac{p}{D}\frac{\Delta D}{\Delta p}. \qquad (12.7)$$

With a continuous dependence ΔD on Δp the difference ratio in the expression (20.7) is replaced by the limit at $\Delta p\to 0$:

$$E(D)=p\frac{D'(p)}{D(p)}. \qquad (12.8)$$

Due to the fact that the demand function $D=D(p)$ is a decreasing function of price (see Fig. 7.14), its derivative is negative and the elasticity of demand is also *negative*. (Some authors define elasticity as a positive value, putting a minus sign in front of the right side of formulas (20.6) - (20.8).)

There are **three types of demand**:

1) *elastic*, if $|E(D)|>1$;

2) *neutral*, if $|E(D)|=1$;

3) *inelastic*, if $|E(D)|<1$.

***Example* 12.1.** The demand function has the form $D=\sqrt{240-p}$. Find the elasticity of demand at a price $p=176$.

Solution: $E(D) = p \cdot \frac{D'}{D} = p \frac{1}{-2(240-p)}$. Using $p = 176$, we obtain $E(D) = -\frac{176}{128} = -1{,}375$;

$|E(D)| = 1{,}375 > 1$ – demand is elastic.

Similarly, the concept of **elasticity of supply** is introduced as the ratio of the percentage change in supply to the percentage change in price. Since the function of the proposal $S = S(p)$ is increasing (see. Fig. 7.15), then

$$E(S) = p \frac{S'}{S}$$

there is a *positive* value.

12.3. Optimization of taxation

Let t be the tax per unit of output, $S = S(x)$ is the cost function, $D = D(x)$ is the income function, $P = P(x)$ is the profit function. Then the profit function has the form:

$$P(x) = D(x) - S(x) - tx.$$

For instance, let the price of products $v(x) = a - bx$, i.e. linearly decreases with increasing volume of production, and the cost function has the form $S(x) = x^2 + c$. Here *a, b, c* are some positive constants. The profit function in this case has the form:

$$P(x) = x(a - bx) - x^2 - c - tx.$$

To maximize profits, the company needs the optimal output. The condition for maximum profit:,

$P'(x)=0$, or. $a-t-2bx-2x=0$, next

$$x_0=\frac{a-t}{2b+2}.$$

With this value of the volume of production, the total tax *T* has the form $T=\frac{t\,(a-t)}{2\,(b+1)}$. The interests of the state are that the value of *T* be maximum. We differentiate *T* and, equating the derivative to zero: $T'=0$, $a-2t=0$, we obtain

$$t_0=\frac{a}{2}.$$

We consider this problem for specific numerical values of the constants *a*, *b*, and *c*. Let be $a=80$, $b=1$, $c=10$. Then $t_0=40$, $x_0=10$. With these values, the maximum value of profit $P_0=190$, and state revenue $T_0=t_0x_0=400$. (Note that in the absence of taxes, maximum profit would be achieved with twice as much production $x_0=20$ and would be $P_0=790$.)

Questions

1. At what ratio between marginal cost and marginal revenue the maximum profit is achieved?
2. What is the rate of change of function?
3. What is called function elasticity? What is the connection between the elasticity and the rate of change of function?
4. How to determine the elasticity of demand relative to price?
5. When the demand is considered elastic?
6. How to define the concept of elasticity of supply? Is a supply elasticity positive or negative?

INTEGRAL.

Chapter 13. Indefinite integral. Integration methods.

13.1. Antiderivative and indefinite integral.

Definition. Function $y = F(x)$ is called an antiderivative for function $f(x)$ in between *X* if for any $x \in X$ holds true:

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

For example:

Function $F(x) = \ln x$ is an antiderivative for function $f(x) = \frac{1}{x}$ on an infinite interval $(0, 0, +\infty)$, so for any *x* from this interval the equality will be $(\ln x)' = \frac{1}{x}$;

Function $F(x) = \arcsin x$ is an antiderivative for function $f(x) = \frac{1}{\sqrt{1-x^2}}$ $f(x)$ in between $(-1, 1)$, so at each point of this interval, the equality is: $(\arcsin x)' = \frac{1}{\sqrt{1-x^2}}$.

Obviously, if for this function $f(x)$ there exists an antiderivative, then this antiderivative is not the one; if $F(x)$ – antiderivative for $f(x)$ and *C* – random constant, then $F(x)+C$ is antiderivative for $f(x)$.

Theorem 13.1. If $f(x)$ is differentiable between X and if $f'(x)=0$ throughout this gap, then $f(x)$ is constant along *X* segment.

Proof. It is enough to prove that for any two different points x_1, $x_2 \in X$ equality $f(x_1)=f(x_2)$ is true. As function $f(x)$ is differentiable on X, a $(x_1,x_2)\subset X$, then $f(x)$ differentiable (and is continuous) over the entire interval $[x_1,x_2]$, and we can apply the Lagrange theorem to function $f(x)$ for $[x_1, x_2]$ (see. § 9.1), according to which inside this segment there is a point *c*: $f(x_2)-f(x_1)=f'(c)(x_2-x_1)$.

$f'(x)=0$ at any point, then in particular $f'(c)=0$, consequently, $f(x_1)=f(x_2)$, Q.E.D.

***Theorem* 13.2.** If $F(x)$ is antiderivative for a function $f(x)$ in between *X*, then any other antiderivative for $f(x)$ at *X* can be represented as $F(x)+C$, where *C* is a number.

Proof. Let $F'(x)=f(x)$ and $\Phi'(x)=f(x)$ for all $x\in X$. Then $[\Phi(x)-F(x)]'=0$.

From the above proved Theorem 13.1 it follows that $\Phi(x)-F(x)=C=\text{const}$, i.e. $\Phi(x)=F(x)+C$, Q.E.D..

Definition. If $F(x)$ is an antiderivative for $f(x)$, equation $F(x)+C$ where *C* is a random constant is called an indefinite integral for function $f(x)$ and signed as $\int f(x)dx$.

So, if $F'(x)=f(x)$, then

$$\int f(x)dx = F(x)+C. \qquad (13.1)$$

The indefinite integral for function $f(x)$ is the totality of all primitives for function $f(x)$.

In equation (13.1), the sign $\int$ is called an integral sign, $f(x)$ – *integrand function* $f(x)dx$ - an integrand, *C – constant integration.*

The operation of seeking the indefinite integral of a given function is called the **integration** of this function. Integration is the inverse of differentiation. The correctness of integration is verified by differentiation.

Properties of the indefinite integral:

1. $\left(\int f(x)dx\right)' = f(x)$,

i.e. *the derivative of the indefinite integral is equal to the integrand.*

2. $d\int f(x)dx = f(x)dx$,

i.e. *the differential from the indefinite integral is equal to the integrand.*

3. $\int dF(x) = F(x)+C$.

These properties automatically follow from the definition of the integral.

Integration Rules:

I. If $C = const \neq 0$, then $\int cf(x)dx = c\int f(x)dx$, *i.e. the constant factor can be taken out from under the integral sign.*

II. $\int [f(x) \pm g(x)]dx = \int f(x)dx \pm \int g(x)dx$, *i.e. the indefinite integral of the algebraic sum of two functions is equal to the algebraic sum of the integrals of these functions.*

III. If $F'(x) = f(x)$,then $\int f(ax+b)dx = \frac{1}{a}F(ax+b) + C$.

Rules I and II follow from the corresponding differentiation rules (see § 8.4). We verify the validity of equality III:

$\left(\frac{1}{a}F(ax+b) + C\right)' = \frac{1}{a}f(ax+b) \cdot (ax+b)' =$

$= \frac{1}{a}f(ax+b) \cdot a = f(ax+b).$

Integral table

1. $\int 0 \cdot dx = C.$

2. $\int x^{\alpha}dx = \frac{x^{\alpha+1}}{\alpha+1} + C(\alpha \neq -1).$

3. $\int \frac{dx}{x} = ln\ |x| + C.$

4. $\int a^x dx = \frac{a^x}{ln\, a} + C.$ **4′.** $\int e^x dx = e^x + C.$

5. $\int cos\, x\, dx = sin\, x + C.$ **6.** $\int sin\, x\, dx = -cos\, x + C.$

7. $\int \frac{dx}{cos^2 x} = \text{tg}\ x + C.$ **8.** $\int \frac{dx}{\sin^2 x} = -ctg\ x + C.$

9. $\int \frac{dx}{\sqrt{1-x^2}} = arcsin\, x + C.$ **9′.** $\int \frac{dx}{\sqrt{a^2-x^2}} = arcsin\frac{x}{a} + C.$

10. $\int \frac{dx}{1+x^2} = arctg\ x + C\ .$ **10′.** $\int \frac{dx}{a^2+x^2} = \frac{1}{a} \cdot arctg\ \frac{x}{a} + C.$

11. $\int \text{ctg}\ xdx = ln\ |sin\, x| + C.$

12. $\int \text{tg}\ xdx = -ln\ |cos\, x| + C.$

13. $\int \frac{dx}{\sqrt{x^2+A}} = ln\ \left|x + \sqrt{x^2+A}\right| + C.$

14. $\int \frac{dx}{x^2-a^2} = \frac{1}{2a}ln\ \left|\frac{x-a}{x+a}\right| + C(a \neq 0).$

Most of the formulas given in the table directly follow from the table of derivatives (see § 8.5). The validity of any of the formulas is easily verified by differentiating the right-hand side.

Check the formula 10 '. By the rule of differentiation of a complex function $\left(\frac{1}{a}\operatorname{arctg}\ \frac{x}{a}+C\right)'=\frac{1}{a}\frac{1}{1+\left(\frac{x}{a}\right)^2}\left(\frac{x}{a}\right)'=\frac{1}{a}\frac{1}{1+\frac{x^2}{a^2}}\frac{1}{a}=\frac{1}{a^2+x^2}$.

Formula 9 ′ is verified in a similar way.

Check formula 11. If $\sin x>0$, then

$$\left(\ln\left|\sin x\right|+C\right)'=\left(\ln\sin x\right)'=\frac{1}{\sin x}\left(\sin x\right)'=\frac{\cos x}{\sin x}=\operatorname{ctg} x$$

.

If $\sin x<0$, then

$$\left(\ln\left|\sin x\right|+C\right)'=\left(\ln\left(-\sin x\right)\right)'=\frac{1}{-\sin x}\left(-\sin x\right)'=\frac{-\cos x}{-\sin x}=\operatorname{ctg} x$$

.

Formula 12 is checked similarly.

We verify formula 13. Let $x+\sqrt{x^2+A}>0$. Then

$\left(ln\ \ \left|x+\sqrt{x^2+A}\right|\right)'=\left[ln\ \ \left(x+\sqrt{x^2+A}\right)\right]'=$

$=\frac{1}{x+\sqrt{x^2+A}}\left(x+\sqrt{x^2+A}\right)'=\frac{1}{x+\sqrt{x^2+A}}\left(1+\frac{x}{\sqrt{x^2+A}}\right)=$

$\frac{1}{x+\sqrt{x^2+A}}\frac{\sqrt{x^2+A}+x}{\sqrt{x^2+A}}=\frac{1}{\sqrt{x^2+A}}$.

The case is treated similarly$x+\sqrt{x^2+A}<0$. Formula 14 is also verified similarly.

13.2. Basic integration methods

Direct integration

The calculation of integrals based on the application of formulas 1–14 and rules I – III is called direct integration. Consider the following **examples**:

1. $\int \left(6x^2 + 2\cos x - \frac{3}{x}\right) dx = 6\int x^2 dx + 2\int \cos x\, dx - 3\int \frac{dx}{x} =$
$= 2x^3 + 2\sin x - 3\ln\ |x| + C.$

It should be noted that at the end of the solution one general constant C is written, without writing out the constants from the integration of the individual terms.

2. $\int x\sqrt{x}dx = \int x^{\frac{3}{2}}dx = \frac{x^{\frac{5}{2}}}{\frac{5}{2}} + C = \frac{2x^2\sqrt{x}}{5} + C.$

3. $\int \frac{3x^2+1}{x^2(x^2+1)}dx = \int \frac{x^2+1+2x^2}{x^2(x^2+1)}dx = \int \frac{x^2+1}{x^2(x^2+1)}dx + \int \frac{2x^2}{x^2(x^2+1)}dx =$
$= \int \frac{dx}{x^2} + 2\int \frac{dx}{x^2+1} = -\frac{1}{x} + 2\,\text{arctg}\ x + C.$

Substitution method (variable replacement method).

Replacing the integration variable is one of the most common and effective methods of reducing an indefinite integral to a combination of tabular ones.

Let the integral be given $\int f(x)dx$. We introduce a new variable by the formula $x = \varphi(t)$, where $\varphi(t)$ – differentiable function/ Then we substitute these expressions into the integral:

$$\int f(x)dx = \int f(\phi(t))\phi'(t)dt. \qquad (13.2)$$

Formula (13.2) is called the variable **replacement** formula.

***Example* 13.1.** Find: a) $\int \frac{\sqrt{x}}{x+1}dx$; б) $\int x(x-1)^9 dx$.

Decision. a) Let's make a variable change: $x = t^2$, $t = \sqrt{x}$. Then $dx = 2tdt$.

$$\int \frac{\sqrt{x}}{x+1}dx = \int \frac{t \cdot 2t\, dt}{t^2+1} = 2\int \frac{t^2}{t^2+1}dt = 2\int \frac{t^2+1-1}{t^2+1}dt = 2\left(\int dt - \int \frac{dt}{t^2+1}\right) =$$

$$= 2(t - \operatorname{arctg} t) + C = 2\left(\sqrt{x} - \operatorname{arctg}\sqrt{x}\right) + C.$$

b) Put $x = t+1$. Then $x-1 = t$, $dx = dt$.

$$\int x(x-1)^9 dx = \int (t+1)t^9 dt = \int (t^{10} + t^9)\, dt =$$

$$= \frac{t^{11}}{11} + \frac{t^{10}}{10} + C = \frac{(x-1)^{11}}{11} + \frac{(x-1)^{10}}{10} + C.$$

Often, changing a variable is not done in the form $x = \varphi(t)$ but $t = \psi(x)$.

***Example* 13.2.** Find: a) $\int \sin^2 x \cos x\, dx$; б) $\int \frac{2x\, dx}{1+x^4}$.

Decision. a) Make a replacement $t = \sin x$. Then $dt = \cos x\, dx$. We get

$$\int \sin^2 x \cos x\, dx = \int t^2 dt = \frac{t^3}{3} + C = \frac{\sin^3 x}{3} + C.$$

b) Put $t = x^2$. Then $dt = 2x\, dx$. We get

$$\int \frac{2xdx}{1+x^4} = \int \frac{dt}{1+t^2} = \operatorname{arctg} t + C = \operatorname{arctg} x^2 + C.$$

Note that a new variable can and not be written out explicitly. In such cases, they talk about summing up under the sign of the differential.

In particular, the calculation of the integral in Example 13.2 can be written in the following form:

$$\int \frac{2x\,dx}{1+x^4} = \int \frac{d\left(x^2\right)}{1+\left(x^2\right)^2} = \operatorname{arctg} x^2 + C$$

.

***Example* 13.3.** $\int \frac{\ln x\,dx}{x} = \int \ln x\, d\left(\ln x\right) = \frac{\ln^2 x}{2} + C$.

Part Integration

Let $u = u(x)$ and $v = v(x)$ – differentiable functions. Then $d(uv) = vdu + udv$, or

$$udv = d(uv) - vdu.$$

Integrating both sides of the last equality, we obtain the **integration formula by parts**:

$$\int udv = uv - \int vdu \quad (13.3)$$

***Example* 13.4.** Find: a) $\int xe^x dx$; б) $\int (2x+3)\cos x\,dx$; в) $\int x \ln x dx$

.

Decision. a) Let $x = u$, $e^x dx = dv$. Then $du = dx$, $u = e^x$. By the formula (21.3) we obtain.

$$\int xe^x dx = xe^x - \int e^x dx = xe^x - e^x + C = e^x(x-1) + C$$

.

b) Let $u = 2x+3$, $\cos x\,dx = dv$. Then $du = 2dx$, $v = \sin x$. We get

$$\int (2x+3)\cos x\,dx = (2x+3)\sin x - 2\int \sin x\,dx = (2x+3)\sin x + 2\cos x + C$$

.

c) Let $u = \ln x$, $dv = x\,dx$. Then $du = \frac{dx}{x}$, $v = \frac{x^2}{2}$. And get

$$\int x \ln x\, dx = \frac{x^2}{2}\ln x - \int \frac{x^2}{2}\cdot\frac{1}{x}dx = \frac{x^2}{2}\ln x - \frac{1}{2}\int x dx = \frac{x^2}{2}\ln x - \frac{x^2}{4} + C$$

.

In some cases, the integration formula is applied in parts several times, gradually simplifying the integrand.

***Example* 13.5.** Find $\int x^2 \cos x\, dx$.

Decision. Let $u = x^2$, $\cos x\, dx = dv$. Then $du = 2xdx$, $v = \sin x$. Получаем

$$\int x^2 \cos x dx = x^2 \sin x - \int 2x \sin x dx.$$

The resulting integral is not tabular, but it is simpler than the original, the degree of the variable x has decreased. We reuse the integration formula in parts by setting $u = x$, $\sin x\, dx = dv$. Then $du = dx$, $v = -\cos x$. We get

$$\int x^2 \cos x\, dx = x^2 \sin x - \int 2x \sin x\, dx = x^2 \sin x - 2\left(-x\cos x + \int \cos x\, dx\right) =$$

$$= x^2 \sin x + 2x\cos x - 2\sin x + C.$$

We indicate the most common types of integrals, for finding which the integration formula by parts is applied.

1. $\int P_n(x)\mathrm{e}^{ax}dx$, $\int P_n(x)a^x dx$, $\int P_n(x)\sin ax\, dx$, $\int P_n(x)\cos ax\, dx$

.

2. $\int P_n(x)\ln x\, dx$, $\int P_n(x)\arcsin x\, dx$, $\int P_n(x)\arccos x\, dx$, $\int P_n(x)\operatorname{arctg} x\, dx$, $\int P_n(x)\operatorname{arcctg} x\, dx$.

Here $P_n(x)$ is polynomial of degree n.

For finding the integrals of the *first* group $P_n(x) = u$ (and the remaining factors are dv). For finding the integrals of the second group, $P_n(x)\,dx = dv$ (the remaining factors are taken as u).

Obviously, the integrals a), b) in Example 13.4 and the integral in Example 13.5 refer to the first type and the integral c) in Example 13.4 to the second.

Example 13.6. Calculate integral $\int e^x \cos x dx.$

Decision. Note that this integral does not apply to any of the two types mentioned. Let $u = e^x$. Then $dv = \cos dx.$ We have

$du = \left(e^x\right)' dx = e^x dx;\ \ v = \sin x.$

Using formula 13.3, we obtain

$$\int e^x \cos x\, dx = e^x \sin x - \int e^x \sin x\, dx. \qquad (*)$$

Applying formula 13.3 to the integral on the right-hand side, we again apply the method of integration by parts setting $u = e^x, dv = \sin x dx$ and $du = e^x, v = -\cos x.$ We get

$$\int e^x \sin x\, dx = -e^x \cos x + \int e^x \cos x\, dx. \qquad (**)$$

Knowing (*) and (**), we get

$\int e^x \cos x\, dx = e^x \sin x - (-e^x \cos x + \int e^x \cos x\, dx) =$

$= e^x \sin x + e^x \cos x - \int e^x \cos x\, dx.$

Move the integral from the right to the left so we get

$2\int e^x \cos x\, dx = e^x(\sin x + \cos x) + C.$

We divide both sides of the last equality at 2 and, given that C is a random constant, we obtain

$$\int e^x \cos x\, dx = \frac{e^x}{2}(\sin x + \cos x) + C.$$

It is easy to see that if we took u not as e^x, we would get the same result.

13.3. Integration of rational shots

Equation $\frac{P(x)}{Q(x)}$, where $P(x)$ and $Q(x)$ are polynomials that called r**ational fraction**. A rational fraction is called **correct** if the degree of the numerator is less than the degree of the denominator. If the degree of the numerator is greater than or equal to the degree of the denominator, then the fraction is called **incorrect**.

An irregular fraction can be represented as the sum of a polynomial and a regular fraction dividing the numerator by the denominator:

$$\frac{P(x)}{Q(x)} = R(x) + \frac{S(x)}{Q(x)}.$$

Here $R(x)$ – some polynomial, and the second term is a regular fraction.

For Example, $\frac{x^5 + 3x^4 + x^3 + 2x^2 + 3}{x^3 + x^2 - x + 1} = x^2 + 2x + \frac{3x^2 - 2x + 3}{x^3 + x^2 - x + 1}$.

In order to integrate the right fraction, it is decomposed at simple fractions, having previously expanded the denominator at the elementary factors.

Without proof, we give a **decomposition formula** for a regular fraction. Let the denominator $Q(x)$ be factorized $(x-a)^{\alpha}\left(x^2 + px + q\right)^{\beta}$.

Here $x = a$ – valid root $Q(x)$ multiplicities α, $x^2 + px + q$ – square trinomial with negative discriminant. Then the correct fraction is

decomposed at the sum of the elementary fractions using the so-called method of indefinite coefficients as follows:

$$\frac{P(x)}{(x-a)^{\alpha}\left(x^2+px+q\right)^{\beta}}=\frac{A_1}{(x-a)^{\alpha}}+\frac{A_2}{(x-a)^{\alpha-1}}+\ldots+\frac{A_{\alpha}}{x-a}+\ldots+$$

$$+\frac{M_1x+N_1}{\left(x^2+px+q\right)^{\beta}}+\frac{M_2x+N_2}{\left(x^2+px+q\right)^{\beta-1}}+\ldots+\frac{M_{\beta}x+N_{\beta}}{x^2+px+q},$$

where the coefficients are to be clarified in the process of fraction decomposition.

In connection with the above decomposition, it is necessary to consider the so-called **simple fractions**:

I. $\frac{A}{x-a}$.

II. $\frac{A}{(x-a)^{\alpha}}$.

III. $\frac{Mx+N}{x^2+px+q}$.

IV. $\frac{Mx+N}{\left(x^2+px+q\right)^{\beta}}$,

where a, p, q, A, M, N – real numbers; α, β – integers; in addition, it is assumed that the denominators of fractions III and IV do not have valid кщщеы, i.e. $\left(\frac{p}{2}\right)^2-q<0$.

Consider **the integrals of these simple fractions.**

Fractions of types I and II are easy to integrate:

$$\int\frac{A}{x-a}dx=A\ln|x-a|+C,$$

$$\int\frac{A}{(x-a)^{\alpha}}dx=-\frac{1}{\alpha-1}\cdot\frac{A}{(x-a)^{\alpha-1}}+C.$$

For calculating the integral of a fraction of type III from the trinomial in the denominator, a full square is extracted:

$$x^2+px+q=\left(x+\frac{p}{2}\right)^2+\left(q-\frac{p^2}{4}\right)=\left(x+\frac{p}{2}\right)^2+a^2;$$

$$\int\frac{Mx+N}{x^2+px+q}dx=\int\frac{\frac{M}{2}(2x+p)+\left(N-\frac{Mp}{2}\right)}{x^2+px+q}dx=\frac{M}{2}\int\frac{2x+p}{x^2+px+q}dx+$$

$$+\left(N-\frac{Mp}{2}\right)\int\frac{dx}{\left(x+\frac{p}{2}\right)^2+\left(q-\frac{p^2}{4}\right)}=\frac{M}{2}\ln\left(x^2+px+q\right)+$$

$$+\frac{2N-Mp}{\sqrt{4q-p^2}}\operatorname{arctg}\frac{2x+p}{\sqrt{4q-p^2}}+C.$$

***Example* 13.6.** Find $\int\frac{3x+4}{x^2+2x+5}dx$.

Decision. You can use the formula of the integral of a fraction of type III derived above, but we will once again repeat the process of its derivation at this specific example

$$\int\frac{3x+4}{x^2+2x+5}dx=\frac{3}{2}\int\frac{2x+2}{x^2+2x+5}dx+\int\frac{dx}{(x+1)^2+4}=\frac{3}{2}\int\frac{d\left(x^2+2x+5\right)}{x^2+2x+5}+$$

$$+\int\frac{dx}{(x+1)^2+4}=\frac{3}{2}\ln\left(x^2+2x+5\right)+\frac{1}{2}\operatorname{arctg}\frac{x+1}{2}+C.$$

IV. We proceed to the calculation of the integrals of a fraction of type IV, i.e. integrals $\int\frac{Mx+N}{\left(x^2+px+q\right)^\beta}dx$ when $\beta\geq 2$.

Apply the same substitution $x+\frac{p}{2}=t$ which is in the case of a fraction of type III:

$$\int \frac{Mx+N}{\left(x^2+px+q\right)^{\beta}}dx = \int \frac{Mt+\left(N-\frac{Mp}{2}\right)}{\left(t^2+a^2\right)^{\beta}}dt = \frac{M}{2}\int \frac{2tdt}{\left(t^2+a^2\right)^{\beta}} +$$
$$+\left(N-\frac{Mp}{2}\right)\int \frac{dt}{\left(t^2+a^2\right)^{\beta}}.$$

The first of the obtained integrals is taken by substitution $t^2+a^2=z$, $2tdt=dz$:

$$\int \frac{2tdt}{\left(t^2+a^2\right)^{\beta}} = \int \frac{dz}{z^{\beta}} = -\frac{1}{\beta-1}\frac{1}{z^{\beta-1}}+C = -\frac{1}{\beta-1}\frac{1}{\left(t^2+a^2\right)^{\beta-1}}+C.$$

The calculation of the second of the remaining integrals requires some effort. So, we need to calculate the integral

$$J_{\beta} = \int \frac{dt}{\left(t^2+a^2\right)^{\beta}} \quad (\beta = 1, 2, 3, \ldots).$$

We apply the integration formula in parts.

Let $u = \frac{1}{\left(t^2+a^2\right)^{\beta}}$, $dv = dt$; then $du = -\frac{2\beta tdt}{\left(t^2+a^2\right)^{\beta+1}}$, $v = t$.

We get

$$J_{\beta} = \int \frac{dt}{\left(t^2+a^2\right)^{\beta}} + 2\beta\int \frac{t^2}{\left(t^2+a^2\right)^{\beta+1}}dt. \qquad (*)$$

Convert last integral:

$$\int \frac{t^2}{(t^2+a^2)^{\beta+1}}dt = \int \frac{t^2+a^2-a^2}{(t^2+a^2)^{\beta+1}}dt = \int \frac{dt}{(t^2+a^2)^{\beta}} - a^2\int \frac{dt}{(t^2+a^2)^{\beta+1}} =$$
$$= J_{\beta} - a^2 J_{\beta+1}.$$

Substitute the last expression into equality (*):

$$J_\beta = \frac{t}{\left(t^2+a^2\right)^\beta} + 2\beta J_\beta - 2\beta a^2 J_{\beta+1}.$$

Express from here $J_{\beta+1}$:

$$J_{\beta+1} = \frac{1}{2\beta a^2}\frac{t}{(t^2+a^2)^\beta} + \frac{2\beta-1}{2\beta}\frac{1}{a^2}J_\beta. \qquad (**)$$

The resulting formula reduces the calculation of the integral

$$J_{\beta+1} = \int \frac{dt}{\left(t^2+a^2\right)^{\beta+1}}$$

To calculate

$$J_\beta = \int \frac{dt}{\left(t^2+a^2\right)^\beta}.$$

(Recall that b is a positive integer.)

Obviously

$$J_1 = \int \frac{dt}{t^2+a^2} = \frac{1}{a}\operatorname{arctg}\frac{t}{a} + C.$$

We take one of its values, namely, when $C=0$. By the formula (**) at $\beta=1$ find

$$J_2 = \int \frac{dt}{\left(t^2+a^2\right)^2} = \frac{1}{2a^2}\frac{t}{t^2+a^2} + \frac{1}{2a^3}\operatorname{arctg}\frac{t}{a}.$$

If now take $\beta=2$,then

$$J_3 = \int \frac{dt}{\left(t^2+a^2\right)^3} = \frac{1}{4a^2}\frac{t}{\left(t^2+a^2\right)^2} + \frac{3}{4a^2}J_2 =$$

$$= \frac{1}{4a^2}\frac{t}{\left(t^2+a^2\right)^2} + \frac{3}{8a^4}\frac{t}{t^2+a^2} + \frac{3}{8a^5}\operatorname{arctg}\frac{t}{a}$$

etc. Now it remains only to recall that $t = x + \frac{p}{2}$ and return to the variable x. Note that formulas of the form (**), which allow us to reduce the calculation $J_{\beta+1}$ to the calculation J_{β} with less than one sign, are called recurrence formulas. We will further encounter recurrence relations in linear algebra while studying determinants.

Now consider the integration of rational fractions that are not simple. As mentioned above, a rational fraction if it is not correct is transformed and represented as the sum of a polynomial and a regular fraction. Then the correct fraction is decomposed at simple fractions. The method of uncertain coefficients is used. After the fraction is represented as the sum of simple fractions, the integral of it is calculated as the sum of the integrals of these simple fractions. Consider this with examples.

***Example* 13.7.** Calculate integral $\int \frac{2x^3 - x^2 + 2x + 1}{x^4 - 2x^3 + 2x^2 - 2x + 1} dx$.

Decision. The integral function is the right fraction. We decompose it into simple fractions. For this, first expand the denominator at factors: $x^4 - 2x^3 + 2x^2 - 2x + 1 = (x-1)^2(x^2+1)$. . (This can be done, for example, by finding the root $x = 1$ and divide at $(x-1)$.) We apply the decomposition formula:

$$\frac{2x^3 - x^2 + 2x + 1}{(x-1)^2(x^2+1)} = \frac{A_1}{(x-1)^2} + \frac{A_2}{x-1} + \frac{Mx+N}{x^2+1}.$$

Multiply both sides of this equality by the denominator of the left side:

$$2x^3 - x^2 + 2x + 1 = A_1(x^2+1) + A_2(x-1)(x^2+1) + (Mx+N)(x-1)^2$$

.

After reduction of similar terms, we get

$$2x^3 - x^2 + 2x + 1 = (A_2 + M)x^3 + (A_1 - A_2 - 2M + N)x^2 +$$
$$+ (A_2 + M - 2N)x + A_1 - A_2 + N.$$

We equate the coefficients for the same powers of x on the left and on the right:

$$\left.\begin{array}{l|l} x^3 & A_2 + M = 2, \\ x^2 & A_1 - A_2 - 2M + N = -1, \\ x^1 & A_2 + M - 2N = 2, \\ x^0 & A_1 - A_2 + N = 1. \end{array}\right.$$

Solving the system, we find: $A_1 = 2$, $A_2 = 1$, $M = 1$, $N = 0$. So,

$$\frac{2x^3 - x^2 + 2x + 1}{(x-1)^2(x^2+1)} = \frac{2}{(x-1)^2} + \frac{1}{x-1} + \frac{x}{x^2+1}.$$

We integrate:

$$\int \frac{2x^3 - x^2 + 2x + 1}{x^4 - 2x^3 + 2x^2 - 2x + 1} dx = 2\int \frac{dx}{(x-1)^2} + \int \frac{dx}{x-1} + \int \frac{xdx}{x^2+1} =$$

$$= -\frac{2}{x-1} + \ln|x-1| + \frac{1}{2}\ln(x^2+1) + C.$$

***Example* 13.8. Calculate integral**

$$I = \int \frac{2x^6 - 5x^5 + 6x^4 - 10x^3 + 8x^2 - 3x + 15}{x^5 - 2x^4 + 2x^3 - 4x^2 + x - 2} dx.$$

Decision. The integral function is the wrong rational fraction. Dividing the numerator by the denominator, select the integer part:

$$2x^6 - 5x^5 + 6x^4 - 10x^3 + 8x^2 - 3x + 15 =$$
$$= (2x - 1)\ (x^5 - 2x^4 + 2x^3 - 4x^2 + x - 2) + 2x^2 + 2x + 13.$$

Consequently, the integrand function has the form

$$2x-1+\frac{2x^2+2x+13}{x^5-2x^4+2x^3-4x^2+x-2}.$$

The denominator of the remaining correct fraction is factorized:

$$x^5-2x^4+2x^3-4x^2+x-2=x^4(x-2)+2x^2(x-2)+x-2=$$

$$=(x-2)(x^4+2x^2+1)=(x-2)(x^2+1)^2.$$

(One could have done otherwise - by finding; find the root of the denominator $x=2$, and then divide the denominator by $(x-2)$.)

So,

$$I=\int(2x-1)\,dx+\int\frac{2x^2+2x+13}{(x-2)(x^2+1)^2}\,dx.$$

The first of these two integrals is calculated immediately:

$$\int(2x-1)dx=x^2-x+C,$$

and for computing the second integral, we expand its integrand, which is a regular fraction, at simple fractions:

$$\frac{2x^2+2x+13}{(x-2)(x^2+1)^2}=\frac{A}{x-2}+\frac{M_1x+N_1}{(x^2+1)^2}+\frac{M_2x+N_2}{x^2+1}.$$

Bringing the fractions on the right side to a common denominator, we equate the numerators:

$$2x^2+2x+13=A(x^2+1)^2+(M_1x+N)\ (x-2)+$$

$$+(M_2x+N_2)\ (x^2+1)(x-2).$$

Equating the coefficients at the same degrees x left and right, we arrive at a system of five linear equations:

$$\begin{array}{l|l} x^4 & A + M_2 = 0, \\ x^3 & -2M_2 + N_2 = 0, \\ x^2 & 2A + M_1 + M_2 - 2N_2 = 2, \\ x & -2M_1 - 2M_2 + N_1 + N_2 = 2, \\ x^0 & A - 2N_1 - 2N_2 = 13. \end{array}$$

Solving this system, we find:

$A = 1$, $M_1 = -3$, $N_1 = -4$, $M_2 = -1$, $N_2 = -2$.

We get

$$\frac{2x^2 + 2x - 13}{(x-2)(x^2+1)^2} = \frac{1}{x-2} - \frac{3x+4}{(x^2+1)^2} - \frac{x+2}{x^2+1}.$$

We calculate the integrals of each of the terms to the right:

$$\int \frac{dx}{x-2} = \ln|x-2| + C,$$

$$\int \frac{x+2}{x^2+1}dx = \frac{1}{2}\ln(x^2+1) + 2\operatorname{arctg} x + C,$$

$$\int \frac{3x+4}{(x^2+1)^2}dx = \frac{3}{2}\int \frac{2xdx}{(x^2+1)^2} + 4\int \frac{dx}{(x^2+1)^2} = -\frac{3}{2(x^2+1)} + 4\int \frac{dx}{(x^2+1)^2}$$

.

The last integral is calculated using the recurrence formula (**):

$$\int \frac{dx}{(x^2+1)^2} = \frac{1}{2}\frac{x}{x^2+1} + \frac{1}{2}\operatorname{arctg} x.$$

Taking into account the coefficients after obvious transformations, we obtain

$$\int \frac{2x^6 - 5x^5 + 6x^4 - 10x^3 + 8x^2 - 3x + 15}{x^5 - 2x^4 + 2x^3 - 4x^2 + x - 2} dx =$$

$$= x^2 - x + \frac{1}{2}\frac{3-4x}{x^2+1} + \frac{1}{2}\ln\frac{(x-2)^2}{x^2+1} - 4\operatorname{arctg} x + C.$$

13.4. Integration of irrational functions

Consider the cases when the change of variable allows us to reduce the integrals of irrational functions to the integrals of rational functions (i.e., rationalizes the integral). Denote by $R(u, v)$ a rational function of u and v, i.e. a function that is obtained using only arithmetic operations on the variables u and v.

1. Consider the integrals of the form $\int R\left(x, x^{\frac{m}{n}}, x^{\frac{r}{s}}, \ldots\right) dx$. Let k be the common denominator of fractions $\frac{m}{n}$, $\frac{r}{s}$, … . Let's make a substitution:

$x = t^k$, $dx = kt^{k-1}dt$.

Then , obviously, the integrand function is transformed into a rational function of t.

***Example* 13.9. Calculate integral** $\int \frac{\sqrt{x}dx}{\sqrt[4]{x^3}+1}$.

Decision. The smallest common multiple of the root indices is 4. Therefore, we substitute $x = t^4$, $dx = 4t^3 dt$:

$$\int \frac{\sqrt{x}dx}{\sqrt[4]{x^3}+1} = 4\int \frac{t^2}{t^3+1}t^3dt = 4\int \frac{t^5}{t^3+1}dt = 4\int\left(t^2 - \frac{t^2}{t^3+1}\right)dt =$$

$$= 4\int t^2dt - \frac{4}{3}\int \frac{d\left(t^3+1\right)}{t^3+1} = \frac{4}{3}t^3 - \frac{4}{3}\ln\left|t^3+1\right| + C = \frac{4}{3}\left(\sqrt[4]{x^3} - \ln\left(\sqrt[4]{x^3}+1\right)\right) + C.$$

2. We now consider the integral of the form

$$\int \boldsymbol{R}\left(\boldsymbol{x}, \left(\frac{\boldsymbol{ax+b}}{\boldsymbol{cx+d}}\right)^{\frac{\boldsymbol{m}}{\boldsymbol{n}}}, \left(\frac{\boldsymbol{ax+b}}{\boldsymbol{cx+d}}\right)^{\frac{\boldsymbol{r}}{\boldsymbol{s}}}, \boldsymbol{...}\right)\boldsymbol{dx}.$$

Such integrals are calculated by substitution

$$\frac{ax+b}{cx+d} = t^k,$$

where k – common denominator of fractions $\frac{m}{n}, \frac{r}{s}, \ldots$.

***Example* 13.10.** Calculate $\int \frac{dx}{\sqrt{x+5}\cdot\left(\sqrt[4]{x+5}+1\right)}$.

Decision. Here $a=1,\ b=5,\ c=0,\ d=1;\ \frac{m}{n}=\frac{1}{4},\ \frac{r}{s}=\frac{1}{2}$. We do the substitution $x+5=t^4,\ dx=4t^3dt$. We get

$$\int \frac{dx}{\sqrt{x+5}\left(\sqrt[4]{x+5}+1\right)} = 4\int \frac{t^3dt}{t^2(t+1)} = 4\int \frac{tdt}{t+1} = 4\int\left(1 - \frac{1}{t+1}\right)dt =$$

$$= 4\left(t - \ln\left|t+1\right|\right) + C = 4\left[\sqrt[4]{x+5} - \ln\left(\sqrt[4]{x+5}+1\right)\right] + C.$$

Some integrals of irrational functions are rationalized by trigonometric permutations. In particular, when calculating the integral

$\int R\left(x, \sqrt{a^2 - x^2}\right)dx$ wildcard applies $x = a\sin t$, and an integral of the form $\int R\left(x, \sqrt{a^2 + x^2}\right)dx$ – substitution $x = a\,\mathrm{tg}\,t$.

***Example* 13.11.** Find $\int \frac{dx}{\sqrt{\left(4 - x^2\right)^3}}$.

Decision. Let $x = 2\sin t$. Then $dx = 2\cos t\,dt$,

$$\int \frac{dx}{\sqrt{\left(4 - x^2\right)^3}} = \int \frac{2\cos t\,dt}{\sqrt{\left(4 - 4\sin^2 t\right)^3}} = \int \frac{2\cos t\,dt}{8\cos^3 t} = \frac{1}{4}\int \frac{dt}{\cos^2 t} =$$

$$= \frac{1}{4}\mathrm{tg}\,t + C = \frac{1}{4}\frac{\sin t}{\cos t} + C = \frac{1}{4}\frac{\sin t}{\sqrt{1 - \sin^2 t}} + C = \frac{1}{4}\frac{x}{\sqrt{4 - x}} + C.$$

13.5. Integration of trigonometric functions

Consider an integral of the form $\int R(\sin x, \cos x)\,dx$. We show that it reduces to the integral of a rational function by substituting

$t = \mathrm{tg}\frac{x}{2}$,

called universal trigonometric substitution.

Indeed, expressing $\sin x$, $\cos x$ *and* dx through $\mathrm{tg}\frac{x}{2} = t$,we get :

$\sin x = \frac{2t}{1+t^2}$, $\cos x = \frac{1-t^2}{1+t^2}$, $x = 2\,\mathrm{arctg}\,t$, $dx = \frac{2dt}{1+t^2}$.

Substituting the obtained expressions into the integral, we obtain

$$\int R(\sin x, \cos x)\,dx = \int R\left(\frac{2t}{1+t^2}, \frac{1-t^2}{1+t^2}\right)\frac{2dt}{1+t^2} = \int R_1(t)\,dt$$

.

***Example* 13.16.** Find $\int \frac{dx}{1+\sin x}$.

Decision. Using Universal Substitution $t = \operatorname{tg}\frac{x}{2}$ after obvious transformations we get:

$$\int \frac{dx}{1+\sin x} = 2\int \frac{dt}{(1+t)^2} = -\frac{2}{1+t} + C = -\frac{2}{1+\operatorname{tg}\frac{x}{2}} + C$$

.

It should be noted, however, that universal trigonometric substitution often leads to very complex rational functions. Therefore, in many cases, instead of a universal substitution, other substitutions are used, which make it faster and easier to achieve the goal.

***Example* 13.17.** Find $\int \sin^4 x \cdot \cos^3 x\,dx$.

Decision. Let $t = \sin x$. Then $dt = \cos x\,dx$. We get

$$\int \sin^4 x \cdot \cos^3 x\,dx = \int \sin^4 x \cdot \cos^2 x \cdot \cos x\,dx = \int t^4\left(1-t^2\right)dt =$$

$$= \frac{t^5}{5} - \frac{t^7}{7} + C = \frac{\sin^5 x}{5} - \frac{\sin^7 x}{7} + C.$$

In general, an integral of the form $\int \boldsymbol{\sin^m x \cdot \cos^n x\,dx}$ where m and n are natural numbersis more convenient to calculate using the following substitutions:

a) if m is even, n is odd, then the permutation $t = \sin x$;

b) i f m is odd, n is even, then the permutation $t = \cos x$;

c) if m and n are odd, then any of the substitutions “a” or “b”;

d) if m and n are even, then degree reduction formulas are applied:

$\sin^2 x = \frac{1-\cos 2x}{2}$, $\cos^2 x = \frac{1+\cos 2x}{2}$.

Integrals of the following types: $\int \cos mx \cdot \cos nx\, dx$, $\int \sin mx \cdot \cos nx\, dx$, $\int \sin mx \cdot \sin nx\, dx$ taken using the following formulas known from trigonometry:

$$\cos\alpha\cos\beta = \frac{1}{2}\left[\cos(\alpha+\beta)+\cos(\alpha-\beta)\right],$$

$$\sin\alpha\cos\beta = \frac{1}{2}\left[\sin(\alpha+\beta)+\sin(\alpha-\beta)\right],$$

$$\sin\alpha\sin\beta = \frac{1}{2}\left[-\cos(\alpha+\beta)+\cos(\alpha-\beta)\right].$$

***Example* 13.18.** Find $\int \cos 5x \cdot \cos 3x\, dx$.

Decision. We use the cosine product formula:

$$\int \cos 5x \cdot \cos 3x\, dx = \frac{1}{2}\int (\cos 8x + \cos 2x)\, dx = \frac{\sin 8x}{16} + \frac{\sin 2x}{4} + C$$

.

"Non-countable" integrals

It is known that the differentiation operation does not derive a function from the class of elementary functions. The integration operation is more complicated. Not every integral of an elementary function is expressed in a finite form in terms of elementary functions. Such integrals are called "non-countable". We indicate some of these integrals:

$\int \frac{\sin x}{x} dx$ – integral sine;

$\int \frac{\cos x}{x} dx$ – integral cosine;

$\int \frac{dx}{\ln x}$ – integral logarithm;

$\int e^{-x^2} dx$ – Poisson integral;

$\int \sin x^2 dx$, $\int \cos x^2 dx$ – Fresnel integrals.

These integrals play an important role in applied sciences. For the calculation of "unshifted" integrals, approximate methods are used that allow us to estimate and calculate such integrals with any degree of accuracy.

Questions

1) What is the derivative of the antiderivative for a given function?
2) How can two antiderivatives of the same function differ?
3) How is the antiderivative function different from the indefinite integral from this function?
4) What is the derivative of the indefinite integral?
5) What is the differential from the indefinite integral?
6) What are the main integration methods?
7) On what formula is the method of replacing a variable in an indefinite integral based?
8) What is a piecemeal integration method? What is the integration formula in parts?
9) What is a differential sign?
10) What function is called rational fraction?
11) Which rational fraction is called correct and which incorrect?
12) How to reduce the integration of the wrong rational fraction to the integration of the right rational fraction?

13) What rational fractions are called simple?
14) What is the procedure for integrating a rational fraction that is not simple?
15) What method is used to decompose a regular rational fraction at simple?
16) What is the universal trigonometric substitution?
17) Why does an integration of trigonometric expressions not suffice to own only universal trigonometric substitution?
18) Is the integral of an elementary function always expressed in its final form in terms of elementary functions?
19) What are “non-tilting” integrals? Give Examples of "unshifted" integrals.

Chapter 14. Definite integral and its properties

14.1. The concept of a specific integral

Let function $f(x)$ be defined on the line $[a, b]$and cut the line $[a, b]$ randomly at n parts with dots:

$a = x_0 < x_1 < x_2 < ... < x_{n-1} < x_n = b$.

We choose in each of the partial segments $[x_{i-1}, x_i]$ *random dot* ξ_i:

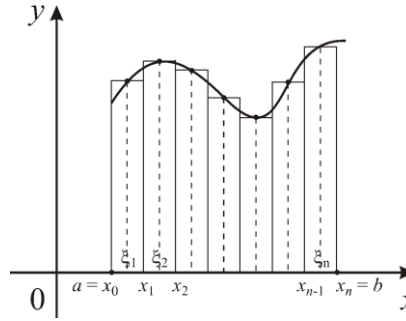
$x_{i-1} \le \xi_i \le x_i$, $1 \le i \le n$.

Denote by Δx_i the length of the i-th partial segment: $\Delta x_i = x_i - x_{i-1}$, $1 \le i \le n$. Consider the amount:

$$\sigma_n = f(\xi_1)\Delta x_1 + f(\xi_2)\Delta x_2 + \ldots + f(\xi_n)\Delta x_n = \sum_{i=1}^{n} f(\xi_i)\Delta x_i. \quad (14.1)$$

This sum (14.1) is called the integral sum of the function $f(x)$ on the line $[a, b]$.

If $f(x) > 0$, then its integral sum is the sum of the areas of the rectangles with bases Δx_i and heights $f(\xi_i)$, i = 1, 2, …, n, i.e., the area of the stepped figure formed by these rectangles (Pic. 14.1).



Pic. 14.1. *Stepped figure*

Denote by λ the length of the largest partial segment of this partition: $\lambda = \max\limits_{1 \le i \le n} \Delta x_i$.

Definition. The final limit of the integral sum (14.1) for $\lambda \to 0$ if it exists and does not depend on either the method of splitting the segment

$[a, b]$ nor from the choice of points ξ_i is called the **definite integral** of a function $f(x)$ on the line $[a, b]$ and denoted as $\int_a^b f(x)dx$:

$$\int_a^b f(x)dx = \lim_{\lambda \to 0} \sigma_n = \lim_{\lambda \to 0} \sum_{i=1}^n f(\xi_i)\Delta x_i. \quad (14.2)$$

If a definite integral (14.2) exists, then function $f(x)$ is called integrable at segment $[a, b]$. The number a in formula (14.2) is called the lower limit of the integral, and the number b is called the upper limit of the integral, $f(x)$ – integrand, x is the integration variable, and the segment $[a, b]$ – segment of integration.

Note the differences in the concepts of definite and indefinite integrals: indefinite integral $\int f(x)dx$ is a family of functions, and a certain integral $\int_a^b f(x)dx$ is a *certain number*.

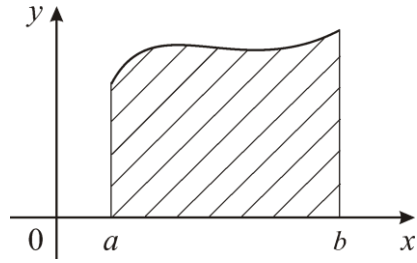
Giving a Definition of the concept of a definite integral, we assumed $a < b$.By definition:

$$\int_a^b f(x)dx = -\int_a^b f(x)dx. \quad (14.3)$$

The geometric meaning of a certain integral

In accordance with the Definitions of the concept of a definite integral, in the case when Function $f(x)$ is non-negative at $[a, b]$, integral $\int_a^b f(x)dx$ is numerically equal to the area S of the figure bounded above

the curve $y = f(x)$, bottom - axis Ox, lateral - straight $x = a$, $x = b$ (Pic. 14.2). This figure is called a curved trapezoid.



Pic. 14.2. *The geometric meaning of a certain integral*

The economic meaning of a certain integral

Let function $z = z(t)$ describe performance versus time t, then the volume v of products produced over the period from the moment $t = t_0$ till the moment is expressed by the integral of $z(t)$ on the segment $[t_0, T]$:

$$v = \int_{t_0}^{\tau} z(t)dt$$

.

Integrable Function Classes

A sufficient condition for the existence of a definite integral is given by the following Theorem (we give it without proof).

***Theorem* 14.1.** If function $y = f(x)$ is continuous on a segment $[a, b]$, then it is integrable on this segment.

As can be seen from Theorem 14.1, the class of integrable functions is wider than the class of differentiable functions. We know that every differentiable function is continuous, but not every continuous function is differentiable. So, the continuity of a function is not enough for its differentiability but enough for integrability. Moreover, there are classes

of functions that are not continuous but are integrable. We give without proof a theorem on these functions.

***Theorem* 14.2.** If function $f(x)$ is bounded at $[a, b]$ and has at it only a finite number of discontinuity points, then it is integrable at this interval.

***Theorem* 14.3.** If function $f(x)$ is monotonically bounded at interval $[a, b]$, then it is integrable at this interval.

Boundedness of integrable function

***Theorem* 14.4.** If function $f(x)$ is integrable on $[a, b]$, then it is limited to $[a, b]$.

Proof. Let function $f(x)$ be unlimited at $[a, b]$. Then it is not limited to at least one of the partial segments $[x_{i-1}, x_i]$. And then, by choosing a point, you can make a product of function $f(\xi_i)\Delta x_i$ arbitrarily large and consequently so the integral sum σ_n ; under these conditions σ_n has no limit. Consequently, $f(x)$ not integrable. From this we conclude that the assumption is false.

14.2. Properties of a specific integral

We first consider the properties of a certain integral expressed by equalities.

1. By definition, we assume

$$\int_a^a f(x)dx = 0, \qquad (14.4)$$

i.e.. a certain integral with the same integration limits is equal to zero.

2. By definition, when rearranging the upper and lower limits of integration, the integral changes to the opposite sign at:

$$\int_a^b f(x)dx = -\int_b^a f(x)dx. \quad (14.5)$$

3. The constant factor can be taken out of the integral sign:

$$\int_a^b cf(x)dx = c\int_a^b f(x)dx. \quad (14.6)$$

4. The integral of the algebraic sum of two functions is equal to the algebraic sum of the integrals of these functions:

$$\int_a^b \big(f(x) \pm g(x)\big)\,dx = \int_a^b f(x)dx \pm \int_a^b g(x)dx. \quad (14.7)$$

5. For any numbers a, b and c, equality holds:

$$\int_a^b f(x)dx = \int_a^c f(x)dx + \int_c^b f(x)dx. \quad (14.8)$$

We state properties 3–5 in more detail and prove them.

***Property** 3. If function* $f(x)$ *is integrable on* $[a,b]$ *and* $c = \text{const}$ *,then* function $cf(x)$ is *integrable on* $[a,b]$ and the following equality holds true:

$$\int_a^b cf(x)dx = c\int_a^b f(x)dx.$$

Proof. For integral sums the following equality holds true:

$$\sum_{i=1}^n cf(\xi_i)\Delta x_i = c\sum_{i=1}^n f(\xi_i)\Delta x_i.$$

This equality is valid for any partition of a segment $[a,b]$ at partial segments and any choice of points ξ_i. Designating, as before, $\lambda = \max \Delta x_i$, we pass to the limit at $\lambda \to 0$:

$$\int_a^b cf(x)dx = \lim_{\lambda\to 0}\sum_{i=1}^{n} cf(\xi_i)\Delta x_i = c\lim_{\lambda\to 0}\sum_{i=1}^{n} f(\xi_i)\Delta x_i = c\int_a^b f(x)dx$$

.

Property 4. If $f(x)$ and $g(x)$ integrable on $[a,b]$then their algebraic sum is also integrable at $[a,b]$ and the equality is valid:

$$\int_a^b (f(x)\pm g(x))dx = \int_a^b f(x)dx \pm \int_a^b g(x)dx$$

.

Proof. For any segment partition $[a,b]$ and any choice of points ξ_i for integral sums, the equality runs:

$$\sum_{i=1}^{n}(f(\xi_i)\pm g(\xi_i))\Delta x_i = \sum_{i=1}^{n} f(\xi_i)\Delta x_i \pm \sum_{i=1}^{n} g(\xi_i)\Delta x_i$$

.

Therefore

$$\int_a^b (f(x)\pm g(x))dx = \lim_{\lambda\to 0}\sum_{i=1}^{n}(f(\xi_i)\pm g(\xi_i))\Delta x_i =$$

$$= \lim_{\lambda\to 0}\left(\sum_{i=1}^{n} f(\xi_i)\Delta x_i \pm \lim_{\lambda\to 0}\sum_{i=1}^{n} g(\xi_i)\Delta x_i\right) = \lim_{\lambda\to 0}\sum_{i=1}^{n} f(\xi_i)\Delta x_i \pm \lim_{\lambda\to 0}\sum_{i=1}^{n} g(\xi_i)\Delta x_i =$$

$$= \int_a^b f(x)dx \pm \int_a^b g(x)dx.$$

Property 5. For any three numbers a, b and c the equality holds:

$$\int_a^b f(x)dx = \int_a^c f(x)dx + \int_c^b f(x)dx$$

,

if all these three integrals exist.

Proof. We first consider the case when point c is located between the points a and b, i.e. $a < c < b$. We compose the integral sum of the function $f(x)$ on the segment $[a,b]$.

For the integrable function the limit of the integral sum does not depend on the method of partitioning the segment at partial segments, then we will divide the segment $[a,b]$ apart, so that one of the division points (one of the ends of the partial segment) is point c. Denote by $\sum_a^b$ integral sum according to the segment $[a,b]$, through $\sum_a^c$ – integral sum according to the segment $[a,c]$, and through $\sum_c^b$ – according to the segment $[c,b]$. Then , obviously:

$$\sum_a^b f(\xi_i)\Delta x_i = \sum_a^c f(\xi_i)\Delta x_i + \sum_c^b f(\xi_i)\Delta x_i .$$

Passing in the last equality to the limit at $\lambda \to 0$, we get

$$\int_a^b f(x)dx = \int_a^c f(x)dx + \int_c^b f(x)dx .$$

Now, let point c be to the right of point b: $a < b < c$. Then on the basis of the proved equality:

$$\int_a^c f(x)dx = \int_a^b f(x)dx + \int_b^b f(x)dx .$$

Thus

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx - \int_b^c f(x)\,dx$$

.

However, in accordance with property 2

$$\int_b^c f(x)dx = -\int_c^b f(x)dx$$

;

therefore

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx$$

.

This property of the case is proved similarly when the point c is to the left of the segment $[a,b]$and in general for another arrangement of points a, b and c.

Comment. We formulated and proved property 5 under the assumption that all three integrals under consideration exist. One could relax this requirement and prove property 5 under the assumption that only for the largest of the three segments under consideration does the integral exist.

Now we consider the properties of a certain integral expressed by inequalities.

6. If function $f(x)$ is integrable on $[a,b]$, $a<b$ and $f(x)\geq 0$, then

$$\int_a^b f(x)dx \geq 0. \qquad (14.9)$$

(This follows from the fact that all terms in the integral sum are non-negative.)

7. If functions $f(x)$ and $g(x)$ integrable on $[a,b]$, $a<b$ and $f(x)\geq g(x)$, then

$\int_a^b f(x)dx \geq \int_a^b g(x)dx.$ (14.10)

Proof. $f(x)-g(x)\geq 0$,then property 6 implies

$\int_a^b (f(x)-g(x))dx \geq 0$.

From here, taking into account property 4, we obtain

$\int_a^b f(x)dx - \int_a^b g(x)dx \geq 0$, или $\int_a^b f(x)dx \geq \int_a^b g(x)dx$.

8. Let function $f(x)$ be integrable on $[a, b]$ and satisfy on $[a, b]$ the condition $m \leq f(x) \leq M$. Then

$m(b-a) \leq \int_a^b f(x)dx \leq M(b-a).$ (14.11)

Proof. By virtue of property 7

$\int_a^b mdx \leq \int_a^b f(x)dx \leq \int_a^b Mdx,$

but $\int_a^b mdx = m\int_a^b dx = m(b-a)$, $\int_a^b Mdx = M\int_a^b dx = M(b-a)$, therefore

$m(b-a) \leq \int_a^b f(x)dx \leq M(b-a).$

9 (Theorem on average). If function $f(x)$ is continuous on $[a, b]$,then $\xi \in [a, b]$ exists, and

$\int_a^b f(x)dx = f(\xi)\cdot(b-a).$ (14.12)

Proof. By the second Weierstrass theorem, continuous function $f(x)$ reaches on $[a, b]$ its largest value M and its lowest value m.

$m \le f(x) \le M$,then inequality (14.11) holds. Dividing this inequality term by $(b-a)$,we get

$$m \le \frac{1}{b-a}\int_a^b f(x)dx \le M.$$

By the second theorem of Bolzano - Cauchy Function, $f(x)$ takes on $[a, b]$ all intermediate values between m and M. In particular, there is such $\xi \in [a, b]$, that

$$f(\xi) = \frac{1}{b-a}\int_a^b f(x)dx.$$

And so we get (14.12).

14.3. Basic formula for integral calculation

Variable upper limit integral

If function is integrable on the segment $[a, b]$,then it is integrable on any segment $[a, x]$, where $x \in [a, b]$.

Note that it does not matter which letter denotes the integration variable in a certain integral:

$$\int_a^b f(x)dx = \int_a^b f(z)dz = \int_a^b f(t)dt = \ldots$$

,

change of notation does not affect the value of the integral.

Consider the argument function x:

$$\Phi(x) = \int_a^x f(t)dt. \qquad (14.13)$$

Call the function $\Phi(x)$ **integral with a variable upper limit.** (In formula (14.13), the integration variable is denoted by *t* in order not to mix it with the upper limit x.)

It was previously established that every differentiable function is continuous (see Theorem 8.2), however, a function continuous at a point may not have a derivative at this point. We prove now that each continuous on a given segment *function* $f(x)$ *has an antiderivative in this segment.*

***Theorem* 14.5.** If function $f(x)$ continuous on $[a, b]$,then function $\Phi(x)$ is antiderivative for $f(x)$*:*

$$\Phi'(x) = \left(\int_a^x f(t)dt \right)' = f(x)$$

.

Proof. Let Δx so $(x + \Delta x) \in [a, b]$. Then

$$\Delta\Phi = \Phi(x + \Delta x) - \Phi(x) = \int_a^{x+\Delta x} f(t)dt - \int_a^x f(t)dt = \int_x^{x+\Delta x} f(t)dt$$

.

By the mean value theorem (see 14.12):

$$\int_x^{x+\Delta x} f(t)dt = f(\xi)\Delta x$$

, $\xi \in [x, x + \Delta x]$.

And

$$\frac{\Delta\Phi}{\Delta x} = \frac{\Phi(x + \Delta x) - \Phi(x)}{\Delta x} = f(\xi)$$

.

When $\Delta x \to 0$, Obviously, $\xi \to x$, and $f(x)$ is continuous at x,then $\lim\limits_{\xi \to x} f(\xi) = f(x)$. We get

$$\Phi'(x) = \lim_{\Delta x \to 0} \frac{\Delta \Phi}{\Delta x} = \lim_{\Delta x \to 0} \frac{\Phi(x + \Delta x) - \Phi(x)}{\Delta x} = \lim_{\Delta x \to 0} f(\xi) = \lim_{\xi \to x} f(\xi) = f(x)$$

.

Q.E.D.

The proved theorem can also be formulated as follows: the derivative of a certain integral with respect to the variable upper limit is equal to the integrand (in which the value of the upper limit is substituted for the integration variable).

Newton-Leibniz Formula

According to Theorem 14.5, the integral for function $f(x)$ continuous on $[a, b]$ is

$$\Phi(x) = \int_a^x f(t)dt$$

and this integral is antiderivative. Let $F(x)$ be any antiderivative for $f(x)$. Then

$$\Phi(x) = F(x) + C.$$

Constant C is found from $x = a$ (Obviously, $\Phi(a) = 0$):

$0 = \Phi(a) = F(a) + C$, when $C = -F(a)$.

Then

$$\Phi(x) = F(x) - F(a).$$

If $x = b$ get $\Phi(b) = F(b) - F(a)$, i.e.

$$\int_a^b f(t)dt = F(b) - F(a),$$

Or, which is the same:

$$\int_a^b f(x)dx = F(b) - F(a). \qquad (14.14)$$

Formula (14.14) is called the Newton-Leibniz formula. This is the basic formula for integral calculus.

Difference $F(b)-F(a)$ is written as $F(x)\Big|_a^b$ (« double substitution from a to b "). Then the formula (14.14) takes the form

$$\int_a^b f(x)dx = F(x)\Big|_a^b \qquad (14.15)$$

***Example* 14.1.**

1) $\int_0^3 x^2 dx = \frac{x^3}{3}\Big|_0^3 = \frac{3^3}{3} - \frac{0^3}{3} = 9;$

2) $\int_1^e \frac{dx}{x} = lnx|_1^e = ln\, e - ln\, 1 = 1;$

3) $\int_0^1 \frac{dx}{1+x^2} = \text{arctg}\ x|_0^1 = \text{arctg}\ 1 - \text{arctg}\ 0 = \frac{\pi}{4}.$

14.4. Change variable and integration by parts in definite integrals

Variable replacement

Let the integral be given $\int_a^b f(x)dx$, where $f(x)$ is a function continuous on the segment $[a, b]$. We introduce a new variable by setting $x = \varphi(t)$.

If $\varphi(\alpha) = a$, $\varphi(\beta) = b$, and the values $\varphi(t)$ don't go beyond $[a, b]$ when *t* differs on $[\alpha, \beta]$. Moreover, let $\varphi(t)$ and $\varphi'(t)$ be continuous $[\alpha, \beta]$. Then

$$\int_a^b f(x)dx = \int_\alpha^\beta f\big(\phi(t)\big)\phi'(t)dt. \qquad (14.16)$$

if $F'(x) = f(x)$, then at the same time

$$\int_a^b f(x)dx = F(x)\Big|_a^b = F(b) - F(a)$$

And

$$\int_\alpha^\beta f\big(\varphi(t)\big)\varphi'(t)\,dt = F\big(\varphi(t)\big)\Big|_\alpha^\beta = F\big(\varphi(\beta)\big) - F\big(\varphi(\alpha)\big) = F(b) - F(a),$$

hence the proved equality follows (14.16).

It should be noted that when calculating a certain integral by replacing a variable, there is no need to return to the old variable, i.e. make a reverse replacement.

***Example* 14.2.** Calculate $\int_{1/2}^{1} \frac{\sqrt{1-x^2}}{x^2}\,dx$.

Decision. Replace $x = \sin t$. Then $dx = \cos t\ dt$, $t = \arcsin x$, $t = \frac{\pi}{6}$ if $x = \frac{1}{2}$ and $t = \frac{\pi}{2}$ if $x = 1$. We get

$$\int_{1/2}^{1} \frac{\sqrt{1-x^2}}{x^2}\,dx = \int_{\pi/6}^{\pi/2} \frac{\sqrt{1-\sin^2 t}}{\sin^2 t}\cos t\ dt = \int_{\pi/6}^{\pi/2} \frac{\cos^2 t}{\sin^2 t}\,dt = \int_{\pi/6}^{\pi/2} \frac{1-\sin^2 t}{\sin^2 t}\,dt =$$

$$= \big(-\operatorname{ctg} t - t\big)\Big|_{\pi/6}^{\pi/2} = -\frac{\pi}{2} + \sqrt{3} + \frac{\pi}{6} = \sqrt{3} - \frac{\pi}{3}.$$

Part Integration

Let functions $u = u(x)$ and $v = v(x)$ have continuous derivatives on the segment $[a, b]$. Then

$(uv)' = u'v + uv'$.

We integrate both sides of this equality:

$$\int_a^b (uv)'dx = \int_a^b u'v\,dx + \int_a^b uv'\,dx$$

.

$\int (uv)'dx = uv + C$,then $\int_a^b (uv)'dx = uv\Big|_a^b$. We get

$$uv\Big|_a^b = \int_a^b v\,du + \int_a^b u\,dv$$

.

Consequently,

$$\int_a^b u\,dv = uv|_a^b - \int_a^b v\,du. \qquad (14.17)$$

Equality (14.17) is called **the integration formula by parts in a certain i**ntegral.

***Example* 14.3.** Calculate $\int_1^e x \ln x\,dx$.

Decision. Let $u = ln\,x,\ dv = xdx$. Then $du = \frac{dx}{x},\ v = \frac{x^2}{2}$. We get

$\int_1^e x\ ln\,x\ dx = \frac{x^2}{2} lnx|_1^e - \int_1^e \frac{x^2}{2}\frac{dx}{x} = \frac{e^2}{2} - \frac{1}{4}x^2|_1^e =$

$= \frac{e^2}{2} - \frac{e^2}{4} + \frac{1}{4} = \frac{e^2+1}{4}$.

14.5. Approximate calculation of definite integrals

When solving a number of applied problems, one often has to deal with certain integrals of functions for which antiderivatives are not elementary functions. For the calculation of such integrals, there are various methods

of **approximate calculation.** Here we give the simplest of them: the rectangle formula, the trapezoid formula and the Simpson formula.

Let there be a function $y = f(x)$ on the segment $[a, b]$. Calculation of definite integral $\int_a^b f(x)dx$ is needed.

1. The formula of the rectangles Divine the segment $[a, b]$ with the dots $a = x_0$, x_1, x_2, …, $x_n = b$ into *n* equal parts long Δx:

$$\Delta x = \frac{b-a}{n}.$$

We denote as: $y_0 = f(x_0)$, $y_1 = f(x_1)$, …, $y_n = f(x_n)$.

Sums are:

$$y_0\Delta x + y_1\Delta x + ... + y_{n-1}\Delta x,$$

$$y_1\Delta x + y_2\Delta x + ... + y_n\Delta x.$$

Each of these sums is an integral sum for $f(x)$ at $[a, b]$. Therefore:

$$\int_a^b f(x)\,dx \approx \frac{b-a}{n}(y_0 + y_1 + y_2 + \ldots + y_{n-1}), \qquad (14.18)$$

$$\int_a^b f(x)\,dx \approx \frac{b-a}{n}(y_1 + y_2 + \ldots + y_n). \qquad (14.18')$$

Each of the formulas (22.18) and (22.18 ') is called a **rectangle formula.**

The error made if calculating the integral by the formula of rectangles will be the smaller, the greater the number n (i.e., the smaller the partial segments at which the segment is divided $[a, b]$).

2. Trapezoid formula.

$$\int_a^b f(x)\,dx \approx \frac{b-a}{n}\left(\frac{y_0+y_n}{2} + y_1 + y_2 + \ldots + y_{n-1}\right). \qquad (14.19)$$

3. Simpson's formula. Divide the line $[a, b]$ into an even number of equal parts $n = 2m$.

$$\int_a^b f(x)\,dx \approx \frac{b-a}{6m}[y_0 + y_{2m} + 2(y_2 + y_4 + \cdots + y_{2m-2}) + 4(y_1 + y_3 + \ldots + y_{2m-1})]. \quad (14.20)$$

Note that for the same step the division of the segment $\Delta x = \frac{b-a}{n}$ the trapezoid formula gives a slightly more accurate value of a certain integral than the rectangle formula, and the Simpson formula gives a much more accurate value than the trapezoid formula.

***Example* 14.4. Consider the well-known integral**

$$\int_0^1 \frac{dx}{1+x^2} = \frac{\pi}{4} = 0{,}785398...$$

Divine the line $[0,1]$ into four equal parts: $x_0 = 0$; $x_1 = 0{,}25$; $x_2 = 0{,}5$; $x_3 = 0{,}75$; $x_4 = 1$. Then $y_0 = 1{,}0000$; $y_1 \approx 0{,}9412$; $y_2 = 0{,}8000$; $y_3 = 0{,}6400$; $y_4 = 0{,}5000$.

By the formulas of the rectangles we have:

$$\frac{1}{4}(1 + 0{,}9412 + 0{,}8 + 0{,}64) = 0{,}8453,$$

$$\frac{1}{4}(0{,}9412 + 0{,}8 + 0{,}64 + 0{,}5) = 0{,}7203,$$

according to the trapezoid formula

$$\frac{1}{4}\left(\frac{1+0{,}5}{2} + 0{,}9412 + 0{,}8 + 0{,}64\right) = 0{,}7828,$$

according to the simpson formula

$$\frac{1}{12}(1+0{,}5+3{,}76471+1{,}6+2{,}56)=0{,}78539.$$

(We took an accuracy of 0.00001 $4y_1 \approx 3{,}76471$.)

We see that the Simpson formula gives a very accurate result: all five signs are correct. The trapezoid formula gives an error already in the third digit. If we split the line $[0,1]$ by 10 parts, then the trapezoid formula would give a result that differs from the true value less than at 0,0005. For then, in order to obtain a satisfactory result using the formula of rectangles, it is necessary to divide the segment into a significantly larger number of parts.

In general, in order to know how many division points you need to take in order to calculate the integral with a given degree of accuracy, you need to use the error estimation formulas. These estimates can be found in more detailed courses in mathematical analysis.

Questions

1) What is the integral sum for a given function at a given interval?
2) What is called a certain integral of the function on the segment?
3) What are the differences in the concepts of definite and indefinite integrals?
4) What is the geometric meaning of a certain integral?
5) What is the economic meaning of a certain integral?
6) Is any integrable function differentiable? Is every differentiable function integrable?
7) What is the derivative of a certain integral equal to its variable upper limit?
8) What is the difference between the application of the method of replacing a variable to calculate a certain integral from the application of the same method for calculating an indefinite integral?

9) What methods are used to calculate certain integrals of functions for which there are no primitives expressed in terms of elementary functions?
10) Which of the approximate formulas gives the more accurate value of a certain integral by the same step of dividing the integration interval: the rectangle formula, the trapezoid formula, or the Simpson formula?

Chapter 15. Applications of the definite integral

15.1. Geometrical and mechanical applications of the definite integral

Area of a plain figure

As we mentioned before, for a continuous on $[a,\ b]$ function $f(x) \geq 0$ the area S of a curvilinear trapezoid, bounded by the lines $y = f(x)$, $y = 0$, $x = a$, $x = b$ (Fig 15.2), is expressed by the integral

$$S = \int_a^b f(x)\ dx. \qquad (15.1)$$

***Example* 15.1.** Evaluate the area of a figure, bounded by the graph of the function $y = \ln x$, axis *Ox* and lines $y = \mathrm{e}$, $x = \mathrm{e}^2$ (Fig. 23.1).

Solution. According to formula (23.1) the area is

$$S = \int_e^{e^2} ln\, x\, dx = x(ln\, x - 1)|_e^{e^2} = e^2.$$

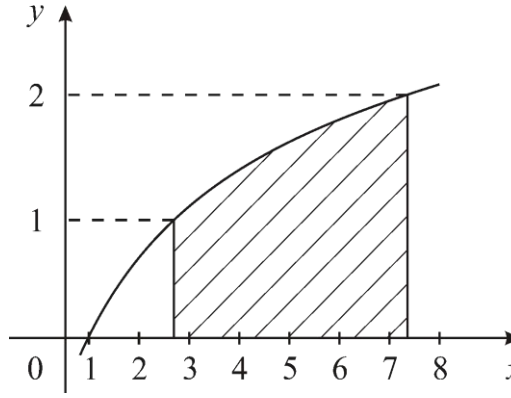


Fig. 15.1

Let $y = f_1(x)$, $y = f_2(x)$ be continuous on $[a, b]$ functions and let $f_2(x) \geq f_1(x)$ be on the specified segment. Then area *S* of the figure bounded by the graphs of the functions $y = f_1(x)$, $y = f_2(x)$ and vertical lines $x = a$, $x = b$ is evaluated by formula:

$$S = \int_a^b \left(f_2(x) - f_1(x)\right)\, dx. \qquad (15.2)$$

Proof. 1. Let $f_1(x) \geq 0$, $f_2(x) \geq 0$. Then formula (15.2) is an obvious consequence of the fact that the area of the figure is equal to the difference of the areas of curvilinear trapezoids (Fig. 15.2):

$$S = \int_a^b f_2(x)\, dx - \int_a^b f_1(x)\, dx = \int_a^b \left(f_2(x) - f_1(x)\right) dx.$$

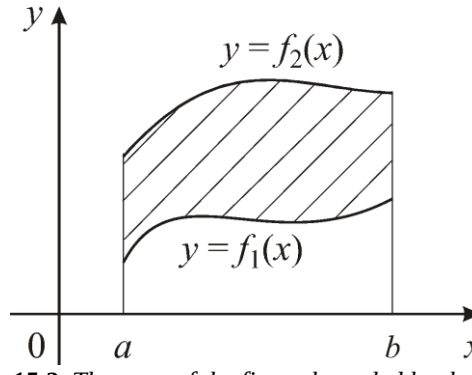


Fig. 15.2. *The area of the figure bounded by the lines* $y = f_1(x)$, $y = f_2(x)$, $x = a$, $x = b$

2. Let the graphs of functions $y = f_1(x)$, $y = f_2(x)$ be fully or partially located below the axis *Ox*. Since these functions are bounded, there exists a number *M* such that $f_1(x) + M \geq 0$, $f_2(x) + M \geq 0$. Obviously, a

figure bounded by the lines $y = f_1(x) + M$, $y = f_2(x) + M$ (located above *Ox*), is obtained by parallel transfer of a figure bounded by the lines $y = f_1(x)$, $y = f_2(x)$, and has the same area:

$$S = \int_a^b \left(f_2(x) + M \right) dx - \int_a^b \left(f_1(x) + M \right) dx = \int_a^b \left(f_2(x) - f_1(x) \right) dx .$$

***Example* 15.2.** Calculate the area of a shape bounded by the lines $y = x^2 - 2x$, $y = 4x - x^2$.

Solution. Equating the right sides of these equations, we find the abscissas of the intersection points of these curves: $x_1 = 0$, $x_2 = 3$. Consequently,

$$S = \int_0^3 \left(4x - x^2 - (x^2 - 2x)\right)\, dx = \int_0^3 (6x - 2x^2)\, dx =$$

$$= \left(3x^2 - \frac{2}{3}x^3\right)\Big|_0^3 = 9.$$

Volume of the body of rotation

Consider a body that is formed by rotation around the axis *Ox* of a curvilinear trapezoid, limited by a graph of a non-negative function $y = f(x)$ continuous on the segment $[a, b]$ and the lines $y = 0$, $x = a$, $x = b$. The volume of this body is expressed by the formula

$$V = \pi \int_a^b f^2(x)\, dx. \qquad (15.3)$$

Proof. Let us divide the segment $[a, b]$ arbitrarily into *n* parts by points $a = x_0 < x_1 < x_2 \ldots < x_n = b$ and on every segment $[x_{i-1}, x_i]$ arbitrarily choose a point ξ_i. When the curve $y = f(x)$ rotates around axis *Ox*, each

rectangle with base $\Delta x_i = x_i - x_{i-1}$ and height $f(\xi_i)$ describes a cylinder of radius $f(\xi_i)$ and height Δx_i (Fig. 15.3).

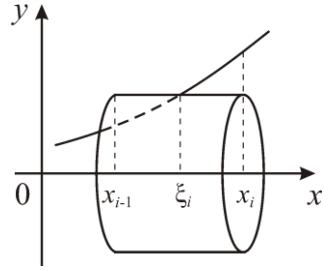


Fig. 15.3. *Evaluation of the volume of a body of rotation*

The sum of the volumes of such cylinders has the form:

$$V_n = \sum_{i=1}^{n} \pi f^2(\xi_i)\Delta x_i. \qquad (15.4)$$

With a smaller partition, this sum gives an approximate value of the desired volume. On the other hand, this sum is an integral sum for the continuous function $y = \pi f^2(x)$. Passing to the limit as $\max \Delta x_i \to 0$, we obtain formula (23.3).

***Example* 15.3.** Evaluate the volume of the cone with radius *R* and height *H*.

Solution. The cone whose volume we evaluate, can be obtained as a result of the rotation of triangle *OAB* around axis *Ox*. Here $AB = R$, $OB = H$ (Fig. 23.4).

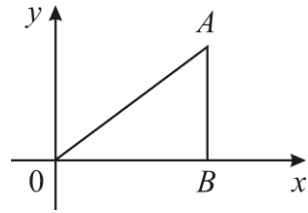


Fig. 15.4

Obviously, tg $AOB = \frac{R}{H}$. Therefore, the equation of the line *OA* has form $y = \frac{R}{H}x$. By formula (15.3) we obtain

$$V = \pi\int_0^H \left(\frac{R}{H}x\right)^2 dx = \frac{\pi R^2}{H^2}\int_0^H x^2 dx = \frac{\pi R^2}{3H^2}\cdot x^3\Big|_0^H = \frac{\pi R^2 H}{3}.$$

This formula is well known from the geometry school course.

***Example* 15.4.** Evaluate the volume of the body formed by the rotation of a figure bounded by lines $y = \mathrm{e}^x$, $y = 0$, $x = -1$, $x = 0$ around axis *Ox* (Fig. 23.5).

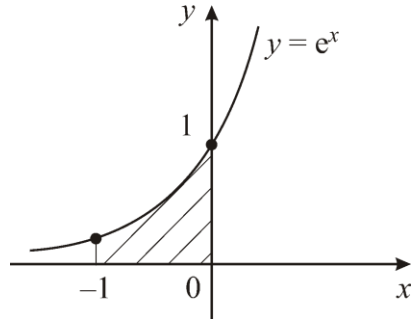


Fig. 15.5

Solution: $$V = \pi\int_{-1}^{0} \left(\mathrm{e}^x\right)^2 dx = \pi\cdot\frac{1}{2}\mathrm{e}^{2x}\Big|_{-1}^{0} = \frac{\pi}{2}\left(1-\frac{1}{\mathrm{e}^2}\right).$$

Arc length of a flat curve

Let the curve on the plane *Oxy* be given by the equation $y = f(x)$ and $f(x)$ have a continuous derivative $f'(x)$ on segment $[a, b]$. Then length *l* of its arc is equal to:

$$l = \int_a^b \sqrt{1 + (y')^2}\,dx, \qquad (15.5)$$

where *a* and *b* are abscissas of the ends of the arc.
(We accept this statement without proof.)

***Example* 15.5.** Find the arc length of curve $y = \ln \sin x$ from $x = \frac{\pi}{3}$ to $x = \frac{2\pi}{3}$.

Solution. Let us evaluate derivative $y' = \operatorname{ctg} x$ and substitute it in formula (23.5):

$$l = \int_{\pi/3}^{2\pi/3} \sqrt{1+\operatorname{ctg}^2 x}\,dx = \int_{\pi/3}^{2\pi/3} \frac{dx}{\sin x} = \ln \operatorname{tg} \frac{x}{2}\Big|_{\frac{\pi}{3}}^{\frac{2\pi}{3}} = \ln 3 .$$

Mechanical and physical applications of an integral

It is well known that the distance travelled is the integral of the speed of motion. This fact is a consequence of the fact that speed is the derivative of the path over time.

Mechanical work is also calculated using the integral. Suppose that under the influence of a certain force *F* a material point moves along the straight line *Os*, and the direction of the force coincides with the direction of motion. It is required to find the work of force F in moving the point from position $s = a$ to position $s = b$. If force *F* is constant, then work *A* is equal to the product of force *F* by the path length: $A = F(b-a)$. If the force continuously changes, i.e. $F = F(s)$ is a continuous function on $[a, b]$, then work *A* is expressed by the formula:

$$A = \int_a^b F(s)\,ds$$

.

Example 15.6. Let there be an inhomogeneous rod of length l and its density at each point is known. Then we can find the mass of any part of it and, in particular, the entire core. To do this, we place the axis Ox along the rod so that its left end is at the origin and denote by $\rho(x)$ its density at point x. Density $\rho(x)$ is the derivative of mass $M(x)$ of the rod segment from 0 to x. Therefore, for any segment $[a,b] \subset [0,l]$ we have

$$M(b) - M(a) = \int_a^b \rho(x)dx.$$

In particular,

$$m = M(l) - M(0) = \int_0^l \rho(x)dx.$$

15.2. Applications of the definite integral in economy

In § 14.1, we noted that knowing the function of labor productivity we can use the definite integral to express the volume of output.

Consider an example.

***Example* 15.7.** Find the daily output *P* for a working day from 8 to 14 hours, if labor productivity is given by an empirical formula

$$P = P(t) = -\frac{t^2}{4} + 5t - 15$$

.

(This formula reflects a process in which productivity rises for the first two hours and then drops.)

Solution. Considering the performance function (Fig. 23.6) on the segment $[8, 14]$, we express the daily output by the integral:

$$A = \int_8^{14} \left(-\frac{t^2}{4} + 5t - 15 \right) dt = \left(-\frac{t^3}{12} + \frac{5t^2}{2} - 15t \right) \Bigg|_8^{14} = 54.$$

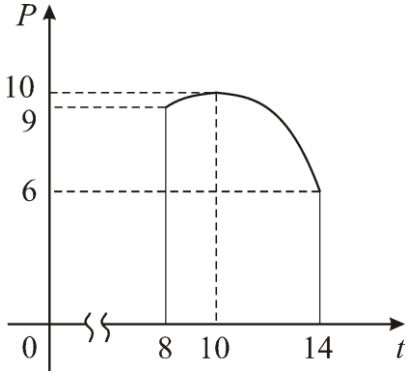


Fig. 15.6

So, over a specified period of time 54 units of production were produced.

In § 12.1, we considered, in particular, the marginal cost given by the derivative of the cost function: $S(x)$: $MS = S'(x)$. This derivative characterizes the cost of producing a unit of additional products. Consider the problem of finding the cost function for a given function of marginal cost.

Example **15.8.** The marginal cost function is given: $MS = 3x^2 - 40x + 125$, $x \in [0, 30]$. Find the cost function $S(x)$ and evaluate the costs in the production of 20 units of production, if the costs for the production of the first unit of production are known to be 100 monetary units.

Solution. The cost function is found by integration:

$$S(x) = \int_1^x MSdx + C$$

.

In this case, constant *C* is determined by condition $S(1) = 100$, so that $C = 100$, since the integral vanishes.

We obtain the cost function:

$S(x) = x^3 - 20x^2 + 125x + 100$.

Substituting $x = 20$ we find:

$S(20) = 2600$.

Another example of the application of a definite integral is a *discontinuity*. Discounting is a determination of the initial monetary amount *S* by its final value S_t after time *t* at an interest rate *p*. The problems of discounting are encountered in determining the economic efficiency of capital investments.

As was established earlier (see § 6.5), with continuous accrual of interest, the final amount is calculated according to formula $S_t = Se^{rt}$, where $r = \frac{p}{100}$. If $S_t = f(t)$, then the discounted amount at time *t* will be equal to $S = f(t)e^{-rt}$.

The total discounted amount S_d over time *T* is calculated by the formula:

$$S_\alpha = \int_0^T f(t)\, e^{-rt} dt \quad . \quad (15.6)$$

15.3 Applications of the definite integral in biology and chemical technique

Let us start with biological applications. We will consider the population size, population biomass, etc. as continuous functions of time.

Population size. The number of individuals in a population changes over time. If the living conditions of the population are favorable, then the birth rate exceeds mortality, and the total number of individuals in the population grows with time. We denote by $v(t)$ the population growth rate, i.e., the increase in the number of individuals per unit of time. In the "old", established populations that have long lived in this area, the growth rate $v(t)$ is low and slowly tends to zero. However, if the population is young, its relationship with other local populations has not yet been established, or there are external causes that change these relationships (for example, conscious human intervention), then $v(t)$ can fluctuate significantly, decreasing or increasing.

If the population growth rate $v(t)$ is known, then we can find the population growth over a period of time from t_1 to t_2. Indeed, it follows from the determination of $v(t)$ that it is a derivative of size $N(t)$ at moment t, and, therefore, size $N(t)$ is the primitive for $v(t)$. Hence

$$N(t_2) - N(t_1) = \int_{t_1}^{t_2} v(t)dt.$$

(15.7)

Under conditions of unlimited nutritional resources, the growth rate of many populations is known to be exponential: $v(t) = ae^{kt}$. The population in this case, as it were, “does not age”. Such conditions can be created, for

example, for microorganisms, replanting the developed culture from time to time in new containers with a nutrient medium. Applying formula (15.7), in this case we obtain

$$N(t_1) = N(t_0) + a\int_{t_0}^{t_1} e^{kt}dt = N(t_0) + \frac{a}{k}(e^{kt_1} - e^{kt_0}).$$

(15.8)

According to a formula similar to (15.8), in particular, the number of cultivated mold fungi that secrete penicillin is calculated.

The biomass of the population. Consider a population in which the weight of an individual changes appreciably throughout life and evaluate the total population mass.

Let τ be the age in various units of time, $N(t)$ be the number of individuals of the population whose age equals τ, and $M(\tau)$ is the biomass of all individuals aged 0 to τ.

Obviously, the product $N(\tau)P(\tau)$ is equal to the biomass of all individuals of age τ. Consider the difference $M(\tau + \Delta\tau) - M(\tau)$. Obviously, this difference, equal to the biomass of all individuals aged τ to $\tau + \Delta\tau$, satisfies inequalities

$$n(\tau)p(\tau)\Delta\tau \le M(\tau + \Delta\tau) - M(\tau) \le \overline{N}(\tau)\overline{P}(\tau)\Delta\tau,$$

(15.9)

where $n(\tau)m(\tau)$ is the smallest, and $\bar{N}(\tau)\bar{M}(\tau)$ is the largest values of function $N(\tau)M(\tau)$ in the segment $[\tau, \tau + \Delta\tau]$.

Let T be the maximum age of an individual in the given population. As

$$M(T) - M(0) = \int_0^T N(\tau)P(\tau)d\tau,$$

and $M(0)$, obviously, equals 0, then

$$M = M(T) = \int_0^T N(\tau)P(\tau)d\tau.$$

Let us move on to chemical engineering. Many chemical reactions and physical processes are characterized by the fact that the rate of change of a variable is proportional to the value of the same variable in the first degree. Such processes are called first-order processes.

These processes are described by the equation:

$$\frac{dx}{d\tau} = kx$$

(15.10)

In the case of a chemical reaction, the values included here mean:

x - amount of substance;

k - constant value (reaction rate constant);

τ - time.

Radioactive decay. Radioactive decay occurs in such a way that the decrease in the number of atoms $-dN$ over time $d\tau$ is proportional to the number N of remaining atoms, i.e.:

$$-dN = \lambda N d\tau,$$

(15.11)

where λ is inherent to the given substance constant called a constant of radioactivity. It is required to calculate the number N of atoms that have not decayed by moment τ, if at the moment $\tau = 0$ there were N_0 atoms.

We divide both sides of (15.11) by N and integrate:

$$\int \frac{dN}{N} = -\lambda \int d\lambda,$$

Whence

$$\ln N = -\lambda\tau + C$$

(15.12)

The value of the integration constant C is found from the condition that $N = N_0$ at $\tau = 0$. Hence $C = \ln N_0$. Substituting this value in (15.12), we obtain:

$$\ln \frac{N}{N_0} = -\lambda\tau \quad ; \qquad N = N_0 e^{\lambda\tau}$$

(15.13)

Of particular interest is the determination of the time $\tau = t$ during which the number of atoms is halved. For this, it is necessary to put in formula (15.13)

$$\frac{N}{N_0} = \frac{1}{2}.$$

Then we have

$$-\lambda t = \ln \frac{1}{2},$$

whence:

$$t = \frac{1}{\lambda} \ln 2 = \frac{0{,}693}{\lambda}$$

(15.14)

Time t is called half-life. For instance, for radon $\lambda = 2{,}084 \cdot 10^{-6}$ sec^{-1}. Substituting this value in (15.14), we obtain the half-life of radon, which is $t = 3$, 15 days.

The average lifetime of an atom of a radioactive substance. Let N_0 be the number of atoms of a radioactive substance at time $\tau = 0$.

Based on the previous example, we calculate the average lifetime of one atom.

The number of atoms that have survived over time τ and decayed in a subsequent period of time $d\tau$, based on equations (15.11) and (15.13) is equal to:

$$dN = N_0 \lambda e^{-\lambda\tau} d\tau .$$

This expression represents the number of atoms having a duration of existence equal to τ. In order to obtain the average duration of the existence of an atom, it is necessary to multiply this number dN of atoms by time τ, during which these atoms existed, integrate over τ in range from $\tau = 0$ to $\tau = \infty$ and divide by the initial number of atoms N_0. We denote the desired average duration by θ. We have:

$$\theta = \frac{\lambda}{N_0} \int_0^\infty N_0 e^{-\lambda\tau} \tau d\tau = \frac{1}{\lambda} .$$

Since for radon $\lambda = 2{,}084 \cdot 10^{-6} sec^{-1}$, the average lifetime of a radon atom is:

$$\theta = 10^6 : 2{,}084 = 5{,}552\ days = 133{,}26\ hours$$

$\theta = 10^6 : 2{,}084 = 5{,}552\ days = 133{,}26\ hours$.

Questions:

1. How is the area of a flat figure expressed using a definite integral?
2. The volume of which bodies and how can be evaluated with the help of definite integrals?
3. Is it possible to express the volume of output in case the function of labor productivity is known and a definite integral is used?
5. How is the full discounted amount expressed with continuous interest accrual?

Chapter 16. Improper integrals

16.1. Improper integrals with infinite integration limits

Introducing the concept of a definite integral, we assumed that the segment of integration is *finite*, and the integrand is *bounded* on this segment. Now we will consider cases when at least one of these conditions is not satisfied.

Definition. Let the function $y = f(x)$ be defined on the infinite interval $[a, +\infty)$ and integrable on any finite segment $[a, b]$, $b > a$, i.e. for any $b > a$ there exists a definite integral

$$\Phi(b) = \int_a^b f(x)\,dx .$$

Then an **improper integral** $\int_a^{+\infty} f(x)\,dx$ of function $f(x)$ on the interval $[a, +\infty)$ is limit

$$\lim_{b\to+\infty} \Phi(b) = \lim_{b\to+\infty} \int_a^b f(x)\,dx .$$

Therefore, by definition

$$\int_a^{+\infty} f(x)\,dx = \lim_{b\to+\infty} \int_a^b f(x)\,dx . \qquad (16.1)$$

If the limit on the right-hand side of equality (16.1) exists and is finite, then an improper integral is called **converging**, otherwise **diverging**. If

improper integral (16.1) converges (i.e., is convergent), then the function is called **integrable** on $[a, +\infty)$.

***Example* 16.1.** Evaluate integrals: а) $\int_1^{+\infty} \frac{dx}{x^2}$; б) $\int_1^{+\infty} \frac{dx}{\sqrt{x}}$.

Solution.

а) $$\int_1^{+\infty} \frac{dx}{x^2} = \lim_{b\to+\infty} \int_1^b x^{-2}dx = \lim_{b\to+\infty}\left(-\frac{1}{x}\right)\Big|_1^b = \lim_{b\to+\infty}\left(-\frac{1}{b}+1\right) = 1,$$

i.e. the integral converges to 1.

б) $$\int_1^{+\infty} \frac{dx}{\sqrt{x}} = \lim_{b\to+\infty} \int_1^b x^{-\frac{1}{2}}dx = 2\lim_{b\to+\infty}\left(\sqrt{b}-1\right) = \infty,$$

i.e. the integral diverges.

***Example* 16.2.** Establish at what values α the integral $\int_1^{+\infty} \frac{dx}{x^\alpha}$ converges, and at what diverges.

Solution. At $\alpha \neq 1$ we have

$$\int_1^b \frac{dx}{x^\alpha} = \frac{1}{1-\alpha} x^{1-\alpha}\Big|_1^b = \frac{1}{1-\alpha}\left(b^{1-\alpha}-1\right);$$

at $\alpha = 1$:

$$\int_1^b \frac{dx}{x} = \ln x\Big|_1^b = \ln b.$$

That's why at $\alpha \neq 1$:

$$\int_1^{\infty} \frac{dx}{x^\alpha} = \lim_{b\to\infty} \frac{1}{1-\alpha}\left(b^{1-\alpha}-1\right),$$

at $\alpha = 1$:

$$\int_1^{+\infty} \frac{dx}{x} = \lim_{b \to +\infty} \ln b$$.

So, we can draw the following conclusions:

if $\alpha > 1$, then $\int_1^{+\infty} \frac{dx}{x^\alpha} = \frac{1}{\alpha - 1}$, the integral converges;

if $\alpha < 1$, then $\int_1^{+\infty} \frac{dx}{x^\alpha} = \infty$, the integral diverges;

if $\alpha = 1$, then $\int_1^{+\infty} \frac{dx}{x} = \infty$, the integral diverges.

Similarly to the improper integral defined by equality (16.1), an *improper integral with an infinite lower limit* is determined, namely:

$$\int_{-\infty}^{b} f(x)\,dx = \lim_{a \to -\infty} \int_a^b f(x)\,dx \quad . \qquad (16.2)$$

Finally, we can consider an *improper integral with infinite lower and upper limits*:

$$\int_{-\infty}^{+\infty} f(x)\,dx$$.

To do so, we take an arbitrary point *c*. It will split the number line into two half-lines. If improper integrals $\int_{-\infty}^{c} f(x)\,dx$ и $\int_c^{+\infty} f(x)\,dx$ exist, then by definition the improper integral $\int_{-\infty}^{+\infty} f(x)\,dx$ exists as well. In this case, we say

$$\int_{-\infty}^{+\infty} f(x)\,dx = \int_{-\infty}^{c} f(x)\,dx + \int_{c}^{+\infty} f(x)\,dx \quad . \qquad (16.3)$$

It can be proved that the right-hand side of equality (16.3) does not depend on the choice of point c.

***Example* 16.3.** Evaluate $\int_{-\infty}^{+\infty} \frac{dx}{1+x^2}$.

Solution: $\int_{-\infty}^{+\infty} \frac{dx}{1+x^2} = \int_{-\infty}^{0} \frac{dx}{1+x^2} + \int_{0}^{+\infty} \frac{dx}{1+x^2}$.

We calculate each of the integrals on the right-hand side of the last equality:

$$\int_{-\infty}^{0} \frac{dx}{1+x^2} = \lim_{a\to-\infty} \int_{a}^{0} \frac{dx}{1+x^2} = \lim_{a\to-\infty} \operatorname{arctg} x\Big|_{a}^{0} = \lim_{a\to-\infty} \left(\operatorname{arctg} 0 - \operatorname{arctg} a\right) = \frac{\pi}{2}$$

,

$$\int_{0}^{+\infty} \frac{dx}{1+x^2} = \lim_{b\to+\infty} \int_{0}^{b} \frac{dx}{1+x^2} = \lim_{b\to+\infty} \operatorname{arctg} x\Big|_{0}^{b} = \lim_{b\to+\infty} \left(\operatorname{arctg} b - \operatorname{arctg} 0\right) = \frac{\pi}{2}$$

.

Consequently,

$$\int_{-\infty}^{+\infty} \frac{dx}{1+x^2} = \frac{\pi}{2} + \frac{\pi}{2} = \pi \quad .$$

Improper integrals with infinite limits, i.e. integrals (16.1), (24.2) and (24.3) are called **improper integrals of the first kind**.

16.2. Improper integrals of unbounded functions

Let function $y = f(x)$ be defined on interval $[a, b)$. The point $x = b$ will be called **singular**, if $f(x)$ is not bounded in any neighborhood of this point, but is bounded and integrable on any segment enclosed in the interval $[a, b)$.

Definition. If $f(x)$ is not bounded on $[a, b)$, but is integrable on any interval $[a, b-\varepsilon]$, $0 < \varepsilon < b - a$, then by an **improper integral** $\int_a^b f(x)\,dx$ of function $f(x)$ over $[a, b]$ the following limit is called:

$$\lim_{\varepsilon \to 0} \int_a^{b-\varepsilon} f(x)\,dx \quad . \qquad (16.4)$$

Thus

$$\int_a^b f(x)\,dx = \lim_{\varepsilon \to 0} \int_a^{b-\varepsilon} f(x)\,dx \quad .(16.5)$$

If the limit (24.4) exists and is finite, then integral (16.5) is called **convergent**, otherwise the integral is called **divergent**.

An improper integral is defined in a similar way when the left end of the interval is a singular point:

$$\int_a^b f(x)\,dx = \lim_{\varepsilon \to 0} \int_{a+\varepsilon}^b f(x)\,dx \quad . \qquad (16.6)$$

Finally, if c is the only internal singular point on $[a, b]$, then by definition

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx \quad . \quad (16.7)$$

Improper integrals of unbounded functions are called **improper integrals of the second kind**.

***Example* 16.4.** Evaluate the improper integral $\int_0^1 \frac{dx}{\sqrt{1-x}}$.

Solution. By formula (24.5) we obtain

$$\int_0^1 \frac{dx}{\sqrt{1-x}} = \lim_{\varepsilon\to 0}\int_0^{1-\varepsilon} \frac{dx}{\sqrt{1-x}} = -2\lim_{\varepsilon\to 0}\sqrt{1-x}\Big|_0^{1-\varepsilon} = -2\lim_{\varepsilon\to 0}\left(\sqrt{\varepsilon}-1\right) = 2 \quad .$$

***Example* 16.5.** Evaluate the improper integral $\int_0^1 \frac{dx}{\sqrt{x}}$.

Solution: $\int_0^1 \frac{dx}{\sqrt{x}} = \lim_{\varepsilon\to 0}\int_\varepsilon^1 \frac{dx}{\sqrt{x}} = \lim_{\varepsilon\to 0} 2\sqrt{x}\Big|_\varepsilon^1 = \lim_{\varepsilon\to 0}\left(2-2\sqrt{\varepsilon}\right) = 2$.

Note that the improper integral $\int_0^1 \frac{dx}{x^\alpha}$, where $\alpha > 0$, *converges* at $0 < \alpha < 1$ (and equals $\frac{1}{1-\alpha}$) and diverges at $\alpha \geq 1$. (Check it out yourself.)

16.3. Improper integrals convergence tests

In the calculation and study of improper integrals, a significant place is occupied by the study of their convergence. In many cases, it is sufficient to establish whether a given integral converges or diverges and evaluate its value. To study convergence, in particular, comparison tests are used,

based on the comparison of a given integral with an integral, the convergence of which is known.

We accept without proof the following statement (**comparison test**).

***Theorem* 16.1.** Let the functions $f(x)$ и $g(x)$ are continuous on the interval $[a, +\infty)$ and satisfy the following condition on it: $0 \le f(x) \le g(x)$. Then:

1) from convergence of $\int_a^{+\infty} g(x)\,dx$ convergence of $\int_a^{+\infty} f(x)\,dx$ follows;

2) from divergence of $\int_a^{+\infty} f(x)\,dx$ divergence of $\int_a^{+\infty} g(x)\,dx$ follows.

Let us look at some examples.

***Example* 16.6.** Investigate convergence of the integral $\int_1^{+\infty} \frac{\sqrt{x^3}+1}{x^2}\,dx$.

Solution. Obviously,

$$\frac{\sqrt{x^3}+1}{x^2} > \frac{\sqrt{x^3}}{x^2} = \frac{1}{\sqrt{x}}.$$

But the integral $\int_1^{+\infty} \frac{dx}{\sqrt{x}}$ diverges (see Example 16.1). Consequently, this integral also diverges.

***Example* 16.7.** Investigate convergence of the integral $\int_3^{+\infty} \frac{dx}{x\sqrt{x}(x-1)}$.

Solution. Compare the integrand $f(x)=\dfrac{1}{x\sqrt{x}(x-1)}$ with the function $g(x)=\dfrac{1}{x\sqrt{x}}$ on $[3,+\infty)$. Obviously, $\dfrac{1}{x\sqrt{x}(x-1)}<\dfrac{1}{x\sqrt{x}}$.

But the integral $\int\limits_3^{+\infty}\dfrac{dx}{x\sqrt{x}}$ converges (see Example 16.2 allowing $\alpha=\dfrac{3}{2}$). Consequently, this integral also converges.

A similar comparison test also holds for **improper integrals of the second kind**: if functions $f(x)$ and $g(x)$ are continuous on interval $[a,b)$ and for all x in some neighborhood of the singular point b the conditions $0\le f(x)\le g(x)$ are satisfied, then

1) from convergence of $\int\limits_a^b g(x)\,dx$ convergence of $\int\limits_a^b f(x)\,dx$ follows;

2) from divergence of $\int\limits_a^b f(x)\,dx$ divergence of $\int\limits_a^b g(x)\,dx$ follows.

***Example* 16.8.** Investigate convergence of the integral $\int\limits_0^1\dfrac{dx}{\sqrt{x}+3x^2}$.

Solution. The singular point here is $x=0$. Let us compare the integrand with the function: $\dfrac{1}{\sqrt{x}}$:

$$\frac{1}{\sqrt{x}+3x^2} < \frac{1}{\sqrt{x}}.$$

The improper integral $\int_0^1 \frac{dx}{\sqrt{x}}$ converges (see Example 16.5). Therefore, this integral converges.

We give, without proof, one more comparison test also commonly used in practice.

***Theorem* 16.2.** If $f(x)$ and $g(x)$ are non-negative functions and there exists a finite limit $\lim\limits_{x\to+\infty} \frac{f(x)}{g(x)} = A \neq 0$, then improper integrals $\int_a^{+\infty} f(x)\,dx$ and $\int_a^{+\infty} g(x)\,dx$ converge or diverge simultaneously.

***Example* 16.9.** The improper integral $\int_3^{+\infty} \frac{dx}{x^2-2x}$ converges. Indeed, the limit of the relation of a function $f(x)=\frac{1}{x^2-2x}$ to a function $g(x)=\frac{1}{x^2}$ is finite: $\lim\limits_{x\to+\infty} \frac{f(x)}{g(x)} = \lim\limits_{x\to+\infty} \frac{x^2}{x^2-2x} = 1$, and the integral $\int_3^{+\infty} \frac{dx}{x^2}$ converges.

Absolute and conditional convergence of improper integrals

We note an important property of improper integrals that distinguishes them from ordinary definite integrals.

For a definite integral, as is known, the following statement is true: if $\int_a^b f(x)\,dx$, then $\int_a^b |f(x)|\,dx$ exists.

In the case of *improper integrals*, the following statement holds: if the improper integral $\int_a^{+\infty} |f(x)|\,dx$ converges, then $\int_a^{+\infty} f(x)\,dx$ converges as well.

We accept this statement also without proof. The converse statement, generally speaking, is not true: the convergence of the integral $\int_a^{+\infty} f(x)\,dx$ does not imply the convergence of the integral $\int_a^{+\infty} |f(x)|\,dx$.

If the improper integral $\int_a^{+\infty} |f(x)|\,dx$ converges, then the integral $\int_a^{+\infty} f(x)\,dx$ is said to **converge absolutely**.

If the integral $\int_a^{+\infty} f(x)\,dx$ converges, but the integral $\int_a^{+\infty} |f(x)|\,dx$ diverges, then the integral $\int_a^{+\infty} f(x)\,dx$ is said to **converge conditionally**.

It follows from the above that the absolute convergence of the integral implies its convergence (in the usual sense): an absolutely convergent integral converges.

***Example* 16.10.** Investigate convergence of the integral $\int_1^{+\infty} \frac{\cos x}{x^2} dx$.

Solution. Obviously, $\left|\frac{\cos x}{x^2}\right| \le \frac{1}{x^2}$. We know that the integral $\int_1^{+\infty} \frac{dx}{x^2}$ converges. Consequently, $\int_1^{+\infty} \left|\frac{\cos x}{x^2}\right| dx$ converges, i.e. the given integral $\int_1^{+\infty} \frac{\cos x}{x^2} dx$ converges absolutely. It follows that it converges.

Questions

How is the improper integral of a function on an infinite half-interval of the form $[a, +\infty)$ defined? In which case is the integral called convergent, and in which –

divergent?

2. How is the improper integral of a function on an infinite interval of the form $(-\infty, b]$ determined?

3. How to determine the improper integral of a function on an infinite interval $(-\infty, \infty)$?

4. What improper integrals are called improper integrals of the first kind?

5. How is the improper integral of an unbounded function determined in the case when the singular point is one of the ends of the integration segment?

6. How is the improper integral of an unbounded function defined when the singular point is the inner point of the integration segment?

7. What are the comparison tests for improper integrals of the first and second kind?

8. Does an absolutely convergent improper integral always converge?

9. Which improper integral is called conditionally convergent?

Chapter 17. Elements of analytical geometry in space

17.1. Vectors

Many physical quantities or characteristics of the phenomena around us are determined by setting a number. For example, the body weight, its temperature, the cost of goods, the number of seats in the classroom, etc. Such values are called **scalar** values, or simply **scalars**. But there are also such quantities, which are determined not only by the number but also by indicating the direction. For example, when studying the action of a force, it is necessary to specify not only the value of this force, but also the direction of its action. Such quantities are called **vector quantities**, or simply **vectors**.

A directed segment on which the beginning, end, and direction are specified is called a **vector**. A vector is denoted either by a symbol $\overline{AB}$, where A is its beginning and B is its end, or by one letter with a line at the top, for example $\bar{a}$. The **length** of a vector (or its **modulus,** or its **magnitude**) is the distance between its beginning and end. Usually, the length of the vector is denoted by $|\overline{AB}|$ or $\bar{a}$.

Vectors $\bar{a}$ and $\bar{b}$ are called **collinear** if they lie on the same line or on parallel lines.

Vectors $\bar{a}$ and $\bar{b}$ are called equal if they are collinear, have the same direction and their magnitudes are equal.

Let a rectangular coordinate system be given in space and let the coordinates of the beginning and end of the vector $\overline{AB}$ are $A(x_1, y_1, z_1)$

and $B(x_2, y_2, z_2)$ respectively. Then the coordinates of this vector are determined by the formulas

$$X = x_2 - x_1, \quad Y = y_2 - y_1, \quad Z = z_2 - z_1.$$

Obviously, the magnitude $|\overline{AB}|$ of the vector $\overline{AB}$ is determined by the formula

$$|\overline{AB}| = \sqrt{X^2 + Y^2 + Z^2}.$$

A vector is called **zero vector** if its beginning and end coincide. The zero vector has no definite direction and has a length equal to zero. We can assume that the zero vector is directed identically with any vector. When writing, we will identify the zero vector $\bar{0}(0,0,0)$ with the real number zero.

Two linear operations are defined over vectors - addition of vectors and multiplication of a vector by a number. Let two vectors $\bar{a} = (a_1, a_2, a_3)$ and $b = (b_1, b_2, b_3)$. be given.

The sum of the vectors is the following vector

$$\bar{a} + \bar{b} = (a_1 + b_2, a_2 + b_2, a_3 + b_3).$$

The product of a vector $\bar{a}$ **by a number** k is the following vector

$$k\bar{a} = (ka_1, ka_2, ka_3).$$

Let us give the basic properties of linear operations (they are easily verified). Here $\bar{a}$, $\bar{b}$, $\bar{c}$ are vectors, k, k_1, k_2 are numbers.

1. $\bar{a} + \bar{b} = \bar{b} + \bar{a}.$
2. $(\bar{a} + \bar{b}) + \bar{c} = \bar{a} + (\bar{b} + \bar{c}).$

3. $k_1\left(k_2\bar{a}\right)=\left(k_1k_2\right)\bar{a}.$

4. $\left(k_1+k_2\right)\bar{a}=k_1\bar{a}+k_2\bar{a}.$

5. $k\left(\bar{a}+\bar{b}\right)=k\bar{a}+k\bar{b}.$

From the definition of collinearity of vectors and the definition of the product of a vector and a number, it follows that two vectors $\bar{a}=\left(a_1,a_2,a_3\right)$ and $\bar{b}=\left(b_{1,}b_2,b_3\right)$ are collinear if and only if their coordinates are proportional:

$$\frac{a_1}{b_1}=\frac{a_1}{b_2}=\frac{a_3}{b_3}.$$

(17.1)

Denoting the general value of relations (17.1) by $k,$ we obtain the collinearity condition in the form

$$\bar{a}=k\bar{b}.$$

Vectors are called **coplanar** if they lie either in the same plane or in parallel planes.

17.2. Scalar product of vectors

Let $\bar{a}$ and $\bar{b}$ be vectors, φ - the angle between them.

The scalar product $\left(\bar{a},\bar{b}\right)$ **of vectors** $\bar{a}$ and $\bar{b}$ is the number defined by the formula

$$\left(\bar{a},\bar{b}\right)=\left|\bar{a}\right|\left|\bar{b}\right|\cos\varphi$$

(17.2)

Let us list the main properties of scalar products (they can easily be deduced from the definition).

1. $(\bar{a},\bar{b})=(\bar{b},\bar{a})$ - commutativity.
2. $(k\bar{a},\bar{b})=k(\bar{a},\bar{b})$ - associativity with respect to multiplication by a number.
3. $(\bar{a},(\bar{b}+\bar{c}))=(\bar{a},\bar{b})+(\bar{a},\bar{c})$ - distributivity.
4. $(\bar{a},\bar{a})=|\bar{a}|^2$ - the formula of a scalar square.
5. $(\bar{a},\bar{b})=0$ for nonzero vectors $\bar{a}$ and $\bar{b}$ if and only if vectors $\bar{a}$ и $\bar{b}$ are mutually perpendicular.

Let the vectors $\bar{i},\ \bar{j},\ \bar{k}$ be the unit vectors of the coordinate axes. Then, obviously, $(\bar{i},\bar{i})=(\bar{j},\bar{j})=(\bar{k},\bar{k})=1,\ \ (\bar{i},\bar{j})=(\bar{i},\bar{k})=(\bar{j},\bar{k})=0$, and the scalar product of vectors $\bar{a}=(a_1,a_2,a_3)$ and $\bar{b}=(b_1,b_2,b_3)$ is expressed via their coordinates as follows:

$$(\bar{a},\bar{b})=a_1b_1+a_2b_2+a_3b_3$$

(17.3)

From formulas (17.2) and (17.3) we obtain the formula for determining the angle between vectors:

$$\cos\varphi=\frac{(\bar{a},\bar{b})}{|\bar{a}|\cdot|\bar{b}|}=\frac{a_1b_1+a_2b_2+a_3b_3}{|\bar{a}|\cdot|\bar{b}|}$$

(17.4)

17.3 Equations of a surface and a line.

Let an equation

$$F(x,y,z) = 0F(x,y,z) = 0$$

(17.5)

be given in a rectangular coordinate system $Oxyz$.

Equation (17.5) is called the equation of the surface *L* if the coordinates of any point lying on the surface *L* satisfy this equation and the coordinates of any point not lying on this surface do not satisfy.

A line in space can be considered as the intersection of two surfaces, i.e. as a set of points located simultaneously on two surfaces. Therefore, a system of two equations of the form (17.5)

$$\begin{cases} F_1(x,y,z) = 0 \\ F_2(x,y,z) = 0 \end{cases}$$

(17.6)

is called the equation of the line in space if this equation satisfies the coordinates of all those and only those points that lie on the line L. In particular, if the equations in system (17.6) are equations of planes, then system (17.5) is an equation of a line.

17.4. Plane in space

Let the coordinate system *Oxyz* be given in space and let the plane Π pass through the point $M_0(x_0, y_0, z_0)$ perpendicularly to the vector $\overline{N} = (A, B, C)$. These two conditions determine the *only* plane in space *Oxyz*. Vector $\overline{N}$ is called the **normal vector** of the plane Π. Let us derive the equation of this plane.

We take an arbitrary point $M(x,y,z)M(x,y,z)$ in plane Π. Then vectors $\overline{M_0M}=(x-x_0,y-y_0,z-z_0)$ and $\overline{N}=(A,B,C)$ will be mutually perpendicular. Therefore, their scalar product is equal to zero: $\left(\overline{N},\overline{M_0M}\right)=0$. Let us write the last equality in scalar form:

$$A\cdot(x-x_0)+B\cdot(y-y_0)+C\cdot(z-z_0)=0. \quad (17.7)$$

This is the **equation of a plane** passing through a point $M_0(x_0,y_0,z_0)$ perpendicularly to a given vector $\overline{N}=(A,B,C)$. From (17.7) we obtain

$$Ax+By+Cz-Ax_0-By_0-Cz_0=0.$$

Denoting $-Ax_0-By_0-Cz_0=D$, we obtain the **general equation of the plane**:

$$Ax+By+Cz+D=0. \quad (17.8)$$

So, the plane equation is a linear equation, or an equation of the first degree with three variables.

It is not difficult to prove the converse statement: *any equation of the first degree with three variables is an equation of the plane*.

Equation (17.8) is called complete if all its coefficients A,B,C, and D are nonzero. Consider the different types of incomplete equations.

If $D=0,$ the plane passes through the origin.

If $A=0,$ the plane is parallel to the axis $Ox.$ The situation is similar with the condition that the plane (17.8) is parallel to other coordinate axes, i.e. if one of the coordinates does not enter into the equation of the plane, then the plane is parallel on the corresponding axis.

If two coordinates in the equation are missing, then the plane is parallel to the corresponding coordinate plane, moreover, if $D=0$, then the

equation $x = 0$ (resp., $y = 0, z = 0$) is the equation of the coordinate plane itself.

Suppose that in equation (17.8) all the coefficients (and the free term) are nonzero, i.e. the equation is complete. Transform this equation:

$$Ax + By + Cz = -D$$

$$\frac{Ax}{-D} + \frac{By}{-D} + \frac{Cz}{-D} = 1$$

Let

$$-\frac{D}{A} = a, \quad -\frac{D}{B} = b, \quad -\frac{D}{C} = c.$$

Then the equation of a plane (17.8) takes the form

$$\frac{x}{a} + \frac{y}{b} + \frac{z}{c} = 1$$

(17.9)

The last equation is called the *equation of the plane in segments.* This name is explained by the fact that the denominators a, b, and c are the segments cut off by the plane from the coordinate axes.

Consider the relative position of two planes. There are two planes given:

$$A_1x + B_1y + C_1z + D_1 = 0,$$

$$A_2x + B_2y + C_2z + D_2 = 0.$$

Their normal vectors are, obviously, $\overline{N}_1 = (A_1, B_1, C_1)$ and $\overline{N}_2 = (A_2, B_2, C_2)$.

The angle between these planes is the angle between $\overline{N}_1$ and $\overline{N}_2$ and is determined by the formula:

$$\cos\varphi = \frac{A_1A_2 + B_1B_2 + C_1C_2}{\sqrt{A_1^2 + B_1^2 + C_1^2}\sqrt{A_2^2 + B_2^2 + C_2^2}}. \quad (17.10)$$

*The condition for the **parallelism** of two planes* is the condition for the proportionality of their normal vectors:

$$\frac{A_1}{A_2} = \frac{B_1}{B_2} = \frac{C_1}{C_2}. \quad (17.11)$$

*The condition for the **coincidence** of the planes* is the following:

$$\frac{A_1}{A_2} = \frac{B_1}{B_2} = \frac{C_1}{C_2} = \frac{D_1}{D_2}. \quad (17.12)$$

*The condition for their **perpendicularity*** is the condition $\cos\varphi = 0$, i.e.

$$A_1A_2 + B_1B_2 + C_1C_2 = 0. \quad (17.13)$$

17.5. Straight line in space. straight line and plane in space

Let a straight line *L* pass through a point $M_0(x_0, y_0, z_0)$ parallel to the vector $\bar{s} = (l, m, n)$. In this case, the vector $\bar{s}$ will be called the **directing vector** of the straight line. Let $M(x, y, z)$ be an arbitrary point of straight line *L*. Obviously, the vectors $\overline{M_0M} = (x - x_0, y - y_0, z - z_0)$ and $\bar{s}$ are proportional. Having written down the condition of their proportionality in coordinate form, we obtain the **canonical equation of a straight line**:

$$\frac{x - x_0}{l} = \frac{y - y_0}{m} = \frac{z - z_0}{n}. \quad (17.14)$$

If the line passes through the points $M_0(x_0, y_{0,}, z_0)$ and $M_1(x_1, y_1, z_1)$, then we can take $\overline{M_0M_1}$ as the directing vector and equation (17.14) will have the form

$$\frac{x-x_0}{x_1-x_0}=\frac{y-y_0}{y_1-y_0}=\frac{z-z_0}{z_1-z_0}.$$

$(17.14')$

From equation (17.14) we obtain:

$$x-x_0=lt,\quad y-y_0=mt,\quad z-z_0=nt,$$

where *t* is the coefficient of proportionality. Hence

$$x=x_0+lt,\quad y=y_0+mt,\quad z=z_0+nt. \qquad (17.15)$$

These are **parametric equations of a line** *L*. (Sometimes they are called in the singular - the parametric equation of the line.)

A straight line in space can also be defined as the line of intersection of two planes, i.e. as a set of points whose coordinates satisfy the system:

$$\begin{cases} A_1x+B_1y+C_1z+D_1=0 \\ A_2x+B_2y+C_2z+D_2=0. \end{cases} \qquad (17.16)$$

Canonical equation (17.13), however, can also be viewed as a pair of plane equations considered together. It is easy to derive the canonical or parametric equation of a line defined in the form (17.16). To do this, it is enough to find some point $M_0(x_0, y_0, z_0)$ belonging to a straight line and a directing vector. Coordinates of point M_0 are easy to find - this is any solution to the system (17.16). For example, setting $z_0=0$, from the system (17.16) we find x_0 and y_0, and obtain $M_0(x_0, y_0, 0)$. Let now $z_1=1$. From the system, we find x_1 and y_1. We obtain $M_1(x_1, y_1, 1)$.

The vector $\overline{M_0M_1}$ is the directing vector of the line (17.16), and we can write its canonical equation.

Example 17.1. Write the canonical equation of a line that is the intersection of the planes

$$2x+3y+5z-3=0,$$
$$x+y+2z-1=0.$$

Solution. Let $z=0.$ Then the previous equations will take the form

$$2x+3y=3,$$
$$x+y=1.$$

Solving this system of equations, we find $x=0,\ y=1.$ Thus, a point $M_0(0,1,0)$ lies on our line. Let now $z=1.$ Then to define x and y we obtain equations

$$2x+3y=-2,$$
$$x+y=-1,$$

from which we find $x=-1,\ y=0.$ Therefore, the other point of our line is the point $M_1(-1,0,1).$ Applying formula (17.14), we obtain the canonical equation

$$\frac{x}{-1}=\frac{y-1}{-1}=\frac{z}{1}.$$

Let us now consider the relative position of a line and a plane in space. Let line L be given as

$$\frac{x-x_0}{l}=\frac{y-y_0}{m}=\frac{z-z_0}{n}$$

and plane Π as

$$Ax+By+Cz+D=0.$$

Obviously, line *L* is parallel to plane Π when the directing vector $\bar{s} = (l, m, n)$ of the line is perpendicular to the normal vector $\bar{N} = (A, B, C)$ of the plane, i.e. the *condition of* ***parallelism*** *of a straight line L and a plane* Π is the following :

$$Al + Bm + Cn = 0 . \quad (17.17)$$

The condition for the proportionality of these vectors is the *condition for the* ***perpendicularity*** *of line L and tlane* Π:

$$\frac{A}{l} = \frac{B}{m} = \frac{C}{n} . \quad (17.18)$$

The angle between a line and a plane is the angle between the line and its projection onto the plane, and this is the angle additional to the angle between the vector director $\bar{s}$ of the line *L* and the normal vector $\bar{N}$ of the plane Π:

$$\sin\varphi = \left|\cos\widehat{\bar{N}, \bar{s}}\right| = \frac{Al + Bm + Cn}{\sqrt{A^2 + B^2 + C^2} \cdot \sqrt{l^2 + m^2 + n^2}} . \quad (17.19)$$

The *distance from a point to a plane* is calculated using a formula similar to the formula for the distance from a point to a line on a plane. Let us show that the distance *d* from point $M_0(x_0, y_0, z_0)$ to the plane

$$Ax + By + Cz + D = 0 . \quad (17.8)$$

is calculated by the formula

$$d = \frac{|Ax_0 + By_0 + Cz_0 + D|}{\sqrt{A^2 + B^2 + C^2}} . \quad (17.20)$$

Let us write the equation of a line passing through point $M_0(x_0, y_0, z_0)$ perpendicularly to the plane (17.8). To do so, we use parametric equations (17.15) (17.15):

$x = x_0 + lt$, $y = y_0 + mt$, $z = z_0 + nt$.

In order for the line (17.8) to be perpendicular to the plane (4.2), it is necessary that its directing vector $\bar{s} = (l, m, n)$ is parallel to the vector $\bar{N} = (A, B, C)$, i.e. so that the coordinates of the vectors $\bar{s}$ and $\bar{N}$ are proportional. The easiest way, of course, is to take as a $\bar{s}$ vector $\bar{N}$, i.e. we take $l = A$, $m = B$, $n = C$. Then parametric equations (17.15) will look like as follows:

$$x = x_0 + At, \quad y = y_0 + Bt, \quad z = z_0 + Ct. \qquad (17.21)$$

The straight line (17.21) is perpendicular to the plane (17.8) and passes through point M_0. Therefore, the distance from point M_0 to the plane (17.8) is the distance between point M_0 and point *M* of the intersection of the line (17.21) with the plane (17.8). Let us find the coordinates of point *M*. To do so, it is necessary to solve equations (4.2) and (4.9 ') together. The easiest way to do it is by substituting the expressions for *x*, *y*, and *z* from (17.21) into (17.8). In turn:

$$A(x_0 + At) + B(y_0 + Bt) + C(z_0 + Ct) + D = 0,$$

$$(A^2 + B^2 + C^2)t + (Ax_0 + By_0 + Cz_0 + D) = 0.$$

Hence, we find *t*:

$$t = -\frac{Ax_0 + By_0 + Cz_0 + D}{A^2 + B^2 + C^2}.$$

This value of *t* determines the coordinates of point *M*, which is the base of the perpendicular dropped from point M_0 to the plane (17.8). Substitute the sought *t* in (17.21):

$$x = x_0 + A\left(-\frac{Ax_0 + By_0 + Cz_0 + D}{A^2 + B^2 + C^2}\right),$$

$$y = y_0 + B\left(-\frac{Ax_0 + By_0 + Cz_0 + D}{A^2 + B^2 + C^2}\right), \qquad (17.22)$$
$$z = z_0 + C\left(-\frac{Ax_0 + By_0 + Cz_0 + D}{A^2 + B^2 + C^2}\right).$$

The distance *d* from the point M_0 to the plane (17.8) is the length of the perpendicular M_0M, or, which is the same, the distance between the points $M_0(x_0, y_0, z_0)$ and $M(x, y, z)$, i.e.

$$d = \sqrt{(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2}.$$

Considering *x*, *y* and *z* are determined by equalities (17.22), we obtain

$$d = \sqrt{(A^2 + B^2 + C^2)\left(-\frac{Ax_0 + By_0 + Cz_0 + D}{A^2 + B^2 + C^2}\right)^2} =$$
$$= \sqrt{A^2 + B^2 + C^2}\,\frac{|Ax_0 + By_0 + Cz_0 + D|}{A^2 + B^2 + C^2},$$

or

$$d = \frac{|Ax_0 + By_0 + Cz_0 + D|}{\sqrt{A^2 + B^2 + C^2}},$$

Q.E.D.

***Example* 17.2.** Find the distance from point $M_0(1, 0, 2)$ to the plane $x + 2y - 2z + 9 = 0$.

Solution.

$$d = \frac{|1 + 2\cdot 0 - 2\cdot 2 + 9|}{\sqrt{1^2 + 2^2 + (-2)^2}} = \frac{6}{3} = 2.$$

***Example* 17.3.** Find the distance from the straight line

$$\frac{x+1}{2}=\frac{y-2}{2}=\frac{z}{1}$$

to the plane $4x-2y-4z+9=0$.

Solution. The line is parallel to the plane. Indeed, the scalar product of its directing vector and the normal plane vector is zero: $2\cdot4+2\cdot(-2)+1\cdot(-4)=0$. Therefore, the distance from a straight line to a plane is equal to the distance from any point M_0 of this straight line to the plane. It is most convenient to take as M_0 the point $(-1,\ 2,\ 0)$, whose coordinates appear in the equation of the line. We obtain

$$d=\frac{\left|4\cdot(-1)-2\cdot2-4\cdot0+9\right|}{\sqrt{4^2+(-2)^2+(-4)^2}}=\frac{1}{6}.$$

***Example* 17.4.** Find the distance from the point $M_0(1,\ 2,\ 3)$ to the straight line

$$\frac{x-6}{2}=\frac{y}{-2}=\frac{z-7}{1}.$$

Solution. We write the equation of the plane that passes through the given point M_0 and is perpendicular to the given line and find the coordinates of point M of the intersection of the line and the plane. Obviously, M_0M is the perpendicular dropped from the point M_0 to a given line. Its length is the desired distance.

The equation of a plane passing through M_0 and perpendicular to the given line is

$$2\cdot(x-1)-2\cdot(y-2)+1\cdot(z-3)=0,$$

or

$$2x-2y+z-1=0. \quad (*)$$

Let us write the equation of this line in parametric form:

$x = 6 + 2t$, $y = -2t$, $z = 7 + t$. (**)

Let us now find the intersection point of the line (**) and the plane (*). To do it, we first substitute *x*, *y* and *z* from (**) into (*) and find *t*:

$2\cdot(6 + 2t) - 2\cdot(-2t) + 7 + t - 1 = 0$,

$9t + 18 = 0$, $t = -2$.

Now, substituting the found value $t = -2$ into (**), we obtain $x = 2$, $y = 4$, $z = 5$. So, the point $M(2, 4, 5)$ is the base of the perpendicular M_0M. Therefore

$$d = M_0M = \sqrt{(2-1)^2 + (4-2)^2 + (5-3)^2} = 3.$$

Note that there is another way to solve this and similar problems, based on the concept of a vector product of vectors, which is not considered here.

17.6. The second order surfaces

Second-order surfaces are surfaces in three-dimensional space, which are determined by algebraic equations of the second degree. A brief study of second-order surfaces is carried out according to their equations by the method of parallel sections. The simplest of these surfaces are *second-order cylindrical surfaces.*

Let a line L lie in plane Oxy. Its equation is

$F(x, y) = 0$. (17.23)

Draw a line parallel to the axis $OzOz$ through each point of line L. The set of these lines forms a certain surface S, called a *cylindrical* one. The mentioned lines are called *generatrices* of surface S, and the initial line L is called its *directrix.* Obviously, equation (17.23) is also the surface

equation of S. So, the equation of a cylindrical surface with generatrix parallel to the axis Oz does not contain coordinate z and coincides with the equation of the directrix.

In particular, if the directrix is an ellipse defined on plane Oxy by the equation

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}=1, \tag{17.24}$$

then the corresponding cylindrical surface is called an **elliptical cylinder** and equation (17.24) is its equation in space $Oxyz$. Similarly, a **parabolic cylinder** is defined as

$$y^2=2px, \tag{17.25}$$

and a **hyperbolic cylinder** as

$$\frac{x^2}{a^2}-\frac{x^2}{b^2}=1. \tag{17.26}$$

Second order cone. *The canonical equation of the cone of the second order:*

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}-\frac{z^2}{c^2}=0. \tag{17.27}$$

In sections of this surface by horizontal planes $z=h$ we obtain ellipses

$$z=h,\quad \frac{x^2}{a^2}+\frac{y^2}{b^2}=\frac{h^2}{c^2}.$$

As $h = 0$ the section degenerates to the point $0(0,0,0)$. In sections of the given surface by coordinate planes, we obtain pairs of intersecting lines

$$\begin{cases} y = 0, \\ \dfrac{x}{a} \mp \dfrac{z}{c} = 0, \end{cases} \qquad \begin{cases} x = 0 \\ \dfrac{y}{b} \pm \dfrac{z}{c} = 0. \end{cases}$$

Fig. 17.1

Ellipsoid. *The canonical equation of an ellipsoid*:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1. \tag{17.28}$$

In sections of the ellipsoid in the planes Oxy and Oxz we obtain ellipses

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1, \quad \frac{x^2}{a^2} + \frac{z^2}{c^2} = 1.$$

Positive numbers $a,\ b, c$ are called the *semi-axes* of the ellipsoid (17.28).

An ellipsoid lies inside a rectangular parallelepiped

$$-a \le x \le a, \quad -b \le x \le b, \quad -c \le x \le c.$$

The general view of the ellipsoid is shown in Fig. 17.2

Fig.17.2

One-sheeted hyperboloid. *The canonical equation for a one-sheeted hyperboloid*:

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}-\frac{z^2}{c^2}=1$$

(17.29)

The view of this surface is shown in Fig. 17.3. In sections of the given surface by coordinate planes Oxz и Oyz hyperbolas are obtained, the equations of which respectively have the form

$$\frac{x^2}{a^2}-\frac{z^2}{c^2}=1, \quad \frac{y^2}{b^2}-\frac{z^2}{c_2}=1.$$

In sections of a given hyperboloid by planes $z=h,$ parallel to the coordinate plane $Oxy,$ ellipses are obtained whose semi-axes increase as moving away from the plane $Oxy.$ The smallest ellipse lies in the plane $Oxy;$ its equation has the form of

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}=1.$$

If $a = b,$ then the surface (17.29) is called a hyperboloid of revolution. Positive numbers a, b and c are called semi-axes of a hyperboloid of one sheet.

Fig.17.3

Hyperboloid of two sheets. *The canonical equation of a two-sheeted hyperboloid***:**

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = -1$$

(17.30)

Positive numbers a, b and c are called the semi-axes of the two-sheeted hyperboloid. In sections of this hyperboloid by coordinate planes Oxz and Oyz hyperbolas are obtained:

$$\frac{x^2}{a^2}-\frac{z^2}{c^2}=-1,\quad \frac{y_2}{b^2}-\frac{z^2}{c^2}=-1.$$

In the sections of this hyperboloid by planes $z=h$ ellipses are obtained:

$$z=h,\quad \frac{x^2}{a^2}+\frac{y^2}{b^2}=\frac{h^2}{c^2}-1.$$

These equations make sense when $|h|>c.$

Thus, the two-sheeted hyperboloid (17.30) is a surface consisting of two separate cavities, having the appearance of convex bowls, which are located symmetrically with respect to plane Oxy, the vertices of which are located at the distance c from their vertices to this plane (Fig. 17.4).

Fig. 17.4.

Elliptical paraboloid. *The canonical equation of an elliptic paraboloid*:

$$\frac{x^2}{p}+\frac{y^2}{q}=2z \quad , \tag{17.31}$$

where $p>0,\ q<0$ are the parameters of an elliptic paraboloid.

In sections of this paraboloid, the coordinate planes Oxz and Oyz give parabolas with the axis of symmetry Oz. Their equations, respectively, have the form

$$x^2=2pz, \quad y^2=2qz.$$

Sections of a given surface by planes $z=h$ lead to ellipses

$$z=h, \quad \frac{x^2}{p}+\frac{y^2}{q}=2h.$$

The point $O(0,0,0)$ is called the top of the elliptic paraboloid (17.31). As $p=q$ equation (17.31) determines the paraboloid of revolution formed by the rotation of a parabola $x^2=2pz$ around axis Oz .

Fig.17.5

Hyperbolic paraboloid. *The canonical equation of a hyperbolic paraboloid*:

$$\frac{x^2}{p} - \frac{y^2}{q} = 2z. \tag{17.32}$$

Here numbers $p > 0,\ \ q > 0$ are the parameters of a hyperbolic paraboloid.

In the section of this paraboloid, plane Oxy produces a parabola

$$x^2 = 2pz.$$

The axis of symmetry of this parabola is the positive axis Oz. Sections $y = h$ also produce parabolas, whose branches are directed upwards.

The section by plane Oyz also gives a parabola

$$y^2 = -2qz,$$

but its axis of symmetry is the negative axis Oz, i.e. the branches of this parabola are directed downward. The sections $x = h$ also produce parabolas, the branches of which are directed downwards.

Finally, in the section of this paraboloid by planes parallel to the plane Oxy hyperbolas are obtained:

$$z = h, \quad \frac{x^2}{p} - \frac{y^2}{q} = 2h.$$

Fig. 17.6

Earlier, we considered the straight-line generatrices of cylindrical surfaces and cones. Let us now consider the **rectilinear generatrices of a hyperbolic paraboloid**. We rewrite equation (17.32) in the form

$$\left(\frac{x}{\sqrt{p}} + \frac{y}{\sqrt{q}}\right)\left(\frac{x}{\sqrt{p}} - \frac{y}{\sqrt{q}}\right) = 2z$$

and consider for each pair of numbers α, β, non-zero at the same time, the equations of two planes:

$$\left.\begin{aligned} \alpha\left(\frac{x}{\sqrt{p}}+\frac{y}{\sqrt{q}}\right)&=2\beta z \\ \beta\left(\frac{x}{\sqrt{p}}-\frac{y}{\sqrt{q}}\right)&=\alpha \end{aligned}\right\}$$

(17.33)

These planes intersect in a straight line lying entirely on the paraboloid (17.32). The straight lines (17.33), each of which is defined by the relation $\beta:\alpha$, form one family of rectilinear generatrices of a paraboloid. We obtain the second family if we consider (for each pair of numbers α', β', not equal to zero at the same time) the system of equations

$$\left.\begin{aligned} \alpha'\left(\frac{x}{\sqrt{p}}-\frac{y}{\sqrt{q}}\right)&=2\beta' z \\ \beta'\left(\frac{x}{\sqrt{p}}+\frac{y}{\sqrt{q}}\right)&=\alpha' \end{aligned}\right\}$$

(17.34)

Through each point of the hyperbolic paraboloid (17.32) passes along one rectilinear generatrix of each family.

Let us return to the one-sheeted hyperboloid. Let a one-sheeted hyperboloid be defined by its canonical equation

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}-\frac{z^2}{c^2}=1$$

(17.29)

Consider the **rectilinear generatrix of a one-sheeted hyperboloid**. Let us rewrite equation (17.29) in the form

$$\frac{x^2}{a^2}-\frac{z^2}{c^2}=1-\frac{y^2}{b^2}$$

or

$$\left(\frac{x}{a}+\frac{z}{c}\right)\left(\frac{x}{a}-\frac{z}{c}\right)=\left(1+\frac{y}{b}\right)\left(1-\frac{y}{b}\right).$$

Let us now consider a pair of real numbers $\alpha,\beta,$ that are not equal to zero simultaneously, and for each such pair we write a system of equations

$$\left.\begin{aligned}\alpha\left(\frac{x}{a}+\frac{z}{c}\right)&=\beta\left(1+\frac{y}{b}\right),\\ \beta\left(\frac{x}{a}-\frac{z}{c}\right)&=\alpha\left(1-\frac{y}{b}\right).\end{aligned}\right\}$$

(*)

For each pair of numbers α,β these equations define a pair of intersecting planes and, therefore, the straight line of their intersection. This line lies entirely on the hyperboloid. In particular, as $\alpha\neq 0,\beta=0,$ $a=b=c=1$ we obtain

$$\left.\begin{aligned}x+z&=0,\\ 1-y&=0,\end{aligned}\right\}$$

and as $\alpha=0,\beta\neq 0,a=b=c=1$

$$\left.\begin{aligned} 1+y=0 \\ x-z=0 \end{aligned}\right\}$$

Similarly to equations $(*)$, for any pair of simultaneously non-zero numbers $\alpha', \beta',$ we can write the system of equations

$$\left.\begin{aligned} \alpha'\left(\frac{x}{a}+\frac{z}{c}\right)=\beta'\left(1-\frac{y}{b}\right), \\ \beta'\left(\frac{x}{a}-\frac{z}{c}\right)=\alpha'\left(1+\frac{y}{b}\right), \end{aligned}\right\}$$

$(**)$

Which defines a straight line lying on the hyperboloid (17.29).

It is easy to verify that through each point of the hyperboloid (17.29) there pass two rectilinear generatrices, one of which belongs to the family $(*)$, and the other to the family $(**)$.

As already noted, with, as $a=b$ hyperboloid (17.29), it is a hyperboloid of revolution - it is obtained by rotating the hyperbola $\frac{x^2}{a^2}-\frac{z^2}{c^2}=1,\ y=0$ (or $\frac{y^2}{a^2}-\frac{z^2}{c^2}=1,\ x=0$) around the axis $0z$. The Shukhov Tower on Shabolovka, built in 1922 by the brilliant engineer Academician V.G.Shukhov, is well known (and is clearly visible from almost any district of Moscow). This television and radio tower consists of six sections, each of which is a hyperboloid of revolution and is made of rectilinear rods. There is no other element in it (except for the rings separating one section from another).

Fig.17.7

FUNCTIONS OF SEVERAL VARIABLES

Chapter 18. Euclidean space.

The concept of the function of several variables. Limit, continuity

18.1. Euclidean space

Definition. The set of all possible ordered collections n of real numbers $(x_1, x_2, ..., x_n)$ is called **the n-dimensional coordinate space** A^n.

Moreover, each ordered collection $(x_1, x_2, ..., x_n)$ is called a point[1] of this space and is denoted by one letter (for example, *M*). The numbers $(x_1, x_2, ..., x_n)$ are called the coordinates of the point. The note $M(x_1, x_2, ..., x_n)$ means that point *M* has coordinates $(x_1, x_2, ..., x_n)$.

Definition. A coordinate space A^n is called an **n-dimensional Euclidean space** $\mathbf{R}^n$, if the distance $\rho(M, M')$ between any two points

[1] It should not be surprising that at the beginning of the book we called the ordered set (ordered collection) of numbers a vector, and here - a point. We know, in particular, that a pair can be considered both as a pair of coordinates of a point on a plane, and as a pair of coordinates of a vector on this plane.

$M(x_1, x_2, ..., x_n)$ and $M'(x_1', x_2', ..., x_n')$ of the space A^n is determined by the formula

$$\rho(M, M') = \sqrt{(x_1' - x_1)^2 + (x_2' - x_2)^2 + ... + (x_n' - x_n)^2}. \qquad (17.1)$$

The specified distance satisfies the following conditions:

1) for any M and M' the following equality holds: $\rho(M, M') = \rho(M', M)$;

2) for any M and M' the following equality holds: $\rho(M, M') \geq 0$, moreover, if $\rho(M, M') = 0$, then points M and M' coincide;

3) for any M, M' and M'' the following inequality holds:

$\rho(M, M'') \leq \leq \rho(M, M') + \rho(M', M'')$.

Note that we have already given the definition of n-dimensional Euclidean space $\mathbf{R}^n$ (see. § 1.6). It is easy to verify that the distance defined by equality (25.1) is the norm of the vector $\overline{MM'} = (x_1' - x_1, x_2' - x_2, ..., x_n' - x_n)$ and the given here definition of the space $\mathbf{R}^n$ is essentially no different from the definition given in chapter 1.

18.2. Sets in euclidean space

Let us consider the simplest sets of points, or domains, in Euclidean space.

1. The set of points $M(x_1, x_2, ..., x_n)$, whose coordinates satisfy the inequality

$$(x_1 - x_1^0)^2 + (x_2 - x_2^0)^2 + ... + (x_n - x_n^0)^2 \leq r^2, \qquad (17.2)$$

is called an **n-dimensional ball** of radius r centered at a point $M_0(x_1^0, x_2^0, ..., x_n^0)$.

Inequality (17.2) can be written in a short form as

$$\rho(M_0, M) \le r. \qquad (17.3)$$

The set of all such points of *M* for which the *inequality* $\rho(M_0, M) < r$ holds is called an **open n-dimensional ball** of radius *r* centered at a point M_0. The set of all such points M for which equality $\rho(M_0, M) = r$ holds is called an **n-dimensional sphere**.

2. The set of points $M(x_1, x_2, ..., x_n)$ whose coordinates satisfy the inequalities

$$\left|x_1 - x_1^0\right| \le d_1, \left|x_2 - x_2^0\right| \le d_2, \ldots, \left|x_n - x_n^0\right| \le d_n, \qquad (17.4)$$

is called **an n-dimensional parallelepiped** centered at the point $M_0(x_1^0, x_2^0, ..., x_n^0)$.

If in relations (17.4) we exclude the equalities:

$$\left|x_1 - x_1^0\right| < d_1, \left|x_2 - x_2^0\right| < d_2, \ldots, \left|x_n - x_n^0\right| < d_n,$$

then this determines an **open n-dimensional parallelepiped**.

We turn to the definition of neighborhoods of a point. Neighborhoods of two types are distinguished in the space $\mathbf{R}^n$: *rectangular* and *spherical*.

As the ε**-neighborhood** of a point $M_0(x_1^0, x_2^0, ..., x_n^0)$ we will call any open n-dimensional ball of radius $\sum$ centered at the point M_0. (That is the so-called spherical neighborhood.)

A **rectangular neighborhood** of a point $M_0(x_1^0, x_2^0, ..., x_n^0)$ is any open n-dimensional parallelepiped centered at a point M_0.

In what follows, when speaking of a neighborhood of a point, we will mean a neighborhood of one of the two types mentioned.

A point M_0 is called an **interior point** of the set *D* if it belongs to the set *D* together with some of its neighborhood. A set *D* is called **open** if each of its points is internal.

Point M_0 is called a **boundary point** of the set *D* if each of its neighborhoods contains both points that belong to set *D* and points that do not belong to it. A set *D* is called **closed** if it contains all its boundary points.

A point M_0 is called a **limit point** of set *D* if in any neighborhood of this point there are points of set *D* other than M_0.

Sequences of points in Euclidean space

If each natural number *n* is associated with a point of the Euclidean space $\mathbf{R}^n$, then the set of points M_1, M_2, …, M_n, … is called a **sequence of points** of the Euclidean space $\mathbf{R}^n$ and denoted $\{M_n\}$.

A sequence of points $\{M_n\}$ is called **converging** to the point M_0, if $\rho(M_n, M_0)$ is an infinitesimal quantity:

$$\lim_{n\to\infty}\rho(M_n, M_0)=0.$$

In this case, the point M_0 is called the **limit of the sequence** $\{M_n\}$.

A set D of points of a Euclidean space is called **bounded** if it is contained in some parallelepiped.

Many **properties**, which were established earlier for numerical sequences, are transferred to the limits of sequences of points in Euclidean space. The most important of them are:

1) *the uniqueness of the limit*;

2) *the boundedness of the convergent sequence*.

18.3. The concept of a function of many variables

Definition. We will say that a function $u = f(M)$ (or $u = f(x_1, x_2, ..., x_n)$) is given in the domain $D \subset \mathbf{R}^n$ if, according to a certain rule or law, one point is assigned to one definite number *u*.

The coordinates of point *M* (i.e., variables $x_1, x_2, ..., x_n$) are called **independent variables**, or **arguments**, *u* is the **dependent variable**, and the symbol f is the **correspondence law**. The set *D* is called the **domain of definition** of the function.

The domain of definition of the function of several variables (as in the case of the function of one variable) is either predefined or is a natural domain of definition, i.e. the set of all such points for which the formula of the functional dependence *f* makes sense.

In case the number of arguments is two, the function is usually denoted as

$$z = f(x, y). \qquad (17.5)$$

The *domain of definition* of such a function is a certain set of points on plane *xOy*. The *∑-neighborhood* of a point $M_0(x_0, y_0)$ is an open circle centered at that point. The *rectangular neighborhood* of point $M_0(x_0, y_0)$ is an open rectangle centered at that point.

***Example* 17.1.** Find the domain of the functions

а) $z = \ln(x + y)$; б) $z = \sqrt{x^2 + y^2 - 4} + \dfrac{1}{\sqrt{9 - x^2 - y^2}}$.

Solution. a) The domain of definition is given by inequality $x+y>0$, i.e $y>-x$. This is the set of all points on the plane above the line $y=-x$.

б) Obviously, the inequalities $x^2+y^2-4\ge 0$ and $9-x^2-y^2>0$ must be satisfied simultaneously. Therefore, the domain of definition is the set of all points of the plane whose coordinates satisfy the double inequality $4\le x^2+y^2<9$.

This area is enclosed between circle $x^2+y^2=4$ and the circle $x^2+y^2=9$. Moreover, the points of the first circle belong to this domain, and the points of the second do not.

Graph of a function of two arguments

The graph of a function of *one* argument is known to be a *line on the plane*. The graph of a function of two variables $z=f(x,y)$ is a surface in three-dimensional space, consisting of points $(x,y,f(x,y))$.

***Example* 17.2.** Let us consider the function $z=\sqrt{9-x^2-y^2}$. The domain of this function is the circle $x^2+y^2\le 9$. The graph is a hemisphere of radius 3 centered at the beginning of the coordinates (see Fig. 25.1).

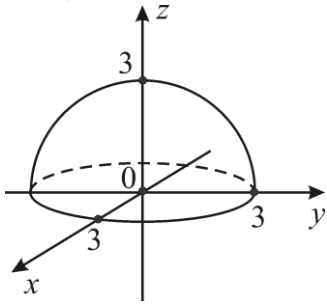


Fig. 17.1. Graph of the function $z = \sqrt{9 - x^2 - y^2}$

***Example* 17.3.** Let us consider the function $z = x^2 + y^2$. It is defined on the entire plane *Oxy*. Its graph is a surface called a paraboloid of revolution. This surface intersects the plane *xOz* in the parabola $z = x^2$, and the plane *yOz* intersects in the parabola $z = y^2$.

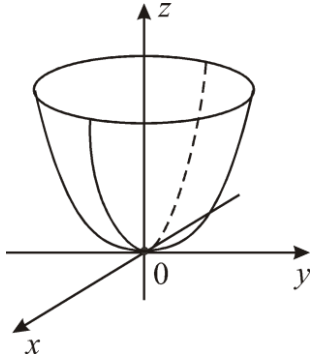


Fig. 17.2. Graph of the function $z = x^2 + y^2$

For the function of *n* arguments as $n \geq 3$ we can formally define the concept of a graph (this is the so-called hypersurface in $(n+1)$-dimensional space), but it is not possible to depict it in the figure. It should be noted that for the function of two arguments the construction of the graph is associated with significant difficulties, and the graph itself is not as clear as the graph of the function of one argument. Therefore, for a figurative representation of the function of two variables, level lines are used. By the **function level line** (17.5) we called the line

$$f(x, y) = C,$$

where *C* is a constant.

For example, for a function $z = x^2 - y^2$ any level line $x^2 - y^2 = C$ at $C \neq 0$ is a hyperbola.

Similarly, for a function of three arguments $w = f(x, y, z)$ a *level surface* is defined:

$f(x, y, z) = C$.

Examples of functions of several variables

Let us look at some examples of frequently encountered functions of several variables.

1. A linear function is a function of the form

$$u = a_1 x_1 + a_2 x_2 + ... + a_n x_n + b, \quad (17.6)$$

where a_1, a_2, …, a_n, b are constant numbers. It can be considered as the sum of n linear functions, each of which depends on one argument.

2. A function of the form

$$u = a_{11} x_1^2 + 2a_{12} x_1 x_2 + a_{22} x_2^2 + 2a_{13} x_1 x_3 + ... + a_{nn} x_n^2$$

or, which is the same,

$$u = \frac{1}{2} \sum_{i,j=1}^{n} a_{ij} x_i x_j, \quad (17.7)$$

where a_{ij} are constant numbers, is called the **quadratic form** of *n* variables x_1, x_2, …, x_n.

3. In §5.4, a utility function has been defined. Its multidimensional analogue is a function $u = f(x_1, x_2, ..., x_n)$ expressing the usefulness of acquiring *n* goods. Most often it occurs in the form of a:

logarithmic function

$$u = \sum_{i=1}^{n} a_i \ln(x_i - c_i), \quad a_i > 0, \quad x_i > c_i \geq 0 \quad (17.8)$$

function of constant elasticity

$$u = \sum_{i=1}^{n} \frac{a_i}{1-b_i}\left(x_i - c_i\right)^{1-b_i}, \quad a_i > 0,\ 0 < b_i < 1,\ x_i > c_i \geq 0. \qquad (17.9)$$

4. Cobb-Douglas Function

$$z = b_0 x_1^{b_1} x_2^{b_2}. \qquad (17.10)$$

Often, other notations are used to write it:

$$Q = AK^{\alpha}L^{\beta}. \qquad (17.11)$$

Wherein α, $\beta \geq 0$, $\alpha + \beta \geq 1$. In particular, as $\alpha + \beta = 1$ it has the form:

$$Q = AK^{\alpha}L^{1-\alpha}. \qquad (17.12)$$

This is a *production function* that expresses the volume of output *Q* (in monetary or natural terms) at the cost of capital *K* and labor *L*. Here *A* is the productivity parameter of a particular technology, 〈 is the share of capital in income ($0 < \alpha < 1$).

In this section, we will present the material mainly for the functions of two* variables. Moreover, almost all concepts and statements formulated for a function of two variables can easily be transferred to the case when the function depends on any number of variables.

18.4. Limit and continuity

Let us consider a function $z = f(x, y)$, defined on a set *D*. Let $M_0(x_0, y_0)$ be the limit point of the set *D*.

Definition. The number b is called the **limit** of a function $f(x, y)$ as the point $M(x, y)$ tends to the point $M_0(x_0, y_0)$, if for any number $\varepsilon > 0$

there exists a number $\delta > 0$ such that for all points $M(x, y)$ satisfying the condition $\rho(M, M_0) < \delta$ the following inequality holds:

$$|f(x, y) - b| < \varepsilon.$$

The fact that number b is the limit for $f(x, y)$ as $M(x, y)$ $\square$ $M_0(x_0, y_0)$, is written as follows:

$$\lim_{\substack{x \to x_0 \\ y \to y_0}} f(x, y) = b, \quad \text{or} \quad \lim_{M \to M_0} f(M) = b \qquad (17.13)$$

Definition. A function $z = f(x, y)$ is called **continuous** at the point $M_0(x_0, y_0)$, if it is defined at this point, has a finite limit as $M(x, y)$ $\square$ $M_0(x_0, y_0)$ and if the following equality holds:

$$\lim_{\substack{x \to x_0 \\ y \to y_0}} f(x, y) = f(x_0, y_0). \qquad (17.14)$$

By the **total increment** of the function we call the difference $\Delta z = f(M) - f(M_0) = f(x, y) - f(x_0, y_0) f(x, y) - f(x_0, y_0)$. If we denote $x - x_0 = \Delta x$, $y - y_0 = \Delta y$, then equality (17.14) can be rewritten as follows

$$\lim_{\substack{\Delta x \to 0 \\ \Delta y \to 0}} \Delta z = 0, \qquad (17.15)$$

i.e., infinitesimal increments of the arguments correspond to an infinitesimal increment of the function.

A function continuous at each point of a given domain *D* is called ***continuous in domain*** *D*.

The geometric meaning of the continuity of the function of two arguments is obvious: the graph of a continuous function $z = f(x, y)$ is a continuous surface that does not have gaps.

Questions

1. What is an *n*-dimensional Euclidean space?
2. How is an n-dimensional ball defined in Euclidean space?
3. What is an open *n*-dimensional ball? *n*-dimensional sphere?
4. How is the n-dimensional parallelepiped determined in Euclidean space?
5. What is the -neighborhood of a point in n-dimensional Euclidean space?
6. What points are called the interior points of a set in Euclidean space?
7. What is a closed set?
8. Can a set be non-open and not closed at the same time?
9. What is called the level line of the function of two arguments?
10. What is called the level surface of a function of three arguments?
11. What is the Cobb-Douglas function? What does this function express?
12. What is a total increment of a function of two arguments?
13. How is the concept of continuity of a function of two arguments defined?

Chapter 19. Partial derivative and their economic meaning. Total differential

19.1. Partial increment and и partial derivative

Let a function $z = f(x, y)$ be defined in some neighborhood of a point $M_0(x_0, y_0)$. (For brevity, we carry out the arguments for the function of two variables). We give the argument *x* an increment Δx (that is, we move from value x_0 to value $x_0 + \Delta x$) for a fixed $y = y_0$, so that the point $M(x_0 + \Delta x, y_0)$ belongs to the specified neighborhood. Then the function *z* changes by

$$\Delta_x z = f(x_0 + \Delta x, y_0) - f(x_0, y_0).$$

This difference is called **the partial increment** of the function *z* with respect to *x*. Particular increment is determined with respect to *y*:

$$\Delta_y z = f(x_0, y_0 + \Delta y) - f(x_0, y_0).$$

Definition. **The partial derivative** of a function of several variables with respect to one of these variables is the limit of the ratio of the partial increment of the function to the increment of the considered independent variable when the latter tends to zero (if this limit exists).

The partial derivatives of a function $z = f(x, y)$ at a point $M_0(x_0, y_0)$ are denoted as follows:

z'_x, $\dfrac{\partial z}{\partial x}$, $f'_x(x_0, y_0)$ (derivative with respect to *x*);

z'_y, $\dfrac{\partial z}{\partial y}$, $f'_y(x_0, y_0)$ (derivative with respect to *y*).

Thus, by definition

$$z'_x = \lim_{\Delta x \to 0} \frac{\Delta_x z}{\Delta x} = \lim_{\Delta x \to 0} \frac{f(x_0 + \Delta x, y_0) - f(x_0, y_0)}{\Delta x}, \quad (18.1)$$

$$z'_y = \lim_{\Delta y \to 0} \frac{\Delta_y z}{\Delta y} = \lim_{\Delta y \to 0} \frac{f(x_0, y_0 + \Delta y) - f(x_0, y_0)}{\Delta y}. \quad (18.2)$$

From the definition of partial derivatives it follows that to find the partial derivative $f'_x(x, y)$ it is necessary to consider the function $z = f(x, y)$ as a function of one argument *x* with constant *y*. Similarly, to find $f'_y(x, y)$ the constant should be considered *x*.

***Example* 18.1.** Find partial derivatives of functions:

а) $z = x^5 y^2 + x^3 y^4$; б) $z = y^x$.

Decision.

а) Counting $y = \text{const}$, we find $z'_x = 5x^4 y^2 + 3x^2 y^4$. Counting now $x = \text{const}$, we find $z'_y = 2x^5 y + 4x^3 y^3$.

б) Counting $y = \text{const}$, we find z'_x as a derivative of an exponential function: $z'_x = y^x \ln y$. Counting now $x = \text{const}$, we find z'_y as a derivative of an exponentiation function: $z'_y = xy^{x-1}$.

Second - and Higher - order partial derivatives

Partial derivatives $z'_x = f'_x(x, y)$ and $z'_y = f'_y(x, y)$ are functions of *x* and *y*, therefore, partial derivatives of them can be found. These derivatives are called **second-order partial derivatives** of the function $z = f(x, y)$:

$$z''_{xx} = (z'_x)'_x,\ z''_{xy} = (z'_x)'_y,\ z''_{yx} = (z'_y)'_x,\ z''_{yy} = (z'_y)'_y.$$

The partial derivatives of the third and higher orders are defined in a similar way. For example, $z'''_{xxy} = (z''_{xx})'_y$.

***Example* 18.2.** Find second-order partial derivatives of a function $z = x^5y^2 + x^3y^4$.

Decision. In example 18.1 partial derivatives of the first order have already been found:

$$z'_x = 5x^4y^2 + 3x^2y^4;\qquad z'_y = 2x^5y + 4x^3y^3.$$

Now we find the second-order partial derivatives:

$$z''_{xx} = 20x^3y^2 + 6xy^4;\qquad z''_{xy} = 10x^4y + 12x^2y^3;$$

$$z''_{yx} = 10x^4y + 12x^2y^3;\qquad z''_{yy} = 2x^5 + 12x^3y^2.$$

Partial derivatives z''_{xy} and z''_{yx} are called **mixed partial derivatives**.

***Example* 18.3.** $z = x^2e^y$. Find z''_{xy} and z''_{yx}.

Decision. $z'_x = 2xe^y,\ z''_{xy} = 2xe^y;\ z'_y = x^2e^y,\ z''_{yx} = 2xe^y$.

We see that for the functions considered in examples 26.2 and 26.3, the mixed derivatives coincide: $z''_{xy} = z''_{yx}$. This is no coincidence. The following statement is true:

***Theorem* 18.1.** If the partial derivatives of the second order of the function $z = f(x, y)$ are continuous at a point (x_0, y_0), then at this point $z''_{xy} = z''_{yx}$.

We will not prove this theorem. We only note that usually functions used in economics have continuous second-order partial derivatives.

The economic meaning of partial derivative

Consider the Cobb–Douglas production function as an example [see formula (17.12)]:

$$Q = AK^{\alpha}L^{1-\alpha}.$$

Let us find the rate of change in the volume of production Q when one of the factors changes: capital expenditures K or workforce L. The partial derivatives of the function Q solve this problem:

$$Q'_K = A\alpha K^{\alpha-1}L^{1-\alpha},\ Q'_L = A(1-\alpha)K^{\alpha}L^{-\alpha}.$$

The partial derivative $Q'_K = A\alpha K^{\alpha-1}L^{1-\alpha}$ is called the **marginal fixed-asset turnover**, and the partial derivative $Q'_L = A(1-\alpha)K^{\alpha}L^{-\alpha}$ – **marginal workforce productivity.**

Recall that in the case of a function of one variable $y = f(x)$ *the elasticity of the function with respect to the argument* is the quantity

$$E_x(y) = x\frac{y'}{y}$$

[see formula (12.2)].

For a function of several variables, the ordinary derivative is replaced by the partial derivative. For the Cobb-Douglas function, the elasticity of the output of the product by capital expenditure $E_K(Q) = K\dfrac{Q'_K}{Q} = \alpha$.

Similarly, the elasticity of workforce $E_L(Q) = L\dfrac{Q'_L}{Q} = 1-\alpha$.

So, in the Cobb-Douglas function, the exponents α and $1-\alpha$ are respectively the elasticity coefficients for each of its arguments.

19.3. Total increment and total differential

The total increment of the function $z = f(x, y)$ at the point $M_0(x_0, y_0)$, corresponding to the increment of the arguments Δx and Δy is called the difference

$$\Delta z = f(x_0 + \Delta x, y_0 + \Delta y) - f(x_0, y_0). \quad (18.3)$$

Definition. A function $z = f(x, y)$ called **differentiable** at a point $M_0(x_0, y_0)$, if its total increment (26.3) at this point can be represented as

$$\Delta z = A_1\Delta x + A_2\Delta y + \alpha_1\Delta x + \alpha_2\Delta y, \quad (18.4)$$

where A_1, A_2 – constants independent of Δx, Δy, and α_1, α_2 – are infinitesimal for $\Delta x \to 0$, $\Delta y \to 0$.

If at least one of the numbers A_1, A_2 is nonzero, then the amount

$$L = A_1\Delta x + A_2\Delta y \quad (18.5)$$

is the main linear Δx, Δy part of the increment of the differentiable function. This main part of the increment of the function is called **the total differential**.

***Theorem* 18.1.** If function $z = f(x, y)$ is differentiable at a point $M_0(x_0, y_0)$, it has partial derivatives at this point with respect to *x* and *y*.

Proofs. From equality (26.4) it follows that $\Delta_x z = A_1 \Delta x + \alpha_1 \Delta x$, whence

$$\frac{\Delta_x z}{\Delta x} = A_1 + \alpha_1 . \quad (*)$$

Since $\alpha_1 \to 0$ for $\Delta x \to 0$, then it follows from (*), that there is a limit

$$\lim_{\Delta x \to 0} \frac{\Delta_x z}{\Delta x} = A_1 ,$$

i.e. z'_x exists (and $z'_x = A_1$).

It is similarly proved that there exists z'_y (and $z'_y = A_2$).

The main linear part of the increment of the function $z = f(x, y)$ has the form $z'_x \Delta x + z'_y \Delta y$.

Now we can formulate the concept of a total differential in the following form.

***Definition.* The total differential** dz of a function $z = f(x, y)$ is the sum of the products of the partial derivatives of this function by the increments of the corresponding independent arguments, i.e.

$$dz = z'_x \Delta x + z'_y \Delta y . \quad (18.6)$$

Consider, in particular, a function $z = f(x, y) = 0$. Obviously, $dz = dx = 1 \cdot \Delta x$, i.e. $dx = \Delta x$. Similarly, considering $z = f(x, y) = y$, we get $dy = \Delta y$. Therefore, formula (26.6) can be written as

$dz = z'_x dx + z'_y dy$, (18.7)

or, which is the same,

$$dz = \frac{\partial z}{\partial x} dx + \frac{\partial z}{\partial y} dy.$$

Note that the differential, being the main part of the increment of the function, is used in approximate calculations.

***Example* 18.4.** Calculate approximately $\sqrt{(1{,}02)^3 + (1{,}97)^3}$.

Decision. The desired number will be considered as the value of the function $z = f(x, y) = \sqrt{x^3 + y^3}$ for $x = x_0 + \Delta x$, $y = y_0 + \Delta y$, where $x_0 = 1$, $y_0 = 2$, $\Delta x = 0{,}02$, $\Delta y = -0{,}03$. We have:

$$f(1, 2) = \sqrt{1^3 + 2^3} = 3,$$

$$\Delta z \approx dz = \frac{3x^2 dx + 3y^2 dy}{2\sqrt{x^3 + y^3}},$$

$$\Delta z(1, 2) \approx \frac{3 \cdot 1 \cdot 0{,}02 + 3 \cdot 2^2 (-0{,}03)}{2 \cdot 3} = \frac{0{,}06 - 0{,}36}{6} = -0{,}05.$$

Hence, $\sqrt{(1{,}02)^3 + (1{,}97)^3} \approx 3 - 0{,}05 = 2{,}95$.

***Theorem* 18.2.** If a function $z = f(x, y)$ is differentiable at a given point, then it is continuous at this point.

Proofs. From the differentiability condition (18.4) it follows that

$$\lim_{\substack{\Delta x \to 0 \\ \Delta y \to 0}} \Delta z = 0,$$

and this means the continuity of the function [see formula (17.15)].

Recall that for a function of one argument $y = f(x)$ the existence of a derivative is equivalent to differentiability.

However, for a function of several arguments, a similar statement, generally speaking, is not true. It does not follow from the existence of partial derivatives with respect to all arguments that the function is differentiable, and it does not even follow that it is continuous. It can be shown (we will not do this) that the function

$$z = \begin{cases} \dfrac{xy}{x^2 + y^2}, & \text{если } x^2 + y^2 \neq 0, \\ 0, & \text{если } x = 0,\ y = 0, \end{cases}$$

is not differentiable (and is not continuous) at the point $O(0, 0)$. Nevertheless, at this point (and at all other points), this function has partial derivatives with respect to x and y.

So, the existence of partial derivatives, generally speaking, is not enough for the differentiability of the function of several variables.

Sufficient conditions for differentiability are given by the following theorem.

***Theorem* 18.3.** If the function $z = f(x, y)$ has partial derivatives $f'_x(x, y)$ and $f'_y(x, y)$ in a neighborhood of the point M and these derivatives are *continuous* at the point M, then this function is differentiable at the point M.

Higher - order differentials

Let a function $z = f(x, y)$ having continuous first-order partial derivatives be given in the region D. Then, as we already know, the differential dz is called the following expression:

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

where dx, dy – the differentials of the independent arguments *x* and *y*, or, which is the same, arbitrary increments of these arguments.

Obviously, dz is also a function of *x*, *y*. If the function *z* has continuous second-order partial derivatives, then the differential dz has first-order continuous partial derivatives, and we can raise the question of the differential of this differential dz, i.e. about $d(dz)$. This latter is called **a second-order differential** (or **second differential**) and is denoted by d^2z :

$$d^2z = d(dz).$$

When calculating the second differential, the differentials of the independent arguments (i.e., their increments) dx and dy are considered as constants. Therefore, $d^2x = d^2y = 0$. So,

$$d^2z = d(dz) = d\left(\frac{\partial z}{\partial x}dx + \frac{\partial z}{\partial y}dy\right) = d\left(\frac{\partial z}{\partial x}\right)dx + d\left(\frac{\partial z}{\partial y}\right)dy.$$

From here

$$d^2z = \left(\frac{\partial^2 z}{\partial x^2}dx + \frac{\partial^2 z}{\partial x \partial y}dy\right)dx + \left(\frac{\partial^2 z}{\partial y \partial x}dx + \frac{\partial^2 z}{\partial y^2}dy\right)dy,$$

and given that $\frac{\partial^2 z}{\partial x \partial y} = \frac{\partial^2 z}{\partial y \partial x}$, we obtain

$$d^2z = \frac{\partial^2 z}{\partial x^2}dx^2 + 2\frac{\partial^2 z}{\partial x \partial y}dxdy + \frac{\partial^2 z}{\partial y^2}dy^2.$$

Of course, this expression could be written in a slightly different form, given that we can denote the same partial derivatives in different ways, for example, $\frac{\partial^2 z}{\partial x \partial y} = \frac{\partial^2 z}{\partial y \partial x}$ we can denote both how $\frac{\partial^2 f}{\partial x^2}$, and how z''_{xx}, etc.

Second-order differentials for functions of three or more arguments are written similarly, i.e. for functions $w = f(x, y, z)$, $w = f(x_1, x_2, ..., x_n)$.

The higher-order differentials are determined by the same rule:

$d^3 z = d\left(d^2 z\right)$, …, $d^n z = d\left(d^{n-1} z\right)$.

Here we looked at the functions of independent arguments. For the case when the arguments themselves are functions, for example, $x = \varphi(s,t)$, $y = \psi(s,t)$, it can be shown (just as it was done for the functions of one argument) that the form of the first-order differential is *invariant*, and for the second and higher-order differentials it is *not invariant*. We will not dwell on this in detail.

Remark. Despite the fact that for differentials of the second and higher orders, the invariance of the form does not take place at all, in some particular simple cases the shape of the differential of any order can remain unchanged. In particular, in the case when the arguments *x*, *y* of the function $z = f(x, y)$ *linearly* depend on the independent argument *t*: $x = a_1 t + a_2$, $y = b_1 t + b_2$, the shape of the differential of any order remains unchanged. This is easy to verify.

19.4. Directional derivative. Gradient

Let a function $z = f(x, y)$ be defined in some neighborhood of a point $M_0(x_0, y_0)$. Consider a certain direction defined by a unit vector

$\bar{l} = (\cos\alpha, \cos\beta)$, where $\cos^2\alpha + \cos^2\beta = 1$ (fig. 18.1). On a line passing in this direction through a point M_0, take a point $M(x_0 + \Delta x, y_0 + \Delta y)$. Denote by Δl the length of the segment M_0M. Obviously,

$$\Delta l = \sqrt{\Delta x^2 + \Delta y^2}.$$

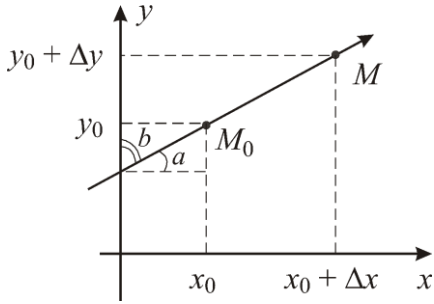


Fig. 18.1. Direction $\bar{l} = (\cos\alpha, \cos\beta)$

Consider the increment of the function $f(x, y)$:

$$\Delta z = f(x_0 + \Delta x, y_0 + \Delta y) - f(x_0, y_0),$$

where Δx and Δy are related by the relations $\Delta x = \Delta l \cos\alpha$, $\Delta y = \Delta l \cos\beta$.

Definition. The limit of the ratio $\frac{\Delta z}{\Delta l}$ at $\Delta l \to 0$ is called **the derivative** of the function $z = f(x, y)$ at a point $M_0(x_0, y_0)$ **in the direction** $\bar{l}$ and is denoted $\frac{\partial z}{\partial l}$ (or z'_l):

$$\frac{\partial z}{\partial l} = \lim_{\Delta l \to 0} \frac{\Delta z}{\Delta l}.$$

The derivative $\frac{\partial z}{\partial l}$ characterizes *the rate* of change of function in the direction $\bar{l}$.

Obviously, ordinary partial derivatives $\frac{\partial z}{\partial x}$ and $\frac{\partial z}{\partial y}$ are derivatives in directions parallel to the axes *Ox* and *Oy* respectively. It is easy to verify that for a differentiable function $z = f(x, y)$:

$$\frac{\partial z}{\partial l} = \frac{\partial z}{\partial x}\cos\alpha + \frac{\partial z}{\partial y}\cos\beta \quad (\text{or } z'_l = z'_x \cos\alpha + z'_y \cos\beta). \qquad (18.8)$$

Definition. **The gradient** of a function $z = f(x, y)$ at a point M is a vector whose coordinates are equal respectively to the partial derivatives $\frac{\partial z}{\partial x}$ and $\frac{\partial z}{\partial y}$ at this point.

The gradient of the function is denoted by $\text{grad } z$:

$$\text{grad } z = \left(\frac{\partial z}{\partial x}, \frac{\partial z}{\partial y}\right), \quad \text{or} \quad \text{grad } z = \left(z'_x, z'_y\right). \qquad (18.9)$$

Comparing equalities (18.6) and (18.7), we see that derivative in the direction $\bar{l}$ is the scalar product of vectors $\bar{l}$ and $\text{grad } z$:

$$\frac{\partial z}{\partial l} = \bar{l}\ \text{grad } z.$$

It is known that the scalar product of two vectors has a maximum value when the angle between them is zero. Therefore, the derivative of the

function $z = f(x, y)$ in the direction $\bar{l}$ takes on the maximum value when the directions of the vectors $\bar{l}$ and $\text{grad } z$ coincide.

Thus, the gradient of the function characterizes the direction in which the function changes most rapidly.

Consider **the geometric meaning of the gradient**. The level line of the function $z = f(x, y)$, passing through the point (x_0, y_0), is given by the equation $f(x, y) = C$, where $C = f(x_0, y_0)$. Under certain conditions, this equation can be solved with respect to y i.e. express y in the $y = g(x)$ (if this is not possible, then by solving the equation for x, $x = h(y)$, we can repeat all the arguments for this case). We know that the angular coefficient of the tangent is $g'(x)$, i.e. the tangent direction vector has coordinates $(1, g'(x))$ or, which is the same $\left(1, \dfrac{dg}{dx}\right)$. Therefore, the vector (dx, dy) is also the direction vector of the tangent.

Taking the differential from the left and right sides of the equation defining the level line, we get:

$$dz = \frac{\partial z}{\partial x} dx + \frac{\partial z}{\partial y} dy = 0,$$

i.e. the scalar product of the gradient and the directing vector of the tangent is zero, therefore, these vectors are *perpendicular*.

The concept of a function gradient is generalized to the case of any number of variables. In particular, for the function $u = f(x, y, z)$:

$$\text{grad } u = \left(\frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}, \frac{\partial u}{\partial z} \right).$$

In the case of the three-argument function, all of the above remains valid, the only thing is the level surface will act instead of the level line, and the tangent plane to the level surface will appear instead of the tangent to the level line, i.e. plane.

$$f'_x(x_0, y_0, z_0)(x - x_0) + f'_y(x_0, y_0, z_0)(y - y_0) + f'_z(x_0, y_0, z_0)(z - z_0) = 0$$

.

***Example* 18.5.** Find the gradient of a function $z = x^2 + \frac{y^2}{4}$ at a point $M_0(2, 6)$ and its modulus.

Decision: $\text{grad } z = \left(\frac{\partial z}{\partial x}, \frac{\partial z}{y}\right) = \left(2x, \frac{y}{2}\right)$. When $x = 2$, $y = 6$ we get:

$$\text{grad } z\big|_{(2, 6)} = (4, 3); \ |\text{grad } z| = \sqrt{4^2 + 3^2} = 5.$$

***Example* 18.6.** Find the gradient of a function $u = x^2 + \frac{y^2}{2} - z^2$ at a point $M_0(1, 1, 1)$ and its module.

Decision. $\text{grad } u = \left(\frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}, \frac{\partial u}{\partial z}\right) = (2x, y, -2z)$.

$$\text{grad } u\big|_{(1, 1, 1)} = (2, 1, -2); \ |\text{grad } u| = \sqrt{2^2 + 1^2 + (-2)^2} = 3.$$

19.5. Taylor formula

Let a function $z = f(x, y)$ in a neighborhood of a point $M_0(x_0, y_0)$ have continuous derivatives of all orders – until (n + 1)-th inclusive. We

add x_0 and y_0 some increments Δx and Δy respectively so that the straight line segment connecting the points (x_0, y_0) and $(x_0 + \Delta x, y_0 + \Delta y)$, belongs entirely to the neighborhood of the point under consideration (x_0, y_0).

We show that in this case the following equality holds:

$$\Delta f(x_0, y_0) = f(x_0 + \Delta x, y_0 + \Delta y) - f(x_0, y_0) = df(x_0, y_0) + \\ + \frac{1}{2!} d^2 f(x_0, y_0) + ... + \frac{1}{n!} d^n f(x_0, y_0) + \frac{1}{(n+1)!} d^{n+1} f(\xi, \eta),$$

(18.10)

where $\lfloor$ – between x_0 and $x_0 + \Delta x$, $\rfloor$ – between y_0 and $y_0 + \Delta y$ ($\xi = x_0 + \theta \Delta x$, $\eta = y_0 + \theta \Delta y$, $0 < \theta < 1$).

For proof, we make a replacement:

$$x = x_0 + t\Delta x, \quad y = y_0 + t\Delta y, \quad t \in [0, 1]. \qquad (18.11)$$

Substituting these values of *x* and *y* into the function $f(x, y)$, we obtain a complex function from one argument *t*:

$$F(t) = f(x_0 + t\Delta x, y_0 + t\Delta y).$$

Formulas (26.11) geometrically express a straight line segment connecting the points $M_0(x_0, y_0)$ and $M_1(x_1, y_1)$ (in this case, the point $M_0(x_0, y_0)$ corresponds to the value $t = 0$, and the point $M_1(x_1, y_1)$ – to the value $t = 1$).

Now we can replace the increment $\Delta f(x_0, y_0) = f(x_0 + \Delta x, y_0 + \Delta y) - f(x_0, y_0)$ with an equal increment $\Delta F(0) = F(1) - F(0)$. But the function $F(t)$ is a function of one variable

and has continuous derivatives until (n + 1)-th order, inclusive. Therefore, it can be decomposed according to the Taylor formula. We write this expansion in the form (17.13″):

$$\Delta F(t_0) = dF(t_0) + \frac{1}{2!}d^2F(t_0) + ... + \frac{1}{n!}d^nF(t_0) + \frac{1}{(n+1)!}d^{n+1}F(t_0 + \theta\Delta t),\ 0$$

.

From here we get

$$\Delta F(0) = F(1) - F(0) = dF(0) + \frac{1}{2!}d^2F(0) + ... + \frac{1}{n!}d^nF(0) +$$

$$+ \frac{1}{(n+1)!}d^{n+1}F(\theta), \qquad 0 < \theta < 1.$$

(18.12)

In this case, we note that differential dt, which appears in various degrees on the right-hand side of (26.12) (i.e. contained in the expressions $dF(0) = F'(0)dt$, $d^2F(0) = F''(0)dt^2$, …), is equal to the increment $\Delta t = 1 - 0 = 1$.

Now, taking into account the (linear) replacement (26.11), and also considering the remark made earlier on the invariance of a differential of any order with respect to a linear change of variables (see § 26.2), we obtain

$$dF(0) = f'_x(x_0, y_0)dx + f'_y(x_0, y_0)dy = df(x_0, y_0),$$

$$d^2F(0) = f''_{xx}(x_0, y_0)dx^2 + 2f''_{xy}(x_0, y_0)dxdy + f''_{yy}(x_0, y_0)dy^2 = d^2f(x_0, y_0)$$

,

…

$$d^nF(0) = d^nf(x_0, y_0).$$

Finally, for the differential (n + 1)-th order, we obtain

$$d^{n+1}F(\theta) = d^{n+1}f(x_0 + \theta\Delta x, y_0 + \theta\Delta y).$$

Note that the differentials dx and dy (as for independent arguments) are equal to increments and, respectively. Indeed, taking into account (18.11) and the fact that $dt = 1$, we have:

$$dx = x_t' dt = (x_0 + t\Delta x)_t' dt = \Delta x dt = \Delta x,$$

$$dy = y_t' dt = (y_0 + t\Delta y)_t' dt = \Delta y dt = \Delta y.$$

Now we substitute the expressions for $dF(0)$, $d^2F(0)$, … in (18.12):

$$\Delta f(x_0, y_0) = f(x_0 + \Delta x, y_0 + \Delta y) - f(x_0, y_0) = df(x_0, y_0) +$$

$$+ \frac{1}{2!}d^2 f(x_0, y_0) + \ldots + \frac{1}{n!}d^n f(x_0, y_0) + \frac{1}{(n+1)!}d^{n+1} f(x_0 + \theta\Delta x, y_0 + \theta\Delta y),$$

$$0 < \theta < 1.$$

(18.13)

We have obtained for the function $z = f(x, y)$ **the Taylor formula in a differential form.** In expanded form, it (even for the considered case of the function of two arguments) looks much more complicated.

We write formula (18.13), taking into account the fact that $dx = \Delta x$, $dy = \Delta y$, in expanded form, restricting ourselves to only two terms of the expansion (i.e., for *n* = 2):

$$f(x,y) = f(x_0,y_0) + f'_x(x_0,y_0)(x-x_0) + f'_y(x_0,y_0)(y-y_0) +$$
$$+ \frac{1}{2!}\left[f''_{xx}(x_0,y_0)(x-x_0)^2 + 2f''_{xy}(x_0,y_0)(x-x_0)(y-y_0) +\right.$$
$$\left. + f''_{yy}(x_0,y_0)(y-y_0)^2\right] + \frac{1}{3!}\left[f'''_{xxx}(x_0+\theta\Delta x, y_0+\theta\Delta y)(x-x_0)^3 +\right.$$
$$+ f'''_{xxy}(x_0+\theta\Delta x, y_0+\theta\Delta y)(x-x_0)^2(y-y_0) +$$
$$+ f'''_{xyy}(x_0+\theta\Delta x, y_0+\theta\Delta y)(x-x_0)(y-y_0)^2 +$$
$$\left. + f'''_{yyy}(x_0+\theta\Delta x, y_0+\theta\Delta y)(y-y_0)^3\right].$$

(18.14)

This is Taylor's formula for $z = f(x,y)$ at $n = 2$. As you can see, it looks unwrapped in a cumbersome form, although the function depends on only two variables, and we took only two terms of the expansion.

Questions

1. What is called a partial function increment? What is the difference between a partial increment and a total increment?
2. What is called the partial derivative of a function of several arguments with respect to one of the arguments?
3. Does the process of finding the partial derivative differ fundamentally from the process of differentiating the function of one argument?
4. What are mixed partial derivatives?
5. What property do continuous mixed partial derivatives possess?
6. What is the economic meaning of the partial derivatives of the Cobb-Douglas function?

7. What is the elasticity of workforce for the Cobb-Douglas function? And what is the elasticity of output for capital expenditure for the same function?

8. What is the differentiable function of two arguments?

9. What is called the total differential of a function of two arguments?

10. Is the existence of partial derivatives with respect to both arguments sufficient for the function of two arguments to be differentiable?

11. Is it possible to say that the function of two arguments, which has partial derivatives of both arguments at a given point, is continuous at this point?

12. What is the basis for the use of the total differential in approximate calculations?

13. How is the derivative determined in this direction? What characterizes the directional derivative? Is a scalar or vector quantity a directional derivative?

14. What is the gradient of a function of two arguments? Is the gradient a scalar or vector?

15. In which case does the directional derivative take on the greatest value?

16. What is the geometric meaning of the gradient?

17. What does the Taylor formula for the function of two arguments in differential form look like?

Chapter 20. Extremum. Conditional extremes

20.1. The local extremum of a function of multiple variables

As already noted, we carry out the arguments for the function of two arguments.

Let a function $z = f(x, y)$ be defined in some neighborhood of a point $M_0(x_0, y_0)$.

Definition. A point $M_0(x_0, y_0)$ is called **a point of local maximum (minimum)** of the function $z = f(x, y)$, if there is a neighborhood of the point M_0, such that for all points $M(x, y)$ from this neighborhood inequality holds $f(x_0, y_0) > f(x, y)$ (respectively $f(x_0, y_0) < f(x, y)$).

If $M_0(x_0, y_0)$ – is the point of the local maximum (minimum) of the function $f(x, y)$, then the value $f(x_0, y_0)$ is called **the local maximum (minimum)** of the function. The general term for a local maximum and minimum is **a local extremum**.

Necessary condition for extremum

***Theorem* 20.1.** If the function $z = f(x, y)$ has partial derivatives at the point of local extremum $M_0(x_0, y_0)$, then

$$f'_x(x_0, y_0) = f'_y(x_0, y_0) = 0. \qquad (20.1)$$

Proofs. We fix $y = y_0$. We get the function of one variable. $f(x_0, y_0)$. Its derivative coincides with the partial derivative $f'_x(x, y_0)$, and the function has a local extremum at a point x_0. According to Fermat's theorem $f'_x(x_0, y_0) = 0$. Similarly, fixing $x = x_0$ and considering $f(x_0, y)$, we prove that $f'_y(x_0, y_0) = 0$.

It should be noted that condition (20.1) is not a sufficient condition for the extremum. Consider, for example, a function $z = xy$. Its partial derivatives are equal to zero at a point $O(0, 0)$, however, at this point the function has no extremum. Indeed, $f(0, 0) = 0$, but in any neighborhood of the point *O* there are both positive and negative values of the function.

Points at which the necessary conditions for an extremum are satisfied (i.e., partial derivatives z'_x and z'_y are equal to zero), are called **critical** or **stationary points**. The stationary points of the function $f(x, y)$ can be found by solving the system of equations:

$$\begin{cases} f'_x(x, y) = 0 \\ f'_y(x, y) = 0. \end{cases} \qquad (20.2)$$

***Example* 20.1.** Find stationary function points $z = x^3 + 8y^3 - 6xy + 1$.

Decision. $z'_x = 3x^2 - 6y$, $z'_y = 24y^2 - 6x$. We get the system:

$$\begin{cases} 3x^2 - 6y = 0, \\ 24y^2 - 6x = 0, \end{cases} \quad \text{or} \quad \begin{cases} x^2 - 2y = 0, \\ 4y^2 - x = 0. \end{cases}$$

From the first equation we find $2y = x^2$, $4y^2 = x^4$. Substituting into the second equation, we obtain $x^4 - x = 0$, i.e. $x(x^3 - 1) = 0$. This equation has two real roots $x_1 = 0$, $x_2 = 1$. From the first equation we find $y_1 = 0$, $y_2 = \frac{1}{2}$. Therefore, there are two stationary points $(0, 0)$ and $\left(1, \frac{1}{2}\right)$.

Sufficient extremum conditions

***Theorem* 20.2.** Let a function $z = f(x, y)$ has second-order continuous partial derivatives in some neighborhood of a stationary point $M_0(x_0, y_0)$, let $f''_{xx}(x_0, y_0) = A$, $f''_{xy}(x_0, y_0) = f''_{yx}(x_0, y_0) = B$, $f''_{yy}(x_0, y_0) = C$, $D = AC - B^2$. Then: 1) if $D > 0$, then the function has a local extremum at the point (x_0, y_0), and if $A < 0$ – a local maximum, and if $A > 0$ – a local minimum; 2) if $D < 0$, then at the point (x_0, y_0) there is no extremum.

Proofs. Consider the difference $\Delta f = f(x, y) - f(x_0, y_0)$. We use the Taylor formula (19.13), restricting ourselves to $n = 1$ i.e., the expansion will contain only the first term and the remainder term R_1):

$$\Delta f = \Delta f(x_0, y_0) = f'_x(x_0, y_0)\Delta x + f'_y(x_0, y_0)\Delta y + \frac{1}{2!}\left[f''_{xx}(\xi, \eta)\Delta x^2 + \right.$$
$$\left. + 2 f''_{xy}(\xi, \eta)\Delta x \Delta y + f''_{yy}(\xi, \eta)\Delta y^2\right].$$

(20.3)

(Here $\xi = x_0 + \theta\Delta x$, $\eta = y_0 + \theta\Delta y$, $0 < \theta < 1$.)

Since the point $M_0(x_0, y_0)$ – is stationary, the first terms of the expansion vanish, and we get a simpler expression for function increment at a point (x_0, y_0):

$$\Delta f = \frac{1}{2!}\left[f''_{xx}(\xi,\eta)\Delta x^2 + 2f''_{xy}(\xi,\eta)\Delta x\Delta y + f''_{yy}(\xi,\eta)\Delta y^2\right]. \quad (20.4)$$

In accordance with our designations

$f''_{xx}(x_0, y_0) = A$, $f''_{xy}(x_0, y_0) = B$, $f''_{yy}(x_0, y_0) = C$.

Since the second derivatives are continuous, then

$f''_{xx}(\xi,\eta) = f''_{xx}(x_0 + \theta\Delta x, y_0 + \theta\Delta y) = A + \alpha_{11}$,

$f''_{xy}(\xi,\eta) = B + \alpha_{12}$, $f''_{yy}(\xi,\eta) = C + \alpha_{22}$,

where α_{11}, α_{12}, α_{22} – are infinitesimal for $\Delta x \to 0$, $\Delta y \to 0$.

Now we can rewrite Δf in the form:

$$\Delta f = \frac{1}{2}\left[A\Delta x^2 + 2B\Delta x\Delta y + C\Delta y^2 + \alpha_{11}\Delta x^2 + 2\alpha_{12}\Delta x\Delta y + \alpha_{22}\Delta y^2\right]$$

.

We are interested in the sign of difference Δf. We will see that the sign Δf depends on the sign of the expression $D = AC - B^2$. Denote the distance between the points $M_0(x_0, y_0)$ and $M(x, y)$ by *r*. Obviously, $r = \sqrt{\Delta x^2 + \Delta y^2}$. Now $\Delta x = r\cos\varphi$, $\Delta y = r\sin\varphi$ (where ϕ – is the angle between the segment M_0M and *Ox*). Once again, we rewrite the expression Δf:

$$\Delta f = \frac{r^2}{2}\left[A\cos^2\varphi + 2B\cos\varphi\sin\varphi + C\sin^2\varphi + \alpha_{11}\cos^2\varphi + \right.$$
$$\left. + 2\alpha_{12}\cos\varphi\sin\varphi + \alpha_{22}\sin^2\varphi\right]. \quad (20.5)$$

1. Let $\boldsymbol{AC - B^2 > 0}$.

In this case, $AC > 0$, therefore, $A \neq 0$, and *the first* trinomial in parentheses expression (20.5) can be transformed as follows:

$$A\cos^2\varphi + 2B\cos\varphi\sin\varphi + C\sin^2\varphi =$$
$$= \frac{1}{2}\left[\left(A\cos\varphi + B\sin\varphi\right)^2 + \left(AC - B^2\right)\sin^2\varphi\right]. \quad (20.6)$$

From this it is clear that the expression in square brackets (under our assumption $AC - B^2 > 0$) is always positive. Therefore, the mentioned trinomial for all values of ϕ is nonzero and has the same sign as the coefficient A. This trinomial is a function of the argument ϕ, that is continuous on the interval $[0, 2\pi]$. This function, according to the second Weierstrass theorem, reaches at $[0, 2\pi]$ its smallest value. This smallest value is nonzero. Therefore, the modulus of this square trinomial has a positive smallest value m:

$$\left| A\cos^2\varphi + 2B\cos\varphi\sin\varphi + C\sin^2\varphi \right| \geq m > 0.$$

Now we consider the second trinomial in parentheses on the right-hand side of equality (20.5). Obviously,

$$\left| \alpha_{11}\cos^2\varphi + 2\alpha_{12}\cos\varphi\sin\varphi + \alpha_{22}\sin^2\varphi \right| \leq \left|\alpha_{11}\right| + 2\left|\alpha_{12}\right| + \left|\alpha_{22}\right|$$

.

Since α_{11}, α_{12}, α_{22} – are infinitesimal for $\Delta x \to 0$, $\Delta y \to 0$, then for sufficiently small Δx and Δy the inequality will be fulfilled

$|\alpha_{11}|+2|\alpha_{12}|+|\alpha_{22}|<m$.

Therefore, the expression in brackets on the right-hand side of equality (20.5) will retain the same sign as the first of the trinomials, i.e. the sign of *A*. Consequently, the left side $\Delta f = f(x,y)- f(x_0,y_0)$ also retains the sign of *A*.

So, if $A > 0$, then and $\Delta f > 0$, i.e. at a point (x_0, y_0) the function has *a minimum*; in the case $A < 0$ will be $\Delta f < 0$, i.e. there is *a maximum*.

2. Let now $\boldsymbol{AC - B^2 < 0}$.

We consider separately the cases when $A \neq 0$ and when $A = 0$.

1) $\boldsymbol{A \neq 0}$.

In this case, we can use the transformation (20.6). Let us make sure that in this case, in an arbitrarily small proximity to the point under consideration $M_0(x_0, y_0)$ the difference Δf can be both positive, and negative, i.e at point $M_0(x_0, y_0)$ there is no extremum.

Let $\varphi = \varphi_1 = 0$. Then, on the right-hand side of equality (20.6), the expression in square brackets will be positive (and equal to A^2).

If $\varphi = \varphi_2$ we determine from the condition $A\cos\varphi + B\sin\varphi$ (i.e. $\varphi_2 = -\operatorname{arctg}\dfrac{A}{B}$), then the expression mentioned will be negative (and equal to $(AC - B^2)\sin^2\varphi_2$).

As already noted, the second trinomial on the right-hand side of equality (20.5) for sufficiently small *r* does not affect the sign Δf.

Obviously, we can take a point $M_1(x_1, y_1)$ as close to $M_0(x_0, y_0)$ as needed so that the segment M_0M_1 forms an angle $\varphi = \varphi_1 = 0$ with *Ox*. For it $\Delta f > 0$. In the same way, we can arbitrarily close to $M_0(x_0, y_0)$ take a point $M_2(x_2, y_2)$ so, that the segment M_0M_2 forms an angle $\varphi = \varphi_2$ with *Ox*. For this point will be $\Delta f < 0$.

So, in the case under consideration $AC - B^2 < 0$, $A \neq 0$ in any proximity to the point under consideration (x_0, y_0) the difference Δf can be both positive and negative. Therefore, at this point there is no extremum.

2) $\boldsymbol{A = 0}$.

In this case

$A\cos^2\varphi + 2B\cos\varphi\sin\varphi + C\sin^2\varphi = 2B\cos\varphi\sin\varphi + C\sin^2\varphi =$

$\sin\varphi(2B\cos\varphi + C\sin\varphi)$.

Obviously, $B \neq 0$ (otherwise $AC - B^2 = 0$).

In this case, we can choose such an angle $\tilde{\varphi}$, that

$|C\sin\tilde{\varphi}| < |2B\cos\tilde{\varphi}|$.

Then with $\varphi = \tilde{\varphi}$ and $\varphi = -\tilde{\varphi}$ the trinomial (20.6) will have opposite signs. Therefore (repeating the above reasoning) we are convinced that there is no extremum at the point $M_0(x_0, y_0)$.

The theorem is proved.

In the case, when $AC - B^2 = 0$, the question of the extremum remains open and to solve it requires additional research (for example, involving higher derivatives).

We also note that

$$D = \begin{vmatrix} f''_{xx}(x_0, y_0) & f''_{xy}(x_0, y_0) \\ f''_{yx}(x_0, y_0) & f''_{yy}(x_0, y_0) \end{vmatrix}.$$

***Example* 20.2.** Explore extremum function

$z = 3x^2 - x^3 + 3y^2 + 4y$

Decision. $z'_x = 6x - 3x^2 = 3(2x - x^2)$, $z'_y = 6y + 4 = 2(3y + 2)$. We get the system:

$$\begin{cases} 2x - x^2 = 0, \\ 3y + 2 = 0. \end{cases}$$

Solving the system, we find two stationary points:

$M_1\left(0, -\frac{2}{3}\right)$ и $M_2\left(2, -\frac{2}{3}\right)$.

Find the second-order partial derivatives:

$z''_{xx} = 6 - 6x$, $z''_{xy} = 0$, $z''_{yy} = 6$.

We calculate *A*, *B*, *C* and *D* for each stationary point.

for a point $M_1\left(0, -\frac{2}{3}\right)$

$A_1 = 6$, $B_1 = 0$, $C_1 = 6$, $D_1 = 6 \cdot 6 - 0 = 36 > 0$ – there is an extremum;

$A_1 = 6 > 0$, therefore, a minimum;

$z_{\min} = f\left(0, -\frac{2}{3}\right) = -\frac{4}{3}$;

for a point $M_2\left(2, -\frac{2}{3}\right)$

$A_2 = -6$, $B_2 = 0$, $C_2 = 6$, $D_2 = -36 < 0$ – there is no extremum.

***Example* 20.3.** Explore extremum function:

$z = x^3 + 8y^3 - 6xy + 1$.

Decision. The first partial derivatives and stationary points $(0, 0)$ and $\left(1, \frac{1}{2}\right)$ of this function were found in Example 20.1. Since $z''_{xx} = 6x$, $z''_{xy} = 6$, $z''_{yy} = 48y$, then at the point $(0, 0)$ there will be $A = 0$, $B = -6$, $C = 0$, $D = -36 < 0$, therefore, there is no extremum. At the point $\left(1, \frac{1}{2}\right)$ we have $A = 6$, $B = -6$, $C = 24$, $D = 108 > 0$ – there is an extremum; $A > 0$, therefore minimum;

$z_{\min} = f\left(1, \frac{1}{2}\right) = 0$.

20.2. Largest and lowest values of functions in a closed area

The largest and smallest values (i.e. global maximum and minimum) of a function continuous on some closed set can be reached either *at extremum points* or at *the boundary of the set*.

***Example* 20.4.** Find the largest and smallest values of the function $z = x^2 + y^2$ in a circle of radius 2 centered at a point $(0, 1)$.

Decision. Obviously, the boundary of the area has an equation $x^2 + (y-1)^2 = 4$.

Find the partial derivatives: $z'_x = 2x$, $z'_y = 2y$. Equating them to zero, we find the only stationary point $O(0, 0)$.

We study the function at the boundary of the area. Substituting from the boundary equation $x^2 = 4 - (y-1)^2$ into the function $z = x^2 + y^2$, we obtain the function of one variable $z = 4 - (y-1)^2 + y^2$, i.e. $z = 2y + 3$. Obviously, $y \in [-1, 3]$. This function does not have stationary points, therefore, its largest and smallest values can be reached only at the ends of a segment $[-1, 3]$.

The value of the function $z = x^2 + y^2$ at the stationary point $(0, 0)$ is 0. The value of the function $z = 2y + 3$ for $y = -1$ is 1, and for $y = 3$ is 9. Comparing these three values, we find $z_{max} = f(0, 3) = 9$, $z_{\min} = f(0, 0) = 0$.

20.3. Conditional extremes

Consider the problem of finding the extrema of a function of *several* arguments in the presence of *additional conditions* relating the values of the arguments. Such extremes are called **conditional**.

For example, let it be necessary to find the extrema of the function

$z = x^2 y$, (*)

if its arguments satisfy the condition:

$2x + y - 1 = 0$. (**)

In this case, the extrema are not sought on the entire *Oxy* plane but only on the line $2x + y - 1 = 0$. We substitute in (*) the expression $y = -2x + 1$ from the condition (**), and the problem of the conditional extremum of the function (*) reduces to the problem of finding the non-conditional extremum of the function $z = x^2(-2x + 1) = x^2 - 2x^3$. So,

$z' = 2x - 6x^2 = = 2x(1-3x)$. This function has a minimum at $x = 0$ and a maximum at $x = \frac{1}{3}$, i.e. the function $z = x^2 y$ in the presence of connection (**) has a conditional minimum $z = 0$ at a point $(0, 1)$ and a conditional maximum $z = \frac{1}{27}$ at a point $\left(\frac{1}{3}, \frac{1}{3}\right)$.

Definition. A function $u = f(M) = f(x_1, x_2, ..., x_n)$ has **a conditional maximum** (**conditional minimum**) at a point $M_0(x_1^0, x_2^0, ..., x_n^0)$ if there is a neighborhood of the point M_0, such that for all points $M(x_1, x_2, ..., x_n)$ of this neighborhood satisfying *m* equations $(m < n)$:

$$\begin{cases} g_1(x_1, ..., x_n) = 0, \\ \ldots\ldots\ldots\ldots\ldots \\ g_m(x_1, ..., x_n) = 0, \end{cases} \quad (20.7)$$

inequality holds $f(M_0) \geq f(M)$ (respectively $f(M_0) < f(M)$).

Equations (20.7) are called **coupling equations**.

The problem of finding a conditional extremum is reduced to a study of a function's ordinary extremum

$$L(x_1, ..., x_n, \lambda_1, ..., \lambda_m) = f(x_1, ..., x_n) + \lambda_1 g_1(x_1, ..., x_n) + ... + \lambda_m g_m(x_1, ..., x_n)$$

.

The function *L* is called **the Lagrange function**, and the numbers λ_1, …, λ_m – are called **the Lagrange multipliers.**

The necessary *conditions for a conditional extremum* are expressed by a system of $m+n$ equations:

$$\begin{cases} \dfrac{\partial L(M)}{\partial x_i} = 0, \quad i = 1, 2, ..., n, \\ g_k(M) = 0, \quad k = 1, 2, ..., m, \end{cases} \qquad (20.8)$$

from which unknowns can be found $x_1,...,x_n;\ \ \lambda_1,...,\lambda_m$, where $x_1,...,x_n$ – are the coordinates of the point at which a conditional extremum is possible.

Sufficient *conditions for the conditional extremum* are associated with the study of the second differential of the Lagrange function d^2L, namely, if the inequality holds $d^2L < 0$ at the point of a possible extremum M_0, then at this point there *is a conditional maximum*, if $d^2L > 0$, then, *a conditional minimum*.

In the case of the function of *two* variables $z = f(x, y)$ in the coupling equation $g(x, y) = 0$, the Lagrange function has the form:

$$L(x, y, \lambda) = f(x, y) + \lambda g(x, y).$$

System (20.8) consists of three equations:

$$\frac{\partial L}{\partial x} = 0, \quad \frac{\partial L}{\partial y} = 0, \quad g(x, y) = 0.$$

Let $(x_0, y_0; \lambda_0)$ – be any of the solutions of this system, $M_0(x_0, y_0)$ – be the point of a possible extremum and

$$\Delta = -\begin{vmatrix} 0 & g'_x(M_0) & g'_y(M_0) \\ g'_x(M_0) & L''_{xx}(M_0,\lambda_0) & L''_{xy}(M_0,\lambda_0) \\ g'_y(M_0) & L''_{yx}(M_0,\lambda_0) & L''_{yy}(M_0,\lambda_0) \end{vmatrix}.$$

If $\Delta < 0$, then the function $z = f(x, y)$ has *a conditional maximum* at a point $M_0(x_0, y_0)$, if $\Delta > 0$, then *a conditional minimum*.

***Example* 20.7.** Find the conditional extremum of the function $z = 2x + y$ for $x^2 + y^2 = 1$.

Decision. We compose the Lagrange function:

$$L(x, y, \lambda) = 2x + y + \lambda(x^2 + y^2 - 1).$$

We have $\dfrac{\partial L}{\partial x} = 2 + 2\lambda x$, $\dfrac{\partial L}{\partial y} = 1 + 2\lambda y$.

The system of equations (20.8) has the form:

$$\begin{cases} 2 + 2\lambda x = 0, \\ 1 + 2\lambda y = 0, \\ x^2 + y^2 - 1 = 0. \end{cases}$$

We find solutions to this system: $x_1 = -\dfrac{2}{\sqrt{5}}$, $y_1 = -\dfrac{1}{\sqrt{5}}$, $\lambda_1 = \dfrac{\sqrt{5}}{2}$;

$x_2 = \dfrac{2}{\sqrt{5}}$, $y_2 = \dfrac{1}{\sqrt{5}}$, $\lambda_2 = -\dfrac{\sqrt{5}}{2}$.

We have: $g(x, y) = x^2 + y^2 - 1$, $g'_x = 2x$, $g'_y = 2y$,

$g'_x\left(-\dfrac{2}{\sqrt{5}}, -\dfrac{1}{\sqrt{5}}\right) = -\dfrac{4}{\sqrt{5}}$, $g'_y\left(-\dfrac{2}{\sqrt{5}}, -\dfrac{1}{\sqrt{5}}\right) = -\dfrac{2}{\sqrt{5}}$,

$L''_{xx} = \sqrt{5}$, $L''_{xy} = 0$, $L''_{yy} = \sqrt{5}$ при $\lambda = \dfrac{\sqrt{5}}{2}$.

Consequently,

$$\Delta = -\begin{vmatrix} 0 & -\dfrac{4}{\sqrt{5}} & -\dfrac{2}{\sqrt{5}} \\ -\dfrac{4}{\sqrt{5}} & \sqrt{5} & 0 \\ -\dfrac{2}{\sqrt{5}} & 0 & \sqrt{5} \end{vmatrix} = 4\sqrt{5} > 0,$$

i.e. the function has a conditional minimum at the point $M_1\left(-\dfrac{2}{\sqrt{5}}, -\dfrac{1}{\sqrt{5}}\right)$, $z_{\min} = -\sqrt{5}$.

Similarly for the point $M_2\left(\dfrac{2}{\sqrt{5}}, \dfrac{1}{\sqrt{5}}\right)$:

$$\Delta = -\begin{vmatrix} 0 & \dfrac{4}{\sqrt{5}} & \dfrac{2}{\sqrt{5}} \\ \dfrac{4}{\sqrt{5}} & -\sqrt{5} & 0 \\ \dfrac{2}{\sqrt{5}} & 0 & -\sqrt{5} \end{vmatrix} = -4\sqrt{5} < 0,$$

i.e. at the point $M_2\left(\dfrac{2}{\sqrt{5}}, \dfrac{1}{\sqrt{5}}\right)$ there is a conditional maximum, $z_{\max} = \sqrt{5}$.

20.4. Least squares

The least-squares method is one of the methods of the theory of errors. It refers to the so-called approximation methods, i.e. methods for the approximate expression of any mathematical objects through other, simpler ones.

In practice, we often encounter the need to "smooth out" the dependencies identified as a result of observations. Usually, the problem is formulated as follows: there are observational data at *n* points M_1, M_2, …, M_n of some quantity *u* and the corresponding values of this quantity u_1, u_2, …, u_n; it is necessary to select a function $u = f(M)$, so that it most accurately expresses the total dependence of the measured quantity on the parameters of the measurement points M_1, M_2, …, M_n.

Formulas analytically representing experimental data (or measurement results) are called **empirical formulas**.

For simplicity, we consider the case when the points M_i, at which measurements are taken have the same coordinate x_i, i.e. the relationship between the variables *x* and *y* is represented as a set x_1, x_2, …, x_n and the corresponding values y_1, y_2, …, y_n. These pairs of values are represented on the coordinate plane by points (x_1, y_1), (x_2, y_2), …, (x_n, y_n). The polyline that connects these points is called **the experimental curve** (Fig. 20.1).

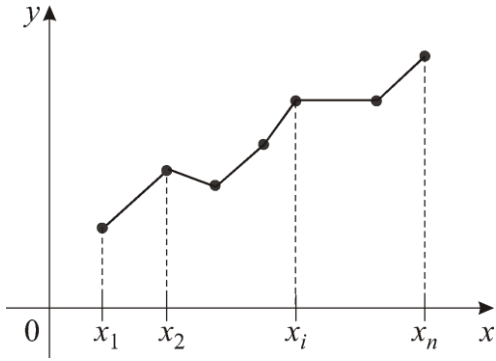


Fig. 20.1. The experimental curve

It is necessary to find an analytical representation of the relationship between x and y in the form of a formula $y = f(x)$. The type of function $y = f(x)$ is determined by economic or other considerations. Typically, the following are used as such functions:

$y = ax + b$ – *linear;*

$y = ax^2 + bx + c$ – *parabolic*;

$y = \frac{a}{x} + b$ – *hyperbolic*;

$y = ae^{bx}$ – *exponential*.

(*Logarithmic*, *power*, and other functions are also used.)

The problem of finding empirical formulas is usually solved in two stages.

At *the first* stage, the general form of the dependence is determined $y = f(x)$, i.e. it must be decided whether it is linear, quadratic, exponential, or some other.

Suppose that the measurement results (experimental data) are plotted on a grid (Fig. 20.2).

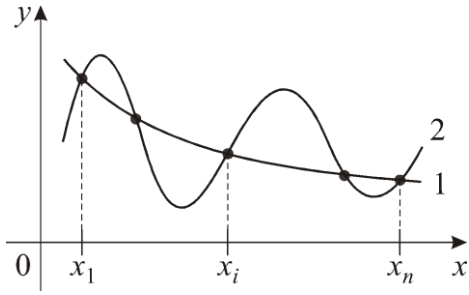


Fig. 20.2. Determination of the general form of dependence $y = f(x)$

Obviously, there are many different curves passing through these points. In the case shown in fig. 20.2, curve *1* is preferable for the researcher in curve *2*. It should be emphasized that the first stage - the stage of selecting the type of empirical function is very important. We see that curve *2* in Fig. 20.2, although it passes through the corresponding points, it does not provide a satisfactory representation of the dependence between *x* and *y*.

In practice, to verify the correctness of the choice of function $y = f(x)$ additional studies are conducted, i.e. a number of additional measurements of *x* and *y* are made, additional points are applied to the coordinate plane. If they find themselves at a fairly close distance from the selected curve, then they consider that the type of curve is established, i.e. set the type of function $y = f(x)$. After choosing the type of function, they go to the second stage.

At *the second* stage, *the parameters* of the selected empirical function $y = f(x)$ are determined. In the above functions, the parameters are unknown numbers *a*, *b* и *c*. The parameters should be chosen so that the values of the empirical function are less likely to deviate at points x_1, x_2, …, x_n from the measured values.

The least-squares method (proposed by K. Gauss) is to minimize the sum of the squares of the deviations of the "theoretical" values $f(x_i)$, found by the empirical formula $y = f(x)$ from the corresponding experimental values y_i. In other words, the quantity

$$S = \sum_{i=1}^{n} \delta_i^2 = \sum_{i=1}^{n} \left(f(x_i) - y_i \right)^2$$

should be *minimal* (see Fig. 20.3).

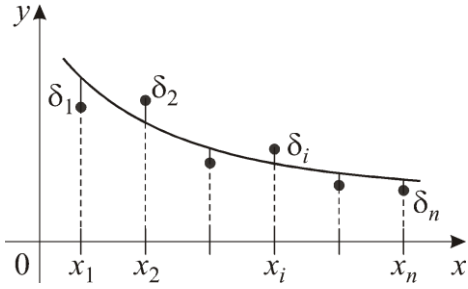


Fig. 20.3. Least Squares Illustration

We illustrate the general least-squares method with an example of a linear function. So, let a function $y = ax + b$ be taken as a function $y = f(x)$ and it is necessary to find such values of unknown parameters *a* and *b*, for which the function

$$S = \sum_{i=1}^{n} \left(ax_i + b - y_i \right)^2$$

takes the smallest value. Here x_i and y_i – are the constants found experimentally, and the function *S* is a function of the parameters *a* and *b*: $S = S(a,b)$.

So, find the critical points of the function $S(a,b)$, and then examine them. To find critical points, it is necessary to solve the system

$$\begin{cases} S'_a(a,b)=0, \\ S'_b(a,b)=0, \end{cases}$$

or, which is the same

$$\begin{cases} \sum_{i=1}^{n} 2(ax_i+b-y_i)x_i = 0, \\ \sum_{i=1}^{n} 2(ax_i+b-y_i) \quad = 0. \end{cases}$$

After the obvious elementary transformations, we get an equivalent system called **the system of normal equations**,

$$\begin{cases} \left(\sum_{i=1}^{n} x_i^2\right) a + \left(\sum_{i=1}^{n} x_i\right) b = \sum_{i=1}^{n} x_i y_i, \\ \left(\sum_{i=1}^{n} x_i^2\right) a \qquad\quad + nb = \sum_{i=1}^{n} y_i. \end{cases} \tag{20.9}$$

This system is linear with respect to the unknowns a and b. The determinant of this system is nonzero:

$$d = \begin{vmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & n \end{vmatrix} = n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 \neq 0 \quad . \tag{20.10}$$

(It is possible to prove that this determinant is positive.)

Therefore, the system has the only solution that can be found by the Cramer rule:

$$a^* = \frac{1}{d}\begin{vmatrix} \sum x_i y_i & \sum x_i \\ \sum y_i & n \end{vmatrix}, \qquad b^* = \frac{1}{d}\begin{vmatrix} \sum x_i^2 & \sum x_i y_i \\ \sum x_i & \sum y_i \end{vmatrix}. \tag{20.11}$$

So, we have found a single critical point (a^*, b^*). Make sure that it achieves a minimum of function $S(a,b)$. To do this, we calculate the second partial derivatives:

$$S''_{aa} = 2\sum_{i=1}^{n} x_i^2 = A \quad S''_{ab} = 2\sum_{i=1}^{n} x_i = B \quad , \quad S''_{bb} = 2n = C \quad .$$

We have

$$D = AC - B^2 = 4n\sum_{i=1}^{n} x_i^2 - 4\left(\sum_{i=1}^{n} x_i\right)^2 = 4d \quad .$$

Above, we noted that $d > 0$. Therefore, $D > 0$, so, according to the sufficient condition for an extremum, there is an extremum at the point under consideration (a^*, b^*). Since $A = 2\sum_{i=1}^{n} x_i^2 > 0$, then this extremum is a minimum. From the foregoing, we conclude that the function $S = S(a,b)$ has a single minimum point (a^*, b^*) determined from the system of normal equations. It should be noted that at this point there is not only a local, but also a global minimum, i.e. smallest function value.

***Example* 20.8.** The following data were obtained on the value of fixed assets *x* (thousand conventional units) and profit of the enterprise *y* (thousand conventional units):

x_i	110	132	154	176	198	220
y_i	40	43,2	52,8	67,2	64	78,4

Assuming a linear relationship exists between the *x* and *y* variables, find the empirical formula using the least squares method.

Decision. To determine the unknown parameters a^* and b^* of the empirical formula $y = a^* x + b^*$ we apply formulas (20.11). We need to pre-calculate the sums $\sum_{i=1}^{n} x_i$, $\sum_{i=1}^{n} y_i$, $\sum_{i=1}^{n} x_i y_i$, $\sum_{i=1}^{n} x_i^2$ (here n = 7). For convenience, we summarize the calculations in a table:

i	x_i	y_i	x_i^2	
1	110	40	12100	4
2	132	43,2	17424	5
3	154	52,8	23716	8
4	176	67,2	30976	1
5	198	64	39204	1
6	220	78,4	48400	1
7	242	96	58564	2
©	1232	441,6	230384	8

The system of normal equations 20.8) has the form:

$$\begin{cases} 230384a + 1232b = 83212{,}8, \\ 1232a + 7b = 441{,}6. \end{cases}$$

We find:

$$d = \begin{vmatrix} 230384 & 1232 \\ 1232 & 7 \end{vmatrix} = 2303847 - 1232^2 = 94864,$$

$$a^* = \frac{1}{94864} \begin{vmatrix} 83212{,}8 & 1232 \\ 441{,}6 & 7 \end{vmatrix} = \frac{38438{,}4}{94864} = 0{,}405,$$

$$b^* = \frac{1}{94864} \begin{vmatrix} 230384 & 83212{,}8 \\ 1232 & 441{,}6 \end{vmatrix} = \frac{-780595{,}2}{94864} = -8{,}229.$$

Thus, the desired dependence has the form:

$y = 0{,}405x - 8{,}229$.

Questions

1. What is the local maximum (minimum) function of two variables?
2. If $f'_x(x_0, y_0) = 0$, then can it be argued that (x_0, y_0) – is the extremum point for $f(x, y)$?
3. What is a critical (stationary) point for a function of two variables?
4. What is the sufficient condition for the extremum for the function of two variables?
5. What is the conditional extremum of a function of *n* variables?
6. What is the Lagrange function?
7. What are empirical formulas? Which line is called the experimental curve?
8. How many stages usually consists of solving the problem of finding empirical formulas? What are these steps?
9. What is the least-squares method?

Chapter 21. Optimization tasks

21.1. Basic concepts

To simplify the presentation, we identify the points of the Euclidean space as $\mathbf{R}^n$ and the vectors of their coordinates, i.e. the point $M(x_1, x_2, ..., x_n)$ will be written as $\bar{x}$, where $\bar{x} = (x_1, x_2, ..., x_n)$. Moreover, all operations on vectors and their properties are transferred to points in space $\mathbf{R}^n$. Let $\bar{x}$ and $\bar{y}$ – be two points of Euclidean space.

Definition. A line passing through points $\bar{x}$ and $\bar{y}$ of Euclidean space $\mathbf{R}^n$ is called the set of points

$$\bar{x} + t(\bar{y} - \bar{x}) \qquad (21.1)$$

of this space where. $t \in (-\infty, \infty)$. **The segment** $\overline{xy}$, connecting these points is called the set (21.1), where. $t \in [0, 1]$.

Obviously, the points $\bar{x}$ and $\bar{y}$ are obtained from (21.1) for $t = 0$ and $t = 1$.

Definition. A set *D* of points in Euclidean space is called convex if, together with any two points $\bar{x}$ and $\bar{y}$ all points of the segment $\overline{xy}$ also belong to this set.

Examples of convex planar sets are shown in Fig. 21.1.

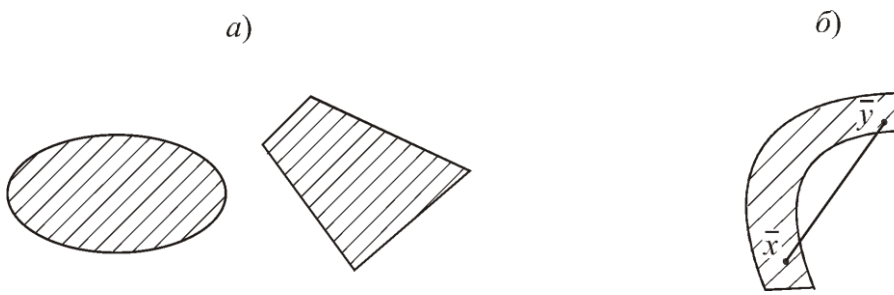


Fig. 21.1. Sets: convex (*a*) and non-convex (*b*)

Definition. A function $f(\bar{x})$, defined on a convex set $D \subset \mathbf{R}^n$ is called **convex** on *D*, if for any two points $\bar{x}$ and $\bar{y}$ from *D* and any $\alpha \in [0,1]$ the inequality holds:

$$f(\alpha\bar{x} + (1-\alpha)\bar{y}) \le \alpha f(\bar{x}) + (1-\alpha) f(\bar{y}). \quad (21.1а)$$

A function $f(\bar{x})$ is called **concave** on a convex set *D*, if, for any two points $\bar{x}$ and $\bar{y}$ from *D* and any $\alpha \in [0,1]$ the inequality holds

$$f(\alpha\bar{x} + (1-\alpha)\bar{y}) \ge \alpha f(\bar{x}) + (1-\alpha) f(\bar{y}). \quad (21.1б)$$

If inequalities (28.2) are replaced by strict inequalities, then we obtain the definition of **strictly convex** and **strictly concave** functions, respectively.

Note that the graph of a convex function of one variable is convex downward, and the graph of a concave function is convex upward.

The following important statement holds (we give it without proof).

***Theorem* 21.1.** If a function $f(\bar{x})$ is differentiable and strictly concave (strictly convex) on a convex set *D*, then it has a local extremum at only one point of this set.

21.2. The biggest value of a concave function. Kuna-Taker conditions

We know that a linear function of *n* variables is a function $l(\bar{x}) = a_1 x_1 + a_2 x_2 + ... + a_n x_n + b$, where a_1, a_2, …, a_n, b – are constants. It is easy to verify that the linear function is simultaneously a convex and concave function in $\mathbf{R}^n$.

An inequality of the form $l(\bar{x}) \geq 0$, where $l(\bar{x})$ – where is a linear function, is also called **linear**.

***Theorem* 21.2.** Let a convex set $D \subset \mathbf{R}^n$ be given by a system of linear inequalities:

$$\begin{cases} l_1(\bar{x}) \geq 0, \\ \cdots\cdots\cdots \\ l_m(\bar{x}) \geq 0, \end{cases} \quad (21.3)$$

D' – some convex subset in *D*; $f(\bar{x})$ – is a function concave on D', and $f(\bar{x})$ is differentiable at a point $\bar{x}^0 \in D'$. Then:

1) if for some numbers λ_1, …, λ_m the conditions are satisfied:

a) $\bar{x}^0$ – function critical point

$L(\bar{x}) = f(\bar{x}) + \lambda_1 l_1(\bar{x}) + ... + \lambda_m l_m(\bar{x})$; (Kuna-Taker co

b) $\lambda_i l_i(\bar{x}^0) = 0$ and $\lambda_i \geq 0$, $i = 1, ..., m$,

then $f(\bar{x}^0)$ – is *the largest* value $f(\bar{x})$ on *D*;

2) on the contrary, if $D' = D$ and $f(\bar{x}^0)$ – is the largest value $f(\bar{x})$ on *D*, then *there are* numbers $\lambda_1, \ldots, \lambda_m$, for which conditions "a" and "b" are fulfilled.

The function $L(\bar{x})$ – is the Lagrange function that we already know (see § 20.3), and the numbers $\lambda_1, \ldots, \lambda_m$ – are the Lagrange multipliers.

Theorem 21.2 is also accepted without proof.

***Example* 21.1.** Find the point of greatest value (global maximum) of the function $u = \ln x_1 + \ln x_2 + \ln x_3$ given that $x_1 + 4x_2 + 9x_3 \le 108$.

Decision. First of all, we note that the function *u* is concave. Indeed, the logarithmic function of one variable is concave, and the sum of concave functions, as is easy to verify, is a concave function.

We compose the Lagrange function:

$$L(\bar{x}) = \ln x_1 + \ln x_2 + \ln x_3 + \lambda \cdot (108 - x_1 - 4x_2 - 9x_3).$$

The Kuhn-Tucker conditions are as follows:

$$\begin{cases} \dfrac{1}{x_1} - \lambda = 0, \\ \dfrac{1}{x_2} - 4\lambda = 0, \\ \dfrac{1}{x_3} - 9\lambda = 0, \\ \lambda(108 - x_1 - 4x_2 - 9x_3) = 0, \\ \lambda \ge 0. \end{cases}$$

From the first three conditions we find: $x_1 = \frac{1}{\lambda}$, $x_2 = \frac{1}{4\lambda}$, $x_3 = \frac{1}{9\lambda}$. At the same time, obviously, $\lambda \neq 0$. Substituting the found values x_1, x_2 and x_3 into the fourth equation, we obtain

$$108 - x_1 - 4x_2 - 9x_3 = 108 - \frac{3}{\lambda} = 0.$$

From here $\lambda = \frac{1}{36}$, $x_1 = 36$, $x_2 = 9$, $x_3 = 4$. As already noted, the function u is concave, so the point $\bar{x}^0 = (36, 9, 4)$ is the point of global maximum.

Profit maximization

Let $F(K, L)$ – be the production function (where *K* and *L* – are the costs of capital and labor, respectively), *P* – is the price of production. The profit function Π is usually calculated by the formula:

$$\Pi(K, L) = P \cdot F(K, L) - WL - RK, \qquad (21.4)$$

where *W* and *R* – accordingly, factor prices for labor and capital expenditures, *W* and *R* – are positive numbers.

A point (K_0, L_0) is called **an optimal plan** if the if-function (21.4) in it assumes the maximum value.

Consider the problem: find the marginal rate of substitution of the production function *F*:

$$\mu = -\frac{F'_L}{F'_K}$$

with the optimal plan.

At the point of local maximum, the first partial derivatives of the profit function $\Pi(K,L)$ are zero. System (21.2) in this case has the form

$$\begin{cases} P \cdot F'_K(K_0, L_0) - R = 0, \\ P \cdot F'_L(K_0, L_0) - W = 0. \end{cases}$$

From here $\mu = -\dfrac{W}{R}$.

Now we consider the problem of maximizing the profit function.

***Example* 21.2.** Find the optimal plan and maximum profit function (28.4) if the production function has the form $F(K,L) = 3K^{1/3}L^{1/3}$.

Decision. The profit function in this case has the form

$$\Pi(K,L) = 3P \cdot K^{1/3}L^{1/3} - WL - RK.$$

We calculate the first partial derivatives with respect to *K* and *L* and equate them to zero:

$$\begin{cases} P \cdot K^{-2/3}L^{1/3} - R = 0, \\ P \cdot K^{1/3}L^{-2/3} - W = 0. \end{cases}$$

From here we find the coordinates of the optimal plan:

$$K_0 = \frac{P^3}{R^2 W}, \qquad L_0 = \frac{P^3}{RW^2}.$$

Substituting these values in the profit function, we get:

$$\Pi_{\max} = \frac{P^3}{RW}.$$

Demand optimization

Consider the task of optimizing the utility function with restrictions on consumer income.

***Example* 21.3.** Find the demand *x* and *y* for two varieties of goods at prices of *p* and *q*, respectively, if the consumer's income is equal to *M*, the utility function has the form $U(x,y)=x^{\frac{p}{p+q+1}}y^{\frac{q}{p+q+1}}$ and the consumer seeks to maximize the utility function.

Decision. It follows from the condition that a consumer can only buy such sets (x,y), whose value does not exceed his income, i.e.

$$px+qy\le M,\ x\ge 0,\ y\ge 0. \qquad (21.5)$$

Constraints (21.5) define a closed area in the form of a triangle on the plane (Fig. 21.2). It is necessary to find the maximum point of the function $U(x,y)$. We calculate the partial derivatives of the utility function:

$$U'_x(x,y)=\frac{p}{p+q+1}x^{-\frac{q+1}{p+q+1}}y^{\frac{q}{p+q+1}};$$

$$U'_y(x,y)=\frac{q}{p+q+1}x^{\frac{p}{p+q+1}}y^{-\frac{p+1}{p+q+1}}.$$

We see that there are no critical points inside the area. Therefore, the maximum can only be achieved at the border. On the lines $x=0$, $y=0$ the utility function is zero: $U(0,y)=U(x,0)=0$, therefore, we must look for the maximum point on the line $px+qy=M$. From here

$$y=\frac{M-px}{q}. \qquad (*)$$

Substituting from this equation the expression *y* into $U(x,y)$, we obtain the function of one variable *x*:

$$U\left(x, \frac{M-px}{1}\right) = f(x) = q^{-\frac{q}{p+q+1}} x^{\frac{p}{p+q+1}} (M-px)^{\frac{q}{p+q+1}}.$$

We calculate $f'(x)$:

$$f'(x) = q^{-\frac{q}{p+q+1}} \left[\frac{p}{p+q+1} x^{-\frac{q+1}{p+q+1}} (M-px)^{\frac{q}{p+q+1}} + \frac{q}{p+q+1} x^{-\frac{p+1}{p+q+1}} (-p) \right].$$

Equating $f'(x)$ to zero, after transformations we get $M - px - qx = 0$,

whence $x = \frac{M}{p+q}$ and taking into account (*) $y = \frac{M}{p+q}$

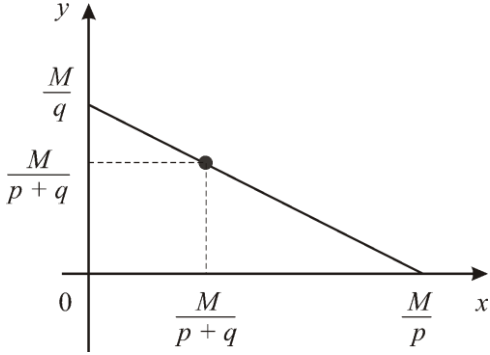


Fig. 21.2. Maximum Utility Function

Note that this problem could be solved by writing out the Lagrange function and the Kuhn-Tucker conditions, but this was not necessary for such a simple case.

Questions

1. What is a convex set in Euclidean space?

2. What function defined on a convex set is called convex (concave)?
3. How many local extrema does a strictly convex function have on a convex set?
4. Can a function be convex and concave at the same time?
5. What are the Kuhn-Tucker conditions?
6. What is the profit function? How is it calculated?
7. What is called an optimal plan?

ELEMENTS OF LINEAR ALGEBRA

Chapter 22. Vectors and operations. Linear spaces

22.1. Linear operations on vectors

It is well known, that if a rectangular coordinate system is set, then every vector $\overline{a}$ is represented by its coordinates a_1, a_2: $\overline{a} = (a_1, a_2)$. In a three-dimensional space, vector $\overline{a}$ is represented by three coordinates $\overline{a} = (a_1, a_2, a_3)$.

Definition. Any set of *n* real numbers ($a_1, a_2, ..., a_n$) is called an ***n*-dimensional vector** $\overline{a}$. These numbers are called coordinates or components of vector $\overline{a}$. For example, $\overline{a} = (4, 3, 2, 0, -7)$ is a five-dimensional vector. In particular, its third component is 2 and the fifth component is –7.

Note, that coordinates $\overline{a}$ can be presented as a row

$$\overline{a} = (a_1, a_2, \ldots, a_n) \qquad (22.1)$$

or column:

$$\bar{a} = \begin{pmatrix} a_1 \\ a_2 \\ \dots \\ a_n \end{pmatrix}. \quad (22.2)$$

The vector in form (22.1) is called a row vector and the vector in form (22.2) is a row column.

The number of vector coordinates is called the dimension of the vector. Two *n*-dimensional vectors $\bar{a} = (a_1, a_2, ..., a_n)$ and $\bar{b} = (b_1, b_2, ..., b_n)$ are **equal** if their corresponding coordinates are equal: $a_1 = b_1$, $a_2 = b_2$, …, $a_n = b_n$. In this case, we denote in form $\bar{a} = \bar{b}$.

The sum of two n-dimensional vectors $\bar{a} = (a_1, a_2, ..., a_n)$ and $\bar{b} = (b_1, b_2, ..., b_n)$ is the following vector

$$\bar{a} + \bar{b} = (a_1 + b_1, a_2 + b_2, ..., a_n + b_n).$$

Vector, which components are equal to zero, is called the zero vector:

$$\bar{0} = (0, 0, ..., 0).$$

Vector $(-a_1, -a_2, ..., -a_n)$ is opposite to vector $\bar{a} = (a_1, a_2, ..., a_n)$ and denoted as $-\bar{a}$:

$$-\bar{a} = (-a_1, -a_2, ..., -a_n).$$

The difference of vectors is defined as: $\bar{a} - \bar{b} = \bar{a} + (-\bar{b})$.

Product of a vector $\bar{a} = (a_1, a_2, ..., a_n)$ **by number *k*** is vector $k\bar{a} = (ka_1, ka_2, ..., ka_n)$.

Addition of vectors and multiplication of a vector by a number are linear operations.

Let us note the following properties of linear operations, which are easy to prove.

1. $\bar{a}+\bar{b}=\bar{b}+\bar{a}$.

2. $(\bar{a}+\bar{b})+\bar{c}=\bar{a}+(\bar{b}+\bar{c})$.

3. $\bar{a}+\bar{0}=\bar{a}$.

4. $\bar{a}+(-\bar{a})=\bar{0}$.

5. $k\cdot(\bar{a}+\bar{b})=k\bar{a}+k\bar{b}$.

6. $(k_1+k_2)\cdot\bar{a}=k_1\bar{a}+k_2\bar{a}$.

7. $k_1(k_2\bar{a})=(k_1k_2)\cdot\bar{a}$.

8. $1\cdot\bar{a}=\bar{a}$.

Definition. The set of all *n*-dimensional vectors, in which operations of addition of vectors and multiplication of a vector by a number are defined, is called ***n*-dimensional vector space** and denoted as $\mathbf{R}^n$.

The space $\mathbf{R}^n$ is a **linear.**

22.2. Dot product of vectors.

Dot product of two vectors $\bar{a}=(a_1, a_2, ..., a_n)$ and $\bar{b}=(b_1, b_2, ..., b_n)$ is a number

$$(\bar{a},\bar{b})=a_1b_1+a_2b_2+...+a_nb_n. \quad (22.3)$$

Let us illustrate the dot product with the following example.

***Example* 22.1.** A housewife buys 0,5 kg of bread, 5 kg of potatoes, 3 kg of cucumbers, 2 kg of tomatoes and 1,5 kg of meat at prices of 12, 11, 15, 30,

80 rubles per kilogram respectively. If we consider a vector of goods $\bar{a}$ = (0,5; 5; 3; 2; 1,5) and vector of prices $\bar{b}$ = (12; 11; 15; 30; 80), then the total sum of money is expressed as dot product: $(\bar{a},\bar{b})$ = 0,5·12 + 5·11 + 3·15 + 2·30 + 1,5·80 = 286 rubles.

***Example* 22.2.** The amount of 3 000 000 rubles is placed at interest for a year at four banks: 500 000 – at 6%, 500 000 – at 8%, 1 000 000 – at 5% and 1000000 – at 10%.

Here $\bar{a}$ = (500 000, 500 000, 1 000 000, 1 000 000) is a deposit vector, and $\bar{b}$ = (0,06; 0,08; 0,05; 0,10) is an interest rate vector.

The initial amount increases by the amount expressed by a dot product $(\bar{a},\bar{b})$ = 500 000·0,06 + 500 000·0,08 + 1 000 000·0,05 + 1 000 000·0,10 = 220 000 rubles.

Let us list the main properties of a dot product:

1. $(\bar{a},\bar{b})=(\bar{b},\bar{a})$.

2. $(k\cdot\bar{a},\bar{b})=k\cdot(\bar{a},\bar{b})$.

3. $(\bar{a},\bar{b}+c)=(\bar{a},\bar{b})+(\bar{a},\bar{c})$.

4. $(\bar{a},\bar{a})\geq 0$; herewith $(\bar{a},\bar{a})$ = 0 if and only if $\bar{a}$ is a zero vector.

22.3. Linear dependence of vectors

***Definition*.** Vector $\bar{a}$ is a **linear combination** of vectors $\bar{a}_1$, $\bar{a}_2$, …, $\bar{a}_s$ from $\mathbf{R}^n$ if

$\bar{a}=\lambda_1\vec{a}_1+\lambda_2\vec{a}_2+...+\lambda_s\bar{a}_s$,

where $\lambda_1, \lambda_2, ..., \lambda_s$ are real numbers. In this case, vector $\bar{a}$ is expressed in terms of its coordinates $\bar{a}_1$, $\bar{a}_2$, …, $\bar{a}_s$.

For example, let $\bar{a}_1$ = (2, 1, 3, –2), $\bar{a}_2$ = (3, 1, 2, 2), $\bar{a}_3$ = (2, 1, 2, 0), then $3\bar{a}_1 + 2\bar{a}_2 - 4\bar{a}_3$ = (6, 3, 9, –6) + (6, 2, 4, 4) – (8, 4, 8, 0) = (4, 1, 5, –2).

Vector $\bar{a}$ = (4, 1, 5, –2) is a linear combination of vectors $\bar{a}_1$, $\bar{a}_2$, $\bar{a}_3$: $\bar{a} = 3\bar{a}_1 + 2\bar{a}_2 - 4\bar{a}_3$.

Let us call any set of vectors from $\mathbf{R}^n$ a **system of vectors.** In the example above the system consists of four vectors: $\bar{a}_1$, $\bar{a}_2$, $\bar{a}_3$ and $\bar{a}$. Herewith vector $\bar{a}$ is a linear combination of other vectors of this system.

Definition. The system of vectors $\bar{a}_1$, $\bar{a}_2$, …, $\bar{a}_m$ is called linearly dependent if there exist numbers $\lambda_1, \lambda_2, ..., \lambda_m$, such that they are not equal to zero at the same time, or

$$\lambda_1\bar{a}_1 + \lambda_2\bar{a}_2 + ... + \lambda_m\bar{a}_m = \bar{0}. \quad (22.4)$$

Otherwise, vectors $\bar{a}_1$, $\bar{a}_2$, …, $\bar{a}_m$ are called linearly independent. In other words, vectors $\bar{a}_1$, $\bar{a}_2$, …, $\bar{a}_m$ are linearly dependent if it follows from an equality (22.4) that $\lambda_1 = \lambda_2 = ... = \lambda_m = 0$.

Let us prove, that the system, which consist of more than one vector $\bar{a}_1$, $\bar{a}_2$, …, $\bar{a}_m$, is linearly dependent if and only if at least one of the vectors is a linear combination of the others.

1. Let equality (22.4) be verified and at least one of the coefficients is not equal to zero, (for instance $\lambda_m \neq 0$). Then

$$\bar{a}_m = -\frac{\lambda_1}{\lambda_m}\bar{a}_1 - \frac{\lambda_2}{\lambda_m}\bar{a}_2 - \ldots,$$

i.e. vector $\bar{a}_m$ is a linear combination of the other vectors.

2. Let one of vectors (for example $\bar{a}_2$) be a linear combination of the others:

$$\bar{a}_2 = \lambda_1\bar{a}_1 + \ldots + \lambda_m\bar{a}_m.$$

Then

$$\lambda_1\bar{a}_1 + (-1)\cdot\bar{a}_2 + \ldots + \lambda_m\bar{a}_m = \bar{0},$$

and in the last equality, there is a coefficient, which is not equal to zero ($\lambda_2 = -1$). Thus, the system of vectors $\bar{a}_1$, $\bar{a}_2$, …, $\bar{a}_m$ is linearly dependent.

The geometric meaning of the linear dependence of vectors is evident for the case of two-dimensional vectors on the plane and three-dimensional vectors in space:

a) the system which consists of two vectors is linearly dependent if and only if the vectors are collinear;

b) the system which consists of three vectors is linearly dependent if and only if these three vectors are collinear.

Let's note that some **properties of vectors in space** $\mathbf{R}^n$.

1. *If the system* $\bar{a}_1$, $\bar{a}_2$, …, $\bar{a}_m$ *contains a zero vector, then it is linearly dependent*.

To verify this, it is enough to take the zero vector with one coefficient equal to one and the rest coefficient equal to zero on the left-hand side of equality (22.4).

2. If a part of vectors of the system $\bar{a}_1, \bar{a}_2, \ldots, \bar{a}_m$ is linearly dependent[1], then all these vectors are linearly dependent.

Indeed, for example, let vectors $\bar{a}_2, \bar{a}_3, \bar{a}_5$ of the system $\bar{a}_1, \bar{a}_2, \bar{a}_3, \bar{a}_4, \bar{a}_5$ be linearly dependent: $\lambda_2\bar{a}_2 + \lambda_3\bar{a}_3 + \lambda_5\bar{a}_5 = \bar{0}$ and, for example, $\lambda_3 \neq 0$. Then

$$0 \cdot \bar{a}_1 + \lambda_2\bar{a}_2 + \lambda_3\bar{a}_3 + 0 \cdot \bar{a}_4 + \lambda_5\bar{a}_5 = \bar{0},$$

and in this equality at least one coefficient is not equal to zero ($\lambda_3 \neq 0$).

Thus, this system $\bar{a}_1, \bar{a}_2, \bar{a}_3, \bar{a}_4,$ is linearly dependent.
Without proof, we give the following important theorem.

***Theorem* 22.1.** Any system which contains $(n+1)$ vectors of space $\mathbf{R}^n$ is linearly dependent.
In particular, any four vectors in three-dimensional space are linearly dependent.

22.4. Basis and rank of vector system

Let us consider:

$$\bar{a}_1, \bar{a}_2, \ldots, \bar{a}_k. \quad (22.5)$$

Any subsystem of vector system (22.5) is called the **basis** of this system if it satisfies the following properties:
1) this subsystem is linearly independent;
2) any vector of system (22.5) is expressed linearly in terms of vectors of this system.

[1] In other words, this system of vectors contains a linearly dependent subsystem.

A system of vectors can have several bases. It is possible to show that all bases of the system of vectors consist of the same number of vectors. This number is called the **rank** of the system.

Obviously, the set $\mathbf{R}^n$ is a system that contains all *n*-dimensional vectors. The concept of a basis extents to $\mathbf{R}^n$.

Definition. The system of vectors is called basis of the space $\mathbf{R}^n$ if:
1) this system is linearly independent;

2) any vector of space $\mathbf{R}^n$ is expressed linearly in terms of vectors of this system.

An example of a system of vectors in $\mathbf{R}^n$ is a system which consists of *n* unit vectors

$\bar{e}_1 = (1, 0, \ldots, 0),$

$\bar{e}_2 = (0, 1, \ldots, 0),$

……

$\bar{e}_n = (0, 0, \ldots, 1).$

Indeed, on the one hand, this system is linearly independent (as from $\lambda_1\bar{e}_1 + \lambda_2\bar{e}_2 + \ldots + \lambda_n\bar{e}_n = \bar{0}$ follows $\lambda_1 = \lambda_2 = \ldots = \lambda_n = 0$), on the other hand, any vector $\bar{a} = (a_1, a_2, \ldots, a_n)$ is presented as:

$\bar{a} = a_1\bar{e}_1 + a_2\bar{e}_2 + \ldots + a_n\bar{e}_n$,

i.e. it is a linear combination of vectors $\bar{e}_1, \bar{e}_2, \ldots, \bar{e}_n$.

In the previous example, the basis consisted of *n* vectors. The following theorem takes place (we give it without proof).

***Theorem* 22.2.** A linearly independent system of vectors in $\mathbf{R}^n$ is a basis if and only if the number of these vectors is equal to *n*.

The number of basis vectors of a space, i.e. the maximal number of its linearly independent vectors, is called the dimension of the space. Space $\mathbf{R}^n$, previously called as n-dimensional for other reasons, is also n-dimensional in the sense that its dimension is equal to n.

22.5. Decomposition of the vector in the basis

Let the system of vectors

$$\bar{a}_1 = (a_{11}, a_{12}, ..., a_{1n}),\ \bar{a}_2 = (a_{21}, a_{22}, ..., a_{2n}),\ \ldots,\ \bar{a}_m = (a_{m1}, a_{m2}, ..., a_{mn}) \quad (22.6)$$

be a basis[1] and let vector $\bar{x}$ be decomposed in vectors (22.6):

$$\bar{x} = x_1\bar{a}_1 + x_2\bar{a}_2 + ... + x_m\bar{a}_m. \quad (22.7)$$

The question arises: are the coefficients x_1, x_2, …, x_m of decomposition (22.7) uniquely determined?

***Theorem* 22.3.** The decomposition of vector $\bar{x}$ in the basis vectors is unique.

Proof. We suppose that vector $\bar{x}$ is presented in the form of a linear combination of vectors (22.6) in two different ways:

$$\bar{x} = x_1\bar{a}_1 + x_2\bar{a}_2 + ... + x_m\bar{a}_m,$$

$$\bar{x} = x_1'\bar{a}_1 + x_2'\bar{a}_2 + ... + x_m'\bar{a}_m.$$

Subtracting the second equality from the first, we get

[1] Here components of vectors must be provided with double indices: the first one indicates the number of vectors and the second one indicates the number of the component.

$$(x_1 - x_1') \cdot \bar{a}_1 + (x_2 - x_2') \cdot \bar{a}_2 + ... + (x_m - x_m') \cdot \bar{a}_m = \bar{0}.$$

However, system (22.6) is linearly independent, thus,

$$x_1 - x_1' = 0, \quad x_2 - x_2' = 0, \quad ..., \quad x_m - x_m' = 0.$$

From here we get:

$$x_1 = x_1', \quad x_2 = x_2', \quad ..., \quad x_m = x_m'.$$

So, decomposition (22.7) is unique. Decomposition coefficients (22.7) are called *coordinates of vector* $\bar{x}$ *in basis* (22.6).

22.6. Normed vector spaces. Euclidean space

Definition. **Linear space** is a set *V of arbitrary elements*, called vectors, for which operations of addition and multiplication by a real number are defined, i.e. for any two vectors $\bar{u}_1$ and $\bar{u}_2$ from *V* vector $\bar{u}$, called a sum of vectors $\bar{u}_1$ and $\bar{u}_2$, is defined and denoted as $\bar{u}_1 + \bar{u}_2$ and for any vector $\bar{u}$ and any real number λ vector $\lambda\bar{u}$, called the multiplication of vector $\bar{u}$ *by a number* λ, is defined and the following conditions are satisfied:

1) $\bar{u}_1 + \bar{u}_2 = \bar{u}_2 + \bar{u}_1$;

2) $(\bar{u}_1 + \bar{u}_2) + \bar{u}_3 = \bar{u}_1 + (\bar{u}_2 + \bar{u}_3)$;

3) in the set *V*, there is an element $\bar{0}$ that is called a zero element which satisfies for any $\bar{u}$ the following condition:

$\bar{u} + \bar{0} = \bar{u}$;

4) for any vector $\bar{u}$ there is a vector $-\bar{u}$ that is called the opposite vector $\bar{u}$, which satisfies condition

$\bar{u}+(-\bar{u})=\bar{0}$;

5) $\lambda(\bar{u}_1+\bar{u}_2)=\lambda\bar{u}_1+\lambda\bar{u}_2$;

6) $(\lambda_1+\lambda_2)\bar{u}=\lambda_1\bar{u}+\lambda_2\bar{u}$;

7) $\lambda_1(\lambda_2\bar{u})=(\lambda_1\lambda_2)\bar{u}$;

8) $1\cdot\bar{u}=\bar{u}$.

(The conditions listed above are called axioms of linear space.)

It is necessary to note that set *V* can consist of elements of any kind of nature.

Examples of linear spaces are

1) *n*-dimensional vector space;

2) the set of all polynomials $P_n(x)$ of degree not higher than *n* with ordinary addition and multiplication by numbers.

Let us note that the set of all polynomials which degree is equal to n is not a linear space with respect to the usual operations of addition and multiplication by numbers. This is due to the fact that the algebraic sum of polynomials of degree n can be a polynomial of a degree less than n.

Definition. Linear space *V* is called a normed space if for any vector $\bar{u}$ a norm $\|\bar{u}\|$ with the following properties is defined

1) $\|\bar{0}\|=0$;

2) any $\bar{u}\neq\bar{0}$ satisfies inequality $\|\bar{u}\|>0$;

3) any real number λ satisfies equality $\|\lambda\bar{u}\|=\lambda\|\bar{u}\|$;

4) for any $\bar{u}$ and $\bar{v}$ from *V* the triangle inequality holds:
$\|\bar{u}+\bar{v}\|\leq\|\bar{u}\|+\|\bar{v}\|$.

Let us note that if V is a set of vectors of an ordinary plane, then V is a normed space with the norm $\|\bar{u}\| = |\bar{u}|$, where $|\bar{u}|$ is a vector length.

One way to define a norm in space V is to define a dot product.

We say that the dot product is given in space V if for each pair of vectors $\bar{u}$ and $\bar{v}$ there is a number $(\bar{u}, \bar{v})$, so as the following conditions are satisfied:

1) $(\bar{u}, \bar{v}) = (\bar{v}, \bar{u})$;

2) $(\lambda\bar{u}, \bar{v}) = \lambda(\bar{u}, \bar{v})$;

3) $(\bar{u}, \bar{v}_1 + \bar{v}_2) = (\bar{u}, \bar{v}_1) + (\bar{u}, \bar{v}_2)$;

4) $(\bar{u}, \bar{u}) \geq 0$; herewith $(\bar{u}, \bar{u}) = 0$ if and only if $\bar{u}$ is a zero vector: $\bar{u} = \bar{0}$.

If a dot product is set, then the norm is defined as following:

$$\|\bar{u}\| = \sqrt{(\bar{u}, \bar{u})}. \quad (22.8)$$

Make sure that the norm defined by equality (22.8) has all the properties listed above. First three properties are evident, it is necessary to check only the triangle inequality. For this purpose we previously prove the Cauchy-Bunyakovsky inequality:

$$(\bar{u}, \bar{v})^2 \leq (\bar{u}, \bar{u})(\bar{v}, \bar{v}). \quad (22.9)$$

Let us consider vector $\bar{w} = t\bar{u} + \bar{v}$, where t is an arbitrary number. We have

$$(\bar{w}, \bar{w}) = (t\bar{u} + \bar{v}, t\bar{u} + \bar{v}) = t^2(\bar{u}, \bar{u}) + 2t(\bar{u}, \bar{v}) + (\bar{v}, \bar{v}).$$

We denote $(\bar{u}, \bar{u}) = \alpha$, $(\bar{u}, \bar{v}) = \beta$, $(\bar{v}, \bar{v}) = \gamma$. We get

$$(\bar{w}, \bar{w}) = \alpha t^2 + 2\beta t + \gamma.$$

As $(\overline{w},\overline{w})\geq 0$, then $\alpha t^2+2\beta t+\gamma\geq 0$ for all *t*. Thus, the discriminant of this square trinomial is less than or equal to zero. So, $\beta^2-\alpha\gamma\leq 0$, i.e. $\beta^2\leq\alpha\gamma$ or $(\overline{u},\overline{v})^2\leq(\overline{u},\overline{u})(\overline{v},\overline{v})$.

Let us prove the triangular inequality

$\sqrt{\overline{u}+\overline{v},\overline{u}+\overline{v}}\leq\sqrt{(\overline{u},\overline{u})}+\sqrt{(\overline{v},\overline{v})}$.

Using the Cauchy-Bunyakovsky inequality we get

$\|\overline{u}+\overline{v}\|^2=(\overline{u}+\overline{v},\overline{u}+\overline{v})=(\overline{u},\overline{u})+(\overline{v},\overline{v})+2(\overline{u},\overline{v})\leq\|\overline{u}\|^2+\|\overline{v}\|^2+2\|\overline{u}\|\,\|\overline{v}\|=$

$=(\|\overline{u}\|+\|\overline{v}\|)^2$,

that proves it.

Previously (see. § 22.2) we considered a dot product of *n*-dimensional vectors $\overline{a}=(a_1,a_2,...,a_n)$ and $\overline{b}=(b_1,b_2,...,b_n)$ defined by formula (22.3):

$(\overline{a},\overline{b})=a_1b_1+a_2b_2+...+a_nb_n$.

Definition. *N-dimensional vector space* $\mathbf{R}^n$, in which a dot vector product is set, is called Euclidean space.

The length (norm) of vector $\overline{a}=(a_1,a_2,...,a_n)$ in Euclidean space $\mathbf{R}^n$ is a square root of its dot product

$$|\overline{a}|=\sqrt{(\overline{a},\overline{a})}=\sqrt{a_1^2+a_2^2+...+a_n^2}. \qquad (22.10)$$

The angle between two vectors $\overline{a}$ and $\overline{b}$ is defined by an equality

$$\cos\varphi=\frac{(\overline{a},\overline{b})}{|\overline{a}|\cdot|\overline{b}|}. \qquad (22.11)$$

From the Cauchy-Bunyakovsky it follows that cos ϕ, defined by formula (22.11), satisfies a condition $|\cos\varphi| \le 1$, so the definition of angle between vectors is correct.

Vectors $\bar{a}$ and $\bar{b}$ are called **orthogonal** if $(\bar{a}, \bar{b}) = 0$. It follows from the definition that if two nonzero vectors are orthogonal, then the angle between them is equal to $\frac{\pi}{2}$.

We say that vectors $\bar{e}_1$, $\bar{e}_2$, ..., $\bar{e}_m$ form an orthonormal system in *n*-dimensional Euclidean space $\mathbf{R}^n$ if these vectors are pairwise orthogonal, i.e. $(\bar{e}_i, \bar{e}_j) = 0$ for $i \ne j$, (*i*, *j* = 1, 2, ..., *m*), and a norm of each of them is equal to one: $(\bar{e}_i, \bar{e}_i) = 1$.

Make sure that every orthonormal system is linearly independent, i.e. from equality

$$\lambda_1\bar{e}_1 + \lambda_2\bar{e}_2 + ... + \lambda_m\bar{e}_m = \bar{0} \qquad (22.12)$$

follows that $\lambda_1 = \lambda_2 = ... = \lambda_m = 0$.

Let *i* be an arbitrary number that satisfies condition $1 \le i \le m$. Let us multiply equality (22.12) by $\bar{e}_i$:

$$\lambda_1(\bar{e}_1, \bar{e}_i) + \lambda_2(\bar{e}_2, \bar{e}_i) + ... + \lambda_m(\bar{e}_m, \bar{e}_i) = 0.$$

As $(\bar{e}_i, \bar{e}_j) = 0$ for $i \ne j$, so the last equality is equal to equality $\lambda_i(\bar{e}_i, \bar{e}_i) = 0$.

Hence, as $(\bar{e}_i, \bar{e}_i) \ne 0$, we get $\lambda_i = 0$. So, $\lambda_i = 0$ for all *i* = 1, 2, ..., *m*.

Since any orthonormal system of vectors is linearly independent, it follows that the orthonormal system containing *n* vectors forms a basis in space $\mathbf{R}^n$ called orthonormal. The following statement holds.

***Theorem* 22.4.** In any *n*-dimensional Euclidean space, there exists an orthonormal basis.

An example of orthonormal basis is a system of n unit vectors $\bar{e}_1$ = (1, 0, ..., 0), $\bar{e}_2$ = (0, 1, ..., 0), ..., $\bar{e}_n$ = (0, 0, ..., 1).

Questions

1. Can two vectors be equal if one of them is four-dimensional and the other is five-dimensional?
2. What vectors are obtained from vector $\bar{a}$ by multiplying it by numbers 0 and –1?
3. What vectors are called linearly independent?
4. Is the system of vectors $\bar{a} = (1, 2, 3)$, $\bar{b} = (2, 3, 4)$, $\bar{c} = (3, 4, 5)$, $\bar{d} = (4, 5, 6)$ linearly independent?
5. Do vectors $\bar{e}_1 = (1, 0, 0, 0)$, $\bar{e}_2 = (0, 2, 0, 0)$, $\bar{e}_4 = (0, 0, 0, 4)$ form a basis in space $\mathbf{R}^4$?
6. What numbers are called the coordinates of a vector in the basis?
7. For which value of *y* is the dot product of vectors $\bar{a} = (1, 2, 3)$ and $\bar{b} = (1, y, 3)$ equal to zero?
8. For which values *y* do vectors $\bar{a} = (1, 2, 3)$ and $\bar{b} = (-3, y, -9)$ form a linearly independent system?

Chapter 23. Matrices and operations on them

23.1. Basic concepts

Definition. **A matrix** (a numerical matrix) of dimension $m \times n$ is a rectangular table of numbers that contains *m* rows and *n* columns:

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots. & \ldots. & \ldots & \ldots. \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{pmatrix}. \quad (23.1)$$

Numbers that form a matrix are called **elements**.

To denote elements of matrix double indexed letters are used: a_{ij}, where *i* is a row number and *j* is a column number. Matrix is also written in a short form:

$A = (a_{ij})$, $i = 1,2, \ldots, m$; $j = 1, 2, \ldots, n$. (23.2)

In case the number of matrix rows is equal to the number of its columns, i.e. $m = n$, it is called a square matrix of order *n*.

Matrix can consist of one row or one column

$$A = (a_{11}, a_{12}, \ldots, a_{1n}), \qquad B = \begin{pmatrix} b_{11} \\ b_{12} \\ \ldots. \\ b_{1m} \end{pmatrix}.$$

Thus, row-vector or column-vector are special cases of matrices.

Elements of matrix a_{ij}, for which the number of rows is equal to the column number, i.e. $i = j$, are called diagonal. For square matrix elements $a_{11}, a_{22}, ..., a_{nn}$ form the main diagonal.

Matrix is called **symmetrical** if its elements, which are symmetrical to each other against the main diagonal are equal to each other $a_{ij} = a_{ji}$.

A square matrix is called diagonal if all its elements outside the main diagonal are equal to zero

$$A = \begin{pmatrix} a_{11} & 0 & ... & 0 \\ 0 & a_{22} & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & a_{nn} \end{pmatrix}.$$

A diagonal matrix is called **identity** if all its diagonal elements are equal to one

$$\begin{pmatrix} 1 & 0 & ... & 0 \\ 0 & 1 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & 1 \end{pmatrix}.$$

A matrix of any dimension $m \times n$ is called **zero matrix** or **null matrix** if all its elements are equal to zero

$$0 = \begin{pmatrix} 0 & 0 & ... & 0 \\ 0 & 0 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & 0 \end{pmatrix}.$$

Two matrices $A=(a_{ij})$ and $B=(b_{ij})$ are **equal** ($A=B$) if they have equal dimensions and their corresponding elements are equal: $a_{ij}=b_{ij}$.

23.2. Linear operations on matrices. Transposition of matrices

For matrices operations of addition and multiplication are defined.

A sum of matrices $A=(a_{ij})$ and $B=(b_{ij})$ of the equal order is matrix $C=(c_{ij})$ whose elements have a form: $c_{ij}=a_{ij}+b_{ij}$, i = 1,2, …, m; j = 1, 2, …, n. In this case, we write $C=A+B$.

Example **23.1.** Let $A=\begin{pmatrix}1 & 2\\ 3 & 0\\ 4 & 5\end{pmatrix}$, $B=\begin{pmatrix}3 & 4\\ 2 & 6\\ 5 & 3\end{pmatrix}$. Then $C=A+B=\begin{pmatrix}4 & 6\\ 5 & 6\\ 9 & 8\end{pmatrix}$.

A matrix product or matrix multiplication $A=(a_{ij})$ **by a real number** λ is matrix $\lambda A=(\lambda a_{ij})$.

Example **23.2.** Let $A=\begin{pmatrix}2 & 3 & 4\\ 3 & 2 & 1\end{pmatrix}$, λ = 4. Then $\lambda A=\begin{pmatrix}8 & 12 & 16\\ 12 & 8 & 4\end{pmatrix}$.

Addition of matrix and matrix multiplication by a product are called **linear operations on matrices.**

Properties of linear operations (directly follow from the definition):

1. $A+B=B+A$.

2. $(A+B)+C=A+(B+C)$.

3. $\lambda\cdot(A+B)=\lambda A+\lambda B$.

4. $(\lambda_1+\lambda_2)\cdot A=\lambda_1 A+\lambda_2 A$.

5. $(\lambda_1\lambda_2)\cdot A=\lambda_1(\lambda_2 A)=\lambda_2(\lambda_1 A)$.

6. $A+0=A$ (0 is a zero matrix).

7. If $\lambda=0$, then $\lambda A=0$ is a zero matrix.

The transposition of a matrix is an operation of replacing the matrix rows with its columns while preserving their order.

Denoting matrix A' which was obtained by transposition of matrix $A=(a_{ij})$ we can write: $A'=(a_{ji})$.

In particular, if matrix *A* is a row-vector, then matrix A' is a column-vector and vice versa.

***Example* 23.3.** If $A=\begin{pmatrix}2\\5\\8\end{pmatrix}$, then $A'=\begin{pmatrix}2 & 5 & 8\end{pmatrix}$.

If $A=\begin{pmatrix}2 & 3 & 4 & 5\\1 & 5 & 7 & 9\end{pmatrix}$, then $A'=\begin{pmatrix}2 & 1\\3 & 5\\4 & 7\\5 & 9\end{pmatrix}$.

Let us note the evident properties of transposition operations:

1. $A''=A$.

2. If *A* is a symmetric matrix, then $A'=A$.

23.3. Matrix multiplication

Multiplication of matrix *A* by matrix *B* is defined only for the cases when the *column number* of *A* is **equal** to *row number* of matrix *B*.

The multiplication of matrix A with dimension $m \times k$ by matrix B with dimension $k \times n$ is matrix $C = (c_{ij})$ with elements

$$c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + ... + a_{ik}b_{kj} = \sum_{s=1}^{k} a_{is}b_{sj};$$

$i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, n$.

It easy to note that element c_{ij} of matrix *C* is a dot product of *i*-th row-vector of matrix *A* by *j*-th column-vector of matrix *B*.

***Example* 23.4.** Calculate matrix multiplication *AB* where

$$A = \begin{pmatrix} 2 & 0 & 1 \\ -1 & 3 & 2 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 2 & 1 \\ 0 & 1 & 2 \\ -2 & 0 & 3 \end{pmatrix}.$$

Solution. Make sure that the column number of A is equal to the row number of B (and equal to 3). Thus, the multiplication is possible. The matrix $C = A \cdot B$ has dimension 2·3:

$$AB = \begin{pmatrix} 2\cdot 1 + 0\cdot 0 + 1\cdot(-2) & 2\cdot 2 + 0\cdot 1 + 1\cdot 0 & 2\cdot 1 + 0\cdot 2 + 1\cdot 3 \\ -1\cdot 1 + 3\cdot 0 + 2\cdot(-2) & -1\cdot 2 + 3\cdot 1 + 2\cdot 0 & -1\cdot 1 + 3\cdot 2 + 2\cdot 3 \end{pmatrix} =$$

$$= \begin{pmatrix} 0 & 4 & 5 \\ -5 & 1 & 11 \end{pmatrix}.$$

Let us list the **properties of matrix multiplication.** Let *A*,*B* and *C* be such a matrix that the matrix multiplication is defined. Then:

1. $(AB)\cdot C = A\cdot(BC)$.

2. $(A+B)\cdot C = AC + BC$.

3. $A\cdot(B+C) = AB + AC$.

4. $\lambda\cdot(AB) = (\lambda A)\cdot B$.

5. $AE = EA = A$.

Let us note that there is no commutative property ($AB = BA$) among the property of matrix multiplication. Moreover, if multiplication *AB* exists, then permutation of factors is not always possible, i.e. the multiplication *BA* may not exist.

In case *AB* and *BA* exist, these products may not coincide (and can be matrices of different orders).

***Example* 23.5.** Let $A = \begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix}, \quad B = \begin{pmatrix} 2 & 0 \\ 1 & 2 \end{pmatrix}$. Then

$$AB = \begin{pmatrix} 4 & 4 \\ 8 & 4 \end{pmatrix}, \quad BA = \begin{pmatrix} 2 & 4 \\ 7 & 6 \end{pmatrix},$$

i.e. $AB \neq BA$.

***Example* 23.6.** Let $A = \begin{pmatrix} 2 & 1 & 2 \\ 1 & -1 & 3 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 1 \\ 2 & 0 \\ 1 & -1 \end{pmatrix}$. Then

$$AB = \begin{pmatrix} 4 & 0 \\ 1 & -2 \end{pmatrix}, \quad BA = \begin{pmatrix} 1 & -1 & 3 \\ 4 & 2 & 4 \\ 1 & 2 & -1 \end{pmatrix},$$

i.e. here not only does $AB \neq BA$ have different dimensions, but *AB* and *BA* also do.

Let us consider one more property of matrix multiplication connected with the operation of the transposition.

If matrices *A* and *B* are such that their multiplication is defined, then the following equality holds:

6. $(AB)' = B'A'$.

In other words, the matrix obtained by transposing the product is equal to the product of the matrices obtained by transposing the factors taken in the inverse order.

Proof. First of all, we make sure that the product AB is defined, then the product $B'A'$ is also defined. Indeed, if the product *AB* is defined, then the columns number of matrix *A* is equal to the rows number of matrix *B.*

But the rows number of matrix B is equal to the columns number of matrix B' and the columns number of matrix A is equal to the rows number of matrix A', thus, the product $B'A'$ is defined. Further, the element of matrix $(AB)'$, placed at its *i-th* row and *j-th* column is an element of matrix *AB*, placed at its *j-th* row and *i-th* column.

Thus, it is equal to the dot product of *j-th* row of matrix A and *i-th* column of matrix B, i.e. it is equal to the sum of products of corresponding elements of *j-th* column of matrix A' and *i-th* row of matrix B'. That means that the element of matrix $B'A'$ placed at its *i-th* row and *j-th* column is also equal to the dot product of *j-th* row of matrix *A* and *i-th* product of matrix *B*. The equality is proved.

23.4. Inverse of a matrix

There exists no operation for the matrices division. However, for square matrices, it is possible to define an operation inverse to multiplication under certain conditions. Before doing this, we introduce some necessary concepts. We can consider any matrix as a system of its row-vectors and column-vectors. It is possible to prove that the rank of a system of row-

vectors of a matrix is equal to the rank of a system of its column-vectors (i.e. the maximal number of linearly independent row-vectors of a matrix is equal to the maximal number of its linearly independent column-vectors).

Definition. **The rank** of a matrix is the rank of the system of its row-vectors (or column-vectors).

A square matrix *A* of dimension n is called **nondegenerate** if its rows are linearly independent (i.e. its rank is equal to *n*). Otherwise, matrix *A* is called **degenerate**.

Before defining the concept of an inverse matrix, let us note that for every number $a \neq 0$ there exists a number inverse to it: $a^{-1} = \frac{1}{a}$, such as $aa^{-1} = 1$.

Definition. Let *A* be a square matrix. Matrix A^{-1} is **inverse** with respect to matrix *A* if their product is equal to the unit matrix:

$AA^{-1} = E$.

It is easy to make sure that the multiplication of matrix *A* and A^{-1} is commutative:

$AA^{-1} = A^{-1}A = E$.

Further, we will show that the inverse matrix exists only for a *nondegenerate square* matrix.

Elementary matrix transformations are:

permutation of rows (columns);

multiplication of a row (column) by a nonzero number;

adding to the elements of a row (column) the corresponding elements of another row (column) multiplied by a number.

It is easy to show that, as a result of an elementary transformation of the non degenerate matrix, we obtain again a nondegenerate matrix.

Inverse of a matrix using elementary transformation

1. Let A be a nondegenerate square matrix. Let us attach to it a unit matrix of the same size. Then we obtain a dual matrix ($A|E$).

2. Then we do elementary transformations on the rows of matrix ($A|E$) to obtain a unit matrix *E* at the place of matrix *A*. Then at the place of the attached matrix *E* matrix A^{-1} is obtained.
(Let us note that in practical use there is no need to check nondegeneracy of matrix *A*. It follows from the possibility of reducing *A* to *E*.)

***Example* 23.7.** Given a matrix

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 2 & 3 \\ 3 & 3 & 4 \end{pmatrix}.$$

Find the inverse matrix A^{-1}.

Solution. Let us compose matrix ($A|E$) and apply the method of elementary transformations. Here l_i is *i*-th row (*i* = 1, 2, 3):

$$(A|E)=\left(\begin{array}{ccc|ccc}1&2&3&1&0&0\\2&2&3&0&1&0\\3&3&4&0&0&1\end{array}\right)\xrightarrow{l_2-2l_1,\ l_3-3l_1}\left(\begin{array}{ccc|ccc}1&2&3&1&0&0\\0&-2&-3&-2&1&0\\0&-3&-5&-3&0&1\end{array}\right)$$

$$\xrightarrow{l_1+l_2,\ 2l_3-3l_2}\left(\begin{array}{ccc|ccc}1&0&0&1&1&0\\0&-2&-3&-2&1&0\\0&0&-1&0&-3&2\end{array}\right)\longrightarrow$$

$$\xrightarrow{l_2-3l_3}\left(\begin{array}{ccc|ccc}1&0&0&1&1&0\\0&-2&0&-2&10&-6\\0&0&-1&0&-3&2\end{array}\right)\longrightarrow\left(\begin{array}{ccc|ccc}1&0&0&-1&1&0\\0&1&0&1&-5&3\\0&0&1&0&3&-2\end{array}\right).$$

So, $A^{-1}=\begin{pmatrix}-1&1&0\\1&-5&3\\0&3&-2\end{pmatrix}$.

Using the inverse matrix, it is possible to solve matrix equations of the following types:

$AX=B,\ XA=B$.

***Example* 23.8.** Solve the matrix equations

1) $\begin{pmatrix}1&2\\3&4\end{pmatrix}X=\begin{pmatrix}5&6\\11&12\end{pmatrix}$. 2) $\begin{pmatrix}1&0&2\\2&1&4\\1&1&1\end{pmatrix}X=\begin{pmatrix}1&3&5\\4&7&11\\3&3&3\end{pmatrix}$.

3) $X\begin{pmatrix} 1 & 2 & -2 \\ 2 & 5 & -4 \\ -2 & -4 & 5 \end{pmatrix} = \begin{pmatrix} -1 & -2 & 3 \\ 1 & 4 & -1 \\ 5 & 11 & -11 \end{pmatrix}$.

Solution 1. This is an equation of the form $AX = B$. Its solution is $X = A^{-1}B$. Here $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$. Let us find A^{-1}:

$$\left(\begin{array}{cc|cc} 1 & 2 & 1 & 0 \\ 3 & 4 & 0 & 1 \end{array}\right) \to \left(\begin{array}{cc|cc} 1 & 2 & 1 & 0 \\ 0 & -2 & -3 & 1 \end{array}\right) \to \left(\begin{array}{cc|cc} 1 & 0 & -2 & 1 \\ 0 & -2 & -3 & 1 \end{array}\right) \to \left(\begin{array}{cc|cc} 1 & 0 & -2 & 1 \\ 0 & 1 & 3/2 & -1/2 \end{array}\right)$$

,

i.e. $A^{-1} = \begin{pmatrix} -2 & 1 \\ 3/2 & -1/2 \end{pmatrix}$. Then

$$X = \begin{pmatrix} -2 & 1 \\ 3/2 & -1/2 \end{pmatrix} \cdot \begin{pmatrix} 5 & 6 \\ 11 & 12 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 2 & 3 \end{pmatrix}.$$

So, $X = \begin{pmatrix} 1 & 0 \\ 2 & 3 \end{pmatrix}$.

2. This equation has a form $AX = B$. Here $A = \begin{pmatrix} 1 & 0 & 2 \\ 2 & 1 & 4 \\ 1 & 1 & 1 \end{pmatrix}$. Let us find A^{-1}.

$$\left(\begin{array}{ccc|ccc} 1 & 0 & 2 & 1 & 0 & 0 \\ 2 & 1 & 4 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 \end{array}\right) \to \left(\begin{array}{ccc|ccc} 1 & 0 & 2 & 1 & 0 & 0 \\ 0 & 1 & 0 & -2 & 1 & 0 \\ 0 & 1 & -1 & -1 & 0 & 1 \end{array}\right) \to \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 3 & -2 & 2 \\ 0 & 1 & 0 & -2 & 1 & 0 \\ 0 & 0 & -1 & 1 & -1 & 1 \end{array}\right) -$$

$$\to \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 3 & -2 & 2 \\ 0 & 1 & 0 & -2 & 1 & 0 \\ 0 & 0 & -1 & 1 & -1 & 1 \end{array}\right).$$

i.e. $A^{-1} = \begin{pmatrix} 3 & -2 & 2 \\ -2 & 1 & 0 \\ -1 & 1 & -1 \end{pmatrix}$. Then

$$X = \begin{pmatrix} 3 & -2 & 2 \\ -2 & 1 & 0 \\ -1 & 1 & -1 \end{pmatrix}\begin{pmatrix} 1 & 3 & 5 \\ 4 & 7 & 11 \\ 3 & 3 & 3 \end{pmatrix} = \begin{pmatrix} 1 & 1 & -1 \\ 2 & 1 & 1 \\ 0 & 1 & 3 \end{pmatrix}.$$

3. This equation has a form $XA = B$. Its solution is $X = BA^{-1}$. Here $A = \begin{pmatrix} 1 & 2 & -2 \\ 2 & 5 & -4 \\ -2 & -4 & 5 \end{pmatrix}$. Let us find A^{-1}.

$$\left(\begin{array}{ccc|ccc} 1 & 2 & -2 & 1 & 0 & 0 \\ 2 & 5 & -4 & 0 & 1 & 0 \\ -2 & -4 & 5 & 0 & 0 & 1 \end{array}\right) \to \left(\begin{array}{ccc|ccc} 1 & 2 & -2 & 1 & 0 & 0 \\ 0 & 1 & 0 & -2 & 1 & 0 \\ 0 & 0 & 1 & 2 & 0 & 1 \end{array}\right) \to$$

$$\to \left(\begin{array}{ccc|ccc} 1 & 0 & -2 & 5 & -2 & 0 \\ 0 & 1 & 0 & -2 & 1 & 0 \\ 0 & 0 & 1 & 2 & 0 & 1 \end{array}\right) \to \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 9 & -2 & 2 \\ 0 & 1 & 0 & -2 & 1 & 0 \\ 0 & 0 & 1 & 2 & 0 & 1 \end{array}\right),$$

i.e. $A^{-1} = \begin{pmatrix} 9 & -2 & 2 \\ -2 & 1 & 0 \\ 2 & 0 & 1 \end{pmatrix}$. Then

$$X = \begin{pmatrix} -1 & -2 & 3 \\ 1 & 4 & -1 \\ 5 & 11 & -11 \end{pmatrix} \begin{pmatrix} 9 & -2 & 2 \\ -2 & 1 & 0 \\ 2 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 \\ -1 & 2 & 1 \\ 1 & 1 & -1 \end{pmatrix}.$$

***Example* 23.9.** Solve the matrix equation

$$\begin{pmatrix} 1 & 2 & -2 \\ 2 & 5 & -4 \\ -1 & 4 & 5 \end{pmatrix} X \begin{pmatrix} 1 & 0 & 4 \\ 1 & 1 & 5 \\ 0 & 2 & 3 \end{pmatrix} = \begin{pmatrix} 6 & 1 & 24 \\ 15 & 4 & 62 \\ 11 & 1 & -40 \end{pmatrix}.$$

Solution. This equation has a form $AXB = C$. It is possible, for example, in the following order: let us find A^{-1}, multiply this matrix on the left by both sides of the equation. We obtain $XB = A^{-1}C$. Then we obtain B^{-1} and multiplying it on the right side of the resulting equality we find $X = A^{-1}CB^{-1}$. We can solve this problem in reverse order. So,

$$A = \begin{pmatrix} 1 & 2 & -2 \\ 2 & 5 & -4 \\ -1 & 4 & 5 \end{pmatrix},$$

Matrix A^{-1} has already been found in example 2.8:

$$A^{-1} = \begin{pmatrix} 9 & -2 & 2 \\ -2 & 1 & 0 \\ 2 & 0 & 1 \end{pmatrix}$$

$$XB = \begin{pmatrix} 9 & -2 & 2 \\ -2 & 1 & 0 \\ 2 & 0 & 1 \end{pmatrix} \begin{pmatrix} 6 & 1 & 24 \\ 15 & 4 & 62 \\ 11 & 1 & -40 \end{pmatrix} = \begin{pmatrix} 2 & 3 & 12 \\ 3 & 2 & 14 \\ 1 & 3 & 8 \end{pmatrix},$$

$B=\begin{pmatrix}1&0&4\\1&1&5\\0&2&3\end{pmatrix}$. Let us find B^{-1}:

$$\left(\begin{array}{ccc|ccc}1&0&4&1&0&0\\1&1&5&0&1&0\\0&2&3&0&0&1\end{array}\right)\to\left(\begin{array}{ccc|ccc}1&0&4&1&0&0\\0&1&1&-1&1&0\\0&2&3&0&0&1\end{array}\right)\to\left(\begin{array}{ccc|ccc}1&0&4&1&0&0\\0&1&1&-1&1&0\\0&0&1&0&-2&1\end{array}\right)\to$$

$$\to\left(\begin{array}{ccc|ccc}1&0&0&-7&8&-4\\0&1&0&-3&3&-1\\0&0&1&2&-2&1\end{array}\right),$$

i.e. $B^{-1}=\begin{pmatrix}-7&8&-4\\-3&3&-1\\2&-2&1\end{pmatrix}$. Then

$$X=\begin{pmatrix}2&3&12\\3&2&14\\1&3&8\end{pmatrix}\begin{pmatrix}-7&8&-4\\-3&3&-1\\2&-2&1\end{pmatrix}=\begin{pmatrix}1&1&1\\1&2&0\\0&1&1\end{pmatrix}.$$

Questions

1. Where is element a_{52} located in matrix $A=(a_{ij})$?
2. Can matrix consist of а) one row; б) one column; в) one row and one column?
3. Can any element a_{ii} of the diagonal matrix be equal to zero?
4. Can two matrices be equal if one of them is of the third-order and the other is of the fourth-order?

5. Is it possible to find a sum of two matrices if one of them has dimension 3·4 and another one has dimension 4·3?

6. Does product *AB* exist if matrix *A* has dimension 3·4 and matrix B has dimension 3·4? Does product *BA* exist?

7. Is it possible to find a product of two matrices if one of them is a square matrix and the other is not a square one?

8. Let products AB and BA exist for matrices A and B. Is it possible to claim that matrices A and B have the same dimensions?

9. Can the product of two nonzero matrices be a zero matrix?

10. What is a square of a matrix? Can the square of a nonzero matrix be a zero matrix?

11. Does an inverse matrix A^{-1} exist for diagonal matrix $A=\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 4 \end{pmatrix}$? If A^{-1} exists, then what is its form?

Chapter 24. Determinants

24.1. Basic concepts

There is a rule according to which each square matrix is assigned a *number* characterizing this matrix.

Let us consider a matrix of the second order:

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}. \quad (24.1)$$

Number $a_{11}a_{22} - a_{12}a_{21}$ is a **determinant** of matrix (24.1) and is written in the form:

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}$$

(this is a determinant of the second order).

So,

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}. \quad (24.2)$$

For example,

$$\begin{vmatrix} 2 & 3 \\ 4 & 5 \end{vmatrix} = 2 \cdot 5 - 3 \cdot 4 = -2.$$

The concept of a determinant is associated, in particular, with the solution of the systems of linear equations. Let us consider a system:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 = b_1, \\ a_{21}x_1 + a_{22}x_2 = b_2. \end{cases}$$

To exclude x_2, let us multiply the first equation by a_{22} and the second one by a_{12} and from the first equation subtract the second one:
$(a_{11}a_{22} - a_{12}a_{21}) \cdot x_1 = b_1a_{22} - b_2a_{12}$, or $dx_1 = d_1$,

where $d = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}$, $d_1 = \begin{vmatrix} b_1 & a_{12} \\ b_2 & a_{22} \end{vmatrix}$.

If $d \neq 0$, then we obtain $x_1 = \dfrac{d_1}{d}$. Similarly, we obtain

$dx_2 = d_2$, $x_2 = \dfrac{d_2}{d}$,

where $d_2 = \begin{vmatrix} a_{11} & b_1 \\ a_{21} & b_2 \end{vmatrix}$.

Application of formulas $x_1 = \dfrac{d_1}{d}$ and $x_2 = \dfrac{d_2}{d}$ for the solution of systems of equations is called the **Cramer's rule.**

***Example* 24.1.** Solve the system

$$\begin{cases} 2x_1 + 3x_2 = 7, \\ x_1 + 4x_2 = 6. \end{cases}$$

Solution:

$$d = \begin{vmatrix} 2 & 3 \\ 1 & 4 \end{vmatrix} = 2 \cdot 4 - 3 \cdot 1 = 5, \quad d_1 = \begin{vmatrix} 7 & 3 \\ 6 & 4 \end{vmatrix} = 7 \cdot 4 - 3 \cdot 6 = 10,$$

$$d_2 = \begin{vmatrix} 2 & 7 \\ 1 & 6 \end{vmatrix} = 2 \cdot 6 - 7 \cdot 1 = 5,$$

$$x_1 = \frac{d_1}{d} = \frac{10}{5} = 2, \quad x_2 = \frac{d_2}{d} = \frac{5}{5} = 1.$$

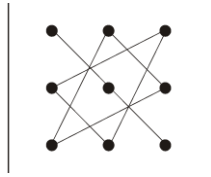
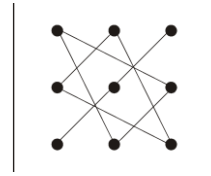
A similar rule for systems with any number of unknowns will be considered later.

For a square matrix of the third order a determinant of the third order is a number defined by the formula:

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31} - a_{12}a_{21}a_{33} - a_{11}a_{23}a_{32}. \quad (24.3)$$

A determinant of the third order is an algebraic sum of six products of elements, taken one from each row and each column.

These products are *terms of determinant.* Formula (24.3) can be schematically depicted as follows:

"+"  "–" 

With a **plus** sign, we take products whose factors are on the *main* diagonal and at the vertices of isosceles triangles with the bases parallel to the main diagonal; with a minus sign - on the minor diagonal and at the vertices of the isosceles triangles with the bases parallel to the minor diagonal.

We turn to the definition of the concept of a determinant of any order. For this purpose, we need some preliminary concepts.

Let us consider the natural numbers from 1 to *n*. These n numbers can be written in one order or another.

Any arrangement of numbers *1,2, ..., n* is called **permutation.**

For example,

3, 1, 5, 4, 2 (24.4)

is a permutation of five numbers.

It is easy to prove that the number of different permutations of *n* numbers is equal to the product of $1 \cdot 2 \cdot \ldots \cdot n$, denoted as $n!$ (factorial of *n*).

Let us consider an arbitrary permutation $\alpha_1, \alpha_2, ..., \alpha_n$, from the first *n* natural numbers. Let us choose two arbitrary numbers α_i and α_j from this permutation. If $i < j$, but $\alpha_i > \alpha_j$, then the numbers α_i and α_j are said to form an **inversion**. (In other words, if in the permutation a larger number precedes a smaller one, then these two numbers form an inversion.)

In particular, in permutation (24.4) numbers $\alpha_3 = 5$ and $\alpha_5 = 2$ form an inversion. If for $i < j$ inequality $\alpha_i < \alpha_j$ holds (i.e. a smaller number precedes a larger one), then the numbers α_i and α_j don't form an inversion. For example, in permutation (24.4) numbers $\alpha_2 = 1$ and $\alpha_4 = 4$ don't form an inversion.

A permutation is called **even** if it has an even number of inversions. Otherwise a permutation is called **odd**.

Let us consider a square matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & ... & a_{1n} \\ a_{21} & a_{22} & ... & a_{2n} \\ .... & .... & ... & .... \\ a_{n1} & a_{n2} & ... & a_{nn} \end{pmatrix}.$$

We compose some product of its *n* elements taken one from each row and each column:

$$a_{1\alpha_1} a_{2\alpha_2} ... a_{n\alpha_n} . \quad (24.5)$$

In product (24.5) factors are written in ascending order of their first indices - rows numbers. The second indices, column numbers, form permutations

$\alpha_1, \alpha_2, ..., \alpha_n$. Product (24.5) is called a *term of a determinant* of matrix *A*. Let us define a **sign rule:** if the second indices form an even permutation, product (24.5) is taken with the *plus* sign; if they form an odd permutation, the product is taken with the *minus* sign. There are $n!$ such permutations (24.5) — as many different permutations form their second indices.

Definition. **A determinant** of a square matrix of order n (a determinant of order n) is an algebraic sum of $n!$ terms, each of which is a product of *n* matrix elements taken one from each row and each column, in accordance with the signs rule.

A determinant of order *n* is denoted as:

$$d = |A| = \begin{vmatrix} a_{11} & a_{12} & ... & a_{1n} \\ a_{21} & a_{22} & ... & a_{2n} \\ .... & .... & ... & .... \\ a_{n1} & a_{n2} & ... & a_{nn} \end{vmatrix}. \quad (24.6)$$

Further, we will talk about elements, rows and columns of the determinant, referring to the elements, rows and columns of the corresponding matrix.

It should be noted that it is difficult to calculate the determinant of order *n* for $n > 3$ based directly on the definition. (For example, to find the determinant of order six it is necessary to calculate a sum of 720 terms, each of which is a product of six elements.)

So, to calculate a determinant, it is previously simplified by transforming considering its properties.

24.2. Properties of determinants

1. Transpose does not change a determinant: $|A| = |A'|$.

Property 1 implies that any statement about the rows of determinant is true for columns and vice versa. In this sense, the rows and columns of the determinant are **equal**. That is why we will formulate properties for rows, meaning that they also hold for columns.

2. If one of the rows of the determinant consists of zeroes, then the determinant is equal to zero.

This property is easy to prove. Let all the elements of the *i-th* row of determinant be equal to zero. Each element of the determinant contains a factor which is an element of this row. That's why every term of determinant is equal to zero. Thus, the determinant is equal to zero.

3. When two rows are replaced, the determinant changes its sign.

(In other words, when two rows in matrix *A* are replaced, we obtain matrix *B* such as $|A| = -|B|$).

4. Determinant which contains two identical rows is equal to zero.

To prove this property let us swap these two identical rows. The determinant will not change but, according to property 3, it will change the sign: $d = -d$. Thus, $d = 0$.

5. If all the elements of any row are multiplied by number *k, then the determinant will multiply by its number k.*

Indeed, in this case, every term of the determinant will multiply by number *k*, thus, the determinant will multiply by its number.

From property 5 follows that the common factor of any row of determinant can be taken out of the determinant sign.

6. A determinant which consists of two proportional rows is equal to zero.

Let us prove this statement.

Let the *i-th* and *j-th* rows be proportional: elements of the *j-th* row are obtained by multiplying the elements of the *i-th* row by number *k*. We take out *k* by the sign of the determinant and obtain a determinant which contains two identical rows. According to property 4 it is equal to zero.

7. If elements of one row of determinant d have a form

$a_{ik} = b_{ik} + c_{ik}$ $(k = 1, 2, \ldots, n)$,

then the determinant is equal to the sum of two determinants d_1 and d_2, in which all the rows, except this one, coincide with the corresponding rows of determinant d. Moreover, at the place of this row, determinant d_1 contains a row which consists of elements b_{ik}, and determinant d_2 contains a row which consists of elements c_{ik} (k = 1, 2, …, n):

$$d = \begin{vmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ \ldots & \ldots & \ldots & \ldots \\ b_{i1} + c_{i1} & b_{i2} + c_{i2} & \ldots & b_{in} + c_{in} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{vmatrix} = d_1 + d_2 =$$

$$= \begin{vmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ \ldots & \ldots & \ldots & \ldots \\ b_{i1} & b_{i2} & \ldots & b_{in} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{vmatrix} + \begin{vmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ \ldots & \ldots & \ldots & \ldots \\ c_{i1} & c_{i2} & \ldots & c_{in} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{vmatrix}.$$

This statement follows easily from the fact that each term of determinant *d* can be represented as a sum of two terms, one of which is a term of determinant d_1 and another one is a term of determinant d_2.

8. If one of the rows of the determinant is a linear combination of the other two rows, then the determinant is equal to zero.

Property 8 is a generalization of property 6.

9. The determinant does not change its sign if the corresponding element of another row multiplied by the same number is added to the elements of any of its rows.

This property is a consequence of properties 4-7.

10. The determinant of a multiplication of two square matrices is equal to multiplication of their determinants:

$|AB| = |A| \cdot |B|$.

In conclusion, we recall once again that all the statements formulated here for the rows of determinant remain true for its columns. (This refers to the columns of the corresponding matrix.)

24.3. Minors and algebraic adjuncts

Let us consider determinant of order *n.* Select an element a_{ij} and cross out the *i-th* row and *j-th* column at the intersection of which this element is located. We obtain a determinant of order ($n-1$) which is called **minor** M_{ij} **of element** a_{ij}.

For example, let us take determinant of order 4:

$$d = \begin{vmatrix} 1 & 0 & 2 & 1 \\ 3 & 1 & 2 & 5 \\ 0 & 2 & 1 & 3 \\ 1 & 0 & 4 & 3 \end{vmatrix}.$$

Minor M_{23} of element a_{23} is obtained by crossing out the second row and the third column at the intersection of which element a_{23} = 2 is placed:

$$M_{23} = \begin{vmatrix} 1 & 0 & 1 \\ 0 & 2 & 3 \\ 1 & 0 & 3 \end{vmatrix} = 4$$

.

Definition. **An algebraic adjunct** of element a_{ij} of determinant (24.6) is number:

$$A_{ij} = (-1)^{i+j} M_{ij}$$

.

In particular, in the above example the algebraic adjunct is:

$$A_{23} = (-1)^5 \cdot 4 = -4$$

.

Minors and algebraic adjuncts play an important role in linear algebra and its applications. One of such applications is the following statement.

Theorem **24.1.** The determinant is equal to the sum of products of any of its rows by its algebraic adjuncts:

$$d = a_{i1}A_{i1} + a_{i2}A_{i2} + ... + a_{in}A_{in}$$

.(24.7)

(We accept this theorem without proof.)

Formula (24.7) is called **decomposition** of a determinant by the *i*-th row. Analogical statement holds for decomposition of a determinant by any column. Formula (24.7) reduces the calculation of the determinant of order n to calculation of *n* determinants of order ($n-1$).

Remark. The sum of pairwise products of the i-th row (column) of the determinant by the corresponding algebraic adjuncts of the *j-th* row (column) for $i \neq j$ is equal to zero, i.e.

$$a_{i1}A_{j1} + a_{i2}A_{j2} + ... + a_{in}A_{jn} = 0$$

$$a_{1i}A_{1j} + a_{2i}A_{2j} + ... + a_{ni}A_{nj} = 0$$

for $i \neq j$.

Let us prove, for example, the last of these two equalities. We decompose determinant

$$d = \begin{vmatrix} a_{11} & \dots & a_{1i} & \dots & a_{1j} & \dots & a_{1n} \\ a_{21} & \dots & a_{2i} & \dots & a_{2j} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ a_{n1} & \dots & a_{ni} & \dots & a_{nj} & \dots & a_{nn} \end{vmatrix}$$

by the *j*-th column:

$$d = a_{1j}A_{1j} + a_{2j}A_{2j} + \dots + a_{nj}A_{nj}.$$

Now we replace the elements of *j-th* column by the elements of *i-th* column (leaving the *i-th* column unchanged). We obtain determinant

$$d' = \begin{vmatrix} a_{11} & \dots & a_{1i} & \dots & a_{1i} & \dots & a_{1n} \\ a_{21} & \dots & a_{2i} & \dots & a_{2i} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ a_{n1} & \dots & a_{ni} & \dots & a_{ni} & \dots & a_{nn} \end{vmatrix},$$

which contains two similar columns at the *i-th* and *j-th* places and, is obviously equal to zero: $d' = 0$. But its decomposition by the *j-th* column has a form:

$$d' = a_{1i}A_{1j} + a_{2i}A_{2j} + \dots + a_{ni}A_{nj}.$$

Thus,

$$a_{1i}A_{1j} + a_{2i}A_{2j} + \dots + a_{ni}A_{nj} = 0,$$

q.e.d.

Usually, a determinant is preliminarily transformed before calculation according to its properties. Usually, it is reduced to a triangular form since the following **statement** holds:

If all the elements of the determinant located on one side of the main diagonal are equal to zero, then this determinant is equal to the product of the element placed on the main diagonal.

We prove this statement using the method of mathematical induction. For the determinant of the second-order, this statement is obvious. Assume that it holds for determinant of the $(n-1)$-th order and consider determinant of *n*-th order:

$$d = \begin{vmatrix} a_{11} & a_{12} & a_{13} & \ldots & a_{1n} \\ 0 & a_{22} & a_{23} & \ldots & a_{2n} \\ 0 & 0 & a_{33} & \ldots & a_{3n} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & a_{nn} \end{vmatrix}.$$

We decompose it by the first column:

$$d = a_{11} \begin{vmatrix} a_{22} & a_{23} & \ldots & a_{2n} \\ 0 & a_{33} & \ldots & a_{3n} \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & a_{nn} \end{vmatrix}.$$

On the right side of the obtained equality is the determinant of the $(n-1)$-th order. For this determinant the following equality holds

$$\begin{vmatrix} a_{22} & a_{23} & \ldots & a_{2n} \\ 0 & a_{33} & \ldots & a_{3n} \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & a_{nn} \end{vmatrix} = a_{22}a_{33}\cdots a_{nn}.$$

Thus, $d = a_{11}a_{22}a_{33}\cdots a_{nn}$.

***Example* 24.2.** Calculate the following determinants:

1) $d_1 = \begin{vmatrix} 1 & 2 & 4 & 3 \\ 0 & 3 & 1 & 0 \\ 0 & 2 & 0 & 0 \\ 1 & 3 & 2 & 1 \end{vmatrix}$; 2) $d_2 = \begin{vmatrix} 1 & 2 & 3 & 2 \\ 1 & 3 & 4 & 5 \\ 1 & 2 & 5 & 3 \\ 1 & 2 & 7 & 8 \end{vmatrix}$.

Solution. 1. It is possible to decompose determinant d_1 by any number. However, the shortest calculation is obtained by decomposition by the row with the largest number of zeroes. We decompose d_1 by the third row and then A_{32} by the second row:

$$d_1 = -2 \cdot \begin{vmatrix} 1 & 4 & 3 \\ 0 & 1 & 0 \\ 1 & 2 & 1 \end{vmatrix} = -2 \cdot 1 \cdot \begin{vmatrix} 1 & 3 \\ 1 & 1 \end{vmatrix} = 4.$$

2. Subtract the first row of determinant d_2 from all the others and then subtract the doubled third row from the fourth:

$$d_2 = \begin{vmatrix} 1 & 2 & 3 & 2 \\ 0 & 1 & 1 & 3 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 4 & 6 \end{vmatrix} = \begin{vmatrix} 1 & 2 & 3 & 2 \\ 0 & 1 & 1 & 3 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 4 \end{vmatrix} = 1 \cdot 1 \cdot 2 \cdot 4 = 8.$$

We used the fact that the obtained triangular determinant is equal to the product of the elements of the main diagonal.

24.4. Application of determinants

Definition. Square matrix A is **nondegenerate** if its determinant is not equal to zero: $|A| \neq 0$. Otherwise, the matrix is called **degenerate**.

Let us note that this definition is obviously equal to the definition given above.

***Theorem* 24.2.** Inverse matrix A^{-1} for matrix *A* exists if and only if matrix *A* is nondegenerate.

Proof. Necessity. Let matrix A have inverse matrix A^{-1}. Then $A^{-1}A = AA^{-1} = E$. So as $|E| = 1 \neq 0$ and the determinant of matrix product is equal to the product of their determinants, then $|A^{-1}A| = |A| \cdot |A^{-1}| = |E| \neq 0$, thus, $|A^{-1}| \neq 0$ and $|A| \neq 0$.

2.*Sufficiency.* Given a nondegenerate matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}$$

and its determinant $|A| = d \neq 0$.

We transpose matrix *A* and then replace its elements by its algebraic adjuncts:

$$A^* = \begin{pmatrix} A_{11} & A_{21} & \dots & A_{n1} \\ A_{12} & A_{22} & \dots & A_{n2} \\ \dots & \dots & \dots & \dots \\ A_{1n} & A_{2n} & \dots & A_{nn} \end{pmatrix}. \quad (24.8)$$

Matrix A^* is called the **adjugate** matrix of matrix *A*.

Let us find product AA^*. Given the decomposition (24.7) and the following remark, we obtain

$$AA^* = \begin{pmatrix} d & 0 & ... & 0 \\ 0 & d & ... & 0 \\ .... & .... & ... & .... \\ 0 & 0 & ... & d \end{pmatrix} = dE \quad . \qquad (24.9)$$

(Write the multiplication AA^* in detail and make sure in equality (24.9) by yourself).

It is also easy to make sure of $AA^* = A^* A$.

So, $AA^* = A^* A = dE$. Hence

$$A\frac{1}{d}A^* = \frac{1}{d}A^* A = E \quad .$$

Thus,

$$A^{-1} = \frac{1}{d}A^* \quad ,$$

or, more particularly

$$A^{-1} = \frac{1}{d} \cdot \begin{pmatrix} A_{11} & A_{21} & ... & A_{n1} \\ A_{12} & A_{22} & ... & A_{n2} \\ .... & .... & ... & .... \\ A_{1n} & A_{2n} & ... & A_{nn} \end{pmatrix} . \qquad (24.10)$$

The theorem is proved.

Calculation of an inverse matrix by the adjugate matrix method.

1. Let us find the determinant of the initial matrix $d = |A|$. If $d = 0$, i.e. matrix A is degenerate, then an inverse matrix does not exist. If $d \neq 0$, we continue the process.

2. We find the algebraic adjuncts of elements of matrix *A* and form an adjugate matrix A^* of its elements.

3. We find an inverse matrix by formula (24.10). In some cases it is useful to check up, i.e. to check multiplications $A^{-1}A$ and AA^{-1} (or one of them) and make sure that we get the identity matrix *E*.

***Example* 24.3.** Find an inverse matrix of matrix *A*:

$$A=\begin{pmatrix}1 & 2 & 3\\ 2 & 3 & 1\\ 1 & 2 & 1\end{pmatrix}.$$

Solution. 1. Let us calculate the determinant

$$d=|A|=(3-2)-2(2-1)+3(4-3)=2.$$

2. Let us find the algebraic adjuncts:

$$A_{11}=(-1)^{1+1}\begin{vmatrix}3 & 1\\ 2 & 1\end{vmatrix}=1, \qquad A_{21}=-\begin{vmatrix}2 & 3\\ 2 & 1\end{vmatrix}=4, \qquad A_{31}=-7,$$

$$A_{12}=(-1)^{1+2}\begin{vmatrix}2 & 1\\ 1 & 1\end{vmatrix}=-1, \qquad A_{22}=\begin{vmatrix}1 & 3\\ 1 & 1\end{vmatrix}=-2, \qquad A_{32}=5,$$

$$A_{13}=(-1)^{1+3}\begin{vmatrix}2 & 3\\ 1 & 2\end{vmatrix}=1, \qquad A_{23}=-\begin{vmatrix}1 & 2\\ 1 & 2\end{vmatrix}=0, \qquad A_{32}=-1.$$

We form an adjugate matrix:

$$A^{*}=\begin{pmatrix}1 & 4 & -7\\ -1 & -2 & 5\\ 1 & 0 & -1\end{pmatrix}.$$

3. We calculate an inverse matrix

$$A^{-1}=\frac{1}{2}\begin{pmatrix}1 & 4 & -7\\ -1 & -2 & 5\\ 1 & 0 & -1\end{pmatrix}=\begin{pmatrix}1/2 & 2 & -7/2\\ -1/2 & -1 & 5/2\\ 1/2 & 0 & -1/2\end{pmatrix}.$$

To test yourself, make sure that $A^{-1}A=AA^{-1}=E$.

Let us note that for calculation of inverse matrices of the higher dimension matrices we use another method - the method of elementary transformations. (see. § 23.4).

24.5. Matrix rank

Let *A* be a matrix with dimension $m\times n$. Pick up *k* rows and *k* columns in an arbitrary way. Elements placed at the intersection of the selected rows and columns form a square matrix of order k; its determinant is called a minor of order k of matrix A. Herewith, obviously, $k\le\min(m,n)$.

Definition. The highest order of minors of matrix A, which are not equal to zero, is called the rank of matrix A.

***Example* 24.4.** Calculate the rank of matrix

$$A=\begin{pmatrix}2 & 3 & 4 & 5 & 6\\ 1 & 2 & 1 & 2 & 1\\ 0 & 0 & 0 & 0 & 0\\ 3 & 6 & 3 & 6 & 3\end{pmatrix}.$$

Solution. It is easy to check that the rank of matrix A is equal to two: $\operatorname{rg} A=2$. Indeed, the second-order minor placed at the upper left corner is not equal to zero:

$$\begin{vmatrix}2 & 3\\ 1 & 2\end{vmatrix}=1\neq 0,$$

however, each of the third-order minors contains either a zero line or two proportional lines and, thus, is equal to zero.

In § 23.4 we defined the matrix rank as the maximal number of its linearly independent vectors.

In this regard, it makes sense to define the matrix rank as the maximal number of its linearly independent rows. This definition is equivalent to the previous one. It is possible to prove (it is done in algebra course) that the maximal number of linearly independent matrix rows is equal to the maximal number of its linearly independent columns and also to the maximal order of the nonzero minors.

Here are the main methods for calculating the rank of a matrix.

1. Bordering minors method. Let a nonzero minor M of order k be found in matrix A. We consider minors of order $k+1$ which contain minor M. If all of them are equal to zero, then the matrix rank is equal to k. Otherwise, the procedure continues.

Example 24.5. Find the matrix rank

$$A=\begin{pmatrix} 2 & 3 & 0 & 2 & 4 \\ 5 & 8 & 0 & 2 & 5 \\ 0 & 0 & 2 & 4 & 6 \\ 0 & 0 & 1 & 2 & 3 \end{pmatrix}.$$

Solution. Let us fix a nonzero second-order minor:

$$M_2=\begin{vmatrix} 2 & 3 \\ 5 & 8 \end{vmatrix}=1\neq 0.$$

One of the bordering third-order minors is also nonzero:

$$M_3=\begin{vmatrix} 2 & 3 & 0 \\ 5 & 8 & 0 \\ 0 & 0 & 2 \end{vmatrix}=2\neq 0.$$

However, both fourth-order minors bordering M_3 are equal to zero since each of them has the proportional third and fourth rows.

Thus, the matrix rank is equal to three: $rg\ A = 3$.

2. The method of elementary transformations. Elementary transformations do not change the matrix rank. Using elementary transformations, we can bring the matrix to such a form when all the elements except a_{11}, a_{22}, …, a_{rr} are zero. The number of nonzero elements of the transformed matrix is obviously equal to the matrix rank.

***Example* 24.6.** Find the matrix rank

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 5 & 7 \\ 0 & -1 & -1 \\ 1 & 0 & 1 \end{pmatrix}.$$

Solution:

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 5 & 7 \\ 0 & -1 & -1 \\ 1 & 0 & 1 \end{pmatrix} \to \begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & 1 \\ 0 & -1 & -1 \\ 0 & -2 & -2 \end{pmatrix} \to \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \to \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

The rank of the transformed matrix is equal to two, thus, $\text{rg}\ A = 2$.

Questions

1. Under which conditions is the determinant of a second-order matrix equal to zero?

2. With which sign does the term $a_{11}a_{23}a_{34}a_{42}$ enter the determinant of the four-order matrix?

3. Can the multiplication $a_{12}a_{23}a_{32}a_{41}a_{54}$, taken with the appropriate sign, be the term of determinant of a five-order matrix?

4. What is the difference between minor M_{54} and algebraic adjunct A_{54}?

5. Let matrix *A* contain a five-order minor which is not equal to zero. What can be concluded about the matrix rank?

6. What is the sum of the products of the elements of a row of the matrix by the algebraic component of the elements of another row of this matrix?

7. Is it possible to calculate the determinant of the product of two square matrices without multiplying these matrices?

8. What is the determinant of a triangular matrix?

9. Which method for calculating the inverse seventh-order matrix is preferable: the adjoint matrix method or the elementary transformations method?

10. Can the rank of matrix A with dimension 7·3 be equal to four?

Chapter 25. Systems of linear equations

25.1. Basic concepts

The system of m linear equations with n unknowns has the form:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n = b_m. \end{cases} \quad (25.1)$$

The matrix composed of the coefficients of the equations of system (25.1), i.e.

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots. & \ldots. & \ldots & \ldots. \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{pmatrix},$$

called **the matrix of the system**.

If we denote by X the matrix column of unknowns, and by B the column matrix of free terms:

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \ldots \\ x_n \end{pmatrix}, \qquad B = \begin{pmatrix} b_1 \\ b_2 \\ \ldots \\ b_m \end{pmatrix},$$

then system (4.1) can be written in the form of a single matrix equation: $AX = B$.

Adding columns of free terms to matrix A, we obtain **an expanded matrix of system** (25.1):

$$\overline{A} = \left(\begin{array}{cccc|c} a_{11} & a_{12} & ... & a_{1n} & b_1 \\ a_{21} & a_{22} & ... & a_{2n} & b_2 \\ .... & .... & ... & .... & ... \\ a_{m1} & a_{m2} & ... & a_{mn} & b_m \end{array}\right).$$

(Usually a column of free members is separated by a vertical bar).

The expanded matrix contains all the information about the system.

A solution of the system (25.1) is a set of numbers

$$x_1 = \alpha_1, x_2 = \alpha_2, ..., x_n = \alpha_n,$$

when substituted into this system, all equations turn into identities.

A system of equations is called **compatible** if it has at least one solution. A system that does not have a single solution is called **incompatible**. A compatible system having a unique solution is called **definite**. If the system has more than one solution, then it is called **indefinite**.

Solving a system means finding many of its solutions. The set of all solutions of the system is called its **general solution**.

Two systems are called **equivalent** if they have the same set of solutions, or, that is equal, the same general solution.

Usually, in order to solve a system, it is first transformed. Moreover, the transformed system should be equivalent to the original.

We list the elementary transformations of system (25.1):

• permutation of equations;

• multiplication of both parts of one equation by any number other than zero;

• adding to both sides of one of the equations of the system the corresponding parts of the other equation multiplied by the same number;

• crossing out equations of the form $0 \cdot x_1 + 0 \cdot x_2 + ... + 0 \cdot x_n = 0$.

As a result of elementary transformations, a system equivalent to the original one is obtained.

25.2. Methods of solving systems of linear equations

1. Gauss method. This is the most convenient method for solving systems of the form (25.1). Let us state its essence.

Suppose for definiteness in system (25.1) $a_{11} \neq 0$ (if $a_{11} = 0$, (if, then we displace in the first place another equation with a nonzero first coefficient).

Multiply the first equation[1] by $-\frac{a_{21}}{a_{11}}$ and add to the second. Then we multiply the first equation by $-\frac{a_{31}}{a_{11}}$ and add to the third, etc. Finally, multiply the first equation by $-\frac{a_{m1}}{a_{11}}$ and add to the last one. As a result of these elementary transformations, we obtain a system that is equivalent to the original, but in the new system none of the equations except the first contains the unknown x_1:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n = b_1, \\ \quad a'_{22}x_2 + a'_{23}x_3 + \ldots + a'_{2n}x_n = b'_2, \\ \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \quad a'_{m2}x_2 + a'_{m3}x_3 + \ldots + a'_{mn}x_n = b'_m. \end{cases} \qquad (25.2)$$

[1] Speaking about the multiplication of the equation by a number, we, of course, mean the multiplication of all members of both sides of this equation by this number.

Note that only coefficients and free terms are transformed; therefore, it is more convenient to write the system transformation as a transformation of its extended matrix:

$$\overline{A}' = \left(\begin{array}{ccccc|c} a_{11} & a_{12} & a_{13} & \ldots & a_{1n} & b_1 \\ 0 & a'_{22} & a'_{23} & \ldots & a'_{2n} & b'_2 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & a'_{m2} & a'_{m3} & \ldots & a'_{mn} & b'_m \end{array}\right). \qquad (25.3)$$

At the second stage, using the *second equation*, we similarly transform all the equations, starting from the third, or, which is the same, multiplying the second row of the matrix $\overline{A}'$ by corresponding numbers ($-\frac{a'_{32}}{a'_{22}}$, …, $-\frac{a'_{m2}}{a'_{22}}$) and adding to the third, ..., m-th lines, we get:

$$\overline{A}'' = \left(\begin{array}{ccccc|c} a_{11} & a_{12} & a_{13} & \ldots & a_{1n} & b_1 \\ 0 & a'_{22} & a'_{23} & \ldots & a'_{2n} & b'_2 \\ 0 & 0 & a''_{33} & \ldots & a''_{3n} & b''_3 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & a''_{m3} & \ldots & a''_{mn} & b''_m \end{array}\right).$$

We continue this process in the same way: further, all lines except the first two will be converted, then except the first three, etc.

We did not investigate the compatibility system in advance. Nevertheless, the Gauss method allows one of the stages to establish the possibility of system incompatibility. Indeed, if, as a result of the transformations, we obtain a row in which all terms except the last are equal to zero and the last is non-zero, then this corresponds to an equation of the form:

$$0 \cdot x_1 + 0 \cdot x_2 + ... + 0 \cdot x_n = b \neq 0,$$

that has no solutions. Therefore, the system containing such an equation is incompatible.

In the process of applying the Gauss method, lines entirely consisting of zeros may also appear, which corresponds to equations of the form

$$0 \cdot x_1 + 0 \cdot x_2 + ... + 0 \cdot x_n = 0.$$

This can happen if the corresponding equations of the original system are linear combinations of other equations of the system.

If system (25.1) is defined, then its matrix as a result of transformations[1] will take the form:

$$\left(\begin{array}{ccccc|c} a_{11} & a_{12} & a_{13} & ... & a_{1n} & b_1 \\ 0 & a'_{22} & a'_{23} & ... & a'_{2n} & b'_2 \\ ... & ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & a_{nn}^{(n-1)} & b_n^{(n-1)} \end{array}\right),$$

I.e. the system will have a "triangular" look:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + ... + a_{1n}x_n = b_1, \\ \qquad a'_{22}x_2 + a'_{23}x_3 + ... + a'_{2n}x_n = b'_2, \\ \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \qquad\qquad\qquad\qquad a_{nn}^{(n-1)}x_n = b_n^{(n-1)}. \end{cases} \tag{25.4}$$

(superscripts and primes indicate how many times the coefficients and free terms have changed during the transformations).

[1] Note that the method of finding the inverse matrix (see § 23.4) is based on similar transformations.

From the last equation of the system (25.4) immediately find $x_n = \frac{b_n^{(n-1)}}{a_{nn}^{(n-1)}}$, then, substituting the found x_n in the penultimate equation (containing only x_n and x_{n-1}), find x_{n-1} and so on. Thus, we successively find all other unknowns. (This process is sometimes called the inverse of the Gauss method.)

Example* 25.1 Solve the system*:

$$\begin{cases} x_1 + x_2 + 2x_3 + x_4 = 3 \\ 2x_1 + 3x_2 + 5x_3 + 3x_4 = 9 \\ x_1 - x_2 + 2x_3 = -3 \\ 2x_1 + 2x_2 + 6x_3 + 4x_4 = 8 \end{cases}$$

Decision. We compose the extended matrix of the system and apply the Gauss method:

$$\left(\begin{array}{cccc|c} 1 & 1 & 2 & 1 & 3 \\ 2 & 3 & 5 & 3 & 9 \\ 1 & -1 & 2 & 0 & -3 \\ 2 & 2 & 6 & 4 & 8 \end{array}\right) \xrightarrow{l_2-2l_1,\ l_3-l_1,\ l_4-2l_1} \left(\begin{array}{cccc|c} 1 & 1 & 2 & 1 & 3 \\ 0 & 1 & 1 & 1 & 3 \\ 0 & -2 & 0 & -1 & -6 \\ 0 & 0 & 2 & 2 & 2 \end{array}\right) \rightarrow$$

$$\xrightarrow{l_3+2l_2} \left(\begin{array}{cccc|c} 1 & 1 & 2 & 1 & 3 \\ 0 & 1 & 1 & 1 & 3 \\ 0 & 0 & 2 & 1 & 0 \\ 0 & 0 & 2 & 2 & 2 \end{array}\right) \xrightarrow{l_4-l_3} \left(\begin{array}{cccc|c} 1 & 1 & 2 & 1 & 3 \\ 0 & 1 & 1 & 1 & 3 \\ 0 & 0 & 2 & 1 & 0 \\ 0 & 0 & 0 & 1 & 2 \end{array}\right).$$

The resulting expanded matrix corresponds to the system:

$$\begin{cases} x_1 + x_2 + 2x_3 + x_4 = 3 \\ \quad x_2 + \ x_3 + x_4 = 3 \\ \qquad 2x_3 + x_4 = 0 \\ \qquad\qquad x_4 = 2, \end{cases}$$

which is equal to the original system. Substitute $x_4 = 2$ to the penultimate equation, find $x_3 = -1$; substitute x_4 и x_3 to the second equation, find $x_2 = 2$; finally, substitute found x_4, x_3, x_2 to the first one find $x_1 = 1$. So, the system has a unique solution.:

$x_1 = 1$, $x_2 = 2$, $x_3 = -1$, $x_4 = 2$.

It is possible to solve such a system, leading it not to a triangular form, but turning it into the so-called allowed system. Let us illustrate this with an example, and then describe the process in a general way.

***Example* 25.2** Solve the system

$$\begin{cases} x_1 + +2x_2 + x_3 + 4x_4 = 7, \\ 3x_1 + 2x_2 + x_3 + \ x_4 = 3, \\ x_1 + x_2 + 2x_3 + 2x_4 = 4, \\ 2x_1 + 5x_2 + 3x_3 + 9x_4 = 15. \end{cases}$$

Solution. We compose an expanded matrix and transform it in such a way that each row and each column of the transformed matrix of the system contain one element equal to one, and the rest equal zero:

$$\left(\begin{array}{cccc|c} 1 & 2 & 1 & 4 & 7 \\ 3 & 2 & 1 & 1 & 3 \\ 1 & 1 & 2 & 2 & 4 \\ 2 & 5 & 3 & 9 & 15 \end{array}\right) \xrightarrow{l_2-3l_1,\ l_3-l_1,\ l_4-2l_1} \left(\begin{array}{cccc|c} 1 & 2 & 1 & 4 & 7 \\ 0 & -4 & -2 & -11 & -18 \\ 0 & -1 & 1 & -2 & -3 \\ 0 & 1 & 1 & 1 & 1 \end{array}\right) \rightarrow$$

$$\xrightarrow{l_3+l_4,\ l_2+4l_4,\ l_1-2l_4} \left(\begin{array}{cccc|c} 1 & 0 & -1 & 2 & 5 \\ 0 & 0 & 2 & -7 & -14 \\ 0 & 0 & 2 & -1 & -2 \\ 0 & 1 & 1 & 1 & 1 \end{array}\right) \rightarrow \left(\begin{array}{cccc|c} 1 & 0 & -1 & 2 & 5 \\ 0 & 0 & 2 & -7 & -14 \\ 0 & 0 & -2 & 1 & 2 \\ 0 & 1 & 1 & 1 & 1 \end{array}\right) \rightarrow$$

$$\xrightarrow{l_4-l_3,\ l_2+7l_3,\ l_1-2l_3} \left(\begin{array}{cccc|c} 1 & 0 & 3 & 0 & 1 \\ 0 & 0 & 12 & 0 & 0 \\ 0 & 0 & -2 & 1 & 2 \\ 0 & 1 & 3 & 0 & -1 \end{array}\right) \rightarrow \left(\begin{array}{cccc|c} 1 & 0 & 3 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -2 & 1 & 2 \\ 0 & 1 & 3 & 0 & -1 \end{array}\right) \rightarrow$$

$$\xrightarrow{l_1-3l_2,\ l_3+2l_2,\ l_4-3l_2} \left(\begin{array}{cccc|c} 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 2 \\ 0 & 1 & 0 & 0 & -1 \end{array}\right).$$

So, gotten system is:

$$\begin{cases} x_1 = 1, \\ x_3 = 0, \\ x_4 = 2, \\ x_2 = -1, \end{cases}$$

i.e. we got a solution $x_1 = 1$, $x_2 = -1$, $x_3 = 0$, $x_4 = 2$, or (1, –1, 0, 2). Obviously, system (25.1) can be reduced to the form (4.4) when the rank of its matrix coincides with the number of unknowns: $\operatorname{rg} A = n$.

In the case when the system is compatible and $\operatorname{rg} A = r < n$, the expanded matrix $\overline{A}$ is converted by the Gauss method to the form:

$$\left(\begin{array}{ccccccc} a_{11} & a_{12} & \ldots & a_{1r} & \ldots & a_{1n} & b_1 \\ 0 & a'_{22} & \ldots & a'_{2r} & \ldots & a'_{2n} & b'_2 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & a_{rr}^{(r-1)} & \ldots & a_{rn}^{(r-1)} & b_r^{(r-1)} \end{array}\right),$$

where $a_{11} \neq 0$, $a'_{22} \neq 0$, …, $a_{rr}^{(r-1)} \neq 0$. The corresponding system has a "trapezoidal" shape:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1r}x_r + \ldots + a_{1n}x_n = b_1, \\ \qquad a'_{22}x_2 + \ldots + a'_{2r}x_r + \ldots + a'_{2n}x_n = b'_2, \\ \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \qquad\qquad a_{rr}^{(r-1)}x_r + \ldots + a_{rn}^{(n-1)}x_n = b_r^{(r-1)}. \end{cases} \tag{25.5}$$

In this case, declare the unknown x_{r+1}, …, x_n free and move to the right side of the equations.

***Example* 25.3.** Solve the system.

$$\begin{cases} x_1 + x_2 - 2x_3 + x_4 = 1, \\ 2x_1 + 3x_2 - 3x_3 = 5, \\ x_1 + 4x_2 + 2x_3 - 6x_4 = 11, \\ x_1 - x_2 - 5x_3 + 6x_4 = -6. \end{cases}$$

Solution. Transform the extended matrix of the system:

$$\left(\begin{array}{cccc|c} 1 & 1 & -2 & 1 & 1 \\ 2 & 3 & -3 & 0 & 5 \\ 1 & 4 & 2 & -6 & 11 \\ 1 & -1 & -5 & 6 & -6 \end{array}\right) \xrightarrow{l_2-2l_1,\, l_3-l_1,\, l_4-l_1} \left(\begin{array}{cccc|c} 1 & 1 & -2 & 1 & 1 \\ 0 & 1 & 1 & -2 & 3 \\ 0 & 3 & 4 & -7 & 10 \\ 0 & -2 & -3 & 5 & -7 \end{array}\right) \rightarrow$$

$$\xrightarrow{l_3-3l_2,\, l_4+2l_2} \left(\begin{array}{cccc|c} 1 & 1 & -2 & 1 & 1 \\ 0 & 1 & 1 & -2 & 3 \\ 0 & 0 & 1 & -1 & 1 \\ 0 & 0 & -1 & 1 & -1 \end{array}\right) \xrightarrow{l_4+l_3} \left(\begin{array}{cccc|c} 1 & 1 & -2 & 1 & 1 \\ 0 & 1 & 1 & -2 & 3 \\ 0 & 0 & 1 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right).$$

Obtain the system:

$$\begin{cases} x_1 + x_2 - 2x_3 + x_4 = 1 \\ x_2 + x_3 - 2x_4 = 3 \\ x_3 - x_4 = 1. \end{cases}$$

Unknown x_4 declare as free: $x_4 = c$. From the last equation get: $x_3 = 1 + c$. Substitute this value to the second equation: $x_2 = 2 + c$. Substitute found x_3 и x_2 to the first equation: $x_1 = 1$.

The system has an infinite number of solutions:

$$\bar{x} = (1, 2 + c, 1 + c, c),$$

where c gets any numerical values. This is a general system solution.

We now consider the Gauss method in a slightly different form. The method consists of several steps. Suppose that the first *k-1* steps are taken, and describe the next *k-th* step.

1. We check if there is at least one *contradictory* equation in the system (obtained after the k-1 previous steps). If such an equation exists in the system, then it is incompatible - work with it stops.

2. If the system has *trivial* equations 0 = 0, then delete them.

3. Let there be no contradictory equations in the system. Then one of the equations is chosen as the *resolving equation*, and one of the unknowns is declared as *resolving unknowns*. The following conditions must be met:
1) in the previous steps this equation was not resolving;
2) in the resolving equation, the coefficient for the resolving unknown must be nonzero[1] (this coefficient is sometimes called the *resolving element*);
3) from all equations except the resolving one, we exclude the resolving unknown. For this, we add a resolving equation multiplied by the corresponding number to each of these equations.

After a finite number of steps, the process will stop, and either the incompatibility of the system will be established, or a general solution of this system will be obtained. This will happen when all the equations are in the role of resolving ones.

Let us look at some examples. As usual, we will transform not the systems themselves, but their extended matrices.

***Example* 25.4.** Find a general solution and one particular solution to a system of equations

$$\begin{cases} x_1 + x_2 - 4x_3 - x_4 + 3x_6 = -1, \\ 2x_1 + x_2 - 5x_3 + x_6 = 7, \\ 3x_1 + 4x_3 + x_5 - 3x_6 = -2, \\ x_1 - x_3 + x_4 - 2x_6 = 8. \end{cases}$$

Solution. We write the extended matrix of the system.

[1] Using elementary transformations, one can resolve the equation such that the coefficient for the unknown sought becomes equal to unity.

$$\left(\begin{array}{cccccc|c} 1 & 1 & -4 & -1 & 0 & 3 & -1 \\ 2 & 1 & -5 & 0 & 0 & 1 & 7 \\ 3 & 0 & 4 & 0 & 1 & -3 & -2 \\ 1 & 0 & -1 & 1 & 0 & -2 & 8 \end{array}\right).$$

Step 1. We make sure that this system does not contain contradictory and trivial equations. We select the first equation as the resolving equation, and the coefficient $a_{12}=1$ – as a resolving element. We do the transformation $l_2 - l_1$:

$$\left(\begin{array}{cccccc|c} 1 & 1 & -4 & -1 & 0 & 3 & -1 \\ 1 & 0 & -1 & 1 & 0 & -2 & 8 \\ 3 & 0 & 4 & 0 & 1 & -3 & -2 \\ 1 & 0 & -1 & 1 & 0 & -2 & 8 \end{array}\right).$$

Step 2. The matrix obtained after the first step is an extended matrix of a system that does not contain contradictory and trivial equations. We take the second equation as the resolving equation, and the coefficient as the resolving element $a_{24}=1$. We do the transformation $l_1 + l_2, l_4 - l_2$:

$$\left(\begin{array}{cccccc|c} 2 & 1 & -5 & 0 & 0 & 1 & 7 \\ 1 & 0 & -1 & 1 & 0 & -2 & 8 \\ 3 & 0 & 4 & 0 & 1 & -3 & -2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right).$$

Step 3. The system of equations obtained after the second step contains a trivial equation - delete it (delete the line consisting of zeros):

$$\left(\begin{array}{cccccc|c} 2 & 1 & -5 & 0 & 0 & 1 & 7 \\ 1 & 0 & -1 & 1 & 0 & -2 & 8 \\ 3 & 0 & 4 & 0 & 1 & -3 & -2 \end{array}\right).$$

In the third equation coefficient $a_{35}=1$. It can be taken as the resolving one, but the remaining equations no longer contain the resolving unknown x_5. Obtain the system:

$$\begin{cases} 2x_1+x_2-5x_3 \qquad\quad +x_6=7, \\ x_1 \quad -x_3+x_4 \quad -2x_6=8, \\ 3x_1 \quad +4x_3 \quad +x_5-3x_6=-2, \end{cases}$$

has unknowns x_2, x_4, x_5 expressed through free unknowns x_1, x_3, x_6:

$$\begin{cases} x_2= \ 7-2x_1+5x_3 \ -x_6, \\ x_4= \ 8 \ -x_1 \ +x_3+2x_6, \\ x_5=-2-3x_1-4x_3+3x_6. \end{cases}$$

Consider $x_1=c_1,\ x_3=c_2,\ x_6=c_3$, get a solution
$\bar{x}=(c_1,7-2c_1+5c_2-c_3,c_2,8-c_1+c_2+2c_3,-2-3c_1-4c_2+3c_3,c_3)$.

Take for example $c_1=1,\ c_2=0,\ c_3=3$, obtain one of the partial solutions:

$x_1=1,\ x_2=1,\ x_3=0,\ x_4=13,\ x_5=4,\ x_6=3$.

***Example* 25.5.** Find a general solution of the equation:

$$\begin{cases} x_1+2x_2 \ +x_3+4x_4 \ +x_5=-4, \\ 2x_1+5x_2+3x_3+8x_4+3x_5=-10, \\ \qquad x_2+2x_3+2x_4+6x_5=-1, \\ 4x_1+4x_2 \qquad +16x_4 \qquad =-3. \end{cases}$$

Solution:

$$\left(\begin{array}{ccccc|c} 1 & 2 & 1 & 4 & 1 & -4 \\ 2 & 5 & 3 & 8 & 3 & -10 \\ 0 & 1 & 2 & 2 & 6 & -1 \\ 4 & 4 & 0 & 16 & 0 & -3 \end{array}\right) \xrightarrow{l_2-2l_1,\ l_4-4l_1} \left(\begin{array}{ccccc|c} 1 & 2 & 1 & 4 & 1 & -4 \\ 0 & 1 & 1 & 0 & 1 & -2 \\ 0 & 1 & 2 & 2 & 6 & -1 \\ 0 & -4 & -4 & 0 & -4 & 13 \end{array}\right) \rightarrow$$

$$\xrightarrow{l_1-2l_2,\ l_3-l_2,\ l_4+4l_2} \left(\begin{array}{ccccc|c} 1 & 0 & -1 & 4 & -1 & 0 \\ 0 & 1 & 1 & 0 & 1 & -2 \\ 0 & 0 & 1 & 2 & 5 & 1 \\ 0 & 0 & 0 & 0 & 0 & 5 \end{array}\right).$$

The system is incompatible.

2. Inverse matrix method. Let the number of equations in the system of equations (25.1) be equal to the number of unknowns $m=n$, the system has the form:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n = b_n. \end{cases} \qquad (25.6)$$

We compose a square matrix A of this system:

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots. & \ldots. & \ldots & \ldots. \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix}.$$

We write system (21.6) in matrix form:

$$AX = B. \qquad (25.7)$$

Let matrix *A* be non degenerate matrix: $|A| \neq 0$. Then there is an inverse matrix A^{-1}.

Multiply both sides of the equation (21.7) by A^{-1} on the left:

$A^{-1}AX = A^{-1}B$,

And get solution of the system (4.6):

$$X = A^{-1}B. \qquad (25.8)$$

***Example* 25.6.** Solve three systems of equations:

$$1)\begin{cases} x_1 + 2x_2 + 3x_3 = 4, \\ 2x_1 + 2x_2 + 3x_3 = 5, \\ 3x_1 + 3x_2 + 4x_3 = 7; \end{cases} \quad 2)\begin{cases} x_1 + 2x_2 + 3x_3 = 0, \\ 2x_1 + 2x_2 + 3x_3 = 1, \\ 3x_1 + 3x_2 + 4x_3 = 1; \end{cases},$$

$$3)\begin{cases} x_1 + 2x_2 + 3x_3 = 14, \\ 2x_1 + 2x_2 + 3x_3 = 15, \\ 3x_1 + 3x_2 + 4x_3 = 21. \end{cases}.$$

Solution. Matrix A^{-1} was found in the example 19.7.

1. $$X = \begin{pmatrix} -1 & 1 & 0 \\ 1 & -5 & 3 \\ 0 & 3 & -2 \end{pmatrix} \cdot \begin{pmatrix} 4 \\ 5 \\ 7 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}.$$

2. $$X = \begin{pmatrix} -1 & 1 & 0 \\ 1 & -5 & 3 \\ 0 & 3 & -2 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}.$$

3. $$X = \begin{pmatrix} -1 & 1 & 0 \\ 1 & -5 & 3 \\ 0 & 3 & -2 \end{pmatrix} \cdot \begin{pmatrix} 14 \\ 15 \\ 21 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}.$$

3. The Cramer Rule. In § 24.1, we already considered the Cramer rule for solving a system of two equations with two unknowns. We generalize it to the case of any number of unknowns.

***Theorem* 25.1 (Kramer's theorem).** Let a system of n linear equations with n unknowns be given $AX = B$. If $|A| \neq 0$, then system has only one solution:

$$x_1 = \frac{|A_1|}{|A|}, \quad x_2 = \frac{|A_2|}{|A|}, \quad ..., \quad x_n = \frac{|A_n|}{|A|}, \tag{25.10}$$

where A_i means the matrix obtained from A by replacing its ith column with a column of free terms B ($i = 1, 2, \ldots, n$).

Evidence. We write in expanded form, taking into account (25.10) the solution of the system $X = A^{-1}B$ of the system $AX = B$:

$$\begin{pmatrix} x_1 \\ x_2 \\ ... \\ x_n \end{pmatrix} = \frac{1}{|A|} \cdot \begin{pmatrix} A_{11} & A_{12} & ... & A_{1n} \\ A_{21} & A_{22} & ... & A_{2n} \\ .... & .... & ... & .... \\ A_{n1} & A_{n2} & ... & A_{nn} \end{pmatrix} \cdot \begin{pmatrix} b_1 \\ b_2 \\ ... \\ b_n \end{pmatrix}.$$

According to the rule of multiplying matrices, we get:

$$x_i = \frac{1}{|A|} \cdot \left(A_{1i} b_1 + A_{2i} b_2 + ... + A_{ni} b_n \right), \quad i = 1, 2, \ldots, n.$$

However, the expression in parentheses is the expansion of the determinant $|A_i|$ in the ith column: $A_{1i}b_1 + A_{2i}b_2 + ... + A_{ni}b_n = |A_i|$.

So, $x_i = \frac{|A_i|}{|A|}$. The theorem is proved..

Formulas (25.9) are called **the Cramer formulas**, and the rule for solving systems using these formulas is called **the Cramer rule**.

Cramer's formulas are mainly of theoretical value. Their application for solving systems with a large number of unknowns would lead to cumbersome calculations. However, these formulas have a very important merit: they give an *explicit* expression of the meanings of all unknowns.

25.3. Compatibility of systems of linear equations

Consider matrix A and the extended matrix $\overline{A}$ of system (25.1). It is known that the rank of a matrix is equal to the largest number of its linearly independent columns. Therefore, attaching a column of free terms to the matrix A, we either obtain a matrix whose rank is one more than rank A or do not increase the rank - if the column of free terms is a linear combination of the remaining columns:

$$\begin{pmatrix} a_{11} \\ a_{21} \\ ... \\ a_{m1} \end{pmatrix} \cdot k_1 + \begin{pmatrix} a_{12} \\ a_{22} \\ ... \\ a_{m2} \end{pmatrix} \cdot k_2 + ... + \begin{pmatrix} a_{1n} \\ a_{2n} \\ ... \\ a_{mn} \end{pmatrix} \cdot k_n = \begin{pmatrix} b_1 \\ b_2 \\ ... \\ b_m \end{pmatrix}. \qquad (25.10)$$

It is easy to see that equality (25.10) is equivalent to the fact that system (25.1) has a solution $x_1 = k_1, x_2 = k_2, ..., x_n = k_n$. The question of

compatibility of a system of linear equations is solved by the following theorem.

***Theorem* 25.2 (Kronecker-Capelli theorem).** The system of linear equations (4.1) is compatible if and only if the rank of its extended matrix $\overline{A}$ is equal to the rank of its matrix A.

Evidence. 1. Let system (25.1) be compatible and lets $k_1, k_2, ..., k_n$ – be its solution. We substitute these numbers for unknowns and obtain a system of identities that is equivalent to equality (25.10), which means that the column of free terms is a linear combination of columns of the matrix A. It follows that

$\text{rg } A = \text{rg } \overline{A}$.

2. Lets given that $\text{rg } A = \text{rg } \overline{A}$. Then any maximal linearly independent column system of the matrix A remains a maximal linearly independent column system in the matrix $\overline{A}$. Therefore, in particular, the column of free terms is a linear combination of columns of this system, and it follows that the column of free terms is a linear combination of all columns of the matrix A, i.e. equality of the form (25.10) holds. This, in turn, means that numbers $k_1, k_2, ..., k_n$ (among which, of course, some may be zeros) constitute a solution to the system (25.1). We proved that the compatibility of system (25.1) follows from the equality $\text{rg } A = \text{rg } \overline{A}$. The proof is over.

Now take a look at the application of the Kronecker-Capelli theorem: Let $\text{rg } A = \text{rg } \overline{A} = r$. In this case, we say that the rank of system (25.1) is r. Then system (25.1) is compatible.

If $r = n$, then the system is defined. Its unique solution can be calculated either according to the Kramer rule or by reduction to the form (25.4). If

$r < n$, the system is reduced to the form (25.5). Unknowns $x_{r+1}, ..., x_n$ declare free and transfer to the right side of the equations. In the left parts remain terms containing unknowns x_1, x_2, …, x_r called the **basis**:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + ... + a_{1r}x_r = b_1 - a_{1,r+1}x_{r+1} - ... - a_{1n}x_n, \\ a'_{22}x_2 + ... + a'_{2r}x_r = b'_2 - a'_{2,r+1}x_{r+1} - ... - a'_{2n}x_n, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{rr}^{(r-1)}x_r = b_r^{(r-1)} - a_{r,r+1}^{(r-1)}x_{r+1} - ... - a_{rn}^{(r-1)}x_n. \end{cases}$$

From the last equation find x_r, substitute its value (depending on free unknowns) in the penultimate equation, find x_{r-1} and so on. This way find unknowns x_1, x_2, …, x_r, expressed by free variables, which can have any values. Therefore, system (25.1) in this case has an infinite number of solutions.

Giving free unknowns arbitrary values and obtaining the corresponding values of basic unknowns, we obtain all solutions of system (25.1).

***Example* 25.7.** Solve the system of equations:

$$\begin{cases} x_1 + 2x_2 + x_4 = 5, \\ 3x_1 + 5x_2 - x_3 + 2x_4 = 13, \\ x_1 + 3x_2 + x_3 + 2x_4 = 7, \\ x_1 + x_2 - x_3 = 3. \end{cases}$$

Solution. We compose an extended matrix and apply the Gauss method:

$$\left(\begin{array}{cccc|c} 1 & 2 & 0 & 1 & 5 \\ 3 & 5 & -1 & 2 & 13 \\ 1 & 3 & 1 & 2 & 7 \\ 1 & 1 & -1 & 0 & 3 \end{array}\right) \to \left(\begin{array}{cccc|c} 1 & 2 & 0 & 1 & 5 \\ 0 & -1 & -1 & -1 & -2 \\ 0 & 1 & 1 & 1 & 2 \\ 0 & -1 & -1 & -1 & -2 \end{array}\right) \to \left(\begin{array}{cccc|c} 1 & 2 & 0 & 1 & 5 \\ 0 & -1 & -1 & -1 & -2 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right)$$

.

We see that $\operatorname{rg} A = \operatorname{rg} \overline{A} = 2$, in particular, the minor composed of the coefficients of and in the first two equations is nonzero:

$$\begin{vmatrix} 1 & 2 \\ 0 & -1 \end{vmatrix} = -1 \neq 0$$

,

and all third-order minors are zero. Unknowns x_1, x_2 take as basis ones, the others, i.e. x_3, x_4 declare as free and put to the right side of equation:

$$\begin{cases} x_1 + 2x_2 = 5 - x_4 \\ -\ x_2 = -2 + x_3 + x_4 \end{cases}$$

Giving free variables arbitrary values $x_3 = c_1$, $x_4 = c_2$, we find an infinite number of solutions to the system: $x_1 = 1 + 2c_1 + c_2$, $x_2 = 2 - c_1 - c_2$, $x_3 = c_1$, $x_4 = c_2$.

25.3. Homogeneous equation systems

A system of linear equations is called **homogeneous** if in all its equations the free terms are equal to zero:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n = 0, \\ a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n = 0, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{m1}x_1 + a_{m2}x_2 + ... + a_{mn}x_n = 0. \end{cases} \quad (25.11)$$

Obviously, such a system always has a zero (or trivial) solution: $x_1 = 0, x_2 = 0, ..., x_n = 0$. Therefore, a homogeneous system is always compatible. (It is also clear that $\text{rg}\, A = \text{rg}\, \overline{A}$.)

Of interest is the question of the existence of nonzero solutions to this system. If the determinant of the system (25.11) is nonzero, then the system has only a zero solution (this follows from Cramer's theorem).

The following statement is true (we give it without proof).

***Theorem* 25.3.** A homogeneous system has a nonzero solution if and only if the rank of this system is less than the number of unknowns.

In particular, with a system has a nonzero solution if and only if its determinant is equal to zero.

We will write down every solution to the system (25.11):

$x_1 = c_1, x_2 = c_2, \quad ..., x_n = c_n$ as a row vector $\overline{c} = (c_1, c_2, ..., c_n)$.

We note the ***properties of solutions of homogeneous systems***:

1. If row vector $\overline{c} = (c_1, c_2, ..., c_n)$ – the solution of the system (25.11), then for each number *k* vector $k\overline{c} = (kc_1, kc_2, ..., kc_n)$ is a solution of the system (25.11) as well.

2. If vector $\overline{c} = (c_1, c_2, ..., c_n)$ and vector $\overline{c}' = (c_1', c_2', ..., c_n')$ are solutions of the system (4.11), then the sum of these vectors $\overline{c} + \overline{c}'$ is a solution of the same system.

The validity of these properties is verified by direct substitution of the indicated solutions into the equations of the system. (We invite the reader to do this on their own.)

From the formulated properties it follows that *each linear combination of solutions of a homogeneous system is also a solution to this system.*

Obviously, if a homogeneous system has a nonzero solution, then it has an infinite number of solutions. From the set of solution vectors of the homogeneous system (25.11), one can choose a basis. This basis is called ***the fundamental system of solutions of the homogeneous system*** (25.11). The system (25.11) in this case has many different fundamental systems of solutions.

Theorem 25.4. If the rank r of the system of linear homogeneous equations (25.11) is less than the number of unknowns n, then any fundamental system of solutions to the system (25.11) consists of solutions.

(We accept Theorem 25.4 without proof.)

We indicate ***a method for finding fundamental systems of solutions*** to the system (25.11). We must take any system of linearly independent *(n-r)* - dimensional vectors, take the components of each of these vectors for the values of free unknowns $x_{r+1}, \ldots, x_n$ and find the corresponding values for r basis unknowns. We obtain *n-r* solutions of the system of equations (25.11) that make up the fundamental system.

Usually, it is most convenient to take *(n-r)* -dimensional unit vectors (1, 0, ..., 0), (0, 1, ..., 0), ..., (0, 0, ..., 1) as value vectors for free unknowns.

***Example* 25.8.** To solve a system:

$$\begin{cases} x_1 + 2x_2 + x_3 + x_4 - 3x_5 = 0, \\ 2x_1 + 3x_2 + x_3 + 2x_4 + x_5 = 0, \\ 2x_1 + 5x_2 + 3x_3 + 2x_4 - 13x_5 = 0, \\ x_1 + x_2 + x_4 + 4x_5 = 0, \\ x_1 - x_2 - 2x_3 + x_4 + 18x_5 = 0. \end{cases}$$

Solution. Transform the system matrix:

$$\begin{pmatrix} 1 & 2 & 1 & 1 & -3 \\ 2 & 3 & 1 & 2 & 1 \\ 2 & 5 & 3 & 2 & -13 \\ 1 & 1 & 0 & 1 & 4 \\ 1 & -1 & -2 & 1 & 18 \end{pmatrix} \xrightarrow[l_4 - l_1,\, l_5 - l_1]{l_2 - 2l_1,\, l_3 - 2l_1,} \begin{pmatrix} 1 & 2 & 1 & 1 & -3 \\ 0 & -1 & -1 & 0 & 7 \\ 0 & 1 & 1 & 0 & -7 \\ 0 & -1 & -1 & 0 & 7 \\ 0 & -3 & -3 & 0 & 21 \end{pmatrix} \rightarrow$$

$$\rightarrow \begin{pmatrix} 1 & 2 & 1 & 1 & -3 \\ 0 & -1 & -1 & 0 & 7 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

We get a system equivalent to the original:

$$\begin{cases} x_1 + 2x_2 + x_3 + x_4 - 3x_5 = 0 \\ \quad - x_2 - x_3 \qquad + 7x_5 = 0. \end{cases}$$

We choose as basic unknowns x_1 и x_2 (the coefficients in them form a minor other than zero):

$$\begin{cases} x_1 + 2x_2 = -x_3 - x_4 + 3x_5 \\ \quad - x_2 = \quad x_3 \qquad - 7x_5. \end{cases}$$

We get a fundamental system of solutions:

$\bar{e}_1 =$ (1, –1, 1, 0, 0), $\bar{e}_2 =$ (–1, 0, 0, 1, 0), $\bar{e}_3 =$ (–11, 7, 0, 0, 1).

25.5. Heterogeneous systems.

Structure of the general solution of the system of linear heterogeneous equations

Consider a system of linear inhomogeneous equations (25.1):

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n = b_1, \\ a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n = b_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{m1}x_1 + a_{m2}x_2 + ... + a_{mn}x_n = b_m \end{cases}$$

and its corresponding homogeneous system (25.11)

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n = 0, \\ a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n = 0, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{m1}x_1 + a_{m2}x_2 + ... + a_{mn}x_n = 0, \end{cases}$$

obtained from system (25.1) by replacing the free terms with zeros.

We establish a connection between the solutions of systems (25.1) and (25.11). The following ***statements*** are true:

1. The sum of any solution to system (25.1) with any solution to the corresponding homogeneous system (25.11) is again a solution to system (25.1).

2. The difference of any two solutions of the system (25.1) is the solution of the corresponding homogeneous system (25.11).

The proofs of these statements are very simple. We prove, for example, the second.

Let $\bar{c} = (c_1, c_2, ..., c_n)$ and $\bar{c}' = (c_1', c_2', ..., c_n')$ are two solutions of the system (25.1). We substitute the difference of these solutions $\bar{c} - \bar{c}' = (c_1 - c_1', c_2 - c_2', ..., c_n - c_n')$ to left part of any *i*-th equation of the system (25.11) and regroup summands:

$$\alpha_{i1}(c_1 - c_1') + \alpha_{i2}(c_2 - c_2') + ... + \alpha_{in}(c_n - c_n') = (\alpha_{i1}c_1 + \alpha_{i2}c_2 + ... + \alpha_{in}c_n) -$$
$$-(\alpha_{i1}c_1' + \alpha_{i2}c_2' + ... + \alpha_{in}c_n') = b_i - b_i = 0.$$

The equation turns into the identity 0 = 0.

From these two statements it follows that, *finding one solution to the system of linear inhomogeneous equations (25.1) and adding it to each of the solutions to the corresponding homogeneous system (25.11), we obtain all solutions to the system (25.1).*

In other words, the general solution of the system of linear equations (25.1) is the sum of any particular solution to this system with the general solution of the corresponding homogeneous system (25.11).

***Example* 25.9.** To solve a system:

$$\begin{cases} x_1 + 2x_2 + x_3 + x_4 - 3x_5 = 4 \\ 2x_1 + 3x_2 + x_3 + 2x_4 + x_5 = 6 \\ 2x_1 + 5x_2 + 3x_3 + 2x_4 - 13x_5 = 10 \\ x_1 + x_2 + x_4 + 4x_5 = 2 \\ x_1 - x_2 - 2x_3 + x_4 + 18x_5 = -2 \end{cases}$$

Solution. The homogeneous system corresponding to this system is considered in Example 25.8. Transforming the extended matrix of this system, we come to the system:

$$\begin{cases} x_1 + 2x_2 = 4 - x_3 - x_4 + 3x_5 \\ -x_2 = -2 + x_3 - 7x_5. \end{cases}$$

We find one of the particular solutions to this system. The easiest way to do this is by setting $x_3 = x_4 = x_5 = 0$. We get a solution c_0 = (0, 2, 0, 0, 0). (A particular solution in which all the values of free unknowns are equal to zero is sometimes called **basis**.)

The fundamental system of solutions of the corresponding homogeneous system is already found in Example 25.8. We use it and get a general solution to this heterogeneous system:

(0, 2, 0, 0, 0) + $k_1 \cdot$ (1, –1, 1, 0, 0) + $k_2 \cdot$ (–1, 0, 0, 1, 0) + $k_3 \cdot$ (–11, 7, 0, 0, 1).

Giving the coefficients k_1, k_2, k_3 all possible values, we get all the solutions of this system.

Questions

1. Can an indefinite system of linear equations be incompatible?
2. What is called a general solution of a system of linear equations?
3. Can a system containing seven equations with five unknowns be equivalent to a system of four equations with five unknowns?
4. To which system of linear equations does the Cramer rule apply?
5. Is the inverse matrix method applicable to indefinite system of linear equations?
6. Can a homogeneous system of linear equations be incompatible?
7. What is called the fundamental system of solutions of a homogeneous system of linear equations?
8. How many solutions does the 4-ranked fundamental system of solutions of a homogeneous system of equations with six unknowns?
9. What is the structure of the general solution of a system of linear inhomogeneous equations?

Chapter 26. Linear operators

26.1. The concept of a linear operator

Let two vector spaces *U* and *V* be given.

Definition. If a certain rule A is given, according to which each vector $\bar{u}$ of space U is assigned a unique vector $\bar{v} = A(\bar{u})$ of the space V, then an **operator** A acts from U to V. The vector $\bar{v} = A(\bar{u})$ is called the **image** of the vector $\bar{u}$, and the vector $\bar{u}$ is the prototype of the vector $\bar{v}$.

Operator $A(\bar{u})$ is called **linear** if it satisfies the following two conditions:

1) for any two vectors $\bar{u}_1$ and $\bar{u}_2$ of the space *U*
$A(\bar{u}_1 + \bar{u}_2) = A(\bar{u}_1) + A(\bar{u}_2)$;

2) for any $\bar{u}$ from *U* and any number λ
$A(\lambda\bar{u}) = \lambda A(\bar{u})$.

The concept of a linear operator is one of the fundamental concepts of linear algebra.

Then let $U = \mathbf{R}^n$, $V = \mathbf{R}^m$.

Later we will see that if in the space $\mathbf{R}^n$ set some basis $\bar{e}_1$, $\bar{e}_2$, ..., $\bar{e}_n$, and in the space $\mathbf{R}^m$ set some basis $\bar{f}_1$, $\bar{f}_2$, ..., $\bar{f}_m$, then linear operator *A* if defined by matrix size $m \times n$.

So let $\bar{e}_1$, $\bar{e}_2$, ..., $\bar{e}_n$ is some basis of space $\mathbf{R}^n$. Take an arbitrary vector $\bar{x}$ and expand it in terms of basis:
$\bar{x} = x_1\bar{e}_1 + x_2\bar{e}_2 + ... + x_n\bar{e}_n$.

Operator $A(\bar{x})$ is a linear then

$$A(\bar{x}) = x_1 A(\bar{e}_1) + x_2 A(\bar{e}_2) + ... + x_n A(\bar{e}_n). \quad (26.1)$$

However, each of the vectors $A(\bar{e}_i)$ (i = 1, 2, …, n) is a vector from the space $\mathbf{R}^m$, therefore, it can be decomposed into a basis $\bar{f}_1$, $\bar{f}_2$, ..., $\bar{f}_m$:

$$A(\bar{e}_i) = a_{1i}\bar{f}_1 + a_{2i}\bar{f}_2 + ... + a_{mi}\bar{f}_m. \quad (26.2)$$

Substituting the decomposition $A(\bar{e}_i)$ (i = 1, 2, …, n) from (5.2) to (5.1), get:

$$A(\bar{x}) = x_1(a_{11}\bar{f}_1 + a_{21}\bar{f}_2 + ... + a_{m1}\bar{f}_m) + x_2(a_{12}\bar{f}_1 + a_{22}\bar{f}_2 + ... + a_{m2}\bar{f}_m) + \\ + ... + x_n(a_{1n}\bar{f}_1 + a_{2n}\bar{f}_2 + ... + a_{mn}\bar{f}_m).$$

Regrouping the terms and collecting the coefficients for $\bar{f}_1$, $\bar{f}_2$, ..., $\bar{f}_m$, we get

$$A(\bar{x}) = (a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n)\cdot\bar{f}_1 + (a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n)\cdot\bar{f}_2 + \\ + ... + (a_{m1}x_1 + a_{m2}x_2 + ... + a_{mn}x_n)\cdot\bar{f}_m. \quad (26.3)$$

Let y_1, y_2, …, y_m be coordinates of the vector image $\bar{x}$, i.e. coordinates of the vector $\bar{y} = A(\bar{x})$ in basis $\bar{f}_1$, $\bar{f}_2$, ..., $\bar{f}_m$. Then

$$A(\bar{x}) = y_1\bar{f}_1 + y_2\bar{f}_2 + ... + y_m\bar{f}_m. \quad (26.4)$$

Due to the uniqueness of the expansion of the vector along the basis, the right-hand sides of equalities (26.3) and (26.4) coincide. Hence:

$$\begin{cases} y_1 = a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n \\ y_2 = a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_m = a_{m1}x_1 + a_{m2}x_2 + ... + a_{mn}x_n. \end{cases} \quad (26.5)$$

The matrix A of system (26.5) is called the **matrix of the operator** A with respect to the basis $\bar{e}_1$, $\bar{e}_2$, ..., $\bar{e}_n$ and $\bar{f}_1$, $\bar{f}_2$, ..., $\bar{f}_m$:

$$\mathrm{A} = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{pmatrix}.$$

So, *a matrix of size* $m \cdot n$ corresponds to *each linear operator* $A: \mathbf{R}^n \to \mathbf{R}^m$. Obviously, the converse statement is also true: *a linear operator* $A: \mathbf{R}^n \to \mathbf{R}^m$ *corresponds to a square matrix of size* $m \cdot n$.

If we consider the vectors $\bar{x} = (x_1, x_2, ..., x_n)$ and $\bar{y} = A(\bar{x}) = (y_1, y_2, ..., y_m)$ as column matrices:

$$\mathbf{X} = \begin{pmatrix} x_1 \\ x_2 \\ \ldots\ldots \\ x_n \end{pmatrix}, \qquad \mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \ldots\ldots \\ y_m \end{pmatrix},$$

then equality $\bar{y} = A(\bar{x})$, or system (26.5) which is the same, can be written in the form of matrix equality:

$$\mathbf{Y} = \mathbf{AX},$$

where **A** is a matrix of a linear operator. In particular, when spaces $\mathbf{R}^n$ and $\mathbf{R}^m$ match, space $\mathbf{R}^n$ is mapped into itself through operator *A*. In this case, the operator matrix is a square matrix of order *n*.

26.2. Actions with linear operators

For linear operators, the operations of addition and multiplication by a number are defined.

1. **The sum** of two linear operators A_1 and A_2 is called the operator ($A_1 + A_2$), defined by the equality:
$(A_1 + A_2)(\bar{x}) = A_1(\bar{x}) + A_2(\bar{x})$.

2. **The product of the linear operator *A* by a number** λ is the operator λA defined by the equality:
$\lambda A(\bar{x}) = \lambda(A(\bar{x}))$.

It is known that every linear operator $\mathbf{R}^n$ that maps into itself is determined by the corresponding square matrix. Therefore, the described operations correspond to similar operations on operator matrices - addition and multiplication by a number. The operators ($A_1 + A_2$) and λA are also linear.

Zero operator $\tilde{0}$ is defined as an operator that translates every vector of space $\mathbf{R}^n$ in a zero vector $\bar{0}$.

Obviously, $(A + \tilde{0}) = A$ for each operator *A*.

For linear operators, the multiplication operation can also be defined.

3. **The product of linear operators** A_1 and A_2 is the operator (A_1A_2) defined by the equality:
$(A_1A_2)(\bar{x}) = A_1(A_2(\bar{x}))$.

The product of two linear operators is also a linear operator.

The identity operator E is defined as follows:
$E(\bar{x}) = \bar{x}$.

Obviously, $(AE) = (EA) = A$ for each linear operator A.

26.3. Eigenvectors and eigenvalues of linear operator

Definition. Nonzero vector $\bar{x} \in \mathbf{R}^n$ called **eigenvector** of linear operator A, if there is such a number λ that:

$$A(\bar{x}) = \lambda\bar{x}. \tag{26.6}$$

And number λ is an **eigenvalue** of operator A.

If **A** is a matrix of an operator A, then number λ, satisfying equality (26.6), called *eigenvalue* of matrix **A**, and vector $\bar{x}$ is an *eigenvector* of matrix **A**. Equality (26.6) can be written in matrix form:

$$\mathbf{AX} = \lambda\mathbf{X} \text{ or } \mathbf{AX} = \lambda\mathbf{EX}. \tag{26.7}$$

From the last equation:

$$(\mathbf{A} - \lambda\mathbf{E})\mathbf{X} = 0. \tag{26.8}$$

If $\mathbf{A} = (a_{ij})$, $i, j = 1, 2, \ldots, n$, then

$$\mathbf{A}-\lambda\mathbf{E}=\begin{pmatrix} a_{11}-\lambda & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22}-\lambda & \ldots & a_{2n} \\ \ldots. & \ldots. & \ldots & \ldots. \\ a_{n1} & a_{n2} & \ldots & a_{nn}-\lambda \end{pmatrix}. \qquad (26.9)$$

Equation (26.8) is equal to the system of linear homogeneous equations:

$$\begin{cases} (a_{11}-\lambda)\cdot x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = 0, \\ a_{21}x_1 + (a_{22}-\lambda)\cdot x_2 + \ldots + a_{2n}x_n = 0, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + (a_{nn}-\lambda)\cdot x_n = 0. \end{cases} \qquad (26.10)$$

Since the eigenvector is not zero, the homogeneous system (26.10) must have a nonzero solution. As is known, a homogeneous system of *n* equations with *n* unknowns has a nonzero solution if and only if its determinant is zero. Therefore

$$\begin{vmatrix} a_{11}-\lambda & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22}-\lambda & \ldots & a_{2n} \\ \ldots. & \ldots. & \ldots & \ldots. \\ a_{n1} & a_{n2} & \ldots & a_{nn}-\lambda \end{vmatrix} = 0 \quad , \text{ or } \left|\mathbf{A}-\lambda\mathbf{E}\right| = 0 \qquad (26.11)$$

Equation (26.11) is called **the characteristic equation of the matrix**, and the left side of this equation is called **the characteristic polynomial of matrix** A (or operator *A*).

Obviously, equation (26.11) is an algebraic equation of degree *n*. Its roots are the eigenvalues of the matrix A. Substituting each of the roots into the system (26.10) and solving it, we obtain the corresponding eigenvector.

Example 26.1 Find the eigenvalues of the eigenvectors of the linear operator *A* given by the matrix:

$$\mathbf{A}=\begin{pmatrix}4 & 1\\ 2 & 3\end{pmatrix}.$$

Solution. We draw up the characteristic equation $|\mathbf{A}-\lambda\mathbf{E}|=0$ for this matrix:

$$\begin{vmatrix}4-\lambda & 1\\ 2 & 3-\lambda\end{vmatrix}=0.$$

Solving the determinant, we obtain

$$\lambda^2-7\lambda+10=0.$$

The root of the equation are $\lambda_1=5$, $\lambda_2=2$. Substituting $\lambda_1=5$ to the system (26.11) with $n=2$, get

$$\begin{cases}(4-5)\cdot x_1+x_2=0\\ 2x_1+(3-5)\cdot x_2=0,\end{cases} \quad \text{т.е.} \quad \begin{cases}-x_1+x_2=0\\ 2x_1-2x_2=0.\end{cases} \qquad (*)$$

Eigenvector corresponding to the eigenvalue $\lambda_1=5$ is a solution of the system which is equivalent to the equation:

$$x_1-x_2=0.$$

Declare x_2 a free unknown and consider $x_2=c$, get the first eigenvector

$$\bar{x}_1=(c,c)=c(1,1).$$

Then substitute $\lambda_2=2$:

$$\begin{cases}2x_1+x_2=0\\ 2x_1+x_2=0.\end{cases} \qquad (**)$$

This system is also equivalent to one equation. Consider $x_2 = d$, get $\bar{x}_2 = \left(-\frac{d}{2}, d\right) = -\frac{d}{2}(1, -2)$.

I.e. *c* and *d* are arbitrary numbers, then an infinite number of vectors can correspond to one eigenvalue. In particular, assuming, that $c = 1$ and $d = -2$ we obtain eigenvectors that are fundamental solutions of the corresponding homogeneous systems (*) and (**). They have the form, $\bar{x}_1 = (1, 1)$ and $\bar{x}_2 = (1, -2)$.

Questions

1. What defines a linear operator in the basis of space $\mathbf{R}^n$?
2. Does any *n*-th order square matrix define in a linear operator in $\mathbf{R}^n$?
3. Which linear operator is called null?
4. Write the characteristic equation for the matrix $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$.
5. How many different eigenvalues can a third-order matrix have?
6. Is the number $\lfloor = 5$ an eigenvalue of the matrix $\mathbf{A} = \begin{pmatrix} 5 & 3 \\ 0 & 1 \end{pmatrix}$?
7. Is the vector $\bar{x} = (2, 3)$ an eigenvector of the matrix $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 0 & 4 \end{pmatrix}$?

Chapter 27. Quadratic forms

27.1. Basic concepts

Definition. **A quadratic form** $F(x_1, x_2, ..., x_n)$ of n unknowns $x_1, x_2, ..., x_n$ is a sum, each term of which is either a square of one of these variables, or a product of two different variables.

Example **27.1.** Sum $x^2 - 3xy + 2y$ is a quadratic form of two unknowns: *x* и *y*; sum $x_1^2 + 2x_1x_2 - 3x_1x_3 + 4x_2^2 - x_2x_3$ is a quadratic form of three unknowns x_1, x_2, x_3.

(Note that similar terms are already given in the above quadratic forms.)

Each quadratic form can be written **in a standard form**. The following commonly used symbols are used.

Any type sum $a_k + a_{k+1} + ... + a_n$ is written as

$$a_k + a_{k+1} + ... + a_n = \sum_{i=k}^{n} a_i .$$

In particular,

$$a_1 + a_2 + ... + a_n = \sum_{i=1}^{n} a_i .$$

If the sum is considered, the terms of which a_{ij} are provided with two indices *i* and *j*, moreover, *i = 1, 2, ..., m; j = 1, 2, ..., n*, then you can first take the sum of elements with a fixed first index, i.e.

$$\sum_{i=1}^{n} a_{1j}, \ \sum_{j=1}^{n} a_{2j}, \ \ ..., \ \ \sum_{j=1}^{n} a_{mj} ,$$

and then add up all these amounts. Then for the sum of all these elements we get the record

$$\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij} \quad . \tag{27.1}$$

You can also add first the terms a_{ij} with a fixed second index, and then the amounts already received:

$$\sum_{j=1}^{n}\sum_{i=1}^{m} a_{ij} \quad . \tag{27.2}$$

Therefore

$$\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij} = \sum_{j=1}^{n}\sum_{i=1}^{m} a_{ij} \quad , \tag{27.3}$$

i.e. in double amount, you can change the summation order.

The sums (27.1) and (27.3) can be considered as the sum of the elements of matrix $m \times n$:

$$\begin{pmatrix} a_{11} & \dots & a_{1j} & \dots & a_{1n} \\ \dots & \dots & \dots & \dots & \dots \\ a_{i1} & \dots & a_{ij} & \dots & a_{in} \\ \dots & \dots & \dots & \dots & \dots \\ a_{m1} & \dots & a_{mj} & \dots & a_{mn} \end{pmatrix} .$$

If we add the elements of each row in this matrix and then add the sums obtained, we have (27.1); if we first add up the elements of each column and then add up what happened, we have (27.2).

Let us now return to the question of the standard form of a quadratic form. Assuming that similar terms are already given in quadratic form

$F = F(x_1, x_2, ..., x_n)$, we introduce the following notation: the coefficient of the quadratic form x_i^2 for is denoted as a_{ii}, and the coefficient of the product for $x_i x_j$ for $i \neq j$ is denoted as $2a_{ii}$. Since, obviously, $x_i x_j = x_j x_i$ the coefficient in this product could be denoted as $2a_{ji}$, i.e. it is assumed that

$$a_{ij} = a_{ji}. \tag{27.4}$$

Then term $2a_{ij} x_i x_j$ can we written as

$$2a_{ij} x_i x_j = a_{ij} x_i x_j + a_{ji} x_j x_i,$$

and the entire quadratic form $F = F(x_1, x_2, ..., x_n)$ is written as the sum of all possible terms $a_{ij} x_i x_j$, where *i* and *j* independently of each other take all values from 1 to *n*:

$$F(x_1, x_2, ..., x_n) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j. \tag{27.5}$$

(In particular, if $i = j$, then get $a_{ii} x_i^2$.) Note that with double summation, the summation sign is often used. Equality (6.5) can be written as

$$F(x_1, x_2, ..., x_n) = \sum_{i,j=1}^{n} a_{ij} x_i x_j. \tag{27.5′}$$

The coefficients a_{ij} of the quadratic form (27.5) obviously form a square matrix $A = (a_{ij})$ of order n; it is called **a matrix of quadratic form** $F = F(x_1, x_2, ..., x_n)$, and the rank *r* of *A* is called the **rank** of this

quadratic form. If, in particular, i.e. matrix A is non-degenerate, then the quadratic form $F = F(x_1, x_2, ..., x_n)$ is also called **non-degenerate**. Equality (27.4) means that the elements of matrix A, symmetric with respect to the main diagonal, are equal to each other, i.e. matrix A is symmetric. Obviously, for any n-th order symmetric matrix, we can indicate the well-defined quadratic form (6.5) of n unknowns whose coefficients are elements of matrix A.

***Example* 27.2.** Write a quadratic form

$$F = F(x_1, x_2, x_3) = x_1^2 + 4x_1x_2 + 2x_2^2 - 3x_2x_3 + x_3^2 + x_3x_2$$

in standard form and find its matrix.

Solution. After reduction of similar members, we get

$$x_1^2 + 4x_1x_2 + 2x_2^2 - 2x_2x_3x_3^2 = x_1^2 + 2x_1x_2 + 2x_2x_1 + 2x_2^2 - x_2x_3 - x_3x_2 + x_3^2 =$$

$$= x_1x_1 + 2x_1x_2 + 0 \cdot x_1x_3 + 2x_2x_1 + 2x_2x_2 - x_2x_3 + 0 \cdot x_3x_1 - x_3x_2 + x_3x_3.$$

Matrix of quadratic form is:

$$A = \begin{pmatrix} 1 & 2 & 0 \\ 2 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix}.$$

The quadratic form (27.5) can be written in matrix (vector-matrix) form using the product of rectangular matrices.

Note that matrix A is *symmetric* if and only if it coincides with its transposed one, i.e. when

$$A' = A.$$

Denote by X the matrix column of the unknowns:

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix}.$$

We show that in the matrix notation the quadratic form has the following form:

$$F = X'AX. \tag{27.6}$$

Indeed, product *AX* will be a column matrix:

$$AX = \begin{pmatrix} \sum a_{1j}x_j \\ \sum a_{2j}x_j \\ \dots \\ \sum a_{nj}x_j \end{pmatrix}.$$

(Here we write © instead of $\sum_{j=1}^{n}$, to avoid unnecessarily cumbersome recordings.)

Now multiplying this column matrix on the left by matrix $X' = (x_1, x_2, ..., x_n)$, get

$$X'AX = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j = F,$$

which was to be demonstrated.

***Example* 27.3.** Write the quadratic form from Example 6.2 in matrix form.

Solution. Using the matrix *A*, found in Example 6.2, we obtain

$$F = (x_1, x_2, x_3)\begin{pmatrix} 1 & 2 & 0 \\ 2 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

Let us now consider how the quadratic form changes during the linear transformation of unknowns. Let a linear transformation of the unknown $x_1, x_2, ..., x_n$:

$$x_i = \sum_{k=1}^{n} c_{ik} y_k \text{ , i = 1, 2, …, } n \tag{27.7}$$

With matrix $C = (c_{ik})$, in other words, a linear transformation is given

$$X = CY \text{ ,} \tag{27.8}$$

where Y – unknown column $y_1, y_2, ..., y_n$.

We use one of the properties of the matrix transpose operation:

$$(AB)' = B'A' \text{ .} \tag{27.9}$$

According to (27.9) we get from (27.8)

$$X' = (CY)' = Y'C' \text{ .} \tag{27.10}$$

From here

$$F = X'AX = (Y'C')A(CY) = Y'(C'AC)Y \text{ , или } F = Y'\tilde{A}Y \text{ ,}$$

where

$$\tilde{A} = C'AC \text{ .} \tag{27.11}$$

Matrix $\tilde{A}$ will be symmetrical. Indeed, since, then $A' = A$ taking into account property (6.9)

$$\tilde{A}' = (C'AC)' = C'(C'A)' = C'A'C = C'AC = \tilde{A} \text{ .}$$

So, we have proved the following theorem.

***Theorem* 27.1.** A quadratic form with matrix A as a result of a linear transformation with matrix C turns into a quadratic form from new unknowns with matrix $C'AC$.

Suppose now that the transformation matrix C is non-degenerate. Then, obviously, C' is also a non-degenerate matrix. In this case, the product $C'AC$ is the product of the matrix A by non-degenerate matrices, and therefore the rank of this product is equal to the rank of matrix A.

We have obtained the following theorem.

Theorem 27.2. The rank of a quadratic form does not change under a non-degenerate linear transformation.

***Example* 27.4.** There is a quadratic form

$$F = F(x_1, x_2) = x_1^2 + 2x_1x_2 - 3x_2^2.$$

Find a quadratic form $G(y_1, y_2)$, obtained from a given linear transformation

$$x_1 = 2y_1 - y_2, \quad x_2 = y_1 + y_2.$$

Solution. We write the matrix of this quadratic form A and the transformation matrix C:

$$A = \begin{pmatrix} 1 & 1 \\ 1 & -3 \end{pmatrix}, \quad C = \begin{pmatrix} 2 & -1 \\ 1 & 1 \end{pmatrix}.$$

The matrix of the desired quadratic form $\tilde{A}$ according to (6.11) has the form

$$\tilde{A} = C'AC = \begin{pmatrix} 2 & 1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & -3 \end{pmatrix}\begin{pmatrix} 2 & -1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 5 & -4 \\ -4 & -4 \end{pmatrix}.$$

Hence, $G(y_1, y_2) = 5y_1^2 - 8y_1y_2 - 4y_2^2$.

***Example* 27.5.** There is a quadratic form:

$F = F(x_1, x_2, x_3) = x_1^2 + 5x_2^2 - 4x_3^2 + 2x_1x_2 - 4x_1x_3$.

Find a quadratic form $G(y_1, y_2, y_3)$, obtained from a given linear transformation

$$x_1 = y_1 - \frac{1}{2}y_2 + \frac{5}{6}y_3, \quad x_2 = \frac{1}{2}y_2 - \frac{1}{6}y_3, \quad x_3 = \frac{1}{3}y_3.$$

Solution. We write the matrix A of a given quadratic form and the transformation matrix C and calculate $\tilde{A} = C'AC$:

$$A = \begin{pmatrix} 1 & 1 & -2 \\ 1 & 5 & 0 \\ -2 & 0 & -4 \end{pmatrix}, \quad C = \begin{pmatrix} 1 & -1/2 & 5/6 \\ 0 & 1/2 & -1/6 \\ 0 & 0 & 1/3 \end{pmatrix},$$

$$\tilde{A} = \begin{pmatrix} 1 & 0 & 0 \\ -1/2 & 1/2 & 0 \\ 5/6 & -1/6 & 1/3 \end{pmatrix} \begin{pmatrix} 1 & 1 & -2 \\ 1 & 5 & 0 \\ -2 & 0 & -4 \end{pmatrix} \begin{pmatrix} 1 & -1/2 & 5/6 \\ 0 & 1/2 & -1/6 \\ 0 & 0 & 1/3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix}$$

.

Hence, $G = G(y_1, y_2, y_3) = y_1^2 + y_2^2 - y_3^2$.

Example 27.5 shows that with well-chosen linear transformations, the appearance of a quadratic form can be significantly simplified.

27.2. Canonical view of a quadratic form

We state that a quadratic form has a **canonical form** if its matrix is diagonal ($a_{ij} = 0$ where $i \neq j$):

$$A = \begin{pmatrix} a_{11} & 0 & ... & 0 \\ 0 & a_{22} & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & a_{nn} \end{pmatrix}.$$

Obviously, in this case the quadratic form is the sum of squares of unknowns of the form

$$F = a_{11}x_1^2 + a_{22}x_2^2 + ... + a_{nn}x_n^2 = \sum_{i=1}^{n} a_{ii}x_i^2 .$$

In particular, the quadratic form *G* obtained in Example 6.5 has a canonical form.

We have found that the rank of a quadratic form does not change under non-degenerate linear transformations (see § 27.1). Let a quadratic form $F(x_1, x_2, ..., x_n)$ be reduced to a canonical form by a nondegenerate linear transformation

$$b_1y_1^2 + b_2y_2^2 + ... + b_ny_n^2, \qquad (27.12)$$

where $y_1, y_2, ..., y_n$ are new unknowns. Here any coefficients $b_1, b_2, ..., b_n$ can be zeros.

It is easy to prove that *the rank of the quadratic form is equal to the number of non-zero coefficients in the canonical form to which the given quadratic form is reduced.*

Indeed, if a quadratic form of rank r is reduced by a nondegenerate linear transformation to the form (27.12), this means that the matrix of the transformed quadratic form has the form

$$\begin{pmatrix} b_1 & 0 & \dots & 0 \\ 0 & b_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & b_n \end{pmatrix}$$

has rank *r* as well. And this is equivalent to the fact that there are r non-zero elements on the main diagonal.

The following main theorem on quadratic forms is true (we present it without proof).

***Theorem* 27.3.** Every quadratic form can be reduced to a canonical form by some non-degenerate linear transformation.

It should be noted that any quadratic form can be reduced to a canonical form in various ways. Moreover, the canonical form to which this quadratic form is reduced is not uniquely determined for it.

***Example* 27.6.** There is a quadratic form

$F = 2x_1x_2 - 6x_2x_3 + 2x_3x_1$.

Check that it is canonical by a linear transformation:

$x_1 = \frac{1}{2}y_1 + \frac{1}{2}y_2 + 3y_3$,

$x_2 = \frac{1}{2}y_1 - \frac{1}{2}y_2 - y_3$,

$x_3 = \qquad\qquad y_3$.

Solution:

$$C = \begin{pmatrix} ½ & ½ & 3 \\ ½ & -½ & -1 \\ 0 & 0 & 1 \end{pmatrix}.$$

We get

$$\tilde{A}=\begin{pmatrix} ½ & ½ & 0 \\ ½ & -½ & 0 \\ 3 & -1 & 1 \end{pmatrix}\begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & -3 \\ 1 & -3 & 0 \end{pmatrix}\begin{pmatrix} ½ & ½ & 3 \\ ½ & -½ & -1 \\ 0 & 0 & 1 \end{pmatrix}=\begin{pmatrix} ½ & 0 & 0 \\ 0 & -½ & 0 \\ 0 & 0 & 6 \end{pmatrix},$$

i.e. the quadratic form is reduced to

$$\frac{1}{2}y_1^2-\frac{1}{2}y_2^2+6y_3^2.$$

***Example* 27.7.** There is a quadratic form from example 6.6:

$$F=2x_1x_2-6x_2x_3+2x_3x_1.$$

Verify that the linear transformation:

$$x_1=y_1+3y_2+2y_3,$$

$$x_2=y_1+3y_2+2y_3,$$

$$x_3=\qquad y_2$$

also makes this form canonical.

Solution:

$$C=\begin{pmatrix} 1 & 3 & 2 \\ 1 & -1 & -2 \\ 0 & 1 & 0 \end{pmatrix}.$$

Get

$$\tilde{A}=\begin{pmatrix} 1 & 1 & 0 \\ 3 & -1 & 1 \\ 2 & -2 & 0 \end{pmatrix}\begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & -3 \\ 1 & -3 & 0 \end{pmatrix}\begin{pmatrix} 1 & 3 & 2 \\ 1 & -1 & -2 \\ 0 & 1 & 0 \end{pmatrix}=\begin{pmatrix} 2 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & -8 \end{pmatrix},$$

i.e. quadratic form from example 27. 6 is reduced to another canonical form:

$2y_1^2 + 6y_2^2 - 8y_3^2$.

From examples 27.6 and 27.7 we see that the same quadratic form

$F = 2x_1x_2 - 6x_2x_3 + 2x_3x_1$

and as a result of one non-degenerate linear transformation has acquired the form

$G = \frac{1}{2}y_1^2 - \frac{1}{2}y_2^2 + 6y_3^2$,

and as a result of another linear transformation

$G_1 = 2y_1^2 + 6y_2^2 - 8y_3^2$.

Despite the fact that *G* and G_1 are noticeably different from each other, they still have one common property: they contain the same number of positive and negative coefficients (two and one, respectively). This is no coincidence. The following statement holds (we give it without proof).

***Theorem* 27.4 (law of inertia of quadratic forms).** The number of positive and the number of negative coefficients in the quadratic form in the canonical form, to which the given quadratic form is reduced by a non-degenerate linear transformation does not depend on the choice of this transformation.

Along with the canonical form, the **normal form** of a quadratic form is also considered, i.e. the sum of squares of unknowns with coefficients of +1 or - 1.

In particular, in the example 27.5 quadratic form

$F = x_1^2 + 5x_2^2 - 4x_3^2 + 2x_1x_2 - 4x_1x_3$

converted to the form

$G = y_1^2 + y_2^2 - y_3^2$,

which is not only canonical, but also normal.

It is easy to verify that a quadratic form transformed to a canonical form can always be reduced to a normal form by a non-degenerate linear transformation.

Indeed,

$$G = c_1 y_1^2 + ... + c_k y_k^2 - c_{k+1} y_{k+1}^2 - ... - c_r y_r^2,$$

where $c_1, ..., c_k, c_{k+1}, ..., c_r$ are positive.

Then the transformation $z_i = \sqrt{c_i} y_i$ (i = 1, 2, …, r), $z_j = y_j$ (j = r + 1, …, n) leads G to its normal form:

$$G_1 = z_1^2 + ... + z_k^2 - z_{k+1}^2 - ... - z_r^2.$$

The law of inertia of quadratic forms can now be formulated as follows**:** *the number of positive and negative squares in the normal form of a quadratic form does not depend on the choice of a linear non-degenerate transformation by which the quadratic form is reduced to this form.*

27.3. Positive and negative defined quadratic forms

Definition. Quadratic form $F(x_1, x_2, ..., x_n)$ is called a **positive definite** if, for all unknown values, of which at least one is nonzero, the inequality $F(x_1, x_2, ..., x_n) > 0$.

If $F(x_1, x_2, ..., x_n) < 0$ for all unknown values, of which at least one is nonzero, the quadratic form is called a **negative definite**.

It is easy to prove (we will not do this) that a quadratic form of n unknowns is positive definite if and only if it is reduced to a normal form consisting of n positive squares.

We formulate one of the frequently used criteria for a positive (negative) definite quadratic form.

***Theorem* 27.5.** In order for the quadratic form $F = X'AX$ to be positively (negatively) defined, it is necessary and sufficient that all eigenvalues λ_i of matrix *A* are positive (negative).

***Example* 27.8.** Find out if a quadratic form

$$F = 2x_1^2 - 4x_1x_2 + 5x_2^2$$

Is positive defined.

Solution. Matrix *A* of this quadratic form is

$$A = \begin{pmatrix} 2 & -2 \\ -2 & 5 \end{pmatrix}.$$

We draw up the characteristic equation:

$$\begin{vmatrix} 2-\lambda & -2 \\ -2 & 5-\lambda \end{vmatrix} = \lambda^2 - 7\lambda + 6 = 0.$$

The roots of the characteristic equation $\lambda_1 = 6$, $\lambda_2 = 1$ are positive; therefore, the quadratic form is positive definite.

We formulate one more frequently used criterion for positive definiteness of a quadratic form.

Let $A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}$ be a matrix of quadratic form $F = X'AX$.

The main minors of this matrix are the determinants

$$\Delta_1 = a_{11}, \Delta_2 = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}, ..., \quad \Delta_n = \begin{vmatrix} a_{11} & a_{12} & ... & a_{1n} \\ a_{21} & a_{22} & ... & a_{2n} \\ ... & ... & ... & ... \\ a_{n1} & a_{n2} & ... & a_{nn} \end{vmatrix}$$

***Theorem* 27.6 (Sylvester criterion).** In order for the quadratic form $F = X'AX$ to be positive definite, it is necessary and sufficient that all the principal minors of matrix *A* are positive:

$\Delta_1 > 0, \Delta_2 > 0, ..., \quad \Delta_n > 0$.

In order for the quadratic form to be negative definite, it is necessary and sufficient that the signs of the main minors $\Delta_1, \Delta_2, ..., \quad \Delta_n$ alternate, and $\Delta_1 < 0$.

(We also accept this theorem without proof.)

***Example* 27.9.** Using the Sylvester criterion, verify that the quadratic form

$F = 2x_1^2 - 4x_1x_2 + 5x_2^2$

from example 27.8 is positive defined.

Solution:

$$\Delta_1 = a_{11} = 2 > 0, \Delta_2 = \begin{vmatrix} 2 & -2 \\ -2 & 5 \end{vmatrix} = 6 > 0$$

.

The principal minors of matrix A are positive, therefore, according to the Sylvester criterion, the quadratic form is positive definite.

***Example* 27.10.** Find out if a quadratic form

$F = 5x_1^2 + x_2^2 + 6x_3^2 + 4x_1x_2 - 10x_1x_3 - 4x_2x_3$

is positive defined.

Solution. We calculate the main minors:

$$\Delta_1 = 5,\ \Delta_2 = \begin{vmatrix} 5 & 2 \\ 2 & 1 \end{vmatrix} = 1,\ \Delta_3 = \begin{vmatrix} 5 & 2 & -5 \\ 2 & 1 & -2 \\ -5 & -2 & 6 \end{vmatrix} = 1.$$

The major minors are positive; therefore, this quadratic form is positive definite.

Questions

1. Is the expression $x_1x_2 + x_1x_3 + x_1x_4$ a quadratic form?
2. How is a quadratic matrix determined? Is a quadratic matrix always square?
3. What is called the rank of a quadratic form?
4. How is a quadratic matrix A transformed with a non-degenerate linear transformation C?
5. How is the rank of the quadratic form related to the number of non-zero coefficients in the canonical form to which this form is reduced?
6. What is the law of inertia of quadratic forms?
7. Is the normal type of a quadratic form its canonical type?
8. What is the normal form of a positive definite quadratic form?

Chapter 28. Double and triple integrals

28.1. Basic concepts related to double integral

Consider the integration of functions of two arguments. The results that we get here can be generalized to the case of functions of three or more arguments. Let there be a closed bounded area *D* on the plane Oxy, and let a bounded function $z = f(x, y)$ be given in the area *D*. We divide area *D* by a network of some lines into *arbitrary parts* D_1, D_2, …, D_n that do not have common internal points. For each $i = 1, 2, \ldots, n$ we denote by ΔS_i the area of some partial area D_i. Next, we will do the following: 1) select an arbitrary internal point (ξ_i, η_i) in each partial area D_i and calculate the value of the function at this point; 2) we smartly press this value $f(\xi_i, \eta_i)$ by the area ΔS_i of this partial area D_i; 3) we compose the sum of such products:

$$\sigma = \sum_{i=1}^{n} f(\xi_i, \eta_i)\Delta S_i. \qquad (28.1)$$

Expression (28.1) is called **the integral sum** for a function $f(x, y)$ in the area *D*. (Note that different integral sums correspond to different partitions of the area *D* into partial domains and different points (ξ_i, η_i).)

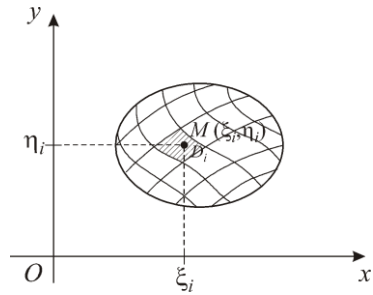


Fig.28.1

For further discussion, we need the concept of the diameter of the area. We call **the diameter** of the area the largest distance between points lying on the boundary of the region. (Note that for a plane closed area bounded by a continuous curve, the diameter is the largest chord). Let denote by $\lambda(\ )$ the largest of all diameters of the partial areas:

$$\lambda = \max diam(D_i),\ i = 1, 2, \ldots, n.$$

Definition. If there exists a finite limit of the integral sum (28.1) for $\lambda \to 0$, which does not depend on the method of dividing the area *D* into partial areas D_i, but on the choice of points (ξ_i, η_i), then this limit is called **the double integral** from the function $f(x,y)$ over the area *D* and is denoted by

$$\iint_D f(x,y)dS \quad \text{или} \quad \iint_D f(x,y)dxdy .$$

A function $f(x,y)$ for which there is an integral over the area D is called **integrable** in the area *D*, and the area *D* – is called **the integration area**.

28.2. Classes of integrable functions

***Theorem* 28.1.** If the function $f(x,y)$ is continuous in a closed bounded area, then it is integrable in this area.

We accept this theorem without proof. Note that a similar theorem was previously formulated in the textbook in chapter 14 for a certain integral of a function of one argument. There, a theorem was formulated on the integrability of a bounded function of a single argument with a finite number of discontinuity points. A similar statement holds for the function of two arguments (only here we are not talking about discontinuity points, but discontinuity lines).

***Theorem* 28.2.** Let a function $f(x,y)$ be bound in a closed bounded area *D* and have discontinuities only on a finite number of lines, which are graphs of continuous functions of the form $y = g(x)$ or $x = h(y)$. Then the function $f(x,y)$ is integrable in the area *D*.

We also accept this theorem without proof.

We see that in the case of two arguments, the class of integrable functions is wider than the class of all continuous functions.

28.3. Geometric sense of double integral

Let a continuous non-negative function $z = f(x, y)$ be defined in the area *D*. Let us consider in space $Oxyz$ the body *V*, bounded below by the area *D*, above by the surface $z = f(x, y)$, and from the sides by a cylindrical surface, with generators parallel to the axis Oz (Fig. 28.2).

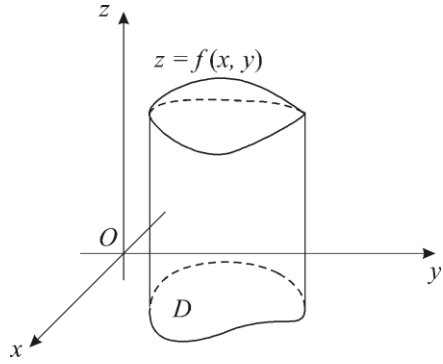


Fig. 28.2

This is **a cylindrical body** (or **a curved cylinder**, or **a cylindroid**).

From the definition of the double integral, it follows that the double integral over the area *D* of a continuous non-negative function $f(x, y)$ is equal to the volume of the cylin-shaped body with base *D* bounded from above by the surface $z = f(x, y)$:

$$V = \iint_D f(x, y)dxdy$$

.

Obviously, the volume of the cylindroid, whose height is equal to unity (i.e., limited by the plane $z = 1$ above) is numerically equal to the area of

the base. Therefore, the double integral of unity is equal to the area of the integration area:

$$S = \iint\limits_D dxdy,$$

where S – is the area of D.

28.4. Double integral properties

One should pay attention to the deep analogy that exists between the concepts of an ordinary (single) definite integral and double integral: in both cases, we consider some function f, only in the first case – the function of one argument $f(x)$ given on the segment $[a,b]$ of the axis Ox, and in the second – the function of two arguments $f(x,y)$ given on a part of the plane Oxy. In both cases, the domain of definition of the function is divided into parts and in each of these parts it is arbitrarily selected at the point at which the value of the function is calculated, and this value is multiplied by the measure of the corresponding partial area. Only in the case of one argument, such a measure was the length Δx_i of the partial segment $[x_{i-1}, x_i]$, and in the case of two arguments – the area ΔS_i of the partial area D_i. Then, in both cases, the integral sum was compiled and the passage to the limit was carried out. Note that the definition of the integral of a function of three or more variables (a triple integral, an n-fold integral) is constructed in the same way.

From the above, it follows that the main properties of the double integral are similar to the properties of a single definite integral (and are proved similarly). Therefore, we restrict ourselves to the wording of some of them.

Property 1. The double integral of the sum of two integrable functions $f(x,y)$ and $g(x,y)$ over the area D exists and is equal to the sum of the double integrals over the area D for each of these functions:

$$\iint\limits_D [f(x,y)+g(x,y)]dxdy = \iint\limits_D f(x,y)dxdy + \iint\limits_D g(x,y)dxdy .$$

Property 2. The constant factor can be taken out of the double integral sign: if $c = \text{const}$, then

$$\iint\limits_D cf(x,y)dxdy = c\iint\limits_D f(x,y)dxdy .$$

Property 3. If the area D is the union of two areas D_1 and D_2 that do not have common internal points, and in each of these regions the function $f(x,y)$ is integrable, then this function is integrable in the area D and the equality holds

$$\iint\limits_D f(x,y)dxdy = \iint\limits_{D_1} f(x,y)dxdy + \iint\limits_{D_2} f(x,y)dxdy .$$

Property 4. If the inequality $f(x,y) \geq 0,$ holds in the whole area of integration, then

$$\iint\limits_D f(x,y)dxdy \geq 0.$$

28.5. Calculation of double integral

To calculate the double integral, it is usually reduced to the repeated one, i.e. to such an integral, which is calculated twice by applying the usual integration process, first according to one of the arguments, then according to the other. This technique is based on the following theorem, which we accept without proof.

***Theorem* 28.3.** Let the function $f(x,y)$ be defined and continuous in the area *D*, which is bounded by the lines $y = y_1(x)$, $y = y_2(x)$, $x = a$, $x = b$, and $y_1(x) \leq y_2(x)$, $a < b$, and the functions $y_1(x)$ and $y_2(x)$ are continuous on the segment $[a,b]$. Then the equality holds

$$\iint\limits_D f(x,y)dxdy = \int\limits_a^b \left(\int\limits_{y_1(x)}^{y_2(x)} f(x,y)\,dy \right) dx \quad . \qquad (28.2)$$

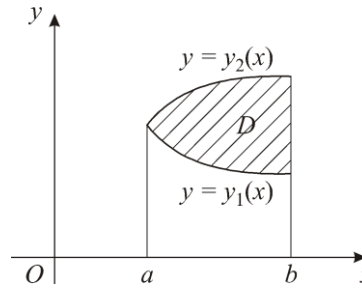


Fig. 28.3

Note that the right-hand side of equality (28.2) is a repeated integral, which is also written in a different form:

$$\int_a^b dx \int_{y_1(x)}^{y_2(x)} f(x,y)dy .$$

Consider the special case of equality (28.2) for $y_1(x) = c = \text{const}$, $y_2(x) = d = \text{const}$:

$$\iint_D f(x,y)dxdy = \int_a^b dx \int_c^d f(x,y)\,dy . \quad (28.3)$$

We also note that if the area *D* is bounded by the lines $x = x_1(y)$, $x = x_2(y)$, $y = c$, $y = d$, where $x_1(y) < x_2(y)$, $c < d$, then the following analogue of equality (2) holds:

$$\iint_D f(x,y)dxdy = \int_c^d \left(\int_{x_1(y)}^{x_2(y)} f(x,y)\,dx \right) dy = \int_c^d dy \int_{x_1(y)}^{x_2(y)} f(x,y)\,dx . \quad (28.4)$$

Let's move on to the examples.

***Example* 28.1.** Calculate the double integral $\iint_D xy\,dxdy$, where the area *D* is bounded by lines $x = 3$, $x = 5$; $y = 0$, $y = 1$, i.e.

$$D = \{(x,y) \mid 3 \le x \le 5;\ 0 \le y \le 1\}.$$

Solution. The area *D* is rectangular. We apply the formula (3):

$$\iint\limits_D xy\,dxdy = \int\limits_3^5 dx \int\limits_0^1 xydy$$.

We calculate the internal integral, assuming x is constant:

$$\int\limits_0^1 xydy = x\frac{y^2}{2}\Big|_{y=0}^{y=1} = \frac{x}{2}$$.

We now calculate the external integral. To do this, integrate the resulting function in the range from 3 to 5:

$$\int\limits_3^5 \frac{x}{2}dx = \frac{x^2}{4}\Big|_3^5 = \frac{25}{4} - \frac{9}{4} = 4$$.

Consequently,

$$\iint\limits_D xy\,dxdy = \int\limits_3^5 dx \int\limits_0^1 xydy = 4$$.

***Example* 28.2.** Calculate the double integral $\iint\limits_D x^2 y\,dxdy$, if the area D is bounded by the lines $x = 0$, $y = 0$, $x^2 + y^2 = 4$, and $x \geq 0$, $y \geq 0$.

Decision. Let's make a figure.

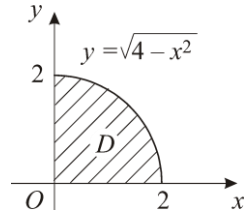


Fig. 4

We see that *D* is a quarter of a circle of radius 2 centered at (0, 0), located in the first quarter. It follows that the domain *D* is bounded on the left and right by the straight lines $x=0$ and $x=1$, and from below and above by the lines $y=0$ and $y=\sqrt{4-x^2}$. Therefore, in accordance with formula (2),

$$\iint\limits_D x^2 y\,dxdy = \int\limits_0^2 dx \int\limits_0^{\sqrt{4-x^2}} x^2 y\,dy.$$

We calculate the internal integral, assuming *x* is constant:

$$\int\limits_0^{\sqrt{4-x^2}} x^2 y\,dy = x^2 \frac{y^2}{4}\Bigg|_{y=0}^{y=\sqrt{4-x^2}} = \frac{x^2\left(4-x^2\right)}{2}.$$

We now calculate the external integral:

$$\int\limits_0^2 \frac{x^2\left(4-x^2\right)}{2}dx = \frac{1}{2}\left(\frac{4x^3}{3}-\frac{x^5}{5}\right)\Bigg|_0^2 = \frac{32}{15}.$$

So,

$$\iint\limits_D x^2 y\,dxdy = \int\limits_0^2 dx \int\limits_0^{\sqrt{4-x^2}} x^2 y\,dy = \frac{32}{15}.$$

***Example* 28.3.** Calculate the double integral $\iint\limits_D (x+y)dxdy$, if the area D is bounded by the lines $y = x$ and $y = 2 - x^2$.

Solution. Find the limits of integration with respect to x. To do this, we find the abscissas of the points of intersection of the lines $y = x$ and $y = 2 - x^2$. The joint solution of these equations gives $x_1 = -2$, $x_2 = 1$. Let's make a figure.

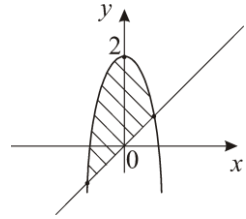


Fig. 28.5

The area D is bounded on the left and on the right by lines $x = -2$ and $x = 1$, on top – by a line $y = 2 - x^2$, and below – by a line $y = x$. We apply the formula (28.2):

$$\iint\limits_D (x+y)dxdy = \int\limits_{-2}^1 dx \int\limits_x^{2-x^2} (x+y)dy.$$

We calculate the integral over dy (assuming x is constant):

$$\int_{x}^{2-x^2}(x+y)dy=\left(xy+\frac{y^2}{2}\right)\Bigg|_{y=x}^{y=2-x^2}=\left[x\left(2-x^2\right)+\frac{\left(2-x^2\right)^2}{2}\right]-$$

$$-\left(x^2+\frac{x^2}{2}\right)=\frac{x^4}{2}-x^3-\frac{7}{2}x^2+2x+2.$$

Now we calculate the external integral:

$$\int_{-2}^{1}\left(\frac{x^4}{2}-x^3-\frac{7}{2}x^2+2x+2\right)dx=\left(\frac{x^5}{10}-\frac{x^4}{4}-\frac{7x^3}{6}+x^2+2x\right)\Bigg|_{-2}^{1}=$$

$$=\frac{101}{60}-\left(-\frac{32}{10}-\frac{16}{4}+\frac{56}{6}+4-4\right)=-\frac{9}{20}.$$

***Example* 28.4.** Calculate the double integral $\iint_D x^2 y\,dxdy$, it the area D is bounded above by an arc of a circle $y=\sqrt{1-x^2}$, and below by segments of lines $y=-x$ at $x<0$ and $y=0$ for $x\geq 0$.

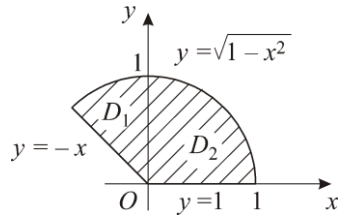


Fig. 28.6

Decision. The lower boundary of the area D consists of segments of two lines $y=-x$ and $y=0$ intersecting at the origin. Therefore, it divides the

area D with the axis Oy into two areas D_1 and D_2 and represents the integral as the sum of two integrals (see property 3):

$$\iint\limits_D x^2 y\,dxdy = \iint\limits_{D_1} x^2 y\,dxdy + \iint\limits_{D_2} x^2 y\,dxdy .$$

Having set the limits of integration in the last two integrals, we obtain

$$\iint\limits_D x^2 y\,dxdy = \int\limits_{-\frac{\sqrt{2}}{2}}^{0} dx \int\limits_{-x}^{\sqrt{1-x^2}} x^2 y\,dy + \int\limits_0^1 dx \int\limits_0^{\sqrt{1-x^2}} x^2 y\,dy =$$

$$= \int\limits_{-\frac{\sqrt{2}}{2}}^{0} \left[\frac{1}{2}x^2\left(1-x^2\right) - \frac{1}{2}x^4\right] dx + \int\limits_0^1 \frac{1}{2}x^2\left(1-x^2\right)dx = \frac{4+\sqrt{2}}{60}.$$

Replacement of variables in double integral

Let the function $f(x,y)$ be continuous in some closed bounded area D, therefore, there exists a double integral

$$\iint\limits_D f(x,y)dxdy ,$$

and let a transition from the variables x, y to the new variables u, v be possible:

$$x = x(u,v),\ \ y = y(u,v). \qquad (28.5)$$

The old variables x, y will be considered the Cartesian coordinates of the current point of one plane Oxy. The new coordinates u, v will be

considered the coordinates of the current point of another plane O^*uv. Thus, we will consider two planes: (x, y) and (u, v), in which the coordinate systems Oxy and O^*uv are given, respectively.

A transformation (28.5) is called **regular** if the following two conditions are satisfied:

1) this transformation of variables establishes a one-to-one correspondence between the points of the area D on the plane Oxy and the points of a certain area D^* on the plane O^*uv:

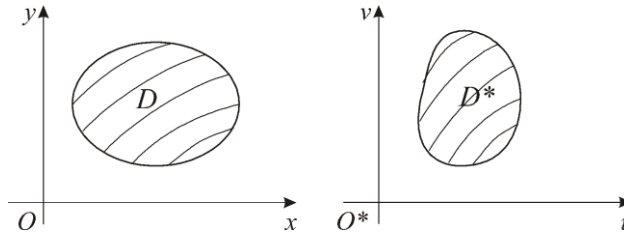


Fig. 28.7

consequently, u and v are determined from formulas (5) uniquely using inverse transformation formulas:

$$u = u(x, y),\ v = v(x, y); \qquad (28.6)$$

2) functions $x = x(u, v),\ y = y(u, v)$ and functions $u = u(x, y),\ v = v(x, y)$ are continuous together with their partial derivatives on the areas D and D^* including their boundaries, and, in addition, their determinant

$$J(u,v)=\begin{vmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} \\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} \end{vmatrix} \qquad (28.7)$$

is nonzero.

We now formulate the rule for converting a double integral to new variables.

***Theorem* 28.4.** Let the function $f(x,y)$ be continuous in a closed bounded area *D*, and let a regular transformation (28.5) be given that maps the area *D* to a closed bounded area *D**. Then the formula for changing variables holds

$$\iint\limits_{D} f(x,y)dxdy = \iint\limits_{D^*} f(x(u,v),y(u,v))\,|J(u,v)|\,dudv ,$$

where $|J(u,v)|$ is the absolute value of the transform determinant

$$J(u,v)=\begin{vmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} \\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} \end{vmatrix}. \qquad (28.7)$$

The proof of this theorem is not given here.

Note that the determinant of transformation (28.5), i.e. $J(u,v)$, is called **Jacobian determinant** or **the Jacobian**.

A conversion of the form called *the transition to polar coordinates*

$x = r\cos\varphi$, $y = r\sin\varphi$ ($0 \le \varphi \le 2\pi$),

is very often used to calculate the double integral. The Jacobian of this transformation is calculated simply:

$$J(r,\varphi)=\begin{vmatrix} \dfrac{\partial x}{\partial r} & \dfrac{\partial x}{\partial \varphi} \\ \dfrac{\partial y}{\partial r} & \dfrac{\partial y}{\partial \varphi} \end{vmatrix} = \begin{vmatrix} \cos\varphi & -r\sin\varphi \\ \sin\varphi & r\cos\varphi \end{vmatrix} = r\left(\cos^2\varphi+\sin^2\varphi\right)=r.$$

Let's look at some examples.

***Example* 5.** Calculate the double integral

$$\iint\limits_D \frac{dxdy}{1+x^2+y^2},$$

if the area D – is a circle: $x^2+y^2\le 1$.

Solution. We apply the transformation $x=r\cos\varphi$, $y=r\sin\varphi$. We get (taking into account the fact that $|J(r,\varphi)|=r$ and $0\le\varphi\le 2\pi$):

$$\iint\limits_D \frac{dxdy}{1+x^2+y^2}=\iint\limits_{D^*}\frac{rdrd\varphi}{1+r^2}=\int\limits_0^{2\pi}d\varphi\int\limits_0^1\frac{rdr}{1+r^2}.$$

Obviously,

$$\int\limits_0^1\frac{rdr}{1+r^2}=\frac{1}{2}\ln\left(1+r^2\right)\Big|_0^1=\frac{1}{2}\ln 2.$$

Therefore, the desired integral is

$$\int_0^{2\pi} \frac{\ln 2}{2} d\varphi = \pi \ln 2$$

.

***Example* 6.** Calculate the integral

$$\iint_D e^{-x^2-y^2} dxdy$$

,

where the area of integration is a circle: $x^2 + y^2 \le R^2$.

Solution. Passing to polar coordinates, we obtain

$$\iint_D e^{-x^2-y^2} dxdy = \int_0^{2\pi} d\varphi \int_0^R e^{-r^2} dr = -\frac{1}{2} \int_0^{2\pi} e^{-r^2} \Big|_0^R d\varphi = \pi\left(1 - e^{-R^2}\right)$$

.

***Example* 7.** Calculate the double integral

$$J = \iint_D e^{-x^2-y^2} dxdy$$

,

if area D – is the entire plane Oxy. Using the result obtained, calculate the integral

$$I = \int_{-\infty}^{+\infty} e^{-x^2} dx$$

.

Solution. We use the result obtained in the previous example, where this double integral is calculated for the case when the area D is a circle

$x^2 + y^2 \le R^2$. If the radius R is unlimitedly increased, then in the limit the area D coincides with the entire plane. Therefore, if we denote the integral calculated in Example 6, by J_R, we obtain

$$J = \lim_{R\to\infty} J_R = \lim_{R\to\infty} \pi\left(1 - e^{-R^2}\right) = \pi .$$

The integral J is an improper double integral. It can be proved that it is equal to the corresponding double integral:

$$J = \int_{-\infty}^{+\infty} dx \int_{-\infty}^{+\infty} e^{-x^2-y^2} dy = \pi .$$

But obviously,

$$\int_{-\infty}^{+\infty} e^{-x^2-y^2} dy = \int_{-\infty}^{+\infty} e^{-x^2} e^{-y^2} dy = e^{-x^2} \int_{-\infty}^{+\infty} e^{-y^2} dy .$$

Since a certain integral does not depend on the designation of the integration variable, then

$$\int_{-\infty}^{+\infty} e^{-y^2} dy = \int_{-\infty}^{+\infty} e^{-x^2} dx = I ,$$

so

$$J = \int_{-\infty}^{+\infty} Ie^{-x^2} dx = I \int_{-\infty}^{+\infty} e^{-x^2} dx = I^2 .$$

Thus,

$I^2 = \pi$.

Consequently,

$$I = \int_{-\infty}^{+\infty} e^{-x^2} dx = \sqrt{\pi}$$

.

This integral[1]* plays a very important role in probability theory and statistics. It should be noted that we could not calculate this integral directly (using the indefinite integral), since the indefinite integral

$$\int e^{-x^2} dx$$

is not expressed in elementary functions.

Questions

1. What is the integral sum for a function $f(x, y)$ in the two-dimensional area *D*?
2. How is the double integral of a function $f(x, y)$ over an area *D* determined?
3. What is the geometric meaning of the double integral?
4. Is it possible to calculate the area of a region using the double integral?
5. What is re-integral?
6. How is the rule for changing variables in the double integral formulated? What is the Jacobian conversion?

[1]* It is usually called the Poisson integral, although it was first calculated by Euler.

7. What is the formula for the transition to polar coordinates in the double integral?

28.6. Triple integrals

The triple integral of a function $f(x,y,z)$ is determined in the same way as the double integral.

Let a bounded function $f(M)=f(x,y,z)$ be given in some closed bounded area V of three-dimensional space. We divide the area V into n arbitrary regions that do not have common points with volumes $\Delta V_1, \Delta V_2, ..., \Delta V_n$. We denote by λ the largest of the diameters of these areas. In each area, we choose an arbitrary point $M_i(\xi_i, \eta_i, \zeta_i)$ and make up the sum

$$\sigma = \sum_{i=1}^{n} f(\xi_i, \eta_i, \zeta_i)\Delta V_i \qquad (28.8)$$

The sum (28.8) is called the integral sum for the function $f(x,y,z)$ over the area V.

Definition. If there is a finite limit I of the sum (1) for $\lambda \to 0$, then this limit is called the triple integral of the function $f(x,y,z)$ over the area V and is denoted by one of the following symbols:

$$I = \iiint_V f(x,y,z)dV = \iiint_V f(x,y,z)dxdydz .$$

In this case, the function $f(x,y,z)$ is called integrable in the area V. Note (without proof) that a continuous function is integrable.

As in the case of double integrals, the calculation of triple integrals reduces to the calculation of integrals of lower multiplicity. Let a three-dimensional area S be bounded, closed, and such that any line parallel to the axis Oz, intersects its boundary at no more than two points whose abscissas $z_1(x,y)$ and $z_2(x,y)$ satisfy the condition $z_1(x,y) \le z_2(x,y)$, and that there is a triple integral for the function $f(x,y,z)$

$$I = \iiint\limits_V f(x,y,z)dxdydz \quad . \qquad (28.9)$$

Suppose, in addition, for any of the x,y from area V_1, that is the projection of the area V onto the plane Oxy, there exists a single integral

$$\int\limits_{z_1(x,y)}^{z_2(x,y)} f(x,y,z)dz \quad .$$

Then there exists a double integral over the area V_1

$$\iint\limits_{V_1} \left[\int\limits_{z_1(x,y)}^{z_2(x,y)} f(x,y,z)dz \right] dxdy$$

And this double integral is equal to the triple integral (28.9).

Now, in the triple integral (28.9), we pass from variables to new variables $u,v,w,$ using the formulas

$$x = x(u,v,w), \quad y = y(u,v,w), \quad z = z(u,v,w). \qquad (28.10)$$

Suppose that the transformations (2) are one-to-one, and denote by S' the region that they translate into S.

Then, if functions (2) have continuous first-order partial derivatives in the area S' and a non-zero determinant

$$J(y,v,w)=\begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} & \frac{\partial x}{\partial w} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} & \frac{\partial y}{\partial w} \\ \frac{\partial z}{\partial u} & \frac{\partial z}{\partial v} & \frac{\partial z}{\partial w} \end{pmatrix},$$

then for the triple integral (1) the formula for changing variables is valid:

$$\iiint\limits_{V} f(x,y,z)dxdydz =$$

$$\iiint\limits_{V'} f\left[x(u,v,w),y(u,v,w),z(u,v,w)\right]\left|J(u,v,w)\right|dudvdw. \tag{28.11}$$

The determinant $J(u,v,w)$ is called the Jacobi determinant or the Jacobian of variables x,y,z in variables $u,v,w.$

The numbers u,v,w are called curvilinear coordinates of a point (x,y,z). In practice, two types of these coordinates are often found.

1. *Cylindrical coordinates.* The cylindrical coordinates of a point (x,y,z) are the numbers $\rho,\varphi,z,$ where ρ and φ - are the polar coordinates of the point (x,y). The transformation is defined by the formulas

$$x=\rho\cos\varphi, y=\rho\sin\varphi, z=z$$

and so the Jacobian of transformation

$$J = J(\rho, \varphi, z) = \begin{vmatrix} \cos\varphi & -\rho\sin\varphi & 0 \\ \sin\varphi & \rho\cos\varphi & 0 \\ 0 & 0 & 1 \end{vmatrix} = \rho$$

and the general formula (28.11) takes the form

$$\iiint\limits_V f(x, y, z)dxdydz = \iiint\limits_{v'} f(\rho\cos\varphi, \rho\sin\varphi, z)\rho d\rho d\varphi dz$$

2. Spherical coordinates. Spherical coordinates r, φ, θ are given by $x = r\cos\varphi\sin\theta$, $y = r\sin\varphi\sin\theta$, $z = r\cos\theta$.

Find the Jacobian

$$J(r, \varphi, \theta) = \begin{vmatrix} \cos\varphi\sin\theta & -r\sin\varphi\sin\theta & r\cos\varphi\cos\theta \\ \sin\varphi\sin\theta & r\cos\varphi\sin\theta & r\sin\varphi\cos\theta \\ \cos\theta & 0 & -r\sin\theta \end{vmatrix} = r^2\sin\theta$$

and formula (3) takes the form

$$\iiint\limits_V f(x, y, z)dxdydz =$$

$$= \iiint\limits_{V'} f(r\cos\varphi\cos\theta, r\sin\varphi\sin\theta, r\cos\theta)r^2\sin\theta drd\varphi d\theta.$$

Chapter 29. First order differential equations and their applications

29.1. Basic definitions

Definition. A differential equation is an equation

$$F\left(x, y, y', ..., y^{(n)}\right) = 0, \tag{29.1}$$

which connects an unknown function y, its independent argument x, and its derivatives y', …, $y^{(n)}$. The **order** of a differential equation is the largest order of the derivative that appears in the equation.

For example, $y' - \frac{2}{x} y = e^x x^2$ is a first-order differential equation and $y'' + 4y = 0$ — a second-order differential equation.

Definition. **The solution of a differential equation** is a function $y = \varphi(x)$ that, when substituted into equation (29.1), turns it into an identity.

In this chapter, we will consider first-order differential equations, i.e. equations of the form

$$F\left(x, y, y'\right) = 0. \tag{29.2}$$

In case it is possible to express y' from equation (29.2), it has the form:

$$y' = f\left(x, y\right). \tag{29.3}$$

Equation (29.3) is called a **first-order equation resolved with respect to the derivative**.

In the theory of differential equations, the main problem is the question of the *existence and uniqueness of the solution*. We present, without proof, a theorem that answers this question.

***Theorem* 29.1 (Cauchy theorem).** Let there be a differential equation (29.3) and let a function $f(x, y)$ and its partial derivative $f'_y(x, y)$ be continuous in some domain *D* of plane *Oxy*. Then in some neighborhood of any inner point $M(x_0, y_0) \in D$, there exists a unique solution of equation (29.3) satisfying the condition $y = y_0$ at $x = x_0$.

The graph of the solution of the differential equation is called the **integral curve**. The domain *D* contains an infinite set of integral curves. The Cauchy's theorem states that, under certain conditions, only one integral curve passes through each inner point of the domain *D*. The conditions that set the value of function *y* at a fixed point x_0 are called **initial conditions** (or **Cauchy conditions**) and are written in the form $y(x_0) = y_0$ or in the form

$$y\big|_{x=x_0} = y_0. \tag{29.4}$$

The problem of finding a solution to equation (29.3) satisfying the condition (29.4) is called the **Cauchy problem**.

Figure 29.1 illustrates theorem 29.1. The entire domain *D* is filled with integral curves, and they can neither intersect nor touch each other.

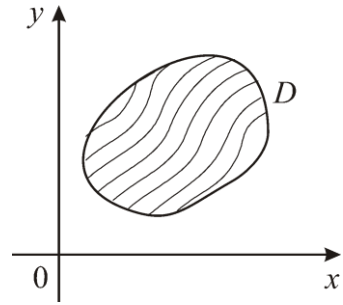


Fig. 29.1. Integral curves

Theorem 29.1 allows us to describe the set of solutions of a differential equation as a general solution.

Definition. The **general solution** of the differential equation (29.2) is the function

$$y = \phi(x, C) y = \phi(x, C) \tag{29.5}$$

depending on x and arbitrary constant *C* if the following conditions hold:
1) for any value of constant C, the function (29.5) is the solution of the differential equation (29.2);

2) no matter what the initial condition (29.4) is, there is a value $C = C_0$, such that function $y = \varphi(x, C_0)$ satisfies this initial condition.

The general solution written in an implicit form: $\Phi(x, y, C) = 0$ is called the **general integral**.

Definition. If constant $C = C_0$ is fixed in the general solution (29.5), then (29.5) is called a **particular solution**.

A particular solution presented implicitly is called a **particular integral**.

To **solve a differential equation** is to find its general solution or general integral.

29.2. Types of first-order differential equations and methods of their solution

Equations with separable variables

A first-order differential equation is called an **equation with separable variables**, if it can be represented as:

$$\frac{dy}{dx} = f(x)g(y) \quad . \tag{29.6}$$

The method for solving this type of equation is called **separation of variables**. We multiply both sides of equation (29.6) by dx and divide by $g(y)$, setting $g(y) \neq 0$:

$$\frac{dy}{g(y)} = f(x)dx \quad . \tag{29.7}$$

This is an **equation with separated variables**. Since the differentials are equal, the indefinite integrals are also equal (more precisely, they differ by constant), therefore

$$\int \frac{dy}{g(y)} = \int f(x)dx + C \quad ,$$

where *C* is an arbitrary constant.

***Example* 29.1.** Solve the equation $xy' - y = 0$.

Solution. Let us separate the variables. To do so, we present the equation in the form

$$x\frac{dy}{dx}=y,$$

we multiply both parts by dx and divide by xy:

$$\frac{dy}{y}=\frac{dx}{x}.$$

Integrating both sides of this equation we obtain

$$\ln|y|=\ln|x|+C_1.$$

We represent an arbitrary constant C_1 in the form $C_1=\ln|C|$, then the general integral will have the form

$$\ln|y|=\ln|x|+\ln|C|.$$

Hence the general solution is $y=\pm Cx$, or, replacing $\pm C$ by C:

$$y=Cx.$$

***Example* 29.2.** Solve the differential equation $y'=-\frac{2xy^2}{x^2-1}$ and find a particular solution that satisfies the initial condition $y(0)=1$.

Solution. Let us separate the variables:

$$\frac{dy}{dx}=-\frac{2xy^2}{x^2-1}$$

$$\frac{dy}{y^2}=-\frac{2xdx}{x^2-1}.$$

By integrating we obtain:

$$\frac{1}{y}=\ln|x^2-1|+C,$$

Hence the general integral is

$$y\left(\ln\left|x^2-1\right|+C\right)=1.$$

Substituting the initial conditions $y(0)=1$: $1\cdot(0+C)=1$ we obtain $C=1$. Therefore,

$$y=\frac{1}{\ln\left|x^2-1\right|+1}.$$

Homogeneous first-order differential equations

Function $f(x,y)$ is called **homogeneous of degree *n* function**, if for any λ

$$f(\lambda x,\lambda y)=\lambda^n f(x,y).$$

For example, the function $f(x,y)=xy-y^2$ is homogeneous of degree 2 since $\lambda x\lambda y-(\lambda y)^2=\lambda^2\left(xy-y^2\right)$, and the function $f(x,y)=\frac{y}{x}$ has degree zero since $\frac{\lambda y}{\lambda x}=\frac{y}{x}$.

A differential equation

$$y'=f(x,y)$$

is called **homogeneous** if $f(x,y)$ is a homogeneous function of degree zero.

This equation can be solved as follows. We transform the right-hand side of the equation by setting $\lambda=\frac{1}{x}$: $f(x,y)=f(\lambda x,\lambda y)=f\left(1,\frac{y}{x}\right)$.

The equation takes the form

$$y' = f\left(1, \frac{y}{x}\right). \qquad (29.8)$$

We do the substitution $u = \frac{y}{x}$, i.e. $y = ux$. Then $y' = u'x + u$. Substituting this expression of the derivative into equation (29.8), we obtain

$$u + x\frac{du}{dx} = f(1, u).$$

This is an equation with separable variables:

$$x\frac{du}{dx} = f(1, u) - u, \quad \text{or} \quad \frac{du}{f(1, u) - u} = \frac{dx}{x}.$$

Hence

$$\int \frac{du}{f(1, u) - u} = \int \frac{dx}{x} + C.$$

Having found the function $u = u(x)$, we must return to the function $y = ux$.

***Example* 29.3.** Solve the equation $y' = \frac{x + y}{x - y}$.

Solution. Let us make sure that the equation is homogeneous: $\frac{\lambda x + \lambda y}{\lambda x - \lambda y} = \frac{x + y}{x - y}$. Changing the variables $\frac{y}{x} = u$ and substituting u into the equation we obtain, given $f(1, u) - u = \frac{1 + u}{1 - u} - u = \frac{1 + u^2}{1 - u}$,

$$\int \frac{1-u}{1+u^2}\,du = \int \frac{dx}{x} + C.$$

Hence

$$\operatorname{arctg} u - \frac{1}{2}\ln\left(1+u^2\right) = \ln|x| + C.$$

Returning to function y, we obtain the general integral:

$$\operatorname{arctg}\frac{y}{x} - \frac{1}{2}\ln\left(1+\left(\frac{y}{x}\right)^2\right) = \ln|x| + C.$$

First-order linear differential equations

The **first-order linear differential equation** is an equation of the form

$$y' + p(x)y = q(x). \tag{29.9}$$

This equation is called linear because the unknown function y and its derivative y' are included into the equation linearly, i.e. they have the first degree, without intermittent multiplication.

One of the methods for solving the linear equation (29.9) is the **Bernoulli method**, which is as follows. We will seek a solution to equation (29.9) in the form $y = u(x)v(x) = uv$. One of these functions can be taken arbitrarily, the other is determined based on equation (29.9). Having made the substitution $y = uv$, we obtain:

$$u'v + uv' + puv = q$$

or

$$u'v + u(v' + pv) = q. \tag{29.10}$$

Choose a function $v = v(x)$, such that

$v' + pv = 0$, т.е. $\frac{dv}{dx} + pv = 0$. (29.11)

Separating the variables, we find

$\frac{dv}{v} = -pdx$.

Integrating, we obtain

$\ln|v| = -\int pdx + \ln|C_1|$, or $v = Ce^{-\int pdx}$,

where $C = \pm C_1$.

Since *any* non-zero solution of equation (29.11) is sufficient for us, we take

$v(x) = \mathrm{e}^{-\int pdx}$ as function $v = v(x)$,

where $\int pdx$ is some primitive.

Substituting the found value $v(x)$ into equation (29.10) we obtain:

$u'v(x) = q(x)$, or $v(x)\frac{du}{dx} = q(x)$.

Hence

$\frac{du}{dx} = \frac{q(x)}{v(x)}$.

We find the general solution for $u = u(x)$:

$u = \int \frac{q(x)}{v(x)}dx + C$.

Substituting u and v into the formula $y = uv$, we finally find

$$y = v(x)\left[\int \frac{q(x)}{v(x)}dx + C\right].$$

Remark. It can be proved that the general solution of equation (29.9) is the sum of any *particular solution* of it and the *general solution* of accompanying it homogeneous equation $y' + p(x)y = 0$.

***Example* 29.4.** Solve the equation $y' - \frac{2}{x}y = 2x^3$.

Solution. Having made a change of variables $y = uv$, we obtain

$$u'v + uv' - \frac{2uv}{x} = 2x^3,$$

$$u'v + u\left(v' - \frac{2v}{x}\right) = 2x^3. \qquad (*)$$

Equate the expression in brackets to zero:

$$v' - \frac{2v}{x} = 0, \qquad \text{whence} \qquad \frac{dv}{v} = \frac{2dx}{x}.$$

We find the function v:

$$\ln v = \ln x^2, \quad v = x^2.$$

Substituting $v = x^2$ into (*), we find *u*:

$$u'x^2 = 2x^3, \qquad \frac{du}{dx} = 2x, \qquad u = x^2 + C.$$

Hence

$$y = \left(x^2 + C\right)x^2.$$

Bernoulli equation

The **Bernoulli equation** is an equation of the form

$$y' + p(x)y = q(x)y^n,$$

where $n \neq 0$, $n \neq 1$.

The Bernoulli equation can also be solved by the Bernoulli method.

***Example* 29.5.** Solve the equation $y' + 2y = y^2 e^x$.

Solution. Let us change the variables as $y = uv$:

$$u'v + uv' + 2uv = (uv)^2 e^x,$$

$$u'v + u(v' + 2v) = (uv)^2 e^x. \qquad (**)$$

Equate the expression in brackets to zero:

$$v' + 2v = 0.$$

We separate the variables:

$$\frac{dv}{v} = -2dx.$$

We find v:

$$\ln v = -2x,$$

$$v = e^{-2x}.$$

Substitute v into the equation (**):

$$u'e^{-2x} = u^2 e^{-4x} e^x,$$

hence

$$u' = u^2 e^{-x}, \qquad \text{i.e.} \qquad \frac{du}{dx} = u^2 e^{-x}.$$

We separate the variables:

$$\frac{du}{u^2} = \mathrm{e}^{-x}dx .$$

We integrate:

$$-\frac{1}{u} = -\mathrm{e}^{-x} + C .$$

We find u (replacing C by $-C$)):

$$u = \frac{1}{\mathrm{e}^{-x} + C} .$$

Therefore, the general solution will be

$$y = \frac{\mathrm{e}^{-2x}}{\mathrm{e}^{-x} + C}, \quad \text{or} \quad y = \frac{1}{\mathrm{e}^{x} + C\mathrm{e}^{2x}} .$$

29.3. Application of differential equations in continuous-time economic models

Consider some examples of applications of differential equations in dynamic problems of economy. The independent variable here is time t. Time in economic dynamics can be considered both continuous and discrete. We consider *continuous* time since in this case, it is possible to use the tools of differential calculus and differential equations.

Let us start with examples of applying the simplest first-order differential equations - *equations with separable variables*.

We consider an equation of the form

$$y' = g(y) \tag{29.12}$$

Obviously, this is a special case of a differential equation with separable variables. Such equations are often found in issues of economic dynamics

(sometimes they are called *autonomous equations*, but in the theory of differential equations this term is not commonly used).

If y^* is a root of the equation $g(y)=0$ ($y^*=\text{const}$), then $y=y^*$ is a solution of equation (29.12). This solution is called **stationary**.

Natural growth model

Let us denote as $y(t)$ the output intensity. It is assumed that the products are sold at a fixed price p and that the market is unsaturated, i.e. all manufactured products are sold out. We call the difference $I=I(t)$ between the total investment and depreciation costs net investment. To increase the output intensity $y(t)$, the net investment *I* must be greater than zero. From the assumption of market unsaturation it follows that as a result of the expansion of production, an increase in income will be obtained, a part of which will again be used to extend the output. This will lead to an increase in output intensity.

It is assumed that the intensity of output y' is directly proportional to the increase in net investment, i.e. the so-called **principle of acceleration** takes place:

$$y'=mI, \tag{29.13}$$

where *m* is the acceleration rate ($m=\text{const}$). Let *a* be the rate of net investment, i.e. part of the income *py*, obtained from the sale of products spent on net investments, $0<a<1$. Then

$$I=apy. \tag{29.14}$$

Substituting the expression I from (29.14) into (29.13), we obtain $y'=\frac{ap}{m}y$. Let us denote $\frac{ap}{m}=k$, then

$$y' = ky \,. \qquad (29.15)$$

Equation (29.15) is an equation with separable variables. We separate the variables:

$$\frac{dy}{y} = kdt \,.$$

Integrating, we find the general solution:

$$\ln|y| = kt + \ln C \,,$$

$$y = Ce^{kt} \,. \qquad (29.16)$$

Let the volume of output y_0 be fixed at the initial moment of time $t = t_0$:

$$y(t_0) = y_0 \,,$$

$$y_0 = Ce^{kt_0} \,.$$

Then we can find the constant C:

$$C = y_0 e^{-kt_0} \,,$$

consequently,

$$y = y_0 e^{k(t-t_0)} \,. \qquad (29.17)$$

Equation (29.17) is called the natural growth equation. This equation also describes the demographic processes, the processes of radioactive decay, the reproduction of bacteria.

Keynes dynamic model

We consider the simplest balance model. Supposing that $Y(t)$ is national income, $E(t)$ is government spending, $S(t)$ is consumption, $I(t)$ is an investment. All these quantities are functions of time *t*.

Let us make up the **balance equations**. First of all, the sum of all *expenses* should be equal to *national income*:

$$Y(t) = S(t) + I(t) + E(t).$$

The total consumption $S(t)$ consists of *domestic consumption* of some of the national income plus *final consumption*. The first term has the form $a(t)Y(t)$, where $a(t)$ is the coefficient of propensity to consume ($0 < a(t) < 1$); the second is denoted by $b(t)$:

$$S(t) = a(t)Y(t) + b(t).$$

Finally, the size of the investment is characterized by the product of the *acceleration rate* $m = m(t)$ and the *marginal national income*:

$$I(t) = K(t)Y'(t).$$

We obtain the system

$$\begin{cases} Y(t) = S(t) + I(t) + E(t), \\ S(t) = a(t)Y(t) + b(t), \\ I(t) = m(t)Y'(t). \end{cases} \qquad (29.18)$$

All functions included into equations (29.18) are positive.

It is assumed that functions $a(t)$, $b(t)$, $m(t)$ and $E(t)$ are given, i.e. they are characteristics of a given state.

It is required to find the dynamics of national income, i.e. find Y as a function of time t.

We substitute the expressions for $S(t)$ from the second equation and $I(t)$ from the third equation of the system (29.18) into the first equation:

$$Y(t) = a(t)Y(t) + b(t) + m(t)Y'(t) + E(t).$$

We express $Y'(t)$:

$$Y'(t) = \frac{1-a(t)}{m(t)}Y(t) - \frac{b(t)+E(t)}{m(t)},$$

or

$$Y'(t) - \frac{1-a(t)}{m(t)}Y = -\frac{b(t)+E(t)}{m(t)}. \qquad (29.19)$$

This is a *linear* differential equation:

$$Y'(t) + p(t)Y = q(t),$$

where $p(t) = -\frac{1-a(t)}{Km(t)}$, $q(t) = -\frac{b(t)+E(t)}{m(t)}$.

We already know the method for finding a general solution to a linear first-order equation (see § 29.2). However, its implementation as applied to equation (29.19) would be very cumbersome. Consider the special case when the main parameters a, b, and m are constant. Then equation (29.19) is simplified:

$$Y' - \frac{1-a}{m}Y = -\frac{b+E}{m}. \qquad (29.20)$$

This is a linear differential equation with *constant* coefficients.

As already noted (see p. 379), the *general solution* of the *inhomogeneous* equation is the sum of its *particular solution* and the *general solution* of the accompanying *homogeneous* equation. As a particular solution $\tilde{Y}$ of equation (29.20), we take the solution obtained for $Y' = 0$, i.e.

$\tilde{Y} = \dfrac{b+E}{1-a}$.

This solution is called **equilibrium**.

Since $E > 0$, $0 < a < 1$, then $\tilde{Y} > 0$.

The general solution of the homogeneous equation $Y' - \dfrac{1-a}{m} Y = 0$ has the form $Y_0 = Ce^{\alpha t}$, where $\alpha = \dfrac{1-a}{m}$. (Obviously, $\alpha > 0$.) Therefore, the general solution of equation (29.20) has the form

$$Y(t) = \frac{b+E}{1-a} + C \exp\left(\frac{1-a}{m} t\right),$$

or

$Y(t) = \dfrac{b+E}{1-a} + Ce^{\alpha t}$, where $\alpha = \dfrac{1-a}{m}$.

If at the initial moment $Y_0 < \tilde{Y}$, then $C = Y_0 - \tilde{Y} < 0$ and the national income decreases with time under the fixed parameters *a*, *b*, *m* and *E*. If $Y_0 > \tilde{Y}$, then $C > 0$ and the national income grows.

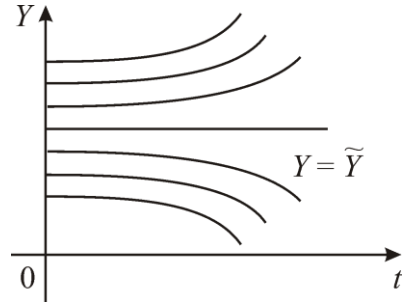


Fig. 29.2. Integral curves of equation (29.20)

Samuelson equation

The **samuelson equation** is the equation:

$$p' = k\left[D(p) - S(p)\right]. \tag{29.21}$$

Here $D(p)$ and $S(p)$ are, respectively, the value of supply and demand at the price $p, k > 0$. Equation (29.21) models the relationship between the change in price *p* and the unmet demand $D(p) - S(p)$.

We consider the simple case when supply and demand are defined by linear functions:

$$D(p) = a - bp, \qquad S(p) = m + np,$$

where *a*, *b*, *m*, *n* are some positive numbers. At the same time, obviously, $a > m$ since at zero price demand exceeds supply. In this case, equation (29.21) has the form

$$p' = k(a - m) - k(n + b)p. \tag{29.22}$$

Equation (29.21), and therefore, (29.22), is a linear differential equation. Find a *solution* to the *homogeneous* equation corresponding to equation (29.22):

$$p' = -k(n+b)p,$$

or

$$\frac{dp}{dt} = -k(n+b)p.$$

We separate variables and integrate:

$$\frac{dp}{p} = -k(n+b)dt,$$

$$\ln|p| = -k(n+b)t + \ln|C|,$$

$$p(t) = Ce^{-k(n+b)t}. \qquad (29.23)$$

Like in the previous case, we can use the *equilibrium* solution $p(t) = \tilde{p} = \text{const}$ as a *particular solution* to equation (29.22), where $\tilde{p}$ is the root of the equation $D(p) - S(p) = 0$ (in this case $p' = 0$). From (29.22) we find

$$\tilde{p} = \frac{a-m}{n+b}.$$

We obtain the general solution of equation (29.22):

$$p(t) = \frac{a-m}{n+b} + Ce^{-k(n+b)t}. \qquad (29.24)$$

Questions

1. What is called a differential equation?
2. What is the order of a differential equation?
3. What is called a solution of a differential equation?
4. How is the Cauchy problem formulated for a first-order differential equation?
5. What is the general solution of a first-order differential equation?
6. What does solving a differential equation mean?
7. What equation is called a differential equation with separable variables?
8. What function is called a homogeneous function of degree n?
9. What kind of a first-order differential equation is called homogeneous?
10. What kind of a first-order differential equation is called linear?
11. What is the Bernoulli method for solving the differential equation? Which first-order differential equations does it usually apply for?
12. What is called the natural growth equation? What processes does this equation describe?
13. What is the Keynes dynamic model?
14. What does the Samuelson equation look like? What is the meaning of its values?

Chapter 30. Differential equations of the second and higher orders

30.1. Basic definitions

A differential equation of the second order has the form:

$$F(x, y, y', y'') = 0 \tag{30.1}$$

or

$$y'' = f(x, y, y'). \tag{30.2}$$

Conditions

$$y(x_0) = y_0,\ y'(x_0) = y_0' \tag{30.3}$$

are called the **initial conditions**.

Definition. Function $y = \varphi(x, C_1, C_2)$ is called the **general solution** of equation (30.1) if it is a solution of equation (30.1) for any values C_1 and C_2, and if for any initial conditions (30.3) there are unique values of constants $C_1 = C_1^0$, $C_2 = C_2^0$, such that the function $y = \varphi(x, C_1^0, C_2^0)$ satisfies these initial conditions.

Definition. Any function $y = \varphi(x, C_1^0, C_2^0)$ obtained from the general solution $y = \varphi(x, C_1, C_2)$ of equation (30.1) for certain constant values $C_1 = C_1^0$, $C_2 = C_2^0$ is called a **particular solution**.

In some cases, solving a second-order differential equation can be reduced to sequential solving of two first-order differential equations.

30.2. Differential equations allowing reduction of order

1. *The equation does not explicitly contain the desired function y*, i.e. has the form $\boldsymbol{F(x, y', y'') = 0}$. In this case, it is sufficient to substitute $y' = z$. Then $y'' = z'$ and the equation takes the form:

$F(x, z, z') = 0$,

i.e. it is a first order equation with respect to *z*.

Let us find the general solution

$z = \varphi(x, C_1)$.

We make the reverse substitution

$y' = \varphi(x, C_1)$.

Hence

$y = \int \varphi(x, C_1)dx + C_2$.

***Example* 30.1.** Solve the equation $xy'' + y' = 0$.

Solution. Let $z = y'$. Then $y'' = z'$, and the original equation has the form

$xz' + z = 0$, or $x\dfrac{dz}{dx} + z = 0$,

whence

$\dfrac{dz}{z} = -\dfrac{dx}{x}$.

Integrating we obtain

$z = \dfrac{C_1}{x}$, or $y' = \dfrac{C_1}{x}$.

Solving the last equation, we obtain:

$y = C_1 \ln|x| + C_2$.

2. *The equation does not explicitly contain argument x*, i.e. has the form $\boldsymbol{F(y, y', y'') = 0}$. In this case, the order of the equation can be reduced letting $y' = z = z(y)$. Then

$$y'' = z'_x = z'_y y'_x = z z'_y = z \frac{dz}{dy}.$$

***Example* 30.2.** Solve the equation $y'' - (y')^2 = 0$.

Solution. By the substitution $y' = z = z(y)$ we reduce this equation to a first-order equation:

$$z \frac{dz}{dy} - z^2 = 0 \quad \text{or} \quad z\left(\frac{dz}{dy} - z\right) = 0.$$

The first solution of this equation is $z = 0$ or $y = C$, where $C = \text{const}$. Next we obtain

$$\frac{dz}{dy} - z = 0.$$

Separating the variables and integrating, we obtain

$z = C_1 e^y$.

We make the reverse substitution:

$$\frac{dy}{dx} = C_1 e^y.$$

Variables are also separated here:

$$e^{-y}dy = C_1 dx,$$

$$-e^{-y} = C_1 x + C_2.$$

Since C_1 and C_2 are arbitrary constants, we can write (taking $-C_1$ instead of C_1, and $-C_2$ instead of C_2):

$$e^{-y} = C_1 x + C_2,$$

$$y = -\ln(C_1 x + C_2).$$

Obviously, this solution also includes the solution $y = C$ obtained above.

3. The equation has the form $y''' = f(y)$. This is a particular case of the equation considered in Sec. 2. Therefore, it is solved by substituting $y' = z = z(y)$, $y''' = z\dfrac{dz}{dy}$. As a result of such substitution, this equation is converted into a first order equation:

$$z\frac{dz}{dy} = f(y).$$

Hence

$$zdz = f(y)dy.$$

Integrating, we obtain

$$\frac{z^2}{2} = \int f(y)dy + C_1.$$

Hence

$$z = \pm\sqrt{2\left(C_1 + \int f(y)dy\right)},$$

i.e.

$$\frac{dy}{dx}=\pm\sqrt{2\left(C_1+\int f(y)dy\right)} \quad \text{or} \quad \frac{dy}{\sqrt{2\left(C_1+\int f(y)dy\right)}}=\pm dx .$$

Integrating the left and right sides of the last equality, we obtain the general integral.

Example **30.3.** Find a particular solution: $2y^3y''=1$, $y\left(\frac{1}{2}\right)=1$, $y'\left(\frac{1}{2}\right)=1$.

Solution. Making the substitution $y''=z\frac{dz}{dy}$, we find

$$y^3z\frac{dz}{dy}=1 .$$

Hence

$$z\frac{dz}{dy}=\frac{1}{y^3}, \qquad zdz=\frac{dy}{y^3} .$$

Integrating, we obtain

$$z^2=-\frac{1}{2y^2}+C_1, \qquad z=\pm\sqrt{-\frac{1}{2y^2}+C_1} .$$

Assuming here $x=\frac{1}{2}$ and considering that at this value x we have $y'=z=1$, we see that we need to take the plus sign in front of the radical, then we find $C_1=\frac{3}{2}$. Thus,

$$z=\sqrt{\frac{3}{2}-\frac{1}{2y^2}}$$

or, which is the same,

$$\frac{dy}{dx}=\sqrt{\frac{3}{2}-\frac{1}{2y^2}}, \quad \text{or} \quad \frac{dy}{dx}=\sqrt{\frac{3y^2-1}{2y^2}}.$$

Hence

$$\frac{\sqrt{2}ydy}{\sqrt{3y^2-1}}=dx.$$

The integral from the left side is taken by the substitution $t=3y^2-1$, $dt=6ydy$.

$$\frac{\sqrt{2}}{3}\sqrt{3y^2-1}=x+C_2.$$

At $x=\frac{1}{2}$ we find $\frac{\sqrt{2}}{3}\cdot\sqrt{2}=\frac{1}{2}+C_2$, or $\frac{2}{3}=\frac{1}{2}+C_2$, whence $C_2=\frac{1}{6}$. We obtain

$$\frac{\sqrt{2}}{3}\sqrt{3y^2-1}=x+\frac{1}{6}.$$

Squaring and making obvious transformations, we finally obtain

$$2y^2 = 3x^2 + x + \frac{3}{4}.$$

30.3. Linear differential equations of order n

A **linear differential equation of order n** is an equation of the form

$$a_0(x)y^{(n)} + a_1(x)y^{(n-1)} + ... + a_{n-1}(x)y' + a_n(x)y = F(x). \quad (30.5)$$

This equation is called linear, because the unknown function *y* and its derivatives y', y'', …, $y^{(n)}$ are included into it linearly, i.e. in the first degree, not multiplying among themselves. Here $a_0(x)$, $a_1(x)$, …, $a_{n-1}(x)$, $a_n(x)$, $F(x)$ are given functions of x (in particular, they can be constant), and for all values of *x* from the domain in which we consider equation (30.4) (otherwise the order of the equation would not be equal to *n*) Therefore, we can divide both sides of the equation by $a_0(x)$ and transform it to the form

$$y^{(n)} + p_1(x)y^{(n-1)} + ... + p_{n-1}(x)y' + p_n(x)y = f(x), \quad (30.5)$$

where $p_1(x) = \frac{a_1(x)}{a_0(x)}$, …, $p_{n-1}(x) = \frac{a_{n-1}(x)}{a_0(x)}$, $f(x) = \frac{F(x)}{a_0(x)}$.

In what follows, we will write a linear differential equation in the form (30.5). The function $f(x)$ in equation (30.6) is called the **free term**. If $f(x)$ identically equals zero, then equation (30.6) is called **homogeneous**; in this case, it obviously has the form

$$y^{(n)} + p_1(x)y^{(n-1)} + \ldots + p_{n-1}(x)y' + p_n(x)y = 0. \quad (30.6)$$

Otherwise, equation (30.5) is called **inhomogeneous**. The function $f(x)$ in (30.5) is called the **right-hand side** (or **free term**) of equation (30.6). In what follows, we will present the theory and carry out the proofs, as a rule, for second-order equations, since here we can study all the main laws of interest to us. So, in what follows we will mainly deal with equations

$$y'' + p(x)y' + q(x)y = f(x). \quad (30.7)$$

First of all, we establish some basic properties of linear homogeneous equations.

Structure of the general solution of a homogeneous linear differential equation

Let us consider an equation of the form

$$y'' + p(x)y' + q(x)y = 0. \quad (30.8)$$

In the future, we will see that in order to be able to solve equation (30.7), in which $f(x) \neq 0$, we must also be able to solve equation (30.8). We consider two simple **properties of solutions of equation** (30.8).

1. If y_0 is a solution to equation (30.8), and C is a constant, then *product* Cy_0 is also a solution to this equation.

Proof. Substitute $y = Cy_0$ into equation (30.8). Since

$y' = Cy_0'$, $y'' = Cy_0''$,

the left side as a result of the substitution looks like

$Cy_0'' + p(x)C\,y_0' + q(x)Cy_0$,

or, which is the same,

$C\left(y_0'' + p(x)y_0' + q(x)y_0\right)$.

Since y_0 is a solution of differential equation (30.8), the expression in brackets is identically equal to zero. Thus, equation (30.8) turned into an identity. The statement is proved.

2. If y_1 and y_2 are solutions of differential equation (30.8), then their *sum* $y_1 + y_2$ is also a solution to this equation.

Proof. Since

$y' = y_1' + y_2'$, $\qquad y'' = y_1'' + y_2''$

and since y_1 and y_2 are solutions of equation (30.8), the following identity equalities hold:

$y_1'' + p(x)y_1' + q(x)y_1 = 0$, $\qquad y_2'' + p(x)y_2' + q(x)y_2 = 0$. (*)

Substituting the sum $y_1 + y_2$ into equation (30.8) and taking into account the identities (*), we obtain

$$\left(y_1 + y_2\right)'' + p(x)\left(y_1 + y_2\right)' + q(x)\left(y_1 + y_2\right) =$$
$$= \left(y_1'' + p(x)y_1' + q(x)y_1\right) + \left(y_2'' + p(x)y_2' + q(x)y_2\right) = 0 + 0 = 0.$$

So, equation (30.8) turned into an identity. The statement is proved.

Definition. Two functions y_1 and y_2 are called **linearly independent** if the identity equality

$$k_1y_1 + k_2y_2 = 0 \qquad (30.8^*)$$

has the only possible solution

$k_1 = k_2 = 0$.

If there exists a nonzero solution (**), then functions y_1 and y_2 are called **linearly dependent**.

Obviously, functions y_1 and y_2 are linearly independent if and only if their relation is not constant:

$$\frac{y_1}{y_2} \neq \text{const}$$

.

Definition. If $y_1 = y_1(x)$, $y_2 = y_2(x)$, then the determinant

$$W(x) = \begin{vmatrix} y_1 & y_2 \\ y_1' & y_2' \end{vmatrix} = y_1 y_2' - y_2 y_1'$$

is called the Wronski determinant of these functions.

***Lemma* 30.1.** If the functions $y_1 = y_1(x)$ and $y_2 = y_2(x)$ are *linearly dependent* on the segment $[a, b]$, then the Wronski determinant, composed of them, is identically equal to zero on this segment; if the functions are linearly independent on $[a, b]$, then the Wronski determinant is nonzero on $[a, b]$.

Proof. Let functions y_1 and y_2 be linearly dependent on the segment $[a, b]$. Then these functions are proportional on $[a, b]$, i.e. $y_1 = ky_2'$. Therefore, the determinant $W(x)$ contains proportional columns, therefore it is equal to zero on the segment $[a, b]$:

$$W(x) = \begin{vmatrix} y_1 & y_2 \\ y_1' & y_2' \end{vmatrix} = \begin{vmatrix} y_1 & ky_1 \\ y_1' & ky_1' \end{vmatrix} = 0$$

.

The first part of the lemma is proved.

We prove the second part of the lemma by contradiction. Let functions $y_1 = y_1(x)$ and $y_2 = y_2(x)$ be linearly independent on the segment $[a, b]$; assume that the determinant $W(x)$ is identically equal to zero on this segment. Then its columns are necessary proportional: $y_2 = ky_1$, $y_2' = ky'$. But this means that functions y_1 and y_2 are proportional, and therefore, linearly dependent, which contradicts the condition of the lemma. The proof is complete.

***Lemma* 30.2.** If the Wronski determinant $W(x)$, composed for solutions y_1 and y_2 of the homogeneous linear equation (30.8) is not equal to zero for some value $x = x_0$ on the segment $[a, b]$, and the coefficients of the equation are continuous on this interval, then $W(x)$ does not vanish at any value x on this interval.

Proof. Since $y_1 = y_1(x)$ and $y_2 = y_2(x)$ are solutions of equation (30.8), the following identities hold:

$$y_1'' + p(x)y_1' + q(x)y_1 = 0, \qquad y_2'' + p(x)y_2' + q(x)y_2 = 0.$$

We multiply both sides of the second equality by y_1, and both sides of the first - by y_2, and subtract the first from the second. We obtain:

$$(y_1y_2'' - y_1''y_2) + p(x)(y_1y_2' - y_1'y_2) = 0. \qquad (30.9)$$

The difference in the second bracket is the Wronski determinant $W(x)$. Indeed, $W(x) = y_1y_2' - y_1'y_2$. We differentiate $W(x)$:

$$W'(x) = (y_1y_2' - y_1'y_2)' = y_1'y_2' + y_1y_2'' - y_1''y_2 - y_1'y_2' = y_1y_2'' - y_1''y_2.$$

As we can see, the difference in the first bracket (30.9) is a derivative of the Wronski determinant, therefore, equation (30.9) can be represented as

$$W' + p(x)W = 0, \tag{30.10}$$

i.e. is a differential equation with separable variables. We find a solution of this equation that satisfies the condition $W(x_0) = W_0$, supposing $W(x_0) \neq 0$. Separating the variables, we obtain

$$\frac{dW}{W} = -p(x)dx.$$

Integrating, we find

$$\ln W = -\int_{x_0}^{x} p(x)dx + \ln C.$$

Hence,

$$\ln \frac{W}{C} = -\int_{x_0}^{x} p(x)dx,$$

and we obtain the general solution of equation (30.10):

$$W = C \exp\left(-\int_{x_0}^{x} p(x)dx \right). \tag{30.11}$$

(Recall that $\exp a$ means e^a for any *a*.)
Formula (30.11) is called the **Liouville formula**.

We now define constant C so that the initial condition $W(x_0) = W_0$ is satisfied. We substitute $x = x_0$ into the left and right sides of the equality

$$\int_{x_0}^{x_0} p(x)dx = 0$$

(30.11). We obtain (given that):

$$W_0 = C.$$

We substitute the found value $C = W_0$ into equality (30.11). So, the solution to equation (30.10), satisfying the initial conditions $W(x_0) = W_0$, has the form

$$W = W_0 \exp\left(-\int_{x_0}^{x_0} p(x)dx \right).$$

Then (since the exponential function does not vanish at any value of the argument, and since $W_0 \neq 0$ by hypothesis) it follows from the last equality that $W \neq 0$ at no value of x. The proof is complete.

From the previously proved properties 1. and 2. of the solutions to the homogeneous linear differential equation (30.8) it follows that the *linear combination of solutions* $y_1(x)$ and $y_2(x)$ equation (30.8), i.e.

$$y = C_1 y_1(x) + C_2 y_2(x),$$

where C_1 and C_2 are constants, is also a solution to equation (30.8).

Let us now formulate and prove the theorem that describes the **structure of the general solution of a homogeneous linear differential equation**.

***Theorem* 30.1.** If $y_1(x)$ and $y_2(x)$ are linearly independent on $[a,b]$ particular solutions of the homogeneous linear differential equation (30.8), then the general solution of this equation has the form

$$y = C_1 y_1(x) + C_2 y_2(x), \tag{30.12}$$

where C_1 and C_2 are arbitrary constants.

Proof. It was previously established that a linear combination (30.12) is a *solution* of equation (30.8); it must be proved that it is the *general solution*, i.e. it must be shown that for any initial conditions there are such values of constants C_1 and C_2, for which this linear combination is a solution satisfying these initial conditions.

Let us take any number $x_0 \in [a,b]$ and any numbers y_0, y_0' and make up the initial conditions:

$$y(x_0) = y_0,\ y'(x_0) = y_0'.$$

The fulfillment of these conditions for the function (30.12) means that

$$\begin{cases} C_1 y_1(x_0) + C_2 y_2(x_0) = y_0, \\ C_1 y_1'(x_0) + C_2 y_2'(x_0) = y_0'. \end{cases}$$

We obtained a linear system of two equations with respect to unknown constants C_1 and C_2. The determinant of this system is Wronski determinant $W(x_0)$, and since y_1 and y_2 are linearly independent, this determinant is not equal to zero. Therefore, the system has a unique solution for any values of the right-hand sides y_0 and y_0':

$$C_1 = C_1^0,\ C_2 = C_2^0.$$

Substituting these values in the solution (30.12), we obtain a particular solution satisfying the given initial conditions. Since the initial conditions are given arbitrarily, we can state that solution (30.12) is a general solution to equation (30.8). The theorem is proved.

Remark. If we discard the condition for the linear dependence of the functions y_1 and y_2, then the function (30.12), although it remains the *solution* of the differential equation (30.8), will no longer be its *general solution.*

Indeed, let, for example,

$$\frac{y_1}{y_2} = 5 .$$

Then $y_1 = 5y_2$ and (30.12) has the form

$$y = C_1 y_1 + C_2 y_2 = 5C_1 y_2 + C_2 y_2 = (5C_1 + C_2) y_2 = C y_2$$

where $5C_1 + C_2 = C$. In other words, in case functions y_1 and y_2 are *linearly dependent*, the number of arbitrary constants in (30.12) can be reduced to one by introducing new notation, and a function containing one arbitrary constant cannot be the general solution of differential equation (30.8).

The meaning of the theorem proved (30.1) is that it reduces the problem of finding the general solution of differential equation (30.8) to a simpler problem, which is the problem of finding two linearly independent *particular* solutions of this equation. We will deal with this last task now, but we will restrict ourselves to the simplest case when the coefficients of the equation are constant: $p(x) = p = \text{const}$, $q(x) = q = \text{const}$.

Homogeneous linear differential equations with constant coefficients

We consider an equation of the form

$$y'' + py' + qy = 0, \tag{30.13}$$

where p and q are some constants.

Let us first find the general solution of equation (30.13). By Theorem 30.1, in order to do this, it is necessary to find two linearly independent particular solutions of equation (30.13).

We will seek for a solution of equation (30.13) in the form

$$y = \mathrm{e}^{kx}. \tag{30.14}$$

Since

$$y' = k\mathrm{e}^{kx},\ y'' = k^2\mathrm{e}^{kx},$$

substituting (30.14) into the left-hand side of (30.13), we obtain

$$k^2\mathrm{e}^{kx} + pk\mathrm{e}^{kx} + q\mathrm{e}^{kx} = 0, \quad \text{or} \quad \mathrm{e}^{kx}\left(k^2 + pk + q\right) = 0.$$

So, for (30.14) to be a solution of equation (30.13), k must be a root of the quadratic equation

$$k^2 + pk + q = 0. \tag{30.15}$$

This equation is called the *characteristic equation* for the differential equation (30.13).

Three cases are possible.

1. The roots (30.15) are *real* and *different*: $k_1 = a$, $k_2 = b$, $a \neq b$. Then (30.13) has two solutions:

$$y_1 = \mathrm{e}^{ax},\ y_2 = \mathrm{e}^{bx}.$$

These solutions are linearly independent since

$\frac{e^{ax}}{e^{bx}} \neq \text{const}$.

The *general solution* of equation (30.13) is as follows:

$y = C_1 e^{ax} + C_2 e^{bx}$.

***Example* 30.4.** Solve the equation $y'' - 5y' + 4y = 0$.

Solution. We compose the characteristic equation:

$k^2 - 5k + 4 = 0$.

We find the roots: $k_1 = 1$, $k_2 = 4$, so the general solution to this equation is

$y = C_1 e^{x} + C_2 e^{4x}$.

2. The roots of equation (30.15) are *real* and *coincident*: $k_1 = k_2 = a$ ($a = -\frac{p}{2}$). One of particular solutions of equation (30.13) will be

$y_1 = C_1 e^{ax}$.

However, we cannot find a second solution y_2 yet, such that y_1 and y_2 are linearly independent.

However, in this case, it turns out that, along with the solution

$y_1 = e^{ax}$,

equation (30.13) has a solution

$y_2 = xe^{ax}$.

We will verify this. In order for the characteristic equation $k^2 + pk + q = 0$ to have equal roots, which are expressed in the form

$k_{1,2} = -\frac{p}{2} \pm \sqrt{\frac{p^2}{4} - q}$, it is necessary for the discriminant to be equal to zero: $\frac{p^2}{4} - q = 0$. In other words, the condition of coincidence of the roots is the equality

$$\frac{p^2}{4} = q.$$

Then $k_1 = k_2 = -\frac{p}{2}$, and one of the solutions of equation (30.13) will be the function

$$y_1 = \mathrm{e}^{-\frac{p}{2}x}.$$

Let us make sure that the solution of equation (30.13) will also be

$$y_2 = x\mathrm{e}^{-\frac{p}{2}x}.$$

We find derivatives y_2' and y_2'':

$$y_2' = \mathrm{e}^{-\frac{p}{2}x} - \frac{p}{2}x\mathrm{e}^{-\frac{p}{2}x}; \qquad y_2'' = -p\mathrm{e}^{-\frac{p}{2}x} + \frac{p^2}{4}x\mathrm{e}^{-\frac{p}{2}x}.$$

We substitute y_2 and its derivatives into the left side of equation (30.13)

$$-pe^{-\frac{p}{2}x} + \frac{p^2}{4}xe^{-\frac{p}{2}x} + p\left(e^{-\frac{p}{2}x} - \frac{p}{2}e^{-\frac{p}{2}x}\right) + qxe^{-\frac{p}{2}x}.$$

Having made the obvious transformations, we obtain

$$\left(q-\frac{p^2}{4}\right)x\mathrm{e}^{-\frac{p}{2}x}.$$

This expression is equal to zero (since $q-\frac{p^2}{4}=0$), and the statement is proved.

So, $y_1=\mathrm{e}^{-\frac{p}{2}x}$ and $y_2=x\mathrm{e}^{-\frac{p}{2}x}$ So, and are two solutions of equation (30.13) in case the roots of the characteristic equation coincide. Linear independence of y_1 and y_2 is obvious. Therefore, the general solution of differential equation (30.13) has the form

$$y=C_1\mathrm{e}^{-\frac{p}{2}x}+C_2x\mathrm{e}^{-\frac{p}{2}x}.$$

So, we note once again that the following **statement** is true: if the characteristic equation has coinciding roots $k_1=k_2=a$, then along with the function

$$y_1=\mathrm{e}^{ax}$$

the solution of differential equation (30.13) is also the function

$$y_2=x\mathrm{e}^{ax}.$$

Then the *general solution* of equation (30.13) is the function

$$y=C_1\mathrm{e}^{ax}+C_2x\mathrm{e}^{ax}. \qquad (30.16)$$

It should be noted that in case the characteristic equation has two *different* roots $k_1=a$, $k_2=b\neq a$, the function $y=x\mathrm{e}^{ax}$ will not be a solution of the differential equation (30.13).

***Example* 30.5.** Solve the equation $y''-4y'+4y=0$.

Solution. The characteristic equation $k^2 - 4k + 4 = 0$ has two coinciding roots $k_1 = k_2 = 2$. Therefore, the general solution of the equation is as follows:

$$y = C_1 \mathrm{e}^{2x} + C_2 x \mathrm{e}^{2x}.$$

3. The roots of the characteristic equation are *complex conjugates*: $k_{1,2} = \alpha \pm \beta i$, where *i* is the imaginary unit, $i^2 = -1$. In this case, we can prove that the *general solution* of equation (30.13) is

$$y = \mathrm{e}^{\alpha x}\left(C_1 \cos \beta x + C_2 \sin \beta x\right).$$

(We accept this statement without proof.)

***Example* 30.6.** Solve the equation $y'' - 4y' + 13y = 0$.

Solution. The characteristic equation $k^2 - 4k + 13 = 0$ has the roots $k_{1,2} = 2 \pm 3i$. The general solution to this equation is

$$y = \mathrm{e}^{2x}\left(C_1 \cos 3x + C_2 \sin 3x\right).$$

30.4. Structure of the general solution of an inhomogeneous linear differential equation

We now move to an inhomogeneous linear differential equation

$$y'' + p(x)y' + q(x)y = f(x). \qquad (30.17)$$

Along with it, we consider the homogeneous linear equation with the same left-hand side, i.e the equation

$$y'' + p(x)y' + q(x)y = 0. \qquad (30.18)$$

Equation (30.18) is called the **corresponding** or **accompanying** equation of the differential equation (30.17).

***Theorem* 30.2.** If $\tilde{y}$ is a particular solution of the differential equation (30.17), and y_0 is the general solution of its accompanying equation (30.18), then their sum

$$y = y_0 + \tilde{y} \tag{30.19}$$

is the general solution of the differential equation (30.17).

(In other words, the *general solution* of an *inhomogeneous* linear differential equation is the sum of its *particular solution* and the *general solution* of the corresponding *homogeneous* equation.)

Proof. 1. Let $\tilde{y}$ be a particular solution of the inhomogeneous equation (30.17), and $y_0 = C_1 y_1 + C_2 y_2$ be the general solution of the accompanying homogeneous equation. First, we make sure that the function

$$y = y_0 + \tilde{y}$$

is a *solution* of equation (30.17). We substitute the function (30.19) into the left side of the equation (30.17):

$$y_0'' + \tilde{y}'' + p(x)(y_0' + \tilde{y}') + q(x)(y_0 + \tilde{y}).$$

Regrouping the terms, we obtain

$$[y_0'' + p(x)y_0' + q(x)y_0] + [\tilde{y}'' + p(x)\tilde{y}' + q(x)\tilde{y}].$$

Since y_0 is a solution of equation (30.18), the expression in the first square brackets is zero. Since $\tilde{y}$ is a solution of equation (30.17), the expression in the second square brackets is equal to $f(x)$. So, substituting (30.19) into equation (30.17), we obtain the identity $f(x) = f(x)$.

Therefore, function (30.19) is indeed a *solution* of differential equation (30.17).

2. Now we need to make sure that function (30.19) is a general solution of the nonhomogeneous equation (30.17). Let y be any solution of the inhomogeneous equation (30.17), and let $\tilde{y}$ be the solution of the same equation (30.17).

Consider the difference $y-\tilde{y}$. We will show that this difference is a solution of the homogeneous equation (30.18). In order to do so, we substitute it into the left side of equation (30.18) and group the corresponding terms:

$$\begin{aligned}&(y-\tilde{y})'' + p(x)(y-\tilde{y})' + q(x)(y-\tilde{y}) = \\ &= [y'' + p(x)y' + q(x)y] - [\tilde{y}'' + p(x)\tilde{y}' + q(x)\tilde{y}] = f(x) - f(x) = 0.\end{aligned}$$

Therefore, this difference is a particular solution of the homogeneous equation (30.18), and this solution can be written in the form

$$y-\tilde{y} = C_1^0 y_1 + C_2^0 y_2,$$

where C_1^0 and C_2^0 are the corresponding values of constants C_1 and C_2 in the formula for the general solution of the homogeneous equation.

We have proved that any solution of equation (30.17) can be obtained by formula (30.19) by appropriate selection of the constants C_1 and C_2. Therefore, function (30.19) is a general solution of the inhomogeneous linear differential equation (30.17). The proof is complete.

While proving Theorem 30.2, we proved the following **properties of solutions of linear differential equations**:

1. If $\tilde{y}$ is a solution of the inhomogeneous differential equation (30.17), and y_0 is a solution of the accompanying homogeneous equation (30.18), then their sum is a solution of the inhomogeneous equation (30.17).

2. If y_1 and y_2 are two solutions of the inhomogeneous differential equation (30.17), then their *difference* $y = y_1 - y_2$ is a solution of the accompanying homogeneous equation (30.18).

The proved Theorem 30.2 indicates a **method for finding the general solution of the inhomogeneous equation** (30.17): it is necessary to find the general solution of the accompanying homogeneous equation (30.18) and some particular solution to the equation (30.17).

We are able to find a general solution of the homogeneous equation (30.18), however, only for the case of constant coefficients: $p(x) = p = \text{const}$, $q(x) = q = \text{const}$. The **problem of finding a particular solution to the inhomogeneous equation** (30.17) in the general case is very complicated. We will consider it only for simple cases, especially for situations when the coefficients of equation (30.17) are constant and the right-hand side $f(x)$ has a special form. So, we now turn to an inhomogeneous linear equation

$$y'' + py' + qy = f(x), \tag{30.20}$$

where $p, q = \text{const}$.

In the future, we will use symbols $P_n(x)$ and $Q_n(x)$ to denote polynomials of degree *n*:

$$P_n(x) = a_0 x^n + a_1 x^{n-1} + \ldots + a_{n-1} x + a_n,$$

$$Q_n(x) = b_0 x^n + b_1 x^{n-1} + \ldots + b_{n-1} x + b_n.$$

We consider three different particular types of function $f(x)$.

A. $f(x) = P_n(x)$.

The right-hand side of equation (30.20) is a polynomial of degree *n*. Since the derivative of the polynomial is a polynomial, we can try to find a particular solution $\tilde{y}$ of equation (30.20) also in the form of a polynomial whose coefficients are not yet known, but to find them there exists the *method of undetermined coefficients* that we already know.

***Example* 30.7.** Find the general solution of the equation

$$y''' - 3y' + 2y = 2x^3 - 7x^2 - 4x + 10.$$

Solution. We see that if we substitute a polynomial of the third degree

$$y = ax^3 + bx^2 + cx + d$$

into the left side of this equation, then a polynomial of the third degree will also appear on the left side. We will try to choose the coefficients *a*, *b*, *c*, *d* in such a way that the equation turns into an identity. We differentiate the polynomial and substitute it into the left side:

$$y' = 3ax^2 + 2bx + c, \qquad y'' = 6ax + 2b,$$

$$6ax + 2b - 3\left(3ax^2 + 2bx + c\right) + 2\left(ax^3 + bx^2 + cx + d\right) = 2x^3 - 7x^2 - 4x + 10$$

,

$$2ax^3 + \left(-9a + 2b\right)x^2 + \left(6a - 6b - 2c\right)x + 2b - 3c + 2d = 2x^3 - 7x^2 - 4x + 10$$

,

$$\begin{array}{c|ll} x^3 & 2a & = 2, \\ x^2 & -9a + 2b & = -7, \\ x & 6a - 6b + 2c & = -4, \\ x^0 & 2b - 3c + 2d & = 10. \end{array}$$

Solving the resulting system, we find

$$a = 1,\ b = 1,\ c = -2,\ d = 1.$$

We obtain a particular solution of the given differential equation:

$$\tilde{y} = x^3 + x^2 - 2x + 1.$$

The accompanying equation has the form

$$y'' - 3y' + 2y = 0.$$

Its characteristic equation

$$k^2 - 3k + 2 = 0$$

has the roots $k_1 = 2$, $k_2 = 1$, and therefore the general solution of the accompanying equation is

$$y_0 = C_1 e^{2x} + C_2 e^x.$$

According to Theorem 30.2, the general solution of this equation is

$$y = C_1 e^{2x} + C_2 e^x + x^3 + x^2 - 2x + 1.$$

In connection with this example, an assumption may arise that a particular solution $\tilde{y}$ of the differential equation

$$y'' + py' + qy = P_n(x)$$

should be sought for in the form of some polynomial of the same degree and we only need to select the coefficients of this polynomial. However, this assumption is erroneous.

Consider the equation

$$y'' - y' = 4x^3 - 9x^2 - 6x - 2.$$

Let us try to find $\tilde{y}$ in the form of a polynomial of the same third degree:

$$\tilde{y} ax^3 + bx^2 + cx + d.$$

Differentiating $\tilde{y}$ and substituting it into the left side of the equation, we obtain:

$$6ax + 2b - \left(3ax^2 + 2bx + c\right) = 4x^3 - 9x^2 - 6x - 2,$$

$$-3ax^2 + \left(6a + 2b\right)x + 2b - c = 4x^3 - 9x^2 - 6x - 2.$$

The identical equality of these polynomials, the degrees of which are different (on the left-hand side there is a polynomial of the second degree, and on the right-hand side there is the third), it is impossible for any *a*, *b*, *c*, and *d*.

Our attempt was unsuccessful because we did not take into account that when differentiating the degree of the polynomial decreases by one, and on the left side of this equation there is no term containing an unknown function *y* (and only terms containing derivatives of this function are present). Therefore, for the left side of this equation to become a polynomial of the same (third) degree, we must take a polynomial of degree one greater than *y*, i.e. polynomial of the fourth degree. However, in this case, the free term of this polynomial will not be taken into account since it will vanish during differentiation. Therefore, we must take a polynomial of the form

$$Q_4(x) = ax^4 + bx^3 + cx^2 + dx,$$

or

$$xQ_3(x) = x\left(ax^3 + bx^2 + cx + d\right).$$

(Recall that we are only looking for a particular solution.)

Taking into account the considerations expressed here, we will seek for a solution of the given equation

$$y'' - y' = 4x^3 - 9x^2 - 6x - 2$$

as $\tilde{y} = xQ_3(x) = x\left(ax^3 + bx^2 + cx + d\right) = ax^4 + bx^3 + cx^2 + dx$. We differentiate $\tilde{y}$ and substitute $\tilde{y}'$ and $\tilde{y}''$ in the left-hand side of this equation:

$\tilde{y}' = 4ax^3 + 3bx^2 + 2cx + d$, $\tilde{y}'' = 12ax^2 + 6bx + 2c$,

$12ax^2 + 6bx + 2c - \left(4ax^3 + 3bx^2 + 2cx + d\right) = 4x^3 - 9x^3 - 6x - 2$,

$-4ax^3 + (12a - 3b)x^2 + (6b - 2c)x + 2c - d = 4x^3 - 9x^3 - 6x - 2$,

$$\begin{array}{c|l} x^3 & -4a = 4, \\ x^2 & 12a - 3b = -9, \\ x & 6b - 2c = -6, \\ x^0 & 2c - d = -2. \end{array}$$

Solving the resulting system, we obtain

$a = -1$, $b = -1$, $c = 0$, $d = 2$.

Hence $\tilde{y} = -x^4 - x^3 + 2x$.

We find now y_0. To do so, we compose a characteristic equation, find its roots, and use them:

$k^2 - k = 0$,

$k(k-1) = 0$.

We obtain the general solution of the accompanying equation:

$y_0 = C_1 + C_2 e^x$.

The general solution of this equation is

$y = C_1 + C_2 e^x - x^4 - x^3 + 2x$.

So, with $q = 0$ and $P_n(x)$ on the right-hand side, we found a particular solution $\tilde{y}$ in the form of a polynomial of degree *n + 1*:

$\tilde{y} = xQ_n(x)$.

Reasoning in a similar way, in the case when not only *q*, but also $p = 0$, the solution $\tilde{y}$ can be found in the form

$\tilde{y} = x^2 Q_n(x)$.

(However, in this case, it is enough to integrate the right-hand side twice.) Let us summarize the first result and indicate methods for finding a particular solution of differential equation (30.20) in case the right-hand side is a polynomial.

If $f(x) = P_n(x)$, where $P_n(x)$ is a polynomial of degree *n*, then a particular solution $\tilde{y}$ must be sought in the form:

- $\tilde{y} = Q_n(x)$, if $q \neq 0$;
- $\tilde{y} = xQ_n(x)$, if $q = 0$, $p \neq 0$;
- $\tilde{y} = x^2 Q_n(x)$, if $q = 0$, $p = 0$.

These rules are also preserved in those cases when we are dealing with higher-order differential equations.

***Example* 30.8.** Solve the equation $y'' + 5y' = 10x + 12$.

Solution. We make up the characteristic equation: $k^2 + 5k = 0$. Its roots are $k_1 = 0$, $k_2 = -5$. Since $q = 0$, we will seek for a particular solution in the form:

$\tilde{y} = x(Ax + B) = Ax^2 + Bx$.

Substituting it in the original equation:

$2A + 5(2Ax + B) = 10x + 12$.

We equate the coefficients at the same degrees of x:

$$\begin{cases} 10A = 10, \\ 2A + 5B = 12. \end{cases}$$

Hence, $A = 1$, $B = 2$, $\tilde{y} = x^2 + 2x$. The general solution of the accompanying equation is $y_0 = C_1 + C_2 e^{-5x}$. Therefore, the general solution of this equation is

$y = C_1 + C_2 e^{-5x} + x^2 + 2x$.

***Example* 30.9.** Find the general solution of the differential equation

$y''' - 2y'' - y' + 2y = 2x^3 - 3x^2 - 12x + 8$

Solution. Here, the coefficient of y on the left-hand side is nonzero; therefore, we seek for $\tilde{y}$ in the form of a polynomial of the same degree as the polynomial on the right-hand side, i.e. as $Q_3(x)$:

$\tilde{y} = ax^3 + bx^2 + cx + d$

We differentiate:

$\tilde{y}' = 3ax^2 + 2bx + c$, $\tilde{y}'' = 6ax + 2b$, $\tilde{y}''' = 6a$.

We substitute $\tilde{y}$ and its derivatives:

$6a - 2(6ax + 2b) - (3ax^2 + 2bx + c) + 2(ax^3 + bx^2 + cx + d) = 2x^3 - 3x^2 - 12x + 8$

$$2ax^3+(-3a+2b)x^2+(-12a-2b+2c)x+6a-4b-c+2d=$$
$$=2x^3-3x^2-12x+8,$$

$$\begin{array}{c|ll} x^3 & 2a & =2, \\ x^2 & -3a+2b & =-3, \\ x & -12a-2b+2c & =-12, \\ x^0 & 6a-4b-c+2d & =8. \end{array}$$

We find the coefficients: $a=1$, $b=0$, $c=0$, $d=1$. So,

$\tilde{y}=x^3+1$.

We solve the accompanying equation:

$y'''-2y''-y'+2y=0$,

$k^3-2k^2-y+2=0$,

$k_1=2$, $k_2=1$, $k_3=-1$.

We find the general solution of the accompanying equation:

$y_0=C_1e^{2x}+C_2e^x+C_3e^{-x}$.

Finally, we obtain:

$y_0=C_1e^{2x}+C_2e^x+C_3e^{-x}+x^3+1$.

Let us consider another example with a differential equation, the order of which is higher than two.

***Example* 30.10.** Find the general solution of the equation

$y^{(4)}+y'''=24x+36$.

Solution. We are seeking for $\tilde{y}$ in the form of $x^3Q_1(x)$:

$\tilde{y} = x^3(ax+b) = ax^4 + bx^3$,

$\tilde{y}' = 4ax^3 + 3bx^2$, $\tilde{y}'' = 12ax^2 + 6bx$, $\tilde{y}''' = 24ax + 6b$, $\tilde{y}^{(4)} = 24a$.

Substituting, we obtain

$24a + 24ax + 6b = 24x + 36$,

$$\begin{cases} 24a = 24, \\ 24a + 6b = 36. \end{cases}$$

Hence, $a = 1$, $b = 2$; $\tilde{y} = x^4 + 2x^3$.

We solve the characteristic equation:

$k^4 + k^3 = 0$,

$k^3(k+1) = 0$,

$k_1 = k_2 = k_3 = 0$, $k_4 = -1$.

We obtain

$y_0 = C_1 + C_2 x + C_3 x^2 + C_4 \mathrm{e}^{-x}$.

We find the general solution of this equation:

$y = C_1 + C_2 x + C_3 x^2 + C_4 \mathrm{e}^{-x} + x^4 + 2x^3$.

We now turn to consideration of equations with a more complex right-hand side.

B. Let $f(x) = \mathrm{e}^{\alpha x} P_n(x)$. The considered earlier case $f(x) = P_n(x)$ is obtained from this at $\alpha = 0$. Formally speaking, it could not be considered separately. So, we need to solve the equation

$$y'' + py' + qy = \mathrm{e}^{\alpha x} P_n(x). \qquad (30.21)$$

We will try to reduce this problem to the previous one, i.e. to the case when the polynomial is on the right-hand side of the equation. We apply the substitution

$$y = \mathrm{e}^{\alpha x} z,$$

where z is a new unknown function $z = z(x)$. We find derivatives:

$$y' = \alpha \mathrm{e}^{\alpha x} z + \mathrm{e}^{\alpha x} z',$$

$$y'' = \alpha^2 \mathrm{e}^{\alpha x} z + \alpha \mathrm{e}^{\alpha x} z' + \alpha \mathrm{e}^{\alpha x} z + \mathrm{e}^{\alpha x} z''.$$

Substituting y, y', and y'' into the left-hand side of the equation, we perform obvious transformations:

$$\alpha^2 \mathrm{e}^{\alpha x} z + 2\alpha \mathrm{e}^{\alpha x} z' + \mathrm{e}^{\alpha x} z'' + p\left(\alpha \mathrm{e}^{\alpha x} z + \mathrm{e}^{\alpha x} z'\right) + q\mathrm{e}^{\alpha x} z = \mathrm{e}^{\alpha x} P_n(x),$$

$$\mathrm{e}^{\alpha x} z'' + \left(2\alpha \mathrm{e}^{\alpha x} + p\mathrm{e}^{\alpha x}\right) z' + \left(\alpha^2 \mathrm{e}^{\alpha x} + p\alpha \mathrm{e}^{\alpha x} + q\mathrm{e}^{\alpha x}\right) z = \mathrm{e}^{\alpha x} P_n(x).$$

Reducing by $\mathrm{e}^{\alpha x}$:

$$z'' + (2\alpha + p) z' + \left(\alpha^2 + p\alpha + q\right) z = P_n(x).$$

We obtained an equation of the kind already considered:

$$z'' + \bar{p} z' + \bar{q} z = P_n(x),$$

where $\bar{p} = 2\alpha + p$, $\bar{q} = \alpha^2 + p\alpha + q$. Therefore, we can apply the same rule:

- $\tilde{z} = Q_n(x)$, if $\bar{q} = \alpha^2 + p\alpha + q \neq 0$;
- $\tilde{z} = xQ_n(x)$, if $\bar{q} = \alpha^2 + p\alpha + q = 0$, $\bar{p} = 2\alpha + p \neq 0$;
- $\tilde{z} = x^2 Q_n(x)$, if $\bar{q} = \alpha^2 + p\alpha + q = 0$, $\bar{p} = 2\alpha + p = 0$.

Note that the condition $\overline{q} \neq 0$, i.e. $\alpha^2 + p\alpha + q \neq 0$, means that the number 〈 is not a root of the characteristic equation $k^2 + pk + q = 0$. Therefore if 〈 is not a root of the characteristic equation, we seek the solution as $\tilde{z} = Q_n(x)$, or (given that $\tilde{y} = \mathrm{e}^{\alpha x} z$) as $\tilde{y} = \mathrm{e}^{\alpha x} Q_n(x)$.

Before considering the remaining conditions, we note that the solution of the characteristic equation $k^2 + pk + q = 0$ has the form:

$$k_{1,2} = -\frac{p}{2} \pm \sqrt{\frac{p^2}{4} - q}.$$

If at the same time $\frac{p^2}{4} - q \neq 0$, i.e. $\frac{p^2}{4} \neq q$, then both roots of the characteristic equation are different and, obviously, none of them is equal $-\frac{p}{2}$: $\alpha \neq -\frac{p}{2}$, i.e. $2\alpha + p \neq 0$.

So, if 〈 is the single root of the characteristic equation, then $2\alpha + p \neq 0$, i.e. $\overline{q} = 0$, $\overline{p} \neq 0$. If 〈 is a root of the characteristic equation and $2\alpha + p = 0$, then $\frac{p^2}{4} - q = 0$, i.e. 〈 is the double root of the characteristic equation. We summarize all of the above (and recall that $\tilde{y} = \mathrm{e}^{\alpha x} z$).

If $\boldsymbol{f(x) = \mathrm{e}^{\alpha x} P_n(x)}$, then $\tilde{y}$ must be sought as:

- $\tilde{y} = \mathrm{e}^{\alpha x} Q_n(x)$, if 〈 is *not a root* of the characteristic equation;
- $\tilde{y} = x\mathrm{e}^{\alpha x} Q_n(x)$, if 〈 is the *single root* of the characteristic equati

- $\tilde{y} = x^2 \mathrm{e}^{\alpha x} Q_n(x)$, if 〈 is the *double root* of the characteristic equat

***Example* 30.11.** Solve the equation $y'' - 3y' + 2y = 2\mathrm{e}^{3x}$.

Solution. We make up the characteristic equation: $k^2 - 3k + 2$. Its roots are $k_1 = 1$, $k_2 = 2$.

Obviously, $\alpha = 3$ is not a root of the characteristic equation. Therefore $\tilde{y} = C\mathrm{e}^{3x}$, $\tilde{y}' = 3C\mathrm{e}^{3x}$, $\tilde{y}'' = 9C\mathrm{e}^{3x}$. We substitute all this into the original equation:

$$9C\mathrm{e}^{3x} - 3 \cdot 3C\mathrm{e}^{3x} + 2C\mathrm{e}^{3x} = 2\mathrm{e}^{3x}.$$

We obtain $C = 1$. Consequently,

$$\tilde{y} = \mathrm{e}^{3x}.$$

Obviously, the general solution of the accompanying homogeneous equation is

$$y_0 = C_1\mathrm{e}^{x} + C_2\mathrm{e}^{2x}.$$

We finally obtain

$$y = C_1\mathrm{e}^{x} + C_2\mathrm{e}^{2x} + \mathrm{e}^{3x}.$$

***Example* 30.12.** Solve the equation $y'' + 2y' = (3x + 7)\mathrm{e}^{x}$.

Solution. Here $\alpha = 1$. We compose and solve the characteristic equation:

$$k^2 + 2k = 0,$$

$$k_1 = 0,\ k_2 = -2.$$

We see that 〈 is not a root of the characteristic equation. Therefore $\tilde{y} = Q_1(x)\mathrm{e}^x = (ax+b)\mathrm{e}^x$. We differentiate $\tilde{y}$ and substitute it into the right-hand part of the equation:

$\tilde{y}' = a\mathrm{e}^x + (ax+b)\mathrm{e}^x = (ax+a+b)\mathrm{e}^x$,

$\tilde{y}'' = a\mathrm{e}^x + (ax+a+b)\mathrm{e}^x = (ax+2a+b)\mathrm{e}^x$.

$(ax+2a+b)\mathrm{e}^x + 2(ax+a+b)\mathrm{e}^x = (3x+7)\mathrm{e}^x$.

We reduce it by e^x and combine the like terms:

$3ax + 4a + 3b = 3x + 7$.

We obtain the system

$$\begin{cases} 3a = 3 \\ 4a + 3b = 7 \end{cases}$$

Hence, $a = 1$, $b = 1$; $\tilde{y} = (x+1)\mathrm{e}^x$.

The general solution of the accompanying homogeneous equation is

$y_0 = C_1 + C_2\mathrm{e}^{-2x}$.

Finally, we obtain

$y = C_1 + C_2\mathrm{e}^{-2x} + (x+1)\mathrm{e}^x$.

(Please note: here $q = 0$, but we did not take this into account, unlike case A, since in this case it is only taken into account whether 〈 is a root of the characteristic equation.)

***Example* 30.13.** Solve the equation $y'' + y' - 6y = (10x+2)\mathrm{e}^{2x}$.

Solution. We solve the characteristic equation:

$k^2 + k - 6 = 0$,

$k_1 = 2$, $k_2 = -3$.

So, 〈 = 2 is the single root.

The general solution of the accompanying homogeneous equation is

$y_0 = C_1 e^{2x} + C_2 e^{-3x}$.

We find $\tilde{y}$. Obviously, $\tilde{y}$ must be sought as

$\tilde{y} = xQ_1(x)e^{2x} = x(ax+b)e^{2x} = (ax^2+bx)e^{2x}$;

$\tilde{y}' = (2ax+b)e^{2x} + 2(ax^2+bx)e^{2x} = (2ax^2+2ax+2bx+b)e^{2x}$;

$\tilde{y}'' = (4ax+2a+b)e^{2x} + 2(2ax^2+2ax+2bx+b)e^{2x} =$

$= (4ax^2+8ax+4bx+2a+3b)e^{2x}.$

We substitute all these expressions in the original equation:

$(4ax^2+8ax+4bx+2a+3b)e^{2x} + (2ax^2+2ax+2bx+b)e^{2x} -$

$-6(ax^2+bx)e^{2x} = (10x+2)e^{2x}.$

Reducing by e^{2x} and combining the like terms, we obtain:

$$\begin{cases} 10a = 10, \\ 2a - 2b = 2, \end{cases}$$

i.e. $a = 1$, $b = 0$. Consequently, $\tilde{y} = x^2 e^{2x}$. We obtain

$y = C_1 e^{2x} + C_2 e^{-3x} + x^2 e^{2x}$.

***Example* 30.14.** Solve the equation $y'' + 2y' + y = (6x-2)e^{-x}$.

Solution. Here 〈 = –1 is a double root of the characteristic equation $k^2 + 2k + 1 = 0$. Therefore, we seek the particular solution $\tilde{y}$ in the form

$\tilde{y} = x^2 Q_1(x) e^{-x} \tilde{y} = x^2 Q_1(x) e^{-x} = x^2 (ax + b) \mathrm{e}^{-x} = (ax^3 + bx^2) \mathrm{e}^{-x}$.

We have:

$\tilde{y}' = (3ax^2 + 2bx) \mathrm{e}^{-x} - (ax^3 + bx^2) \mathrm{e}^{-x} = (-ax^3 + (3a - b)x^2 + 2bx) \mathrm{e}^{-x}$;

$\tilde{y}'' = (-3ax^2 + 2(3a - b)x + 2b) \mathrm{e}^{-x} - (-ax^3 + (3a - b)x^2 + 2bx) \mathrm{e}^{-x} =$

$= (ax^3 + (-6a + b)x^2 + 6ax - 4bx + 2b) \mathrm{e}^{-x}$.

We substitute it into the original equation; after combining the like terms, we obtain

$(6ax + 2b) \mathrm{e}^{-x} = (6x - 2) \mathrm{e}^{-x}$.

Hence , $a = 1$, $b = -1$; $\tilde{y} = (x^3 - x^2) \mathrm{e}^{-x}$.

We also find $y_0 = C_1 \mathrm{e}^{-x} + C_2 x \mathrm{e}^{-x}$.

We obtain the general solution:

$y = C_1 \mathrm{e}^{-x} + C_2 x \mathrm{e}^{-x} + (x^3 - x^2) \mathrm{e}^{-x}$.

Naturally arises the question of finding a particular solution of a differential equation of the form

$y'' + py' + qy = f_1(x) + f_2(x)$,

where $f_1(x)$ and $f_2(x)$ are functions of different kinds (for example, $f_1(x) = ax^2 + bx + c$, $f_2(x) = \mathrm{e}^x$). The following statement holds.

Lemma. If y_1 is a solution of the differential equation

$y'' + py' + qy = f_1(x)$,

and y_2 is a solution of the differential equation

$y'' + py' + qy = f_2(x)$,

then the sum of these solutions $y = y_1 + y_2$ is the solution of the differential equation

$$y'' + py' + qy = f_1(x) + f_2(x).$$

(this lemma is called the "**superposition principle**".)

Proof. The lemma is proved easily - by direct substitution. We substitute the sum $y = y_1 + y_2$ into the left-hand side of the equation. We obtain the expression

$$(y_1 + y_2)'' + p(y_1 + y_2)' + q(y_1 + y_2).$$

Regrouping the terms, we obtain

$$(y_1'' + py_1' + qy_1) + (y_2'' + py_2' + qy_2).$$

But since y_1 is a solution of the equation $y'' + py' + qy = f_1(x)$, the expression in the first brackets is identically equal to $f_1(x)$. For a similar reason, the expression in the second bracket is equal to $f_2(x)$. So, the left-hand side is identically equal to $f_1(x) + f_2(x)$. The equation turns into an identity: $f_1(x) + f_2(x) \equiv f_1(x) + f_2(x)$. The proof is complete.

Note that the assertion proved is also true for the case when the coefficients p and q depend on x: $p = p(x)$, $q = q(x)$.

***Example* 30.15.** Solve the equation $y'' + 5y' = 10x + 12 + 6e^x$.

Solution. Solving the equation $y'' + 5y' = 10x + 12$, we obtain: $y_0 = C_1 + C_2 e^{-5x}$, $\tilde{y}_1 = x^2 + 2x$ (see Example 30.8). Obviously, $\tilde{y}_2 = e^x$. Therefore, the general solution is $y = C_1 + C_2 e^{-5x} + x^2 + 2x + e^x$.

C. Let $f(x) = P_m(x)\mathrm{e}^{\alpha x}\cos\beta x + Q_n(x)\mathrm{e}^{\alpha x}\sin\beta x$. In this case, we can use the technique applied in the previous case if we pass from trigonometric functions to exponential ones. In a more detailed course of mathematics, the *Euler formula* is considered, which expresses an exponential function with an imaginary exponent in terms of trigonometric functions (here i is the imaginary unit, $i^2 = -1$):

$$\mathrm{e}^{ix} = \cos x + i\sin x. \qquad (30.22)$$

Substituting $-x$ instead of x in this formula, we obtain

$$\mathrm{e}^{-ix} = \cos x - i\sin x. \qquad (30.23)$$

From equalities (30.22) and (30.23) it is easy to find $\cos x$ and $\sin x$:

$$\cos x = \frac{\mathrm{e}^{ix} + \mathrm{e}^{-ix}}{2}, \qquad \sin x = \frac{\mathrm{e}^{ix} - \mathrm{e}^{-ix}}{2i}$$

These formulas are also called *Euler formulae*. Applying them, we obtain

$$f(x) = P_m(x)\mathrm{e}^{\alpha x}\frac{\mathrm{e}^{i\beta x} + \mathrm{e}^{-i\beta x}}{2} + Q_n(x)\mathrm{e}^{\alpha x}\frac{\mathrm{e}^{i\beta x} - \mathrm{e}^{-i\beta x}}{2i},$$

or

$$f(x) = \left(\frac{1}{2}P_m(x) + \frac{1}{2i}Q_n(x)\right)\mathrm{e}^{(\alpha+i\beta)x} + \left(\frac{1}{2}P_m(x) - \frac{1}{2i}Q_n(x)\right)\mathrm{e}^{(\alpha-i\beta)x}.$$

Here in square brackets are the polynomials whose degrees are equal to the largest of the degrees of $P_m(x)$ and $Q_n(x)$, i.e. to the largest of the numbers m and n. Thus, we have obtained the right-hand side of the form considered in case B. Moreover, we can prove (we do not give this proof) that we can find a particular solution $\tilde{y}$ that does not contain complex numbers.

So, if $\boldsymbol{f(x)=P_m(x)e^{\alpha x}\cos\beta x+Q_n(x)e^{\alpha x}\sin\beta x}$, then $\tilde{y}$ should be sought in the form:

- $\tilde{y}=u(x)e^{\alpha x}\cos\beta x+v(x)e^{\alpha x}\sin\beta x$, if $\alpha+i\beta$ is *not a root* of the characteristic equation;
- $\tilde{y}=x\left(u(x)e^{\alpha x}\cos\beta x+v(x)e^{\alpha x}\sin\beta x\right)$, if $\alpha+i\beta$ is a *root* of the characteristic equation.

Here $u(x)$ and $v(x)$ are polynomials whose degrees are equal to the largest degree of polynomials $P_m(x)$ and $Q_n(x)$.

Remark. Note that the indicated forms of particular solutions are preserved also in the case when in the right-hand side of the differential equation one of the polynomials, $P_m(x)$ or $Q_n(x)$, is identically equal to zero, i.e. either

$f(x)=P_m(x)e^{\alpha x}\cos\beta x$,

or

$f(x)=Q_n(x)e^{\alpha x}\sin\beta x$.

Let us consider in more detail a simpler case - a special case of case *C*.

C₀. Let $f(x)=M\cos\beta x+N\sin\beta x$. Applying Euler's formulae, we rewrite the differential equation

$$y''+py'+qy=M\cos\beta x+N\sin\beta x$$

as

$$y''+py'+qy=M\frac{e^{i\beta x}+e^{-i\beta x}}{2}+N\frac{e^{i\beta x}-e^{-i\beta x}}{2i}.$$

Letting for brevity

$$\frac{M}{2}+\frac{N}{2i}=M_1,\quad \frac{M}{2}-\frac{N}{2i}=N_1,$$

We obtain

$$y''+py'+qy=M_1\mathrm{e}^{i\beta x}+N_1\mathrm{e}^{-i\beta x}. \qquad (*)$$

According to Lemma 30.3, the solution $\tilde{y}$ of our differential equation is the sum of the solutions $\tilde{y}_1$ and $\tilde{y}_2$ of equations

$$y''+py'+qy=M_1\mathrm{e}^{i\beta x}, \qquad y''+py'+qy=N_1\mathrm{e}^{-i\beta x}.$$

Note that the imaginary number $i\beta$ cannot be the double root of the quadratic equation $k^2+pk+q=0$ whose coefficients *p* and *q* are real numbers.

A particular solution $\tilde{y}_1$ of the equation $y''+py'+qy=M_1\mathrm{e}^{i\beta x}$ has the form:

- $\tilde{y}_1=A\mathrm{e}^{i\beta x}$, if $i\beta$ is not a root of the characteristic equation;
- $\tilde{y}_1=Ax\mathrm{e}^{i\beta x}$, if $i\beta$ is a root of the characteristic equation.

A particular solution $\tilde{y}_2$ of the equation $y''+py'+qy=N_1\mathrm{e}^{-i\beta x}$ has the form:

- $\tilde{y}_2=B\mathrm{e}^{-i\beta x}$, if $-i\beta$ is not a root of the characteristic equation;
- $\tilde{y}_2=xB\mathrm{e}^{-i\beta x}$, if $-i\beta$ is a root of the characteristic equation.

Obviously, the numbers $i\beta$ and $-i\beta$ either both are, or both are not the roots of the characteristic equation (since if $\alpha+i\beta$ is a root of the quadratic equation, then $\alpha-i\beta$ is also a root of this equation).

Therefore, a particular solution of our equation will have the form:

- $\tilde{y} = Ae^{i\beta x} + Be^{-i\beta x}$, if $\pm i\beta$ are not the roots of the characteristic equation;
- $\tilde{y} = x\left(Ae^{i\beta x} + Be^{-i\beta x}\right)$,if $\pm i\beta$ are the roots of the characteristic equation.

We apply the Euler formula [see (30.22) and (30.23)] and obtain

$$Ae^{i\beta x} + Be^{-i\beta x} = (A+B)\cos\beta x + i(A-B)\sin\beta x = a\cos\beta x + b\sin\beta x,$$

where, for brevity, $A+B=a$, $i(A-B)=b$.

Hence we obtain the rule for finding a particular solution of the differential equation

$$y'' + py' + qy = M\cos\beta x + N\sin\beta x.$$

So, if $f(x) = M\cos\beta x + N\sin\beta x$, then a particular solution $\tilde{y}$ must be sought in the form

- $\tilde{y} = a\cos\beta x + b\sin\beta x$, if *i*® is *not a root* of the characteristic equation;
- $\tilde{y} = x(a\cos\beta x + b\sin\beta x)$, if *i*® is a *root* of the characteristic equation.

***Example* 30.16.** Solve the equation $y'' + y' - 2y = 8\sin 2x$.

Solution.. The characteristic equation $k^2 + k - 2 = 0$ has the roots $k_1 = 1$, $k_2 = -2$. Here $\beta = 2$, therefore, $i\beta$ is not a root of the characteristic equation. Therefore, a particular solution $\tilde{y}$ must be sought in the form $\tilde{y} = C\cos 2x + D\sin 2x$,

$\tilde{y}' = -2C\sin 2x + 2D\cos 2x$,

$\tilde{y}'' = -4C\cos 2x - 4D\cos 2x$.

Substitute:

$$-4C\cos 2x - 4D\sin 2x - 2C\sin 2x + 2D\cos 2x -$$
$$-2\left(C\cos 2x + D\sin 2x\right) = 8\sin 2x,$$
$$\left(-6C + 2D\right)\cos 2x + \left(-2C - 6D\right)\sin 2x = 8\sin 2x.$$

We equate the coefficients at $\cos 2x$ and $\sin 2x$:

$$\begin{cases} -6C + 2D = 0, \\ -2C - 6D = 8. \end{cases}$$

Hence $C = -\frac{2}{5}$, $D = -\frac{6}{5}$. Consequently,

$\tilde{y} = -\frac{2}{5}\cos 2x - \frac{6}{5}\sin 2x$.

We find y_0:

$y_0 = C_1 e^{x} + C_2 e^{-2x}$.

We obtain the general solution:

$y = C_1 e^{x} + C_2 e^{-2x} - \frac{1}{5}\left(2\cos 2x + 6\sin 2x\right)$.

***Example* 30.17.** Solve the equation $y'' + 4y = \cos 2x$.

Solution. The characteristic equation $k^2 + 4k = 0$ has the roots $k_1 = 2i$, $k_2 = -2i$. Here $\beta = 2$, therefore, $i\beta$ is the root of the characteristic equation; therefore, a particular solution $\tilde{y}$ must be sought in the form $\tilde{y} = x(C\cos 2x + D\sin 2x)$.

We differentiate:

$$\tilde{y}' = C\cos 2x + D\sin 2x = 2x(-C\sin 2x + D\cos 2x),$$

$$\tilde{y}'' = 2(-C\sin 2x + D\cos 2x) + 2(-C\sin 2x + D\cos 2x) +$$
$$+ 4x(-C\cos 2x - D\sin 2x) = 4(-C\sin 2x + D\cos 2x) +$$
$$+ 4x(-C\cos 2x - D\sin 2x).$$

Substituting into the equation, we obtain:

$$4(-C\sin 2x + D\cos 2x) + 4x(-C\cos 2x - D\sin 2x) +$$
$$+ 4x(C\cos 2x + D\sin 2x) = \cos 2x.$$

Combining the like terms, we obtain

$$-4C\sin 2x + 4D\cos 2x = \cos 2x.$$

Equating the coefficients at $\cos 2x$ and $\sin 2x$, we obtain: $-4C = 0$, $4D = 1$; hence, $C = 0$, $D = \frac{1}{4}$. Thus, a particular solution to this equation is

$$\tilde{y} = \frac{1}{4}\sin 2x.$$

The general solution of the accompanying homogeneous equation is

$$y_0 = C_1\cos 2x + C_2\sin 2x.$$

Finally, we obtain

$$y = C_1 \cos 2x + C_2 \sin 2x + \frac{1}{4} \sin 2x .$$

***Example* 30.18.** Solve the equation

$$y'' - y = \mathrm{e}^{2x}(6\cos x - 2\sin x).$$

Solution. Here $\alpha = 2$, $\beta = 1$. The characteristic equation $k^2 - 1 = 0$ has the roots $k_1 = 1$, $k_2 = -1$. Since $\alpha + i\beta = 2 + i$ is not the root of the characteristic equation, we are seeking a particular solution in the form

$$\tilde{y} = \mathrm{e}^{2x}(\cos x + D\sin x).$$

We find $\tilde{y}'$ and $\tilde{y}''$:

$$\tilde{y}' = 2\mathrm{e}^{2x}(C\cos x + D\sin x) + \mathrm{e}^{2x}(-C\sin x + D\cos x),$$

$$\tilde{y}'' = 4\mathrm{e}^{2x}(C\cos x + D\sin x) + 2\mathrm{e}^{2x}(-C\sin x + D\cos x) +$$
$$+ 2\mathrm{e}^{2x}(-C\sin x + D\cos x) + \mathrm{e}^{2x}(-C\cos x - D\sin x) =$$
$$= \mathrm{e}^{2x}(3C\cos x + 4D\cos x + 3D\sin x - 4C\sin x).$$

Substituting the obtained expressions into the equation and combining the like terms, we obtain (after reduction by e^{2x}):

$$(2C + 4D)\cos x + (2D - 4C)\sin x = 6\cos x - 2\sin x .$$

Equating the coefficients at $\cos x$ and $\sin x$, we obtain the system

$$\begin{cases} 2C + 4D = 6, \\ -4C + 2D = 2. \end{cases}$$

Solving this system, we find $C = 1$, $D = 1$.
A particular solution of this equation is

$\tilde{y} = \mathrm{e}^{2x}(\cos x + \sin x)$.

The general solution to the accompanying equation is

$y_0 = C\mathrm{e}^{x} + C_2\mathrm{e}^{-x}$.

We obtain the general solution of the given equation:

$y = C_1\mathrm{e}^{x} + C_2\mathrm{e}^{-x} + \mathrm{e}^{2x}(\cos x + \sin x)$.

Questions

1. What is the general form of a second-order differential equation?
2. What is a general solution of a second-order differential equation?
3. What differential equations can be reduced in order?
4. What is called a linear differential equation of order *n*?
5. What linear differential equation of order *n* is called homogeneous?
6. What are the properties of solutions of a linear homogeneous differential equation?
7. What system of functions is called linearly independent?
8. What does the Wronski determinant for two functions look like?
9. What is the structure of the solution of a homogeneous linear differential equation?
10. What is the characteristic equation?
11. What does the general solution of a second-order homogeneous linear differential equation look like when the roots of the characteristic equation coincide?
12. What is the structure of the general solution of an inhomogeneous linear differential equation? How can one obtain a general solution of the differential equation $y'' + py' + qy = f(x)$, knowing the solution of the differential equation $y'' + py' + qy = 0$?

13. In what form should a particular solution of the differential equation $y'' + py' + qy = P_n(x)$ be sought when the right-hand side $P_n(x)$ is a polynomial of degree *n*? Is this particular solution always also a polynomial of degree *n*?
14. In what form should a particular solution of the differential equation $y'' + py' + qy = \mathrm{e}^{\alpha x} P_n(x)$ be sought?
15. What is the superposition principle?
16. What is the rule for finding a particular solution of the differential equation $y'' + py' + qy = M\cos\beta x + N\sin\beta x$?

Chapter 31. Difference equations

31.1. Basic definitions

In mathematical applications, among functions of continuous argument, we also have to deal with functions of discrete argument – i.e. with functions defined on a finite (or countable) discrete set. Examples of such functions are functions defined by tables, numerical sequences (see chapter 14), series (see section IX).

Discrete argument functions are usually denoted by $f(x_k)$ or $y(x_k)$. The distance $h_k = x_{k+1} - x_k$, k = 1, 2, ... between adjacent values of the argument can be any positive numbers. However, the most interesting thing is the case where values h_k are the same: $h_k = h$ for all k = 1, 2, ... This number *h* is usually called a **sampling step.** In this case, $x_k = kh$, and

function $f(x_k)$ becomes the number function *k*, i.e. $f(x_k) = f(kh)$, *k* = 1, 2, ...

Definition. **A grid on a segment** $[a,b]$ is any finite set of points of this segment. Grid points are called its **nodes.**

Note that we were already dealing with grids and their nodes — when we defined the concept of a definite integral and when we were engaged in the approximate calculation of definite integrals using the formulas of rectangles and trapezoids and the Simpson formula (see § 22.5).

A grid is called **uniform** if its nodes divide a segment $[a,b]$ into *equal* segments. The length *h* of such partial segment is called the grid *step*. Obviously, $h = \dfrac{b-a}{n}$, where *n* is a number of partial segments.

The set of points in $[a,b]$

$\{x_i = a + kh,\ k = 0,\ 1,\ 2,\ ...,\ n\}$

forms *a uniform grid with step h*.

In case the nodes of the grid divide segment $[a,b]$ into unequal segments, the grid is called **nonuniform**.

Definition. A function defined at grid points is called a **mesh function**.

The corresponding values of the mesh function at grid nodes are usually denoted by y_k or f_k. If the mesh function is defined on a uniform grid, then its values are denoted by $y(k)$, where *k* is the number of a grid node (*k* = 0, 1, 2, ..., *n*). In this case, the mesh function is considered as a function of integer argument.

In order to obtain the corresponding mesh function $y(kh)$ from the function of continuous argument $y(x)$, it is necessary to replace argument *x* with *kh*.

***Example* 31.1.** For function $y = 4x^2 + x$, defined on interval [0, 1], compose uniform grid with *n = 4* and the corresponding mesh function.
Solution. Obviously, the grid step *h = 0.25*. We get the grid {0, 0.25, 0.5, 0.75, 1}. The mesh function is also a set consisting of five numbers: {0, 0.5, 1.5, 3, 5}.
An analogue of *the first derivative* of the continuous argument function is the *first difference* of a grid function.

The first-order difference or **the first difference** of mesh function $y(k)$, denoted by $\Delta y(k)$, is defined as:

$$\Delta y(k) = y(k+1) - y(k). \tag{31.1}$$

The second difference $\Delta^2 y(k)$ of function $y(k)$ is defined as the first difference from its first difference:

$$\Delta^2 y(k) = \Delta y(k+1) - \Delta y(k). \tag{31.2}$$

Substituting the values $\Delta y(k)$ and $\Delta y(k+1)$, determined by formula (31.1), we obtain:

$$\Delta^2 y(k) = y(k+2) - 2y(k+1) + y(k).$$

The difference $\Delta^3 y(k)$ is determined similarly. Generally, the difference of any order is determined in the same way. In this case, the *m*-th order difference $\Delta^m y(k)$ can be represented as a linear combination of values $y(k)$, $y(k+1)$, ..., $y(k+m)$. In particular,

$$\Delta^3 y(k) = \Delta^2 y(k+1) - \Delta^2 y(k) = y(k+3) - 3y(k+2) + 3y(k+1) - y(k).$$

***Example* 31.2.** Find all differences up to the *m*-th order inclusively for function $y(k) = \mathrm{e}^{\alpha k}$.

Solution. $\Delta y(k) = \mathrm{e}^{\alpha(k+1)} - \mathrm{e}^{\alpha k} = \left(\mathrm{e}^{\alpha} - 1\right)\mathrm{e}^{\alpha k}$.

We see that the first difference is proportional to the function itself $\mathrm{e}^{\alpha k}$.

Hence, $\Delta^2 y(k) = \left(\mathrm{e}^{\alpha} - 1\right)^2 \mathrm{e}^{\alpha k}$, $\Delta^3 y(k) = \left(\mathrm{e}^{\alpha} - 1\right)^3 \mathrm{e}^{\alpha k}$, ..., $\Delta^m y(k) = \left(\mathrm{e}^{\alpha} - 1\right)^m \mathrm{e}^{\alpha k}$.

Definition. Equation of form

$$F\left(k,\ y(k),\ \Delta y(k),\ ...,\ \Delta^m y(k)\right) = 0, \tag{31.3}$$

where $y(k)$ is an unknown function of integer argument, and $\Delta y(k)$, ..., $\Delta^m y(k)$ – its differences, is called a **difference equation** or a **finite difference** equation of the *m*-th order.

The solution of a difference equation is any mesh function that turns it into an identity.

Earlier, we made sure that finite differences of various orders can be expressed in terms of original mesh function values. Therefore, equation (31.3) can be represented as:

$$F_1\left(k,\ y(k+m),\ ...,\ y(k+1),\ y(k)\right) = 0. \tag{31.4}$$

Difference equations have numerous applications in discrete-time models of economic dynamics.

31.2. Linear difference equations

Definition. Difference equation of the form

$$a_0(k)\,y(k+m) + a_1(k)\,y(k+m-1) + ... + a_m(k)\,y(k) = f(k), \tag{31.5}$$

where $a_j(k)$ and $f(k)$ are known functions, and $y(k+j)$ is an unknown function from k (j = 0, 1, ..., m), moreover $a_m(k)$ and $a_0(k)$ are not equal at any k, called **the *m*-th order linear difference equation.**

In case coefficients a_0, a_1, ..., a_m are constants, methods for solving such equations are similar to methods for solving linear differential equations with constant coefficients.

Together with an inhomogeneous equation

$$a_0 y(k+m) + a_1 y(k+m-1) + ... + a_m y(k) = f(k) \tag{31.6}$$

the corresponding homogeneous equation is considered

$$a_0 y(k+m) + a_1 y(k+m-1) + ... + a_m y(k) = 0. \tag{31.7}$$

For difference equations (in particular, for linear difference equations), as well as for their differential analogues, the concepts of general and particular solutions are defined.

General solution of equation (31.6) has the form:

$$y(k) = \varphi(k, c_1, ..., c_m),$$

where $c_1, ..., c_m$ are arbitrary constants; their number is equal to the order of the equation.

Particular solution of equation (31.6) is distinguished by setting the values of function $y(k)$ at m arbitrary but consecutive points.

As well as for linear differential equations, the concept of a linearly independent system of solutions is determined, it is proved that the general solution of equation (31.6) has the form

$$y(k) = y_0(k) + \tilde{y}(k), \tag{31.8}$$

where $y_0(k)$ is a general solution of the corresponding homogeneous equation (31.7), and $\tilde{y}(k)$ is some particular solution of the original equation (31.6).

In the future, we will restrict ourselves to the consideration of the second-order difference equations. The results that we will obtain can be extended to difference equations of higher orders.

So, we consider a *second-order* ***homogeneous*** *linear difference equati*on:

$$y(k+2)+py(k+1)+qy(k)=0. \tag{31.9}$$

We will search the solution to this equation in the form:

$$y(k)=\lambda^k.$$

We obtain a *characteristic equation* after obvious simplifications:

$$\lambda^2+p\lambda+q=0. \tag{31.10}$$

Three options are possible.

1. Both roots λ_1 and λ_2 of equation (31.10) are *real* and *distinctive*. In this case, the *general solution* has the form:

$$y_0(k)=c_1\lambda_1^k+c_2\lambda_2^k. \tag{31.11}$$

***Example* 31.3.** Find a general solution for the difference equation

$$y(k+2)-5y(k+1)+6y(k)=0.$$

Solution. Characteristic equation $\lambda^2-5\lambda+6=0$ has two distinctive real roots: $\lambda_1=3$, $\lambda_2=2$. Therefore, according to formula (31.11), the general solution of the given equation is mesh function

$$y_0(k)=c_1\cdot 3^k+c_2\cdot 2^k.$$

2. Both roots are *real* and equal to each other: $\lambda_1 = \lambda_2 = \lambda$. Then the *general solution* has the form

$$y_0(k) = c_1\lambda^k + c_2 k\lambda^k.$$

3. The characteristic equation has *complex conjugate* roots $\lambda_1 = \alpha + \beta i$, $\lambda_2 = \alpha - \beta i$.

Represent the roots in trigonometric form: $\lambda_1 = r(\cos\varphi + i\sin\varphi)$, $\lambda_2 = r(\cos\varphi - i\sin\varphi)$, where the module is $r = \sqrt{\alpha^2 + \beta^2}$, and the argument ϕ is defined by the ratio $\operatorname{tg}\varphi = \dfrac{\beta}{\alpha}$.

The *general solution* has the form

$$y_0(k) = r^k(c_1\cos k\varphi + c_2\sin k\varphi).$$

Let us now turn to the *second-order* ***inhomogeneous*** *linear difference equation*:

$$y(k+2) + py(k+1) + qy(k) = f(k). \qquad (31.12)$$

Its general solution has the form (31.8). To find a particular solution $\tilde{y}(k)$ of equation (31.12) the method of indefinite coefficients is often used.

***Example* 31.4.** Solve equation

$$y(k+2) - 7y(k+1) + 10y(k) = 4\cdot 6^k.$$

Solution. To find a general solution to the corresponding homogeneous equation, we compose the characteristic equation:

$$\lambda^2 - 7\lambda + 10 = 0.$$

Its roots are $\lambda_1 = 5$, $\lambda_2 = 2$. Hence,

$y_0(k) = c_1 \cdot 5^k + c_2 \cdot 2^k$.

To find a particular solution $\tilde{y}(k)$ to the original equation, we use the indefinite coefficients method. We will search $\tilde{y}(k)$ in the form $\tilde{y}(k) = c \cdot 6^k$. Substituting this expression in the equation, we obtain:

$c \cdot 6^{k+2} - 7c \cdot 6^{k+1} + 10c \cdot 6^k = 4 \cdot 6^k$,

$c \cdot (36 - 42 + 10) \cdot 6^k = 4 \cdot 6^k$.

Hence, $c = 1$, which means,

$\tilde{y}(k) = 6^k$.

Adding $y_0(k)$ and $\tilde{y}(k)$, we get the general solution of the equation:

$y(k) = c_1 \cdot 5^k + c_2 \cdot 2^k + 6^k$.

31.3. The Samuelson –Hicks business cycle model

As an example of the difference equations application, we consider **the Samuelson–Hicks business cycle model** known in macroeconomic theory. This model uses the assumption that the volume of investment is directly proportional to the growth of national income. This assumption – *acceleration principle* already known to us – is described by the following equation:

$$I(k) = \chi \left(y(k-1) - y(k-2) \right), \tag{31.13}$$

where $|$ is the proportionality coefficient called the acceleration factor ($| > 0$), $I(k)$ is the amount of investments in the period *k* (in the *k*-th

calendar year), and $y(k-1)$, $y(k-2)$ is the national income in the previous periods – in the $(k-1)$-th and $(k-2)$-th respectively. It is assumed that consumption $C(k)$ in the considered *k*-th period also linearly depends on the value of national income $y(k-1)$ for the previous period:

$$C(k) = ay(k-1) + b. \tag{31.14}$$

It is assumed that income $y(k)$ is divided between producers and consumers. Therefore

$$y(k) = C(k) + I(k). \tag{31.15}$$

We substitute in (31.15) the expression for $I(k)$ from (31.13), as well as the expression for $C(k)$ from (31.14):

$$y(k) = ay(k-1) + b + \chi\left[y(k-1) - y(k-2)\right].$$

We get the so-called **Hicks equation**:

$$y(k) - (a+\chi)y(k-1) + \chi y(k-2) = b. \tag{31.16}$$

If we assume that the values *a* and | are constant over the considered time periods, then equation (31.16) is a *second-order linear inhomogeneous difference equation with constant coefficients*.

If we assume that the value of national income remained *constant* over the considered period, i.e.

$$y(k) = y(k-1) = y(k-2) = \tilde{y},$$

we can find a simple particular solution to equation (31.16):

$$\tilde{y} = (a+\chi)\tilde{y} - \chi\tilde{y} + b.$$

From here

$$\tilde{y} = b(1-a)^{-1}. \qquad (31.17)$$

Expression $(1-a)^{-1}$ in formula (31.17) is called *the Keynes multiplier*.

***Example* 31.5.** Consider the Samuelson-Hicks model, provided a = 0,48, β = 0,72, b = 1,3. Find a general solution to the Hicks equation.

Solution. In this case, equation (31.16) has the form:

$$y(k) - 1{,}2\,y(k-1) + 0{,}72\,y(k-2) = 1{,}3.$$

The particular solution to this equation, according to (31.17), is

$$\tilde{y}(k) = \frac{1{,}3}{1-0{,}48} = 2{,}5.$$

We write the characteristic equation:

$$\lambda^2 - 1{,}2\lambda + 0{,}72 = 0.$$

Its roots are $\lambda_{1,2} = 0{,}6 \pm 0{,}6i = 0{,}6\cdot\sqrt{2}\left(\cos\frac{\pi}{4} \pm i\sin\frac{\pi}{4}\right)$.

The general solution to the corresponding homogeneous equation is

$$y_0(k) = \left(0{,}6\sqrt{2}\right)^k\left(c_1\cos\frac{\pi k}{4} + c_2\sin\frac{\pi k}{4}\right).$$

We get the general solution of this equation:

$$y(k) = 2{,}5 + \left(0{,}6\sqrt{2}\right)^k\left(c_1\cos\frac{\pi k}{4} + c_2\sin\frac{\pi k}{4}\right).$$

In the considered example, dynamics are oscillatory with a damping amplitude. Obviously, with complex conjugate roots of the characteristic equation with the absolute value exceeding one, dynamics would be growing. In general, depending on the values of a and β dynamics can be

growing or damping and at the same time have or not have an oscillatory character.

Questions

1. Which functions are called mesh functions?
2. How are the first, second and subsequent differences of the grid function determined?
3. What equations are called finite-difference?
4. What is a characteristic equation for homogeneous linear difference equation?
5. How to find a general solution to the inhomogeneous difference equation?

Section IX. SERIES

Chapter 32. Number series

32.1. Concept of numeric series

Definition. Consider an arbitrary numerical sequence

$a_1, a_2, \ldots, a_n, \ldots$.

The formally composed infinite sum of all elements of this sequence, i.e. expression of the form

$$a_1 + a_2 + \ldots + a_n + \ldots \qquad (32.1)$$

is called a **numeric series** or simply **series**.

The numbers themselves $a_1, a_2, \ldots, a_n, \ldots$ are called **terms of series**, the *n*-th term of series a_n is the **general term of series**.

A series is considered given if its general term a_n is given. For example, to set series

$$\frac{1}{2} + \frac{1}{4} + \frac{1}{6} + \frac{1}{8} + \ldots,$$

it is necessary to indicate that its terms are given by the formula

$$a_n = \frac{1}{2n.}$$

Series (32.1) is also written as $\sum_{n=1}^{\infty} a_n$.

Note that from the real numbers theory we only know what the sum of a finite number of numbers means. The sum of an infinite number of terms has not yet been determined.

Consider the sum of a finite number with the first terms of series:

$$S_1 = a_1,$$
$$S_2 = a_1 + a_2,$$
$$S_3 = a_1 + a_2 + a_3,$$
$$\ldots\ldots\ldots\ldots\ldots$$
$$S_n = a_1 + a_2 + a_3 + \ldots + a_n,$$

called **partial sums of series** (32.1).

Since the number of series terms is infinite, the partial sums of series form an infinite numerical sequence:

$$S_1,\ S_2,\ S_3,\ \ldots,\ S_n,\ \ldots .$$

Definition. A series is called **convergent** if there is a finite limit to sequence of its partial sums, i.e.

$$\lim_{n\to\infty} S_n = S \quad . \qquad (32.2)$$

Otherwise, the series (32.1) is called **divergent**.

The number S defined by (32.2) is called the **sum** of series.

Let series (32.1) converge, S be its sum. Consider the difference between value S and partial sum S_n of this series: $S - S_n = r_n$. The value r_n is called a **remainder** of series. It's obvious that $\lim\limits_{n\to\infty} r_n = \lim\limits_{n\to\infty}(S - S_n) = 0$.

If series (32.1) converges, its sum is written in the form of symbolic equality

$$S = a_1 + a_2 + \ldots + a_n + \ldots, \qquad \text{or} \qquad S = \sum_{n=1}^{\infty} a_n \quad .$$

***Example* 32.1.** Consider a series composed of infinite geometric progression terms:

$$b + bq + bq^2 + ... + bq^{n-1} + ... \quad . \tag{32.3}$$

Partial sum S_n of this series is the sum of *n* terms of geometric progression:

$$S_n = b + bq + bq^2 + ... + bq^{n-1} \quad .$$

This sum, as known, with $q \neq 1$ has the form

$$S_n = \frac{b\left(1 - q^n\right)}{1 - q} = \frac{b}{1 - q} - \frac{bq^n}{1 - q} \quad .$$

If $|q| < 1$, that $\lim\limits_{n \to \infty} q^n = 0$. Therefore

$$\lim_{n \to \infty} S_n = \lim_{n \to \infty} \left(\frac{b}{1 - q} - \frac{bq^n}{1 - q} \right) = \frac{b}{1 - q} \quad ,$$

i.e. series (32.3) *converges* and its sum $S = \dfrac{b}{1 - q}$.

It is easy to verify that for $|q| \geq 1$ series (32.3) *diverges.*

***Example* 32.2.** Consider series

$$\frac{1}{1 \cdot 2} + \frac{1}{2 \cdot 3} + ... + \frac{1}{n\left(n + 1\right)} + ... \quad .$$

Obviously, $a_n = \dfrac{1}{n\left(n + 1\right)} = \dfrac{1}{n} - \dfrac{1}{n + 1}$. Therefore

$$S_n = \left(1 - \frac{1}{2}\right) + \left(\frac{1}{2} - \frac{1}{3}\right) + ... + \left(\frac{1}{n} - \frac{1}{n + 1}\right) \quad .$$

If you open brackets, then all summands, except the first and last one, will be mutually destroyed. We get

$$S_n = 1 - \frac{1}{n+1}.$$

From here $\lim\limits_{n\to\infty} S_n = 1$. So, the series converges and its sum is equal to 1.

32.2. Basic properties of series

Property 1. If series $a_1 + a_2 + ... + a_n + ...$ converges and its sum is equal to *S*, then for any number λ series $\lambda a_1 + \lambda a_2 + ... + \lambda a_n + ...$ also converges and its sum is equal to λS.

Proof. Let a convergent series be given

$$a_1 + a_2 + ... + a_n + ... \tag{32.1}$$

and let *S* be its sum. Form series

$$\lambda a_1 + \lambda a_2 + ... + \lambda a_n + ... \tag{32.4}$$

Let

$$S_n = a_1 + a_2 + ... + a_n,$$

$$S'_n = \lambda a_1 + \lambda a_2 + ... + \lambda a_n.$$

Then, obviously, $S'_n = \lambda S$. But by condition $\lim\limits_{n\to\infty} S_n = S$. So $\lim\limits_{n\to\infty} S'_n = \lim\limits_{n\to\infty} \lambda S_n = \lambda \lim\limits_{n\to\infty} S_n = \lambda S$, Q.E.D.

Property 2. If series $a_1 + a_2 + ... + a_n + ...$ and $b_1 + b_2 + ... + b_n + ...$ converge and their sums are equal to *S* and S' respectively, then series

$(a_1 \pm b_1)+(a_2 \pm b_2)+...+(a_n \pm b_n)+...$ also converges and its sum is equal to $S \pm S'$.

Proof. Consider series

$$a_1 + a_2 + ... + a_n + ..., \quad (32.1)$$

$$b_1 + b_2 + ... + b_n + ..., \quad (32.5)$$

$$(a_1 + b_1)+(a_2 + b_2)+(a_3 + b_3)+...+(a_n + b_n)+... \quad (32.6)$$

Let S_n, S'_n, S''_n be the partial sums of series (32.1), (32.5), (32.6) respectively:

$$S_n = a_1 + a_2 + a_3 + ... + a_n,$$

$$S'_n = b_1 + b_2 + b_3 + ... + b_n,$$

$$S''_n = (a_1 + b_1)+(a_2 + b_2)+(a_3 + b_3)+...+(a_n + b_n).$$

Let S be the sum of series (32.1): $\lim\limits_{n\to\infty} S_n = S$ and let S' be the sum of series (31.5): $\lim\limits_{n\to\infty} S'_n = S'$. Then there is a limit

$$\lim_{n\to\infty} S''_n = \lim_{n\to\infty}(S_n + S'_n) = \lim_{n\to\infty} S_n + \lim_{n\to\infty} S'_n = S + S',$$

i.e. series (32.6) converges and its sum is equal to $S + S'$.

Property 3. Dropping a finite number of series terms does not affect its convergence (divergence).

Proof. In this case, all partial sums of series, starting from a certain sum, will change by the same constant number, equal to the sum of dropping terms. Therefore, a sequence of original series partial sums has a finite limit if and only if there is a finite limit to the sequence of series partial sums obtained from the original by dropping a finite number of terms.

We distinguish another property as well.

Necessary criterion of series convergence

Property 4. General term a_n of the converging series tends to zero with $n \to \infty$.

Proof. Let the series converge and its sum be equal to *S*, i.e. $\lim\limits_{n\to\infty} S_n = S$.

Obviously, $a_n = S_n - S_{n-1}$. Also obviously that $\lim\limits_{n\to\infty} S_{n-1} = S$. Therefore,

$$\lim_{n\to\infty} a_n = \lim_{n\to\infty}\left(S_n - S_{n-1}\right) = \lim_{n\to\infty} S_n - \lim_{n\to\infty} S_{n-1} = S - S = 0.$$

Note that we established only a necessary criterion for convergence, which is not sufficient. From the fact that $\lim\limits_{n\to\infty} a_n = 0$ it does not follow yet that the series converges.

***Example* 32.3.** Consider series $1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n} + \ldots$, which is called *harmonic*.

Obviously, for harmonic series, the necessary criterion for convergence is satisfied $\lim\limits_{n\to\infty} a_n = \lim\limits_{n\to\infty} \frac{1}{n} = 0$. Despite this let us prove that a harmonic series diverges. Assume the opposite, i.e. that the series converges and $\lim\limits_{n\to\infty} S_n = S$. In this case, obviously, $\lim\limits_{n\to\infty} S_{2n} = S$, therefore:

$$\lim_{n\to\infty}\left(S_{2n} - S_n\right) = \lim_{n\to\infty} S_{2n} - \lim_{n\to\infty} S_n = S - S = 0. \qquad (*)$$

But

$$S_{2n} - S_n = \frac{1}{n+1} + \frac{1}{n+2} + \ldots + \frac{1}{2n} > \frac{1}{2n} + \frac{1}{2n} + \ldots + \frac{1}{2n} = n \cdot \frac{1}{2n} = \frac{1}{2},$$

i.e. $S_{2n} - S_n > \frac{1}{2}$, and this is contrary to equality (*). The resulting contradiction means that our assumption of a harmonic series convergence is incorrect.

Note that we can prove the divergence of a series with the help of the necessary criterion for convergence.

***Example* 32.4.** Investigate the convergence of series

$$\sum_{n=1}^{\infty} \frac{2n-1}{7n+5} = \frac{1}{12} + \frac{3}{19} + \frac{5}{26} + ... + \frac{2n-1}{7n+5} + ...$$

Solution. $\lim\limits_{n\to\infty} a_n = \lim\limits_{n\to\infty} \frac{2n-1}{7n+5} = \frac{2}{7} \neq 0$, therefore, this series diverges.

32.3. Series with non-negative terms

A series with non-negative terms (i.e., series whose terms are all non-negative) are the simplest type of number series. **The main property of a series with non-negative terms**: the sequence of the partial sums of such a series is *non-decreasing*.

Convergence criterion

***Theorem* 32.1.** For convergence of a series with non-negative terms, it is necessary and sufficient that the sequence of its partial sums is bounded.

Proof. 1. *Necessity.* Let the series converge. This means that the sequence of its partial sums converges. A convergent sequence, as you know, is limited.

2. *Sufficiency.* Since the partial sums sequence of series is bounded and monotonic, it converges by Theorem 14.4.

Comparison criterions

***Theorem* 32.2 (the first comparison criterion).** Let two series with positive terms be given:

$$\sum_{n=1}^{\infty} a_n = a_1 + a_2 + ... + a_n + ... \quad , \tag{32.7}$$

$$\sum_{n=1}^{\infty} b_n = b_1 + b_2 + ... + b_n + ... \quad , \tag{32.8}$$

moreover $a_n \le b_n$ for all *n*. Then the convergence of series (32.7) follows from the convergence of series (32.8), and the divergence of series (32.8) follows from the divergence of series (32.7).

Proof. Let S_n be the partial sum of series (32.7), S'_n be the partial sum of series (32.8). From the theorem's conditions it follows that $S_n \le S'_n$. If series (32.8) converges, then sequence $\{S'_n\}$ is bounded. Therefore, sequence $\{S_n\}$ is also bounded and converges by Theorem 14.4, i.e. series (32.7) converges.

If series (32.7) diverges, then series (32.8) also diverges. Indeed, if series (32.8) converges, then series (32.7) should also converge (as it has just been proved above). The theorem is proved.

Note that under the conditions of Theorem 32.2, series (32.8) is called the **majorant** of series (32.7), and series (32.7) is called the **minorant** of series (32.8).

The proved theorem can also be stated in the following convenient form for memorization: if the majorant converges, then the minorant converges; if the majorant diverges, then the majorant diverges.

We consider applications of Theorem 32.2.

***Example* 32.5.** Investigate the convergence of series $\sum_{n=1}^{\infty}\frac{1}{2^n+n}$.

Solution. Obviously, $\frac{1}{2^n+n}<\frac{1}{2^n}$, and series $\sum_{n=1}^{\infty}\frac{1}{2^n}$ converges (sum of infinite decreasing geometric progression). Consequently, this series also converges.

***Example* 32.6.** Investigate the convergence of series $\sum_{n=1}^{\infty}\frac{1}{\sqrt{n}}$.

Solution. Since $\frac{1}{\sqrt{n}}\geq\frac{1}{n}$, and series $\sum_{n=1}^{\infty}\frac{1}{n}$ diverges (this is harmonic series), then this series diverges.

***Theorem* 32.3 (the second comparison criterion).** Let (32.7) be a series with non-negative terms and (32.8) – with positive terms, and let there be a non-zero finite limit

$$\lim_{n\to\infty}\frac{a_n}{b_n}=l$$

.

Then both series (32.7) and (32.8) converge or diverge together.

Proof. According to the definition of the limit for arbitrary $\varepsilon>0$, there exists such *N* that for all $n>N$ that the following inequality is satisfied:

$$l-\varepsilon<\frac{a_n}{b_n}<l+\varepsilon$$

, or $(l-\varepsilon)b_n<a_n<(l+\varepsilon)b_n$.

Let series (32.8) converge. Then by property 1 (see §32.2) series $\sum_{n=1}^{\infty}(l+\varepsilon)b_n$, and by theorem 32.2 series (32.7) converges. Similarly, if

series (32.7) converges, then by theorem 32.2 series $\sum_{n=1}^{\infty}(l-\varepsilon)b_n$ converges and by the same property 1 series (32.8) converges.

The statement of series divergence theorem is proved similarly.

Comment. It can be assumed that the common terms of series (32.7) and (32.8), i.e. a_n and b_n are infinitesimal at $n \to \infty$ (otherwise everything would be clear by itself: series, whose common term does not tend to zero, diverges). Therefore ***theorem* 32.3** can be reformulated as follows: if terms a_n and b_n of two positive series are infinitesimal of the same order, then these series converge or diverge together.

***Example* 32.7.** Investigate the convergence of series $\sum_{n=1}^{\infty}\frac{3n-7}{n^2}$.

Solution. Let us compare this series with divergent harmonic series $\sum_{n=1}^{\infty}\frac{1}{n}$ (see example 32.3). Since

$$\lim_{n\to\infty}\left(\frac{3n-7}{n^2}:\frac{1}{n}\right)=\lim_{n\to\infty}\frac{3n^2-7n}{n^2}=3\neq 0,$$

then this series diverges.

Other convergence criteria

Note that both comparison criterions discussed above (theorems 32.2 and 32.3) have the same *disadvantage*: to investigate the convergence of any positive series with this criterion, for comparison with this series it is necessary to choose some other series, the convergence (or divergence) of which is known. There are no general methods for finding such a series. It all depends on intuition, on how extensive the researcher's stock of such "reference" series is, the convergence or divergence of which is known.

Therefore, it is very useful to have at your disposal such convergence criteria, for which it is not necessary to involve any new series, except for the studied one.

***Theorem* 32.4 (D'Alembert criterion).** Let for series $\sum_{n=1}^{\infty} a_n$ there be a limit

$$\lim_{n\to\infty} \frac{a_{n+1}}{a_n} = l \quad . \tag{32.9}$$

Then, when $l < 1$ the series *converges*, and when $l > 1$ the series *diverges*.

Proof. Due to the definition of limit for any $\varepsilon > 0$, there exists such number *N* that for all $n > N$ inequalities are satisfied:

$$l - \varepsilon < \frac{a_{n+1}}{a_n} < l + \varepsilon \quad . \tag{32.10}$$

1. Let $l < 1$. Then take such ε, that $l + \varepsilon < 1$. Denote $l + \varepsilon = q$. From inequalities (32.7) we have:

$$\frac{a_{n+1}}{a_n} < q \quad , \qquad \text{or} \qquad a_{n+1} < a_n q$$

for all $n > N$. We get a system of inequalities

$$a_{N+2} < a_{N+1} q \, , \; a_{N+3} < a_{N+2} q < a_{N+1} q^2 \, , \ldots .$$

So, the terms of series starting from a_{N+2} are smaller than the corresponding terms of decreasing geometric progression. Therefore, the series converges.

2. Let $l>1$. Then take such ε, that $l-\varepsilon>1$. Then it follows from the left inequality (32.10) that $a_{n+1}>a_n$ for all $n>N$, i.e., the terms of series starting from $(N+1)$-th increase, so the limit of the common term is not equal to zero; hence, the series diverges.

***Example* 32.8.** Investigate the convergence of series $\sum_{n=1}^{\infty}\frac{n^2}{2^n}$.

Solution: $\lim\limits_{n\to\infty}\frac{a_{n+1}}{a_n}=\lim\limits_{n\to\infty}\left(\frac{(n+1)^2}{2^{n+1}}:\frac{n^2}{2^n}\right)=\lim\limits_{n\to\infty}\frac{(n+1)^2}{2n^2}=\frac{1}{2}<1$.

Consequently, the series converges on the basis of D'Alembert criterion.

Comment. When $l=1$, the series can both converge and diverge. In particular, for series $\sum_{n=1}^{\infty}\frac{1}{n}$ and $\sum_{n=1}^{\infty}\frac{1}{n^2}$, as it is easy to see, $l=1$, but the first of them (the harmonic series), as we know, diverges, and the second one, as we learn later, converges.

***Theorem* 32.5 (Cauchy criterion).** If for terms of series $\sum_{n=1}^{\infty}a_n$ there is a limit

$$\lim_{n\to\infty}\sqrt[n]{a_n}=l,$$

that series converges when $l<1$ and diverges when $l>1$.

The Proof of this theorem is also based on the fact that when $l<1$, the terms of the series starting from some number are less than the terms of some infinite decreasing geometric progression and when $l>1$ the total term of series does not tend to zero.

1. Let $l<1$. Take some number q satisfying the relation $l<q<1$.

Since $\lim\limits_{n\to\infty}\sqrt[n]{a_n}=l$, then starting from some number $n=N$ the inequality will be satisfied

$$\left|\sqrt[n]{a_n}-l\right|<q-l.$$

It follows that

$$\sqrt[n]{a_n}<q,$$

or, which is the same,

$$a_n<q^n \qquad (*)$$

for all $n\geq N$.

Compare two series:

$$a_1+a_2+\ldots+a_N+a_{N+1}+a_{N+2}+\ldots, \qquad (32.1')$$

$$q^N+q^{N+1}+q^{N+2}+\ldots. \qquad (**)$$

The terms of series (**) form a decreasing geometric progression (its denominator is *q*, by condition *q <1*). Therefore, series (**) converges. We know that discarding a finite number of series terms does not affect its convergence. It follows from condition (*) that the terms of series (32.1 ') starting with a_N are smaller than the corresponding terms of converging series (**). Therefore, series (32.1) converges.

2. Let $l>1$. In this case, it is easy to verify that the limit of the series general term is not equal to zero; therefore, the series diverges. The theorem is proved.

***Example* 32.9.** Investigate the convergence of series $\sum\limits_{n=1}^{\infty}\left(1-\frac{1}{n}\right)^{n^2}$.

Solution. Apply the Cauchy criterion

$$\lim_{n\to\infty}\sqrt[n]{a_n}=\lim_{n\to\infty}\left(1-\frac{1}{n}\right)^n=\mathrm{e}^{-1}=\frac{1}{\mathrm{e}}<1.$$

Therefore, the series converges.

Note that when $l=1$ **the** Cauchy criterion also does not answer the question of the series convergence.

It should be noted that Cauchy and D'Alembert criteria are effective mainly for finding out the convergence of «rapidly» converging series, whose terms are infinitesimal of the same (or higher) order as the terms of decreasing geometric progressions. We have already noted that the D'Alembert criterion does not answer the question of convergence or divergence of harmonic series

$$1+\frac{1}{2}+\frac{1}{3}+\ldots+\frac{1}{n}+\ldots,$$

which, as we know, diverges, as well as the convergence or divergence of generalized harmonic series

$$1+\frac{1}{2^2}+\frac{1}{3^2}+\ldots+\frac{1}{n^2}+\ldots,$$

which, as we will know soon, converges.

Theorem 32.6 (integral convergence criterion). Let the terms of series $\sum_{n=1}^{\infty}a_n$ not increase, i.e. $a_1\ge a_2\ge\ldots\ge a_n\ge\ldots$, and let $f(x)$ be such continuous non-increasing function defined for $x\ge 1$ that

$f(1)=a_1$, $f(2)=a_2$, …, $f(n)=a_n$, … .

Then for the convergence of series $\sum_{n=1}^{\infty} a_n$, it is necessary and sufficient that the integral

$$\int_1^{\infty} f(x)\,dx$$

converges.

Proof. Since $f(x)$ is monotone, then for $x \in [n, n+1]$ the following inequality is satisfied $f(n) \geq f(x) \geq f(n+1)$, or

$$a_n \geq f(x) \geq a_{n+1} \tag{32.11}$$

for any *n*.

Integrate (32.8) on segment $[n, n+1]$:

$$\int_n^{n+1} a_n dx \geq \int_n^{n+1} f(x)dx \geq \int_n^{n+1} a_{n+1}dx .$$

We have

$$a_n \geq \int_n^{n+1} f(x)dx \geq a_{n+1} . \tag{32.12}$$

Consider series

$$\int_1^2 f(x)\,dx + \int_2^3 f(x)\,dx + \ldots + \int_n^{n+1} f(x)\,dx + \ldots . \tag{32.13}$$

Its *n*-th partial sum S_n has the form

$$S_n = \int_1^2 f(x)\,dx + \int_2^3 f(x)\,dx + \ldots + \int_n^{n+1} f(x)\,dx + \ldots + \int_n^{n+1} f(x)\,dx = \int_n^{n+1} f(x)\,dx$$

. (32.14)

The convergence of series (32.10) means the existence of finite limit of its partial sums sequence (32.11), i.e. the convergence of improper integral $\int_1^{\infty} f(x)\,dx$ since

$$\lim_{n\to\infty} S_n = \lim_{n\to\infty} \int_1^{n+1} f(x)\,dx = \int_1^{\infty} f(x)\,dx$$

.

If the series converges, then, according to theorem 32.2, due to the left inequality (32.12), series (32.13) must also converge, and hence improper integral $\int_1^{\infty} f(x)\,dx$. Conversely, if integral $\int_1^{\infty} f(x)\,dx$ converges, i.e. series (32.13) converges, then by the same theorem 32.2 series

$$\sum_{n=1}^{\infty} a_{n+1} = a_2 + a_3 + \ldots + a_n + a_{n+1} + \ldots$$

,

must converge and therefore, this series $\sum_{n=1}^{\infty} a_n$.

***Example* 32.10.** Find out for what $\alpha > 0$ series $\sum_{n=1}^{\infty} \frac{1}{n^{\alpha}}$ converges (this series is called *generalized harmonic*).

Solution. Consider function $f(x) = \frac{1}{x^{\alpha}}$, $x \geq 1$. This function is monotonously decreasing. Therefore, the convergence of the given series

is equivalent to the convergence of improper integral $\int_1^{\infty} \frac{dx}{x^{\alpha}}$. It was previously established (example 24.2) that this integral converges when $\alpha > 1$ and diverges when $0 < \alpha \le 1$. Therefore, this series converges when $\alpha > 1$ and diverges when $\alpha \le 1$.

32.4. Series with terms of the arbitrary sign

Let us proceed to the study of series containing both positive and negative terms.

Definition. The number series is called **alternating** if it has an infinite number of both positive and negative terms.

Consider series

$$\sum_{n=1}^{\infty} a_n = a_1 + a_2 + ... + a_n + ... \qquad (*)$$

and besides, the series

$$\sum_{n=1}^{\infty} |a_n| = |a_1| + |a_2| + ... + |a_n| + ... \qquad (**)$$

.

It can be proved that from the convergence of series (**) follows the convergence of series (*).

Series (*) is called **absolutely convergent** if series (**) converges.

Series (*) is called **conditionally convergent** if it converges, but series (**) composed of modules of its terms diverges.

Definition. A number series $\sum_{n=1}^{\infty} a_n$ is called **alternating** if for any *n* the terms of series a_n and a_{n+1} have different signs. Assuming $a_1 > 0$, you can write the alternating series in the form

$$\sum_{n=1}^{\infty} (-1)^{n+1} c_n = c_1 - c_2 + c_3 - c_4 + ... + (-1)^{n+1} c_n + ..., \qquad (32.15)$$

where $c_n > 0$.

We formulate and prove sufficient criteria for the convergence of image series.

Theorem 32.7 (Leibniz criterion). If the terms of alternating series (32.15) decrease in absolute value:

$$c_1 > c_2 > ... > c_n > ...$$

and the limit of general term of this series for $n \to \infty$ is zero, i.e. $\lim_{n\to\infty} c_n = 0$, the series converges, and its sum does not exceed the first term: $S \le c_1$.

Proof. Consider a partial sum of series (32.15) with an even number of terms $S_{2m} = c_1 - c_2 + c_3 - c_4 + ... + c_{2m-1} - c_{2m}$. It can be represented as

$$S_{2m} = (c_1 - c_2) + (c_3 - c_4) + ... + (c_{2m-1} - c_{2m}).$$

Due to the theorem condition, all differences in brackets are positive, so sequence $\{S_{2m}\}$ is increasing.

On the other hand, S_{2m} can be represented as

$$S_{2m} = c_1 - (c_2 - c_3) - (c_4 - c_5) - ... - (c_{2m-2} - c_{2m-1}) - c_{2m},$$

hence $S_{2m} < c_1$.

So sequence $\{S_{2m}\}$ increases and is limited, hence it has limit $\lim\limits_{m\to\infty} S_{2m} = S$. From inequality $S_{2m} < c_1$ it follows that $S \le c_1$.

Since $S_{2m+1} = S_{2m} + c_{2m+1}$ and by condition $\lim\limits_{m\to\infty} c_{2m+1} = 0$, then $\lim\limits_{m\to\infty} S_{2m+1} = \lim\limits_{m\to\infty} S_{2m} = S$.

So, for any *n* (both for $n = 2m$ and for $n = 2m+1$) $\lim\limits_{n\to\infty} S_n = S$, i.e. the series converges.

***Example* 32.11. .** Investigate the convergence of series $\sum\limits_{n=1}^{\infty} \frac{(-1)^n}{n}$.

Solution. In this case, $c_n = \frac{1}{n}$ and the conditions of Leibniz criterion are fulfilled. Therefore, this series converges. However, the series composed of terms' modules of given series is harmonic series $\sum\limits_{n=1}^{\infty} \frac{1}{n}$ and, as we know, diverges. Therefore, this series converges conditionally.

***Comment* 1.** In the Leibniz theorem not only is condition $\lim\limits_{n\to\infty} c_n = 0$, but also condition $c_1 > c_2 > \ldots > c_n > \ldots$ is essential.
Consider, for example, series

$$\frac{1}{\sqrt{2}-1} - \frac{1}{\sqrt{2}+1} + \ldots + \frac{1}{\sqrt{n+1}-1} - \frac{1}{\sqrt{n+1}+1} + \ldots .$$

The terms of this series tend to zero, but the monotonicity condition is not satisfied.

Obviously, $\frac{1}{\sqrt{2}-1}-\frac{1}{\sqrt{2}+1}=2$, $\frac{1}{\sqrt{3}-1}-\frac{1}{\sqrt{3}+1}=1$, …, $\frac{1}{\sqrt{n+1}-1}-\frac{1}{\sqrt{n+1}+1}=\frac{2}{n}$, … .

Therefore, this series can be represented as:

$2+1+\frac{2}{3}+...+\frac{2}{n}+...$, i.e. $\sum_{n=1}^{\infty}\frac{2}{n}$.

This series diverges since it is obtained from the harmonic series by doubling all its terms.

***Comment* 2.** In the process of proving the Leibniz theorem, we saw that increasing S_{2m} approaches *S*. S_{2m+1} on the contrary, decreases.

Consider S_{2m+1} more detailed:

$S_1 = c_1$,

$S_3 = c_1 - (c_2 - c_3)$,

$S_5 = c_1 - (c_2 - c_3) - (c_4 - c_5)$,

……

Since each of differences written in brackets according to the condition is positive, then obviously,

$S_1 > S_3 > S_5 > ...$

Thus, if the series satisfies the conditions of Leibniz theorem, then sums S_{2m} are approximate values of sum *S* with the disadvantage and sums S_{2m+1} – with the excess.

***Comment* 3.** Sum *S* of series satisfying the conditions of the Leibniz theorem does not exceed in absolute value its first term and has the same

sign as the first term. Indeed, $S_1 = a_1$, $S_2 = a_1 - a_2 > 0$. Therefore, from inequality

$S_2 < S < S_1$

we get

$0 < S < c_1$.

Further, we note that the remainder of a series satisfying the conditions of Leibniz theorem is itself a series satisfying these conditions, and the just made comment applies to it. If the sum of the *n*-th remainder is equal to r_n, then equality

$S = S_n + r_n$

allows us to make the following ***conclusion***: the error made when replacing the sum of series satisfying the conditions of the Leibniz theorem with its partial sum has the same sign as the first dropping term, and the absolute value is less than it.

Questions

1. What is the common term of series?
2. How is the seventh partial sum of series determined S_7?
3. What number series is called convergent?
4. What is the limit of the convergent series common term?
5. What is the remainder of the series?
6. What number series is called harmonic? Does harmonic series converge or diverge?
7. Will the series with positive terms converge for which the ratio limit of the subsequent term to the previous one is equal to 2?
8. What properties should the function used in the convergence integral criterion, have?
9. What limit expression is used in the Cauchy criterion?

10. For which values is 〈 of generalized harmonic series $\sum_{n=1}^{\infty}\frac{1}{n^{\alpha}}$ convergent?
11. Is it possible to establish the convergence or divergence of harmonic (generalized harmonic) series using D'Alembert criterion?
12. Will the alternating series converge for which the series of its terms modules converge?
13. What series is called conditionally convergent?
14. What conditions are sufficient for the convergence of a signed series?
15. Let the series satisfy conditions of the Leibniz theorem. How to estimate the error made when replacing the sum of this series with its partial sum?

Chapter 33. Functional series

33.1. Basic concepts

Let us consider a series whose members are functions defined in some domain *D*:

$$\sum_{n=1}^{\infty}u_n(x)=u_1(x)+u_2(x)+...+u_n(x)+... \quad . \qquad (33.1)$$

This series is called **functional series**.

By giving *x* specific numerical values, we get different numerical series that can converge or diverge.

The set of all values *x* at which the functional series converges is called **the convergence region of series**. Obviously, if D' is the region of

convergence of the series (33.1), then $D' \subseteq D$. The sum of the series is the function of *x* in the convergence region. It is denoted by $S(x)$.

Let us compose partial sums $S_n(x)$ for (33.1) as well as for number series. If the series (33.1) converges and its sum is equal to $S(x)$, then

$$S(x) = S_n(x) + + r_n(x), \tag{33.2}$$

where $r_n(x)$ is the sum of series $u_{n+1}(x) + u_{n+2}(x) + \ldots$, i.e.

$$r_n(x) = u_{n+1}(x) + u_{n+2}(x) + \ldots . \tag{33.3}$$

The value $r_n(x)$ is called **the remainder of the series** (33.1).

Since for every *x* in the convergence region of series we obtain the equality $\lim\limits_{n\to\infty} S_n(x) = S(x)$, then, taking into account (33.2) we obtain

$$\lim_{n\to\infty} r_n(x) = \lim_{n\to\infty} \left[S(x) - S_n(x)\right] = 0 .$$

Thus, the remainder $r_n(x)$ of the convergent series tends to zero as $n \to \infty$.

The convergence of series (33.1) in D' means that for each $x \in D'$ a sequence of partial sums $\{S_n(x)\}$ converges: $\lim\limits_{n\to\infty} S_n(x) = S_n$. According to the definition of the limit of a numerical sequence, for every $\varepsilon > 0$ there exists a positive integer *N* such that for all numbers $n > N$ holds

$$\left| S(x) - S_n(x) \right| < \varepsilon . \qquad (33.4)$$

Here, N depends on ε. Indeed, for the same $\varepsilon > 0$ but another x, it is necessary to choose another N to provide the inequality (33.4). Consider an example. Let there be a series

$$1 + x + x^2 + \ldots + x^n + \ldots .$$

It obviously converges as $x = \frac{1}{5}$ and $x = \frac{1}{10}$. Let $\varepsilon = 0{,}0004$. If $x = \frac{1}{5}$, we obtain

$$1 + \frac{1}{5} + \frac{1}{5^2} + \ldots + \frac{1}{5^n} + \ldots .$$

Its sum (by the formula of the sum of infinitely decreasing progression) is

$$S = \frac{1}{1 - \frac{1}{5}} = \frac{5}{4} = 1{,}25 .$$

We need to take $N = N_1 = 5$ for given $\varepsilon = 0{,}0004$ to satisfy (33.4). Indeed, $S_5 = 1{,}2496$, $S - S_5 = 0{,}0004$, and if $n > 5$, then $\left| S - S_n \right| < 0{,}0004$.

For $x = \frac{1}{10}$ we have:

$$1 + \frac{1}{10} + \frac{1}{10^2} + \ldots + \frac{1}{10^n} + \ldots .$$

Its sum $S = \frac{10}{9} = 1{,}11111...$

If $N = N_2 = 3$, then for $n > N_2$, in particular for $n = 4$, $\left| S - S_4 \right| = 1{,}11111... - 1{,}111 = 0{,}00011... < 0{,}0004$.

Therefore, for $\varepsilon = 0{,}0004$ we need to take $N = 5$ as $x = \frac{1}{5}$ and $N = 3$ as $x = \frac{1}{10}$. It is clear that inequality (33.4) holds for $x = \frac{1}{5}$ and $x = \frac{1}{10}$ as $N = 5$.

Is it always possible to find a number N for a given $\varepsilon > 0$ such that for any $n > N$ and for all $x \in D'$ inequality (33.4) holds? No. There are functional series for which this is not possible.

Definition. The functional series (33.1) is said to be **a uniformly convergent function series** in domain D' if for any $\varepsilon > 0$ there exists a number N such that for any $n > N$ and for all $x \in D'$

$$\left| S(x) - S_n(x) \right| < \varepsilon.$$

Number N, mentioned above, depends only on ε and does not depend on x: $N = N(\varepsilon)$.

The concept of uniform convergence is a very complex concept. It is not possible for now to study a convergent series in a general form. Consider an important special case of uniformly convergent series - majorizable series.

Definition. A functional series

$$u_1(x)+u_2(x)+...+u_n(x)+...$$

is called **majorizable** in some domain if there exists such a convergent number series

$$c_1+c_2+...+c_n+... \tag{33.5}$$

with positive terms that for all x from a given domain, the inequalities

$$|u_1(x)|\le c_1,\ |u_2(x)|\le c_2,\ \ldots,\ |u_n(x)|\le c_n,\ \ldots \tag{33.6}$$

hold.

(Mind the fact that a series is called majorizable if there exists a precisely *convergent* numerical majorant for it.)

For example, a functional series

$$\frac{\sin x}{2}+\frac{\sin 2x}{2^2}+\frac{\sin 3x}{2^3}+...+\frac{\sin nx}{2^n}+...$$

is a series majorizable on the whole number line since the inequalities

$$\left|\frac{\sin nx}{2^n}\right|\le\frac{1}{2^n}\quad (n=1,\,2,\,\ldots),$$

hold for all x and a series

$$\frac{1}{2}+\frac{1}{2^2}+\frac{1}{2^3}+...+\frac{1}{2^n}+...,$$

being a geometric progression, converges.

According to the definition, if a series is majorizable in a certain domain, then it absolutely converges in this domain. Now we introduce the following theorem.

***Theorem* 33.1 (Weierstrass M-test).** Let the functional series

$$u_1(x)+u_2(x)+...+u_n(x)+... \tag{33.7}$$

be majorizable on $[a,b]$; then it converges uniformly on $[a,b]$.

Proof. Denote by S' the sum of the series (33.5):

$$S' = c_1 + c_2 + ... + c_n + c_{n+1} + ... \tag{33.8}$$

Let S'_n be the n-th partial sum, r'_n be a remainder of the series (33.5) after the n-th term. Then

$$S' = S'_n + r'_n .$$

Since the series (33.5) converges

$$\lim_{n\to\infty} S'_n = S' .$$

Therefore,

$$\lim_{n\to\infty} r'_n = 0 .$$

As already noted, the sum of the functional series (33.1)

$$S(x)=S_n(x)+r_n(x), \tag{33.2}$$

where $S_n(x)$ is the n-th partial sum and $r_n(x)$ is a remainder of the series:

$$r_n(x)=u_{n+1}(x)+u_{n+2}(x)+... .$$

According to condition (33.6)

$$|u_{n+1}(x)| \le c_{n+1}, \; |u_{n+2}(x)| \le c_{n+2}, \ldots,$$

hence,

$$\left| r_n(x) \right| \le r_n'$$

for all $x \in [a,b]$. Since $\lim\limits_{n\to\infty} r_n' = 0$ and $\{r_n'\}$ is numerical sequence, then for any $\varepsilon > 0$ there exists a number *N*, independent on *x*, such that for all $n > N$ $\left| r_n' \right| < \varepsilon$. Therefore,

$$\left| S(x) - S_n(x) \right| < \varepsilon$$

for all $n > N$ and for all $x \in [a,b]$. So, the series (33.7) converges on $[a,b]$ uniformly. That completes the proof.

33.2. Properties of a uniformly convergent series

Continuity of the sum of a series

Consider series

$$u_1(x) + u_2(x) + \ldots + u_n(x) + \ldots , \qquad (33.1)$$

here $u_1(x), u_2(x), \ldots, u_n(x), \ldots$ are continuous functions on $[a,b]$. It is well known that the sum of continuous functions is a continuous function, but this is true for a finite number of terms. Any partial sum of series (33.1)

$$S_n(x) = u_1(x) + u_2(x) + \ldots + u_n(x)$$

is a continuous function on $[a,b]$. Is the sum of the series (33.1) continuous? It turns out that there are series of continuous functions having a *discontinuous sum*.

***Example* 33.1.** Consider the following series

$$1+(x-1)+(x^2-x)+(x^3-x^2)+\ldots+(x^n-x^{n-1})+\ldots . \qquad (*)$$

Members of this series are continuous functions for each *x*.

Let us make sure that this series converges on $[0,1]$ and its sum is a discontinuous function. Indeed, the partial sum of this series $S_n(x)$ has the form $S_n(x)=x^{n-1}$. Obviously,

$$S(x)=\lim_{n\to\infty} S_n(x)=0 \quad \text{as} \quad 0\le x<1,$$

$$S(1)=1.$$

Thus, the point $x=1$ is a point of discontinuity of the sum $S(x)$. This series converges irregularly on $[0,1]$. Indeed, for $u_n(x)$ we obtain $|S(x)-S_n(x)|=x^n$. For every fixed *n*, obviously, $\lim\limits_{x\to 1} x^n=1$. Therefore if $\varepsilon<1$, then it is impossible to provide inequality $|S(x)-S_n(x)|<\varepsilon$ for each $x\in[0,1]$ at the same time.

Thus, the series (*) consists of continuous functions but its sum is discontinuous function. Series converges irregularly on $[0,1]$, therefore, this series is not majorizable.

***Theorem* 33.2.** If functions $u_n(x)$ are defined on $[a,b]$ and are continuous on $[a,b]$, and series

$$u_1(x)+u_2(x)+...+u_n(x)+... \tag{33.1}$$

converges uniformly on $[a,b]$ to the sum $S(x)$, then $S(x)$ is also continuous on $[a,b]$.

Proof. Consider an arbitrary point x_0 on $[a,b]$ and let $S(x)$ be a continuous sum at that point. Since for any *n* and $x \in [a,b]$ equality

$$S(x)=S_n(x)+r_n(x),$$

holds, then, in particular,

$$S(x_0)=S_n(x_0)+r_n(x_0).$$

Hence

$$|S(x)-S(x_0)| \le |S_n(x)-S_n(x_0)|+|r_n(x)+r_n(x_0)|. \tag{**}$$

In order to prove the continuity of $S(x)$ we need to show that for $\varepsilon > 0$ there exists $\delta > 0$, such that for all *x* inequalities $|x-x_0|<\delta$, $|S(x)-S(x_0)|<\varepsilon$ holds. Since this series converges uniformly, then for given $\varepsilon > 0$ there exists a number *N*, such that for all $n > N$

$$|r_n(x)| < \frac{\varepsilon}{3} \tag{33.9}$$

for all $x \in [a,b]$, in particular, for $x = x_0$. This inequality holds, in particular, for $n = N+1$. Moreover, the function $S_n(x)$ is continuous at the point x_0 being the sum of a finite number of continuous functions as $n = N+1$. Therefore, for a given $\varepsilon > 0$ there exists $\delta > 0$, such that $|x - x_0| < \delta$, we obtain

$$|S_n(x) - S_n(x_0)| < \frac{\varepsilon}{3}. \qquad (33.10)$$

It follows from (**), (33.9) and (33.10) that for all *x*, such that $|x - x_0| < \delta$, we have

$$|S(x) - S(x_0)| < \varepsilon.$$

So, we proved the continuity of the sum $S(x)$ at an arbitrary point $x_0 \in [a,b]$. Therefore, $S(x)$ is continuous on $[a,b]$. That completes the proof.

Theorem (33.2) is valid (by virtue of Theorem 33.1) for a majorizable series. In other words, the following statement holds.

***Теорема* 31.3.** If continuous functions are majorizable on $[a,b]$, then the sum of such functions is a continuous function on $[a,b]$.

Term integration and differentiation of series

***Theorem* 33.4.** If $u_n(x)$ (n = 1, 2, …) are continuous functions on $[a,b]$ and the series of such functions converges to $S(x)$ uniformly on $[a,b]$, then the series can be integrated term by term from a to b, where the integral of the sum is equal to the sum of the integrals of the series terms:

$$\int_a^b S(x)dx = \int_a^b u_1(x)dx + \int_a^b u_2(x)dx + \ldots + \int_a^b u_n(x)dx + \ldots \quad . \qquad (33.11)$$

Proof. Denote by $S_n(x)$ the n-th partial sum of the series (33.11). Its uniform convergence means that for any $\varepsilon > 0$ there exists a number N, such that for every $n > N$ and for all $x \in [a,b]$

$$|S(x) - S_n(x)| < \frac{\varepsilon}{b-a}.$$

The sum $S(x)$ is continuous function on $[a,b]$ by virtue of Theorem 33.3. Partial sum $S_n(x)$ is also continuous on $[a,b]$ since it is the sum of a finite number of continuous functions. Therefore, $S(x)$ and $S_n(x)$ are integrable on $[a,b]$ and

$$\left| \int_a^b S(x)dx - \int_a^b S_n(x)dx \right| < \int_a^b |S(x) - S_n(x)|dx < (b-a)\cdot\frac{\varepsilon}{b-a} = \varepsilon .$$

So,

$$\lim_{n\to\infty}\left[\int_a^b S(x)dx-\left(\int_a^b u_1(x)dx+\int_a^b u_2(x)dx+\ldots+\int_a^b u_n(x)dx\right)\right]=0,$$

which means that the series (33.11) converges to the sum $\int_a^b S(x)dx$. That completes the proof.

Remark. Obviously, if the series converges on $[a,b]$ uniformly, then it converges on any segment $[a,x]$ where $a<x<b$. Therefore, under the conditions of Theorem 33.4, the equality

$$\int_a^x S(t)dt=\int_a^x u_1(t)dt+\int_a^x u_2(t)dt+\ldots+\int_a^x u_n(t)dt+\ldots \quad (33.12)$$

holds.

Theorem 33.5. Let functions $u_1(x),u_2(x),\ldots,u_n(x),\ldots$ have continuous derivatives on $[a,b]$. If a series

$$u_1(x)+u_2(x)+\ldots+u_n(x)+\ldots, \quad (33.1)$$

converges to the sum $S(x)$ on $[a,b]$ and a series

$$u_1'(x)+u_2'(x)+\ldots+u_n'(x)+\ldots \quad (33.13)$$

converges on $[a,b]$ uniformly, then the sum $S(x)$ of the series (33.13) has a derivative on $[a,b]$, such that

$$S'(x)=u_1'(x)+u_2'(x)+...+u_n'(x)+...$$.

Proof. Let $\widetilde{S}(x)$ be the sum of the series (33.13). Since the series (33.13) converges uniformly on $[a,b]$, then by virtue of the remark to Theorem 33.4, it can be integrated term by term on any segment $[a,x]$, $a<x<b$:

$$\int_a^x \widetilde{S}(t)dt=\int_a^x u_1'(t)dt+\int_a^x u_2'(t)dt+...+\int_a^x u_n'(t)dt+...$$.

Obviously, for all *n*

$$\int_a^x u_n'(t)dt=u_n(x)-u_n(a)$$.

Therefore,

$$\int_a^x \widetilde{S}(t)dt=[u_1(x)-u_1(a)]+[u_2(x)-u_2(a)]+...+[u_n(x)-u_n(a)]+...$$

.

Due to the conditions of Theorem

$$S(x)=u_1(x)+u_2(x)+...+u_n(x)+...,$$

$$S(a)=u_1(a)+u_2(a)+...+u_n(a)+...$$.

Hence,

$$\widetilde{S}(x)=\left(\int_a^x \widetilde{S}(t)dt\right)'=\left(S(x)-S(a)\right)'=S'(x).$$

Therefore,

$$S'(x)=u_1'(x)+u_2'(x)+...+u_n'(x)+... .$$

That completes the proof.

33.3. Power series

Definition. Functional series

$$\sum_{n=0}^{\infty} a_n x^n = a_0 + a_1 x + a_2 x^2 + ... + a_n x^n + ... \tag{33.14}$$

is called a **power series**. Coefficients a_0, a_1, …, a_n, … are called the **coefficients of the power series** (33.14).

Convergence region of the power series

Since the series (33.14) converges at $x=0$, then the convergence region of this series is always a nonempty set.

***Theorem* 33.6 (Abel theorem).** 1. If series (33.14) *converges* at some point $x=x_0$ ($x_0 \neq 0$), then it absolutely converges for all x, such that $|x|<|x_0|$. 2. If the series (33.14) *diverges* at some point $x=x_1$, then it diverges for all x, such that $|x|>|x_1|$.

Proof. 1. By condition, number series

$$\sum_{n=0}^{\infty} a_n x_0^n = a_0 + a_1 x_0 + a_2 x_0^2 + \ldots + a_n x_0^n + \ldots$$

converges, therefore, its general term $a_n x_o^n$ tends to zero as $n \to \infty$. Hence, sequence $\left\{a_n x_o^n\right\}$ is bounded, i.e. there exists a number $M > 0$, such that for all *n*

$$\left|a_n x_o^n\right| < M. \tag{33.15}$$

Consider a series which consists of the absolute values of the terms of the series (33.14):

$$\sum_{n=0}^{\infty} \left|a_n x^n\right| = \left|a_0\right| + \left|a_1 x\right| + \left|a_2 x^2\right| + \ldots + \left|a_n x^n\right| + \ldots \tag{33.16}$$

Rewrite it in the form

$$\sum_{n=0}^{\infty} \left|a_n x^n\right| = \left|a_0\right| + \left|a_1 x_0\right| \cdot \left|\frac{x}{x_0}\right| + \left|a_2 x_0^2\right| \cdot \left|\frac{x}{x_0}\right|^2 + \ldots + \left|a_n x_0^n\right| \cdot \left|\frac{x}{x_0}\right|^n + \ldots \tag{33.17}$$

Let $x < x_0$. Then $q = \left|\frac{x}{x_0}\right| < 1$. It follows from (33.15) and (33.17) that the terms of series (33.16) are less than the corresponding terms of the convergent series

$$\sum_{n=0}^{\infty} M \left|\frac{x}{x_0}\right|^n = M + M\left|\frac{x}{x_0}\right| + M\left|\frac{x}{x_0}\right|^2 + \ldots + M\left|\frac{x}{x_0}\right|^n + \ldots,$$

being the sum of the infinite decreasing geometric progression with the denominator $q<1$. Therefore, series (33.16) converges due to the direct comparison test, i.e. the series (33.14) converges absolutely.

2. The series (33.14) *diverges* at $x=x_1$ due to the condition. Lets prove that it diverges for all *x* satisfying the condition $|x|>|x_1|$. Assume the opposite, i.e. series (33.14) converges for some *x*, such that $|x|>|x_1|$. Then it converges at $x=x_1$ due to the first statement of Theorem but this is contrary to the condition. That completes the proof.

The Abel theorem allows us to determine the location of the points of convergence and divergence of a power series. It follows from the Abel theorem that if a power series converges at $x=x_0$, then it converges absolutely on the interval $\left(-|x_0|,|x_0|\right)$; if a power series diverges at $x=x_1$, then it diverges everywhere outside the segment $\left[-|x_1|,|x_1|\right]$. It follows that there exists a number *R*, such that a power series converges absolutely on the interval $(-R,R)$ and diverges out of $[-R,R]$.

Number *R* is called the **radius of convergence**, interval $(-R,R)$ is called **an interval of convergence** of the power series. The series can as converge as diverge at the ends of the convergence interval (i.e. at $x=-R$ and $x=R$).

Now we introduce a **method for finding the radius of convergence of a power series**. Consider the series (33.14). Apply the d'Alembert's ratio

test to this series at the fixed point x (Theorem 32.4). The series converges if

$$\lim_{n\to\infty}\left|\frac{a_{n+1}x^{n+1}}{a_n x^n}\right| = \lim_{n\to\infty}|x|\left|\frac{a_{n+1}}{a_n}\right| < 1.$$

Let $\lim\limits_{n\to\infty}\left|\frac{a_{n+1}}{a_n}\right| = L$. The series (33.16) converges according to the d'Alembert's ratio test if $|x|L<1$ and converges if $|x|L>1$. Therefore, the series (33.14) *converges absolutely* as $x<\frac{1}{L}$ and *diverges* as $x>\frac{1}{L}$. Thus, $\frac{1}{L}$ is *the radius of convergence* of the series (33.14), i.e.

$$R = \lim_{n\to\infty}\left|\frac{a_n}{a_{n+1}}\right|. \qquad (33.18)$$

Moreover, in particular, it can be $R=0$ or $R=\infty$, i.e. the region of convergence can consist of one point or coincide with the whole number line.

Example 33.2. Find the region of convergence of the power series $\sum\limits_{n=1}^{\infty}\frac{x^n}{n}$.

Solution. Here $a_n=\frac{1}{n}$, $a_{n+1}=\frac{1}{n+1}$. Find the radius of convergence by the formula (33.18):

$$R=\lim_{n\to\infty}\left|\frac{a_n}{a_{n+1}}\right|=\lim_{n\to\infty}\frac{n+1}{n}=1.$$

The convergence interval is $(-1,1)$.

We now clarify the behavior of the series at the ends of the convergence interval: a) alternating series $\sum_{n=1}^{\infty}\frac{(-1)^n}{n}$ converges on the basis of Leibniz as $x=-1$ (see Theorem 32.7); harmonic series $\sum_{n=1}^{\infty}\frac{1}{n}$ diverges as $x=1$. So, the series converges (conditionally) at the left end of the convergence interval and diverges at the right end.

***Example* 33.3.** Find the region of convergence of the power series $\sum_{n=0}^{\infty}\frac{x^n}{n!}$. (Recall that $0!=1$.)

Solution. Find the radius of convergence by the formula (33.18):

$$R=\lim_{n\to\infty}\left|\frac{a_n}{a_{n+1}}\right|=\lim_{n\to\infty}\frac{(n+1)!}{n!}=\lim_{n\to\infty}(n+1)=\infty.$$

This series converges absolutely on the whole number line.

Properties of power series

Let function $f(x)$ be the sum of a power series:

$$f(x)=\sum_{n=0}^{\infty}a_n x^n, \qquad (33.19)$$

its convergence interval is $(-R, R)$; then the *function* $f(x)$ is said to be decomposable into a *power series* on $(-R, R)$.

It follows from the Abel theorem that a power series is majorizable on any segment lying entirely within the convergence interval. Therefore, power series have a number of properties similar to those of ordinary polynomials.

The proof of **theorems on the properties of power series** is based on the uniform convergence of the power series on any segment contained in the convergence interval.

***Theorem* 33.7.** If $[-r, r]$ lies entirely within a converges interval of a power series

$$a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n + \ldots \tag{33.14}$$

then this series is majorizable on any $[-r, r]$.

Proof. Consider a number series

$$|a_0| + |a_1| r + |a_2| r^2 + \ldots + |a_n| r^n + \ldots . \tag{33.20}$$

We need to prove that for the series (33.20) there is a convergent numerical majorant (with positive terms) as $r < R$, where R is the radius of convergence of the series.

According to the Abel theorem this series converges since $r < R$. Inequalities $|a_n x^n| \le |a_n| r^n$ are satisfied for the terms of the series (33.14) as $|x| \le r$, i.e. $x \in [-r, r]$. Therefore, the series (33.14) is majorizable on $[-r, r]$. That completes the proof.

Corollary. A power series converges uniformly on any segment lying entirely within the convergence interval.

Indeed, according to the Weierstrass theorem, the series converges uniformly on $[-r, r]$ since it is majorizable on $[-r, r]$ as $r < R$.

***Theorem* 33.8.** The sum of a power series is a continuous function on any segment lying entirely within the convergence interval.

This theorem follows from Theorems 33.1, 33.2 and 33.7.

***Theorem* 33.9.** A power series can be integrated on any segment $[0, x]$ term by term as $-R < x < R$. In this case, the integral of the sum of the series is equal to the sum of the integrals of the series members:

$$\int_0^x \left(\sum_{n=1}^{\infty} a_n x^n \right) dx = a_0 x + \frac{a_1}{2} x^2 + \frac{a_2}{3} x^3 + \ldots + \frac{a_n}{n+1} x^{n+1} + \ldots .$$

This theorem directly follows from the previous Theorems 33.4 and 33.6.

Let us now find out the possibility of **differentiation of the power series term by term**.

***Theorem* 33.10.** Let function $f(x)$ be decomposable into a power series on $(-R, R)$

$$f(x) = \sum_{n=1}^{\infty} a_n x^n = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n + \ldots . \qquad (33.19)$$

Then the power series

$$\sum_{n=1}^{\infty} n a_n x^{n-1} = a_1 + 2a_2 x + 3a_3 x^2 + \ldots + n a_n x^{n-1} + \ldots , \qquad (33.21)$$

obtained by differentiation of the series (33.19), has the same convergence interval $(-R, R)$ and the function $f(x)$ has decomposable derivative $f'(x)$ on the whole $(-R, R)$:

$$f'(x) = a_1 + 2a_2x + 3a_3x^2 + \ldots + na_nx^{n-1} + \ldots. \qquad (33.22)$$

Proof. Let us show that series (33.21) is majorizable on any $[-r, r]$ as $r < R$. Take an arbitrary point r_0 satisfying the condition $r < r_0 < R$. The series (33.14) converges at that point, therefore, general term of this series tends to zero as $x = r_0$, and as a result, it is bounded. In other words, $\lim\limits_{n\to\infty} a_n r_0^n = 0$, so there exists number $M > 0$, such that

$$|a_n| r_0^n < M \quad (n = 1, 2, \ldots).$$

Then, as $x \in [-r, r]$ we have

$$|na_nx^{n-1}| \le |na_nr^{n-1}| = n|a_n| r_0^{n-1}\left(\frac{r}{r_0}\right)^{n-1} \le n\frac{M}{r_0}\left(\frac{r}{r_0}\right)^{n-1}.$$

Denote $\frac{M}{r_0} = M_0$, $\frac{r}{r_0} = q$; as $q < 1$.

So, $|na_nx^{n-1}| \le nM_0q^{n-1}$. Therefore, all terms of series (33.21) are not greater than the corresponding terms of the majorant numerical series for the specified x

$$M_0 + 2M_0 q + 3M_0 q^2 + \ldots + nM_0 q^{n-1} + \ldots .$$

The last series converges according to the d'Alembert's ratio test. Denote $a'_n = nM_0 q^{n-1}$, $a'_{n+1} = (n+1)M_0 q^n$. Then

$$\lim_{n\to\infty} \frac{a'_{n+1}}{a'_n} = \lim_{n\to\infty} \frac{(n+1)q^n}{nq^{n-1}} = q < 1 .$$

So, series (33.21) is majorized on the segment $[-r, r]$, then due to Theorem 33.5 we can state that series (33.19) is differentiable term by term and the equality (33.22) is true.

Since for any $x \in (-R, R)$ there exists $r < R$, such that $x \in [-r, r]$, it follows that series (33.21) converges at any inner point of the interval $(-R, R)$.

We proved that the radius of convergence *can not be decreased* after the differentiation of the series.

To complete the proof of the theorem, one must now show that the radius of convergence *can not increase* as a result of differentiation. Assume the opposite, i.e. the series (33.21) converges for some $x_1 > R$. By integrating this series from 0 to x_2, where $R < x_2 < x_1$, we would obtain the convergence of the original series (33.19) at point x_2, and this contradicts the condition of Theorem. Thus, the interval of convergence of series (33.19) is the interval of convergence of series (33.21) obtained by differentiation of series (33.19). That completes the proof.

Theorems 33.9 and 33.10 mean that power series (within their convergence interval) behave like ordinary polynomials with respect to differentiation and integration.

By applying Theorem 33.10, it is easy to verify that a function that decomposes into a power series is infinitely differentiable on the convergence interval of this series.

Power series with an arbitrary center

A power series is a functional series of the form

$$\sum_{n=0}^{\infty} a_n (x-x_0)^n = a_0 + a_1(x-x_0) + a_2(x-x_0)^2 + \ldots + a_n(x-x_0)^n + \ldots$$

. (33.23)

This is a power series in powers of the binomial $x-x_0$. Obviously, the power series (33.14) is a special case of series (33.23).

To determine the region of convergence of series (33.23), we make the substitution:

$$x-x_0 = X$$

.

After this substitution, series (33.23) takes the form:

$$a_0 + a_1 X + a_2 X^2 + \ldots + a_n X^n + \ldots$$

. (33.24)

Let $(-R, R)$ be the interval of convergence of series (33.24). Then (33.23) *converges* as $|x-x_0| < R$ and *diverges* as $|x-x_0| > R$. Therefore, the interval of convergence of series (33.23) is the set of all values *x* satisfying inequality $-R < x-x_0 < R$, i.e. $x_0 - R < x < x_0 + R$

.

Therefore, the *interval* of convergence of series (33.23) is the interval $(x_0 - R, x_0 + R)$ centered at x_0. All properties of the power series (33.14) are fully preserved for the power series (33.23).

Questions

1. What is the region of convergence of the functional series?
2. Which functional series is called uniformly convergent?
3. What kind of functional series is called majorizable? Is every majorizable series uniformly convergent?
4. Is the sum of the functional series consisting of continuous functions always continuous?
5. What kind of functional series is called a power series?
6. What is the radius of convergence of a power series? Can the radius of convergence be equal to zero or infinity?
7. What are the main properties of a power series?

Chapter 34. Taylor and Maclaurin series

34.1. Decomposition of functions into a power series

Assume that function $f(x)$ can be decomposed into a power series on a certain interval $(-R, R)$

$$f(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n + \ldots . \qquad (34.1)$$

It follows from Theorem 33.4 that this series can be differentiated term by term any number of times. Then, differentiating the equality (34.1) *n* times, we obtain:

$$f'(x) = a_1 + 2a_2 x + \ldots + n a_n x^{n-1} + \ldots ,$$

$$f''(x) = 2a_2 + \ldots + n(n-1) a_n x^{n-2} + \ldots ,$$

……

$$f^{(n)}(x) = n! a_n + (n+1) n \cdots 3 \cdot 2 \cdot a_{n+1} x + \ldots .$$

Assuming $x = 0$, we have $f(0) = a_0$, $f'(0) = a_1$, $f''(0) = 2! a_2$, …, $f^{(n)}(0) = n! a_n$. Hence

$$a_n = \frac{f^{(n)}(0)}{n!} \quad (n = 0, 1, 2, \ldots). \qquad (34.2)$$

Substituting the obtained coefficients (34.2) into (34.1), we obtain the decomposition of function $f(x)$ into a power series:

$$f(x)=f(0)+\frac{f'(0)}{1!}x+\frac{f''(0)}{2!}x^2+\ldots+\frac{f^{(n)}(0)}{n!}x^n+\ldots \quad . \qquad (34.3)$$

The series (34.3) is called **the Maclaurin series** for the function $f(x)$.

We have proved the following theorem.

***Theorem* 34.1.** If a function $f(x)$ can be decomposed into a power series on $(-R,\,R)$, then this series is Maclaurin series for $f(x)$.

Theorem 34.1 implies that the decomposition of a function into a power series is *unique*. The coefficients of this decomposition are uniquely determined by the formulas (34.2).

It is known that the Maclaurin formula is valid for any function which has its derivatives up to the $(n+1)$ order (see (17.18)):

$$f(x)=f(0)+\frac{f'(0)}{1!}x+\frac{f''(0)}{2!}x^2+\ldots+\frac{f^{(n)}(0)}{n!}x^n+\frac{f^{(n+1)}(\xi)}{(n+1)!}x^{n+1},$$

here ξ is a point between 0 and $x\,(\xi=\theta x,\ 0<\theta<1)$.

If we denote the remainder term by $R_n(x)$:

$$R_n(x)=\frac{f^{(n+1)}(\xi)}{(n+1)!}x^{n+1},$$

and the partial sum of the Maclaurin series as $S_n(x)$, then the Maclaurin formula can be written as follows:

$$f(x)=S_n(x)+R_n(x). \tag{34.4}$$

Equation (34.4) implies the **criterion for the decomposability of a function in a Maclaurin series**.

***Theorem* 34.2.** A necessary and sufficient condition for an infinitely differentiable function $f(x)$ to be decomposed into a Maclaurin series on $(-R, R)$ is that the residual term of the Maclaurin formula for this function tended to zero at the specified interval as $n\to\infty$.

Proof. 1. *Necessity*. Let $\lim\limits_{n\to\infty} S_n(x)=f(x)$ for all $x\in(-R, R)$. Then for all $x\in(-R, R)$ according to (34.4)

$$\lim_{n\to\infty} R_n(x)=\lim_{n\to\infty}\left[f(x)-S_n(x)\right]=f(x)-f(x)=0.$$

2. *Sufficiency*. Now let $\lim\limits_{n\to\infty} R_n(x)=0$ for all $x\in(-R, R)$. Then from (34.4) we get

$$\lim_{n\to\infty} S_n(x)=\lim_{n\to\infty}\left[f(x)-R_n(x)\right]=f(x)-0=f(x),$$

i.e. the Maclaurin series converges to function $f(x)$. That completes the proof.

Note that if $\lim\limits_{n\to\infty} R_n(x)\neq 0$, then the Maclaurin series does not represent a given function, although it may converge (to some other function).

Maclaurin formula is a special case of the Taylor formula (see (17.7) and (17.12)):

$$f(x)=f(a)+\frac{f'(a)}{1!}(x-a)+\frac{f''(a)}{2!}(x-a)^2+\ldots+\frac{f^{(n)}(a)}{n!}(x-a)^n+R_n(x)$$

,

where $R_n(x)=\frac{f^{(n+1)}(\xi)}{(n+1)!}(x-a)^{n+1}$.

If function $f(x)$ has derivatives of any order in the neighborhood of point $x=a$, then we can obtain an infinite series called **the Taylor series**:

$$f(a)+\frac{f'(a)}{1!}(x-a)+\frac{f''(a)}{2!}(x-a)^2+\ldots+\frac{f^{(n)}(a)}{n!}(x-a)^n+\ldots$$

.

The Maclaurin series is a special case of the Taylor series as $a=0$.

The Taylor series for function $f(x)$ converges to this function if and only if the residual term in the Taylor's formula for this function tends to zero: $\lim\limits_{n\to\infty}R_n(x)=0$.

It is easy to make sure that if all derivatives are bounded: $f^{(n)}(x)\le M$ (n = 1, 2, …), then $\lim\limits_{n\to\infty}R_n(x)=0$.

34.2. Decomposition of some elementary functions in the Maclaurin series

1. Let $f(x)=\mathbf{e}^x$. According to the Taylor formula

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \ldots + \frac{x^n}{n!} + R_n(x),$$

where $R_n(x) = \frac{x^{n+1}}{(n+1)!} e^{\theta}$, $0 < \theta < 1$.

Since for any fixed x the value e^x is bounded, then $\lim\limits_{n\to\infty} R_n(x) = 0$. Therefore, for all $x \in (-\infty, \infty)$

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \ldots + \frac{x^n}{n!} + \ldots . \qquad (34.5)$$

2. Similarly, we obtain the Maclaurin decomposition of functions $f(x) = \sin x$ and $f(x) = \cos x$ (see § 17.3; here also $\lim\limits_{n\to\infty} R_n(x) = 0$ for all x):

$$\sin x = x - \frac{x^3}{3!} + \ldots + (-1)^n \frac{x^{2n+1}}{(2n+1)!} + \ldots , \qquad (34.6)$$

$$\cos x = 1 - \frac{x^2}{2!} + \ldots + (-1)^n \frac{x^{2n}}{(2n)!} + \ldots . \qquad (34.7)$$

3. Consider the function $f(x) = (1+x)^m$, where m is an arbitrary constant number. We have

$$f(x) = (1+x)^m,\ f'(x) = m(1+x)^{m-1},\ f''(x) = m(m-1)(1+x)^{m-2},$$

$$\ldots ,$$

$$f^{(n)}(x) = m(m-1)\cdots(m-n+1)(1+x)^{m-n}.$$

For $x=0$:

$$f(0)=1,\ f'(0)=m,\ f''(0)=m(m-1),\ \ldots,$$
$$f^{(n)}(0)=m(m-1)\cdots(m-n+1).$$

We get a series called **binomial**:

$$1+\frac{m}{1!}x+\frac{m(m-1)}{2!}x^2+\ldots+\frac{m(m-1)\cdots(m-n+1)}{n!}x^n+\ldots.$$

Let us define the radius of convergence of this series $R=\lim\limits_{n\to\infty}\left|\frac{a_n}{a_{n+1}}\right|$. Since

$$a_n=\frac{m(m-1)\cdots(m-n+1)}{n!},\quad a_{n+1}=\frac{m(m-1)\cdots(m-n+1)(m-n)}{(n+1)!},$$

we obtain $R=\lim\limits_{n\to\infty}\left|\frac{n+1}{m-n}\right|=|-1|=1$.

Thus, the binomial series converges for $x\in(-1,1)$ and diverges outside of line segment $[-1,1]$.

Estimation of the remainder of this series is associated with certain difficulties, therefore, we accept without any proof, that $\lim\limits_{n\to\infty}R_n(x)=0$ as $x\in(-1,1)$.

So, for $x\in(-1,1)$

$$(1+x)^m = 1 + \frac{m}{1!}x + \frac{m(m-1)}{2!}x^2 + \ldots + \frac{m(m-1)\cdots(m-n+1)}{n!}x^n + \ldots$$

. (34.8)

Note that if m is a positive integer, then starting with the term containing x^{m+1}, all coefficients are zero and the series turns into a finite polynomial.

4. For $m = -1$ the binomial series has the form

$$\frac{1}{1+x} = 1 - x + x^2 - x^3 + \ldots + (-1)^n x^n + \ldots \quad . \qquad (34.9)$$

Integrate this equality from 0 to x, where $|x| < 1$:

$$\int_0^x \frac{dt}{1+t} = \int_0^x \left(1 - t + t^2 - t^3 + \ldots + (-1)^n t^n + \ldots\right) dt \quad .$$

Hence we get the decomposition of function $\boldsymbol{f(x) = \ln(1+x)}$:

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \ldots + \frac{(-1)^{n+1}x^n}{n} + \ldots \quad . \qquad (34.10)$$

This equality holds on $(-1, 1)$. It can be shown that it is also true for $x = 1$. Thus, the convergence region of the series (34.10) is $(-1, 1]$.

34.3. Application of power series to approximate calculations

It is possible to obtain the values of these functions with any accuracy using the decomposition of elementary function in Maclaurin series. To do this, we need to take a sufficient number of terms of the series. The accuracy of the calculation is determined by the residual term of the Maclaurin formula.

1. Consider the **decomposition of exponential functions** $f(x)=\mathrm{e}^x$.

As already noted,

$$\mathrm{e}^x = 1 + x + \frac{x^2}{2!} + \ldots + \frac{x^n}{n!} + R_n(x),$$

where $R_n(x) = \dfrac{x^{n+1}}{(n+1)!}\mathrm{e}^{\theta}$ and for all x the exponential function decomposes into the series

$$\mathrm{e}^x = 1 + x + \frac{x^2}{2!} + \ldots + \frac{x^n}{n!} + \ldots . \qquad (34.5)$$

The radius of convergence of this series is infinite, i.e. the series converges on the entire number line.

The series (34.5) can be calculated only for small values of x. If the absolute value of x is large, then the series is represented as the sum of integer and fractional parts:

$$x = E(x) + q,$$

here $E(x)$ is integer part of x (i.e. the largest integer not exceeding) and q is a fractional part, $0 \le q < 1$. Then

$$\mathrm{e}^x = \mathrm{e}^{E(x)}\mathrm{e}^q.$$

The first multiplier $\mathrm{e}^{E(x)}$, which is an integer power of a number e, can be found using multiplication. The second, i.e. e^q – using decomposition (34.5).

The residue of the series is estimated as follows:

$$0 < R_n(x) < \frac{x^{n+1}}{n!n}.$$

***Example* 34.1.** Find $\sqrt{\mathrm{e}}$ with accuracy 10^{-6}.

Solution. According to (34.5) we have $u_0 = 1$, …, $u_{k-1} = \dfrac{1}{2^{k-1}(k-1)!}$, $u_k = \dfrac{1}{2^k k!}$ (k = 1, 2, …, n), then $u_k = \dfrac{u_{k-1}}{2k}$. We obtain

$$\mathrm{e}^{\frac{1}{2}} = \sum_{k=0}^{n} u_k + R_n\left(\frac{1}{2}\right).$$

Let us count the terms with two spare signs:

$$u_0 = 1, \qquad u_5 = \frac{u_4}{10} = 0{,}00026042,$$

$$u_1 = \frac{u_0}{2} = 0{,}50000000, \qquad u_6 = \frac{u_5}{12} = 0{,}00002170,$$

$$u_2 = \frac{u_1}{4} = 0{,}12500000, \qquad u_7 = \frac{u_6}{14} = 0{,}00000155,$$

$$u_3 = \frac{u_2}{6} = 0{,}02083333, \qquad u_8 = \frac{u_7}{16} = 0{,}00000010,$$

$$u_4 = \frac{u_3}{8} = 0{,}00260417, \qquad S_8 = 1{,}64872117.$$

Rounding the sum to six decimal places after the decimal point, we obtain

$$\sqrt{\mathrm{e}} = 1{,}648721.$$

2. Consider **decomposition of the logarithm and the calculation of the values of the logarithmic function**. Logarithmic function $f(x) = \ln(1+x)$ decomposes into a power series on $(-1, 1]$:

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \ldots + (-1)^n \frac{x^{n+1}}{n+1} + \ldots . \qquad (34.11)$$

Direct application of this series is complicated, in particular, because of its slow convergence. Replace argument x with $-x$ in (34.11):

$$\ln(1-x) = -x - \frac{x^2}{2} - \frac{x^3}{3} - \frac{x^4}{4} - \ldots - \frac{x^{n+1}}{n+1} - \ldots . \qquad (34.12)$$

Subtract (34.12) from (34.11):

$$\ln\frac{1+x}{1-x} = 2\left(x + \frac{x^3}{3} + \frac{x^5}{5} + \ldots\right), \qquad (34.13)$$

here $|x| < 1$.

The series (34.13) converges faster than a geometric progression with the denominator $q = x^2$. In addition, the expression $\frac{1+x}{1-x}$ can take any positive values at the specified x. Therefore, formula (34.13) is very convenient for calculating logarithms.

Since the terms of the series (34.13) are smaller than the terms of the geometric progression $2\left(x + x^3 + x^5 + \ldots\right)$, the remainder of the series is estimated as follows:

$$R_n < \frac{2x^{2n+1}}{(2n+1)\left(1-x^2\right)}. \qquad (34.14)$$

***Example* 34.2.** Calculate $\ln 8$ with the accuracy 10^{-6}.

Solution. From $\frac{1+x}{1-x} = 8$ we obtain $x = 0{,}777\ldots$, it is better to rewrite as $8 = \mathrm{e}^2 \frac{8}{\mathrm{e}^2}$ to accelerate the convergence of the series. Then $\ln 8 = \ln \mathrm{e}^2 + \ln \frac{8}{\mathrm{e}^2} = 2 + \ln \frac{8}{\mathrm{e}^2}$. Assuming $\frac{1+x}{1-x} = \frac{8}{\mathrm{e}^2}$ we obtain $x = \frac{8-\mathrm{e}^2}{8+\mathrm{e}^2} = 0{,}03969989$. It follows from (34.14), that the remainder is approximately equal to the first of the discarded terms for such x. As in the previous example, we will perform calculations with two positive signs:

$u_1 = 2x = 0{,}07939978$,

$$u_2 = 2 \cdot \frac{x^3}{3} = 0{,}00004171,$$

$$u_3 = 2 \cdot \frac{x^5}{5} = 0{,}00000004.$$

We obtain $\ln 8 = 2{,}079441$.

(Functions $\sin x$ and $\cos x$ can be calculated similarly using their decomposition into Maclaurin series.)

3. Consider **calculating values of integrals that are not expressed through elementary functions using Maclaurin series**. It is known that the integral e^{-x^2} is not an elementary function.

***Example* 34.3.** Find the integral

$$\int_0^a e^{-x^2} dx.$$

Solution. In order to compute the integral we decompose the integrand in a series replacing *x* in the decomposition by $-x^2$:

$$e^{-x^2} = 1 - \frac{x^2}{1!} + \frac{x^4}{2!} - \frac{x^6}{3!} + \ldots + (-1)^n \frac{x^{2n}}{n!} + \ldots.$$

Integrating both parts of this equality, we obtain:

$$\int_0^a e^{-x^2} dx = \left(\frac{x}{1} - \frac{x^3}{1! \cdot 3} + \frac{x^5}{2! \cdot 5} - \frac{x^7}{3! \cdot 7} + \ldots \right) \Bigg|_0^a = \frac{a}{1} - \frac{a^3}{1! \cdot 3} + \frac{a^5}{2! \cdot 5} - \frac{a^7}{3! \cdot 7} + \ldots$$

.

We can compute the given integral with any degree of accuracy and any *a* using this equality. In particular, it is enough to take seven terms of decomposition in order to compute $\int_0^1 e^{-x^2} dx$ with the accuracy up to 10^{-4}. Then we obtain

$$\int_0^1 e^{-x^2} dx \approx 0{,}7468 .$$

Note that modulo of the remainder of the series does not exceed modulo of the first discarded term since the series is alternating.

Questions

1. What is the Maclaurin series? What function can it be defined for?
2. Does the Maclaurin series of a function $f(x)$ necessarily converge to this function?
3. What is the criterion for the decomposability of a function in the Maclaurin series?
4. What is the Taylor series?
5. What is the binomial series? What is its region of convergence?
6. Give examples of using series in approximate calculations.

Chapter 35. Fourier series

The power series considered earlier makes it possible to represent the function $f(x)$ in the form of the sum (with corresponding coefficients) of the simplest functions, which are powers of x:

$1,\ x,\ x^2,\ x^3,\ \ldots,\ x^n,\ \ldots$ (35 .1)

We consider functional series in which instead of degrees x trigonometric functions are selected. Trigonometric function system

$1,\ \cos x,\ \sin x,\ \cos 2x,\ \sin 2x,\ \ldots,\ \cos nx,\ \sin nx,\ \ldots$ (35 .2)

is well studied in elementary mathematics.

Definition 1. Series

$$\frac{a_0}{2} + a_1 \cos x + b_1 \sin x + a_2 \cos 2x + b_2 \sin 2x + a_3 \cos 3x + b_3 \sin 3x + \ldots =$$

$$= \frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos nx + b_n \sin nx)$$

(35.3)

is called **trigonometric series**, and numbers $a_0, a_1, b_1, a_2, b_2 \ldots$ are called the **coefficients of the trigonometric series.**

Note that all functions of system (35.2) are periodic with a total period of 2π. Therefore, any partial sum of the trigonometric series (35.3) is also a periodic function with a period of 2π. Hence, if this series converges on $[-\pi, \pi]$, then it converges on the whole number line and its sum is a periodic function with a period of 2π since it is a limit of a sequence of periodic partial sums.

Orthogonality is an important property of the trigonometric system (35.2).

Definition 2. Functions $f(x)$ and $g(x)$ are mutually orthogonal on $[a, b]$, if

$$\int_a^b f(x)g(x)dx = 0$$

***Theorem* 35.1.** Two functions of the system (.2) are mutually orthogonal on $[-\pi, \pi]$.

Indeed, if $k \neq 0$ and k is an integer then

$$\int_{-\pi}^{\pi} \cos kx dx = \frac{1}{k} \sin kx \Big|_{-\pi}^{\pi} = 0,$$

$$\int_{-\pi}^{\pi} \sin kx dx = -\frac{1}{k} \cos kx \Big|_{-\pi}^{\pi} = 0.$$

It means that function $f(x) = 1$ is orthogonal to functions $\cos nx$ or $\sin nx$ of the system (35.2).

There remains verifying the validity of the next equalities for $k \neq n$:

$$\int_{-\pi}^{\pi} \cos kx \cos nx dx = 0, \quad (*)$$

$$\int_{-\pi}^{\pi} \sin kx \cos nx dx = 0, \quad (**)$$

$$\int_{-\pi}^{\pi} \sin kx \sin nx dx = 0. \quad (***)$$

Consider the first of these three integrals. Since

$$\cos kx \cos nx = \frac{1}{2}[\cos(k+n)x + \cos(k-n)x],$$

then

$$\int_{-\pi}^{\pi} \cos kx \cos nx dx = \frac{1}{2}\int_{-\pi}^{\pi} [\cos(k+n)x + \cos(k-n)x]dx =$$

$$= \frac{1}{2}\left[\frac{\sin(k+n)x}{k+n} + \frac{\sin(k-n)x}{k-n}\right]\Bigg|_{-\pi}^{\pi} = 0.$$

Similarly, applying the corresponding formulas

$$\sin kx \cos nx = \frac{1}{2}[\sin(k+n)x + \sin(k-n)x],$$

$$\sin kx \sin nx = \frac{1}{2}[\cos(k-n)x - \cos(k+n)x],$$

we prove the validity of the remaining two equalities. That completes the proof.

We will need the following two equalities

$$\int_{-\pi}^{\pi} \cos^2 nx dx = \pi, \quad \int_{-\pi}^{\pi} \sin^2 x dx = \pi. \qquad (35.4)$$

They are easily proved using formulas

$$\cos^2 \alpha = \frac{1+\cos 2\alpha}{2}, \ \sin^2 \alpha = \frac{1-\cos 2\alpha}{2}.$$

Indeed,

$$\int_{-\pi}^{\pi} \cos^2 nx dx = \frac{1}{2}\int_{-\pi}^{\pi} (1+\cos 2nx)dx = \frac{1}{2}\left(x + \frac{1}{2n}\sin 2nx\right)\Bigg|_{-\pi}^{\pi} = \pi.$$

The second equality in (35.4) is proved similarly.

Decomposition of functions in a Fourier series.

Theorem 35.2. Let function $f(x)$ be integrable on $[-\pi, \pi]$ and decomposable into a trigonometric series

$$f(x)=\frac{a_0}{2}+\sum_{n=1}^{\infty}(a_n\cos nx+b_n\sin nx), \quad (35.5)$$

where series can be integrated term by term when multiplied by a limited function, then such decomposition is unique.

Proof. To calculate the decomposition coefficients, we use the formulas $(*)-(***)$ and (35.4). We integrate the series (35.5) on $[-\pi,\pi]$. We see that all the integrals on the right-hand side, except the first, vanish. Therefore

$$\int_{-\pi}^{\pi} f(x)dx=\frac{1}{2}a_0\int_{-\pi}^{\pi}dx=a_0\pi.$$

Hence

$$a_0=\frac{1}{\pi}\int_{-\pi}^{\pi}f(x)dx. \quad (35.6)$$

Now multiply the series (35.5) by $\cos nx$, $n>0$ and integrate again on $[-\pi,\pi]$. Then, all terms of the integrated series vanish, except for the term containing a_n due to the orthogonality of the trigonometric system. We obtain

$$\int_{-\pi}^{\pi} f(x)\cos nxdx=\int_{-\pi}^{\pi}a_n\cos^2 nxdx=a_n\pi.$$

Hence

$$a_n=\frac{1}{\pi}\int_{-\pi}^{\pi}f(x)\cos nxdx. \quad (35.7)$$

Similarly, multiplying the equality (35.5) by $\sin nx$ and integrating on $[-\pi, \pi]$, we obtain

$$\int_{-\pi}^{\pi} f(x)\sin nxdx = \int_{-\pi}^{\pi} b_n \sin^2 nxdx = b_n \pi.$$

Hence

$$b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin nxdx. \qquad (35.8)$$

Formulas (35.6) - (35.8) uniquely determine all the decomposition coefficients. That completes the proof.

Numbers a_0, a_n, b_n, determined by formulas (35.6) – (35.8), are called **Fourier coefficients** while the trigonometric series (35.5) is called the **Fourier series of function** $f(x)$.

Convergence of the Fourier series.

Now we introduce the definition of **the periodic continuation of the function** $f(x)$ defined on $[-\pi, \pi]$. We say that periodic function $F(x)$, defined on the whole number line with a period of 2π, is a periodic continuation of the function $f(x)$, if $F(x) = f(x)$ on $[-\pi, \pi]$.

We raise the following question: what properties should a function have to provide the convergence of its Fourier series and the sum of its Fourier series to be equal to the values of a given function at the corresponding points?

Definition. Function $f(x)$ is called **a piecewise monotonic function** on $[a, b]$, if this segment can be divided by a finite number of points x_1, x_2,

. . . , x_{n-1}, on the intervals $(a, x_1), (x_1, x_2), \ldots, (x_{n-1}, b)$ such that the function is monotonic on each of these intervals.

It is easy to verify that a piecewise monotonic function can have only discontinuities of the first type. Indeed, if $x = c$ is a discontinuity point $f(x)$, then, due to the monotonicity of the function, there are finite limits $f(c-0)$, $f(c+0)$, i.e. c is a discontinuity point of the first type.

Now we introduce a theorem that gives sufficient conditions for the representability of a function $f(x)$ by the Fourier series.

***Theorem* 35.3.** Let $f(x)$ be a piecewise monotonic and bounded on $[-\pi, \pi]$ periodic function with a period of 2π; then its Fourier series converges at all points on $[-\pi, \pi]$. The sum $S(x)$ of the series is equal to the value of function $f(x)$ at the points of continuity of the function. The sum of the series at the points of discontinuity is equal to the arithmetic mean of the limits of function $f(x)$ on the right-hand and left-hand sides, i.e. if $x = c$ is a discontinuity point of $f(x)$, then

$$S(x)\big|_{x=c} = \frac{f(c-0) + f(c+0)}{2}.$$

We accept this theorem without proof.

It is easy to understand that the class of functions represented by Fourier series is wide. In particular, it is significantly wider than the class of functions represented by the sum of a power series. Therefore, the Fourier series are widely used in various sections of mathematics and its applications.

Note that there are other sufficient conditions of the decomposability of a function in a Fourier series.

Let $f(x)$ satisfy the conditions of Theorem 35.3. Then the sum of the Fourier series is a periodic function with period of 2π. We can continue it to the whole number line using its graph on $[-\pi,\pi]$. Let $f(x)$ be continuous on $[-\pi,\pi]$. Then the sum of its Fourier series coincides with $f(x)$ on the whole $(-\pi,\pi)$ and, therefore, it is continuous on this interval, as well as on any interval $((2k-1)\pi,(2k+1)\pi), k=0,\pm1,\pm2,...$. Moreover, if $f(-\pi)=f(\pi),$ then the sum of the Fourier series will be a continuous function on the entire axis. If $f(-\pi)\neq f(\pi)$, then points $x_k=(2k+1)\pi,\ k=0,\pm1,\pm2,\ldots$ are discontinuity points of the first type for the sum of the series. The sum of the series at these points is $\dfrac{f(-\pi)+f(\pi)}{2}$.

***Example* 35.1.** Let $f(x)=x$ and $f(-\pi)\neq f(\pi)$. Therefore, the sum of its Fourier series is a discontinuous function. Let us construct this series. According to (35.6) – (35.8) we have

$$a_0=\frac{1}{\pi}\int_{-\pi}^{\pi}x dx=\frac{1}{\pi}\frac{x^2}{2}\Big|_{-\pi}^{\pi}=0,$$

$$a_n=\frac{1}{\pi}\int_{-\pi}^{\pi}x\cos nx dx=\frac{1}{\pi}\left\{\frac{1}{n}x\sin nx\Big|_{-\pi}^{\pi}-\frac{1}{n}\int_{-\pi}^{\pi}\sin nx dx\right\}=0$$

.

$$b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} x\sin nx dx = \frac{1}{\pi}\left\{-\frac{1}{n}x\cos nx\Big|_{-\pi}^{\pi} + \frac{1}{n}\int_{-\pi}^{\pi}\cos nx dx\right\} = -\frac{2}{n}\cos n\pi = (-1)^{n+1}$$

The Fourier series for this function has the form

$$x = 2\left(\frac{\sin x}{1} - \frac{\sin 2x}{2} + \frac{\sin 3x}{3} - \ldots\right) = 2\sum_{n=1}^{\infty}(-1)^{n+1}\frac{\sin nx}{n}.$$

This decomposition is valid for $(-\pi, \pi)$, but if $x = \pm\pi$ the sum of the series is 0 .

Fig. 35.1

- ***Example* 35.2.** Let $f(x) = x^2$ and $f(\pi) = f(-\pi)$. Therefore, the sum of its Fourier series is a continuous function on the entire axis. This function coincides with x^2 on $[-\pi, \pi]$ and equals $(x - 2k\pi)^2$ on any segment $[(2k-1)\pi, (2k+1)\pi]$, $k = \pm 1, \pm 2, \ldots$.

For $n = 0$, obviously

$$a_0 = \frac{1}{\pi}\int_{-\pi}^{\pi} x^2 dx = \frac{2}{\pi}\int_{0}^{\pi} x^2 dx = \frac{2}{\pi}\frac{x^3}{3}\Big|_0^{\pi} = \frac{2\pi^2}{3}.$$

For $n =$ 1, 2, 3, … we have

$$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} x^2\cos nx dx = \frac{2}{\pi}\int_{0}^{\pi} x^2\cos nx dx = \frac{2}{\pi n}\left[x^2\sin nx\Big|_0^{\pi} - 2\int_{0}^{\pi} x\sin nx dx\right] =$$

$$= \frac{4}{n^2\pi}\left[x\cos nx\Big|_0^{\pi} - \int_0^{\pi} \cos nx dx \right] = (-1)^n \frac{4}{n^2}.$$

It is easy to verify that $b_n = 0$ for all $n,$ therefore, the Fourier series decomposition has the form:

$$x^2 = \frac{\pi^2}{3} - 4\left(\frac{\cos x}{1} - \frac{\cos 2x}{4} + \frac{\cos 3x}{9} - ... \right) = \frac{\pi^2}{3} + 4\sum_{n=1}^{\infty} (-1)^n \frac{\cos nx}{n^2}.$$

Fig. 35.2

We can decompose the function defined on an arbitrary segment $[-l, l]$ into a trigonometric series similar to the Fourier series. In this case, the decomposition has the form

$$\frac{a_0}{2} + \sum_{n=1}^{\infty}\left(a_n \cos\frac{n\pi x}{l} + b_n \sin\frac{n\pi x}{l} \right), \quad (35.9)$$

$$a_n = \frac{1}{l}\int_{-l}^{l} f(x)\cos\frac{n\pi x}{l}dx,\ \ b_n = \frac{1}{l}\int_{-l}^{l} f(x)\sin\frac{n\pi x}{l}dx,\ \ n = 0, 1, 2, \ldots .$$

If $f(x)$ is an even function on $[-l, l]$, i.e. if $f(-x) = f(x)$, $x \in [-l, l]$, then its Fourier coefficients b_n in (35.9) are equal to zero. Let us prove it. We have

$$b_n = \frac{1}{l}\int_{-l}^{l} f(x)\sin\frac{n\pi x}{l}dx = \frac{1}{l}\left[\int_{-l}^{0} f(x)\sin\frac{n\pi x}{l}dx + \int_{0}^{l}\sin\frac{n\pi x}{l}dx\right].$$

Make a substitution $x = -t$ in the first integral. Then, using the parity of f and oddness of the sine, we obtain

$$\int_{-l}^{0} f(x)\sin\frac{n\pi x}{l}dx = -\int_{l}^{0} f(-t)\sin\frac{n\pi(-t)}{l}dt = -\int_{0}^{l} f(x)\sin\frac{n\pi x}{l}dx.$$

Our assertion follows from here and from the previous equality.

In this case, coefficients a_n can be calculated by formulas

$$a_0 = \frac{2}{l}\int_{0}^{l} f(x)dx,\ \ a_n = \frac{2}{l}\int_{0}^{l} f(x)\cos\frac{n\pi x}{l}dx,\ \ n = 1, 2, \ldots .$$

It is similarly proved that if $f(x)$ is an odd function, then

$$a_n = 0,\ \ b_n = \frac{2}{l}\int_{0}^{l} f(x)\sin\frac{n\pi x}{l}dx.$$

Thus, if the function is even, then its Fourier series (35.5) contains only cosines, and if it is odd, it contains only sines. In the examples considered above, we saw, in particular, that the decomposition of the odd function

$f(x)=x$ contains only sines, and the decomposition of the even function $f(x)=x^2$ contains only cosines.

Standard deviation.

Representation of a function by an infinite series has the practical meaning that the finite sum of the first n terms of a series is an approximate expression of the decomposable function. In this case, it becomes necessary to evaluate the error.

Consider an arbitrary function $y=f(x)$ on $[a,b]$ and estimate the error when replacing this function with another function $\varphi(x)$. Let $\max|f(x)-\varphi(x)|$ be a measure of an error on $[a,b]$, i.e. **the largest deviation** function $\varphi(x)$ from $f(x)$. However, sometimes the largest deviation is inconvenient to take as a measure of approximation, and not only because the study of this value is difficult, but it is often more important to reduce the error "on average" than to decrease the largest deviation solving the function approximation problem. In such cases, the **mean square deviation** is taken as a measure of error δ, where

$$\delta^2=\frac{1}{b-a}\int_a^b[f(x)-\varphi(x)]^2\,dx$$

Let us find out the nature of the approximate representation of the periodic function $f(x)$ by trigonometric polynomials of the form

$$s_n(x)=\frac{a_0}{2}+\sum_{k=1}^{n}(a_k\cos kx+b_k\sin kx),$$

here $a_0,\ a_1,\ b_1,\ a_2,\ b_2,\ \ldots,\ a_n,\ b_n$ are Fourier coefficients, i. e. the sum of the first $(2n+1)$ members of the Fourier series.

Let $f(x)$ be a periodic function with a period of 2π. Among all trigonometric polynomials of order n

$$\frac{a_0}{2}+\sum_{k=1}^{n}\left(\alpha_k\cos kx+\beta_k\sin kx\right)$$

we need to find the polynomial for which the mean-square deviation

$$\delta_n^2=\frac{1}{2\pi}\int_{-\pi}^{\pi}\left[f(x)-\frac{a_0}{2}-\sum\left(\alpha_k\cos kx+\beta_k\sin kx\right)\right]^2 dx,$$

has the least value by choosing coefficients α_k and β_k.
The answer to this question gives the following Theorem.

Theorem . Among all trigonometric polynomials of order n, the smallest mean-square deviation from function $f(x)$ has the polynomial whose coefficients are the Fourier coefficients of the function $f(x)$.
We accept this theorem without proof. Also, without proof, we note that for any bounded piecewise monotonic function the mean-square deviation obtained by replacing this function with the n-th partial sum of the Fourier series tends to zero as $n\to\infty$, i. e. $\delta_n^2\to 0$ as $n\to\infty$.

Reference list

1) Arkhipov G.I., Sadovnichiy V.A., Chubarikov V.N. (2004). *Lectures on Mathematical Analysis*. Visshaya shkola. (in Russian)
2) Beklemishev D. V. (1998). *Course of analytical geometry and linear algebra.* Visshaya shkola. (in Russian)
3) Bugrov J. S., Nikolskiy S. M. (2016). *Higher mathematics in 3 volumes.* URITE. (in Russian)
4) Kremer N. Sh., Putko B. A., Trishin I. M., Friedman M. N. (2010). *Higher mathematics for economists*. UNITY. (in Russian)
5) Castles S. A., Cheremnykh Yu. N., Tolstopiatenko A. V. (1999). *Mathematical methods in Economics*. Delo I Servis. (in Russian)
6) *Crass M. S.*, Chuprunov. (2008). *Mathematics for economic specialties*. PITER. (in Russian)
7) Malykhin V. I. (2009). *Mathematics in Economics*. INFRA-M. (in Russian)
8) *General course of higher mathematics for economists*. (2007). INFRA-M. (in Russian)
9) Solodovnikov A. S., Babaytsev V. A., Brailov A.V., Shandra I. G. (2013). *Mathematics in Economics*. Finansy i statistica. (in Russian)

10) *Collection of problems in higher mathematics for economists*. (2003). INFRA-M.
11) *Handbook of mathematics for economists*. (1997). Visshaya shkola. (in Russian)
12) Fikhtengolts G. M. (2015). *Fundamentals of mathematical analysis*. Lan'. (in Russian)
13) Shikin E. V., Chkhartishvili A. G. (2000). *Mathematical methods and models in management*. Delo. (in Russian)
14) Anthony M., Biggs N. (1998). *Mathematics for Economics and Finance. Methods and Modeling*: Cambridge University Press.
15) Fuente de la Angel. (2000). *Mathematics Methods and Models for Economics*. Cambridge University Press.

Chapter 36. Basic concepts of linear programming

This book does not provide any systematic presentation of linear programming. Here we consider only some examples of optimization problems with limitations given by linear inequalities.

36.1. Resource problem

During economic activities of a single enterprise or the whole industry, it is often necessary to determine how to use available resources to achieve the maximum output. With a large number of possible solutions to this problem, it makes sense to choose the best one.

Mathematically, this problem is usually reduced to finding the maximum or minimum value of a function on a set defined by a system of inequalities. Let an enterprise produce from *m* types of resources *n* types of products. Let the production of the j-th type of product consume a_{ij} units of the i-th type of the resource. Matrix $A = (a_{ij})$ is called technological.

Let c_j be a specific profit margin from the sale of one unit of the *j*-th product. These specific profit margins form vector $C = (c_1, c_2, ..., c_n)$. Then the product $CX = c_1x_1 + c_2x_2 + ... + c_nx_n$ is an amount of profit, received from the sale of X units of manufactured product, where $X = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix}$. We denote this profit as $f(X)$.

Let b_i be a number of units of the i-th resource available to the enterprise. Then the need to take into account that the limited resources when drawing up production plants is expressed by the system of inequalities:

$$a_{i1}x_1 + a_{i2}x_2 + ... + a_{in}x_n \le b_i, \; i = 1, \dots, m. \qquad (36.1)$$

These resources being provided, it is required to produce such a combination of goods at which an enterprise's profit would be maximum. In other words, it is required to find the maximum value of function $f(X) = c_1x_1 + c_2x_2 + ... + c_nx_n$ under conditions (36.1). The problem defined this way is called **the optimization problem**. It is written as follows:

$$f(X) = c_1x_1 + c_2x_2 + ... + c_nx_n \to \max, \tag{36.2}$$

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n \le b_1, \\ a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n \le b_2, \\ ..... \\ a_{m1}x_1 + a_{m2}x_2 + ... + a_{mn}x_n \le b_m, \end{cases} \tag{36.3}$$

$$x_1 \ge 0,\ x_2 \ge 0,\ ...,\ x_n \ge 0. \tag{36.4}$$

Function $f(X)$ is called a **target function.**

Definition. **A valid solution (a plan)** of this problem is vector *X,* satisfying the constrained system (35.3) and non-negative conditions (35.4).

The set of valid solutions forms **a domain of valid sets.**

Definition. **An optimal solution (plan)** of the problem is a valid solution, such that its target function reaches its maximum (minimum).

36.2. General problem of linear programming

Let us formulate in general terms a **problem of linear programming**: *to find an extremum of linear function under linear constraints on variables.* Moreover, the set of variable values that satisfy all the linear constraints of a problem is called a **valid set** and a linear function, whose extremum is found, is a **target function.**

In practice, it is to apply linear programming for solving such problems where there are hundreds or thousands types of resources and types of products. The most popular algorithm for solving a problem of linear programming is the so-called **simplex-method.**

The presentation of this method is beyond the scope of our course. Without proof, we give only two theorems that contain the theoretical basis of the simplex method (and linear programming in general).

***Theorem* 36.1.** The problem of linear programming has an optimal solution if and only if the target function is bounded on the valid set in the direction of the extremum.

Before formulating the second theorem we note that the valid set, on which the extremum is found, is a polyhedral body (for $n = 2$ is a polygon for $n = 3$ is a polyhedron in three-dimensional space). The vertices of this polyhedral body are called the *corner body.*

***Theorem* 36.2.** If an extremum of the considered function of linear programming solution is reached, then it is reached in the corner point of the valid set.

Note that there is a finite number of corner points. The Simplex-method is a directed enumeration of the corner points of a valid set.

Let us consider a solution of linear programming with two variables with the **graphical method.**

***Example* 36.1.** Solve a linear programming problem:

$$f(X) = 3x_1 + 2x_2 \to \max,$$

$$\begin{cases} x_1 - x_2 + 2 \ge 0, & (1) \\ 3x_1 - 2x_2 - 6 \le 0, & (2) \\ x_2 - 3 \le 0, & (3) \end{cases}$$

$$x_1 \ge 0,\ x_2 \ge 0.$$

Solution. We take a rectangular system Ox_1x_2 on a plane (fig. 36.1).

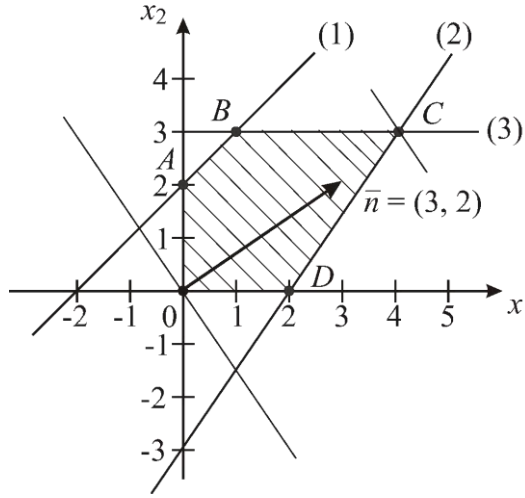


Fig. 36.1

We construct a line $x_1 - x_2 + 2 = 0$ corresponding to the first constraint. To select the desired half-plane, we must substitute coordinates of any point, which doesn't lie on the straight line, for example, $O(0, 0)$: $0 - 0 + 2 > 0$, in inequality (1). We obtain a strict inequality. Thus, point *O* lies in a half-plane of solutions.

Similarly, we construct lines $3x_1 - 2x_2 - 6 = 0$ and $x_2 - 3 = 0$ and choose the corresponding half-planes.

We also take into account the condition of non-negativity $x_1 \geq 0$, $x_2 \geq 0$. The intersection of all five half-planes gives us the desired valid set gives us the pentagon *OABCD*.

We construct a level line, for example, for $l = 0$: $3x_1 + 2x_2 = 0$. We move a level line in the normal direction. The last point along which the level line still crosses the valid set will be the maximum point. In our case, this

is point *C*. We find its coordinates by solving equations of lines which intersect at point *C*:

$$\begin{cases} 3x_1 - 2x_2 - 6 = 0, \\ x_2 - 3 = 0, \end{cases}$$

we obtain $x_1 = 4$, $x_2 = 3$. We calculate $f_{\max} = f(4, 3) = 18$.

It often happens that variables in problems of linear programming are *integers*. Such problems are more difficult to solve and the special methods have been developed for them. But if such a problem has two variables, then it is possible to solve it graphically. It is necessary to move a level line and to find the last *integer* point.

36.3. Elements of duality theory

The central part of linear programming is a dual theory. Any problem of the linear programming can be associated with another problem, which is called **dual (or conjugate)**.

Both problems (initial and dual to it) form a pair of dual problems. Each of the problems is dual to another one of the considered pair.

Let us consider a **resource allocation problem**. Let for production of n types of products P_1, P_2, …, P_n *m* types of resources S_1, S_2, …, S_m are used (this can be various types of raw materials, electricity, semi-finished products, etc.).

The volume of each type of resources is known; in other words, a vector of resources $B = (b_1, b_2, ..., b_m)'$ is known.

The consumption rate a_{ij} of the i-th resource for the production of one unit of the *j-th* type of product, i.e. the technological matrix $A = (a_{ij})$, $i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, n$ is known. Moreover, the profit of the sale of

one unit of each product type is known, i.e. the profit vector $C = (c_1, c_2, ..., c_n)$ is known. (Here vector *B* is a column vector, and C is a row vector.)

A manufacturer draws up a production plan that provides the maximum profit. The mathematical model of this problem, as already noted, in an expanded form is written as:

$$f = c_1 x_1 + c_2 x_2 + ... + c_n x_n \to \max, \tag{36.2}$$

$$\begin{cases} a_{11} x_1 + a_{12} x_2 + ... + a_{1n} x_n \le b_1, \\ a_{21} x_1 + a_{22} x_2 + ... + a_{2n} x_n \le b_2, \\ ..... \\ a_{m1} x_1 + a_{m2} x_2 + ... + a_{mn} x_n \le b_m, \end{cases} \tag{36.3}$$

$$x_1, x_2, \ldots, x_n \ge 0. \tag{36.4}$$

Here x_j, $j = 1, 2, \ldots, n$ is the volume of production of the *j-th* type of product. In the compact form, the target function and constraint system are usually written in the form:

$$f = \sum_{j=1}^{n} c_j x_j \to \max,$$

$$\sum_{j=1}^{n} a_{ij} x_j \le b_i, \; i = 1, 2, \ldots, m,$$

$$x_j \ge 0, \; j = 1, 2, \ldots, n.$$

Suppose that there is a buyer who wants to re-buy partially or fully the resources reserved to complete this task. In market economy, there are no categorical refusals; usually, everything is determined by the price and the terms of sale.

Let us consider the dual problem which solution determines the terms for the sale of resources. Denote the estimate (price) of the *i-th* resource as y_i, then the vector of these estimates will have a form $Y = (y_1, y_2, ..., y_m)'$. The cost of acquiring the *i-th* type of raw materials in quantity b_i is obviously equal to $b_i y_i$. A buyer, obviously, wants to pay less, so for them, the target function has a form

$$\varphi = b_1 y_1 + b_2 y_2 + ... + b_m y_m \to \min. \quad (36.5)$$

However, it is beneficial for the manufacturer acting as a seller to evaluate its resources in such a way that their total cost spent on each product of the *j-th* product is not less than the profit c_j, that the seller would receive from the sale of this product, i.e.

$$a_{1j} y_1 + a_{2j} y_2 + ... + a_{mj} y_m \geq c_j.$$

Thus, the system of constraints of the problem has a form

$$\begin{cases} a_{11} y_1 + a_{21} y_2 + ... + a_{m1} y_m \geq c_1, \\ a_{12} y_1 + a_{22} y_2 + ... + a_{m2} y_m \geq c_2, \\ ..... \\ a_{1n} y_1 + a_{2n} y_2 + ... + a_{nm} y_m \geq c_n. \end{cases} \quad (36.6)$$

Moreover, obviously, estimates of all the types of resources are non-negative:

$$y_i \geq 0, \; i = 1, 2, \ldots, m. \quad (36.7)$$

So, conditions (36.5)–(36.7) define the new problem of linear programming. It is called a **dual** problem to the initial problem (36.2)–(36.4).

Let us consider closely the **connection between the initial and the dual problems:**

1) coefficients c_j of the target function of the initial problem are free terms of the system of constraints (36.6) of the dual problem;

2) free terms b_j of the system of constraints (36.3) of the initial problem are coefficients of the target function of the dual problem;

3) the coefficient matrix of the constraint system of the dual problem is the transposed matrix of coefficients of the constraint system of the initial problem.

Further, it will be clear that if one of the dual problems has an optimal solution, then another also has an optimal solution (see theorem 36.4).

The pair of problems, considered above, refers to the so-called symmetrical problems. In the theory of duality *two pairs of symmetrical dual problems* are considered. We present them in the matrix-vector form (on the left side is an *initial* problem, on the right side is a *dual* one):

1. $f = CX \to \max,$ $\varphi = YB \to \min,$

$AX \le B,$ $YA \ge C,$ (36.8)

$X \ge 0;$ $Y \ge 0.$

2. $f = CX \to \min,$ $\varphi = YB \to \max,$

$AX \ge B,$ $YA \le C,$ (36.9)

$X \ge 0;$ $Y \ge 0.$

Recall that here

$$C = (c_1, c_2, ..., c_n), \qquad Y = (y_1, y_2, ..., y_m),$$

$$A = \begin{pmatrix} a_{11} & a_{12} & ... & a_{1n} \\ a_{21} & a_{22} & ... & a_{2n} \\ ... & ... & ... & ... \\ a_{m1} & a_{m2} & ... & a_{mn} \end{pmatrix}, \qquad B = \begin{pmatrix} b_1 \\ b_2 \\ ... \\ b_m \end{pmatrix}, \qquad X = \begin{pmatrix} x_1 \\ x_2 \\ ... \\ x_n \end{pmatrix}.$$

Note, that in the duality theory non-symmetrical pairs of dual problems are also used, but we will not consider them.

We now formulate more clearly **the rules of constructing the dual problem:**

1. The target function ϕ of the dual problem must be optimized in the opposite way to *f*, i.e. if $f \to \max$, then $\varphi \to \min$ and vice versa.
2. On the right side of the constraints of the initial problems are the coefficients for the variables of the target function of the dual function.
3. Matrices of the coefficients for the unknowns on the left sides of the constraints of both problems (initial and dual) are mutually transposed. Moreover, if the initial problem has a dimension $m \times n$ (m constraints with n unknowns), then the dual problem has a dimension $n \times m$.

***Example* 36.2.** Create a dual problem to the problem:

$$f = x_1 + 4x_2 + x_3 \to \max,$$

$$\begin{cases} -x_1 + 2x_2 + x_3 \le 4, \\ -2x_1 - 3x_2 - x_3 \le -6, \\ 3x_1 + x_2 + 2x_3 \le 9, \end{cases}$$

$x_j \ge 0$, $j = 1, 2, 3$.

Solution. Let us multiply the righthand sides of the constraints by the corresponding variable of the dual problem and construct a target function (which should be minimized, as the target function of the initial task is maximized):

$$4y_1 - 6y_2 + 9y_3 \to \min.$$

We transpose the matrix of coefficients for unknowns in the left-hand sides of the constraints of the initial problem, replace all inequalities with the opposite, and write in the righthand sides the corresponding coefficients of the target function of the initial task:

$$\begin{cases} -y_1 - 2y_2 + 3y_3 \geq 1, \\ 2y_1 - 3y_2 + y_3 \geq 4, \\ y_1 - y_2 + 2y_3 \geq 1. \end{cases}$$

Finally, we obtain the dual problem in the form

$$\varphi = 4y_1 - 6y_2 + 9y_3 \to \min,$$

$$\begin{cases} -y_1 - 2y_2 + 3y_3 \geq 1, \\ 2y_1 - 3y_2 + y_3 \geq 4, \\ y_1 - y_2 + 2y_3 \geq 1, \end{cases}$$

$y_i \geq 0$, $i = 1, 2, 3$.

***Example* 36.3.** Create the dual problem to this problem:

$$f = 2x_1 + 3x_3 + 2x_4 \to \min,$$

$$\begin{cases} x_1 - x_2 + 2x_3 + 3x_4 \geq 9, \\ -x_1 - 2x_2 + x_3 + 2x_4 \leq -8, \end{cases}$$

$x_j \geq 0$, $j = \overline{1, 4}$.

Solution. Since the target function is minimized, so all the inequality constraints should have the form «≥». Therefore, we transform the initial problem by multiplying the second inequality constraint by -1. The initial problem is written as

$$f = 2x_1 + 3x_3 + 2x_4 \to \min,$$

$$\begin{cases} x_1 - x_2 + 2x_3 + 3x_4 \geq 9, \\ x_1 + 2x_2 - x_3 - 2x_4 \geq 8, \end{cases}$$

$$x_j \geq 0, \quad j = \overline{1, 4}.$$

Now we create the dual problem similar to how it was done in the previous example:

$$\varphi = 9y_1 + 8y_2 \to \max,$$

$$\begin{cases} y_1 + y_2 \leq 2, \\ -y_1 + 2y_2 \leq 0, \\ 2y_1 - y_2 \leq 3, \\ 3y_1 - 2y_2 \leq 2, \end{cases}$$

$$y_i \geq 0, \; i = 1, 2.$$

Duality theorems establish a connection between optimal solutions of pairs of dual problems.

Let us consider a symmetrical pair of dual problems (36.8):

I. $f = CX \to \max,$

$AX \leq B,$

$X \geq 0,$

II. $\varphi = BY \to \min$

$YA \geq C,$

$Y \geq 0.$

(Remind that if problem (I) has dimension $m \times n$, then problem (II) has dimension $n \times m$.)

Note **the main inequality of the duality problem.**

***Theorem* 36.3.** Let X be any valid solution of initial problem (I), and Y is any valid solution of the dual problem (II). Then there is an inequality

$$f(X) \le \varphi(Y). \tag{36.10}$$

Proof. Since all the variables in the both problems are non-negative, we obtain (taking into account $YA \ge C$):

$$f(x) = CX \le (YA)X. \tag{*}$$

Due to associativity of matrix multiplication and taking into account $AX \le B$

$$(YA)X = Y(AX) \le YB = \varphi(Y). \tag{**}$$

Combining (*) and (**), we obtain

$f(X) \le \varphi(Y)$, q.e.d.

Let us note, in particular, that as applied to the problem considered in example 35.4, inequality (35.10) means

$$2x_1 + 3x_3 + 2x_4 \le 9y_1 + 8y_2.$$

Consequence forms the main inequality: if a valid set of one of the problems I, II is not empty, then the target function of another problem is bounded in extremum direction on its valid set.

Indeed, for example, let the set D of the initial problem be not empty, i.e. there exists at least one point $X^0 \in D$. Then, according to inequality (35.10), for any point Y from the valid set of problem II inequality

$\varphi(Y) \geq f(X^0)$ holds, i.e. all the values of function ϕ are bounded below by one number $f(X^0)$.

***Theorem* 36.4 (the main duality theorem).** If one of the dual problems I or II has an optimal problem, then another one has an optimal solution, and the extremum values of target functions are equal:

$$\max f = \min \varphi. \tag{36.11}$$

(We accept this theorem without proof).

One of the main consequences of the main duality theorem is a criterion of optimality of valid solutions. Let X^0 and Y^0 are valid solutions of the initial and the valid problems I and II. For these solutions to be optimal, the equality

$$f(X^0) = \varphi(Y^0). \tag{36.12}$$

Proof. 1. *Necessity*. Let X^0 and Y^0 be optimal solutions. Then $f(X^0) = \max f$, $\varphi(Y^0) = \min \varphi$ and equality (36.12) follows from the main duality theorem.

2. *Sufficiency*. Let inequality (36.12) hold and let X^0 be an arbitrary point from a valid set of the initial problem. Then by virtue of the main inequality (36.10), we obtain $f(X) \leq \varphi(Y^0) = f(X^0)$. Thus, $f(X^0) = \max f$, i.e. X^0 is a maximum point.

It is similarly proved that point Y^0, for which inequality (36.12) holds, is a minimum point.

***Theorem* 36.5 (the second duality theorem).** In order for the valid solutions $X = (x_1, x_2, ..., x_n)'$ and $Y = (y_1, y_2, ..., y_m)$ to be optimal

solutions for the pair of dual problems I and II, it is necessary and sufficient that the following equalities hold:

$$x_j\left(\sum_{i=1}^{m} a_{ij}y_i - c_j\right) = 0 \quad , j = 1, 2, \ldots, n; \qquad (36.13)$$

$$y_i\left(\sum_{j=1}^{n} a_{ij}x_j - b_i\right) = 0 \quad , i = 1, 2, \ldots, m. \qquad (36.14)$$

(We accept this theorem without proof.)

We clarify the meaning of equalities (35.13) and (35.14). For example, the second means that if the optimal solution is substituted into the constraint system (35.3), the *i-th* constraint of the initial problem is satisfied as a strict inequality, then the *i-th* coordinate of the optimal solution of the dual problem is equal to zero. Otherwise, if the *i-th* coordinate of the optimal solution of the dual problem is not equal to zero, then the *i-th* constraint of the initial problem when substituting the optimal solution becomes equal. These conditions establish the balance between problems I and II. That is why theorem 36.5 is also called the **equilibrium theorem**.

***Example* 36.4.** Solve the problem:

$$4x_1 + 3x_2 - 30x_3 \to \min,$$

$$\begin{cases} x_1 \quad\quad - 6x_3 \ge 1, \\ \quad x_2 - 5x_3 \ge 2, \end{cases}$$

$$x_1, x_2, x_3 \ge 0.$$

Solution. There are three variables in this problem. It is not possible to solve it graphically such as in example 35.1. Let us create the dual problem, solve it graphically and then solve this problem using the second duality theorem.

So we create the dual problem:

$$y_1 + 2y_2 \to \max,$$

$$\begin{cases} y_1 \le 4, & (1) \\ y_2 \le 3, & (2) \\ -6y_1 - 5y_2 \le -30, & (3) \end{cases},$$

$$y_1, y_2 \ge 0.$$

We solve it graphically. In fig. 36.2 the domain of valid solutions, normal $\overline{n} = (1, 2)$ and an optimal solution — a point (4, 3) are shown.

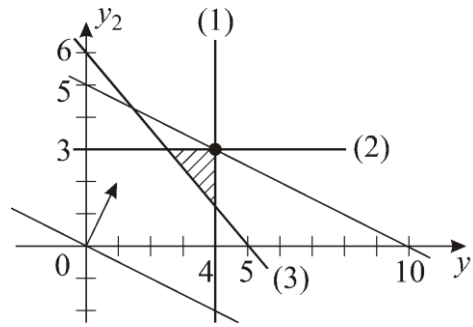


Fig. 36.2

Now we find the solution of the initial problem using the second duality theorem.

Since the third constraint of the dual problem is a strict inequality for $y_1 = 4, y_2 = 3$, then $x_3 = 0$. Then, since $y_1 > 0, y_2 > 0$,

$x_1 - 6x_3 = 1,\ x_2 - 5x_3 = 2$, hence $x_1 = 1,\ x_2 = 2$. Thus, the optimal solution of the initial problem is a point (1, 2, 0).

***Example* 36.5.** Solve the problem using the second duality theorem: $2x_1 + 6x_2 + 3x_3 \to \min$,

$$\begin{cases} 3 + x_1 - 3x_2 \le 0, \\ x_1 - 2x_2 + x_3 \ge 2, \end{cases}$$

$x_1, x_2, x_3 \ge 0$.

Solution. For this problem, we create a dual one. Firstly, we reduce all the inequalities to the form «≥», since the target function is minimized: $2x_1 + 6x_2 + 3x_3 \to \min$,

$$\begin{cases} -x_1 + 3x_2 \ge 3, \\ x_1 - 2x_2 + x_3 \ge 2, \end{cases}$$

$x_1,\ x_2,\ x_3 \ge 0$.

Now we write the dual problem:

$3y_1 + 2y_2 \to \max$,

$$\begin{cases} -y_1 + y_2 \le 2, \\ 3y_1 - 2y_2 \le 6, \\ y_2 \le 3, \end{cases}$$

$y_1, y_2 \ge 0$.

This problem coincides with the problem in example 36.1 (but there are variables x_1, x_2 and here there are variables y_1, y_2). Let us use its solution $y_1 = 4,\ y_2 = 3$.

Since the first constraint of the dual problem is satisfied as a strict inequality, then $x_1 = 0$ Since $y_1 > 0$, $y_2 > 0$, then

$$\begin{cases} -x_1 + 3x_2 = 3, \\ x_1 - 2x_2 + x_3 = 2, \end{cases}$$

hence $x_2 = 1,\ x_3 = 4$.

So, the optimal solution to the initial problem is (0, 1, 4).

Now let us consider a problem with four variables, which we can also reduce to the dual problem, which is solved graphically.

***Example* 36.6.** For the following problem create the dual one, solve it and find a solution to the initial problem, using the second duality theorem:

$$2x_1 + x_2 - 3x_3 + x_4 \to \max,$$

$$\begin{cases} x_1 + 2x_2 - x_4 \le 4, \\ x_1 - x_2 + x_3 + 3x_4 \le 1, \end{cases}$$

$$x_1, x_2, x_3, x_4 \ge 0$$

Solution. Create the dual problem:

$$4y_1 + y_2 \to \min,$$

$$\begin{cases} y_1 + y_2 \ge 2, \\ 2y_1 - y_2 \ge 1, \\ y_2 \ge -3, \\ -y_1 + 3y_2 \ge 1, \end{cases}$$

$$y_1, y_2 \ge 0.$$

Solving it graphically, we find a point (1, 1).

Now we apply the second duality theorem for finding the solution of the initial problem. We see that the third and the fourth constraints of the dual problem hold as strict inequalities.

Thus, $x_3 = 0,\ x_4 = 0$. In addition, since $y_1 > 0,\ y_2 > 0$,

$$\begin{cases} x_1 + 2x_2 \qquad - x_4 = 4, \\ x_1 - x_2 + x_3 + 3x_4 = 1, \end{cases}$$

hence (taking into account that $x_3 = x_4 = 0$) we obtain $x_1 = 2,\ x_2 = 1$. Thus, an optimal solution of the initial problem is a point (2, 1, 0, 0). In this case, obviously, $f(X)_{\max} = \varphi(Y)_{\min} = 5$.

Questions

1. What is a technological problem?
2. What is an optimization problem? How is it written?
3. What is the valid solution of the optimization problem? What solution is called optimal?
4. How is the problem of linear programming formulated?
5. How are the initial and dual problems of linear programming connected?
6. Let the initial problem of linear programming have a dimension $m \times n$. What is the dimension of the dual problem?
7. What is the main inequality in duality theory?
8. How is the main duality theorem formulated?
9. Can a linear programming problem with two variables be dual to a five variable problem?

Chapter 37. Summary of balance analysis

37.1. Leontief model

Effective functioning of diversified economy is possible only if there is a *balance* between sectors. Suppose that the entire production sphere of economy is represented by *n* so-called clean industries.

A clean industry is a conditional concept, a part of economy, relatively integral, producing its own homogeneous product and determined only by the type of product (such as, for example, extraction of raw materials, energy, agriculture, etc.). Some of the products are used for internal production-consumption (both by this industry and other industries), while the other part is intended for consumption in non-production sphere.

Consider the production process for a certain period of time (usually a year is such an interval).

We introduce the following notation:

x_i — a total output of the *i-th* industry (gross output);

x_{ij} — the volume of production of the *i-th* industry consumed by the *j-th* industry in the production process;

y_i — the volume of products of the *i-th* industry, intended for consumption in the non-productive sphere (the volume of final consumption).

Since the gross output of the *i-th* industry is equal to the sum of consumption in manufacturing and non-manufacturing sectors:

$$x_i = \sum_{j=1}^{n} x_{ij} + y_i \quad , i = 1, 2, \ldots, n. \qquad (37.1)$$

Equations (37.1) are called **balance relations.**

Since the products of different industries have different dimensions, we will consider the **value of the interindustry balance**, when all the values included in (37.1) have a value expression.

The mathematical model that allows us to analyze the relationship between industries was developed in 1936 by the American economist W. Leontief. W. Leontief, analyzing the American economy in the period before the Second World War, paid attention to the following important circumstance: for a long time, the values

$$a_{ij} = \frac{x_{ij}}{x_j}, \; i = 1, 2, \ldots, n \qquad (37.2)$$

vary slightly and can be considered as *constant* numbers.

This is because the production technology has remained almost constant for quite some time.

The above allows us to make the following assumption: for the output of products of the *j-th* industry of volume x_j, it is necessary to spend products of the *i-th* industry volume $a_{ij}x_i$, where a_{ij} is a constant coefficient. This assumption is called the **linear hypothesis.** According to this hypothesis

$$x_{ij} = a_{ij}x_j, \; i = 1, 2, \ldots, n. \qquad (37.3)$$

According to the linearity hypothesis, the numbers a_{ij} are constant, they are called **direct cost coefficients.**

Now, equations (37.1) taking into account (37.3) can be written in the form of a system:

$$\begin{cases} x_1 = a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n + y_1, \\ x_2 = a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n + y_2, \\ ..... \\ x_n = a_{n1}x_1 + a_{n2}x_2 + ... + a_{nn}x_n + y_n. \end{cases} \quad (37.4)$$

We introduce the gross output vector $\bar{x}$, the direct cost matrix A and the final consumption (or final product) vector $\bar{y}$:

$$\bar{x} = \begin{pmatrix} x_1 \\ x_2 \\ ... \\ x_n \end{pmatrix}, \quad A = \begin{pmatrix} a_{11} & a_{12} & ... & a_{1n} \\ a_{21} & a_{22} & ... & a_{2n} \\ .... & .... & ... & .... \\ a_{n1} & a_{n2} & ... & a_{nn} \end{pmatrix}, \quad \bar{y} = \begin{pmatrix} y_1 \\ y_2 \\ ... \\ y_n \end{pmatrix}. \quad (37.5)$$

Now the system (37.4) in matrix form has the form

$$\bar{x} = A\bar{x} + \bar{y}. \quad (37.6)$$

Equation (37.6) is called the **linear interindustry balance equation.** Together with the interpretation (37.6) of the vector $\bar{x}$, matrix A and vector $\bar{y}$, this equation is called the **Leontief model**.

The main task of the interindustry balance is to find a gross product vector $\bar{x}$ that, given the known direct cost matrix *A*, provides a given vector of final consumption $\bar{y}$. In other words, how many products of different types should be produced to ensure a given level of final consumption?

Obviously, the problem boils down to solving the equation (system of linear equations) (37.6) with an unknown vector $\bar{x}$, with a known matrix *A* and for a given vector $\bar{y}$.

In short, we agree to call the matrix *A non-negative* if all its components are non-negative. In this case, we write $A \geq 0$. *A nonnegative vector* is defined similarly.

In the problem above, obviously, $A \geq 0$, $\bar{y} \geq 0$ (this directly follows from the economic sense of *A* and $\bar{y}$). The sought vector $\bar{x}$ must also be non-negative: $\bar{x} \geq 0$.

We rewrite equation (37.6) in the form

$$(E - A)\,\bar{x} = \bar{y}. \tag{37.7}$$

If $(E - A)$ is nondegenerate matrix, then there exists a matrix inverse to it $(E - A)^{-1}$ and there is (and, moreover, the only) solution to equation (37.7):

$$\bar{x} = (E - A)^{-1}\bar{y}. \tag{37.8}$$

Matrix $S = (E - A)^{-1}$ called the **total cost matrix.**

Find out the economic meaning of the total cost matrix $S = (s_{ij})$.

Consider the unit vectors of the final product:

$$\bar{y}_1 = \begin{pmatrix} 1 \\ 0 \\ \dots \\ 0 \end{pmatrix}, \quad \bar{y}_2 = \begin{pmatrix} 0 \\ 1 \\ \dots \\ 0 \end{pmatrix}, \dots, \quad \bar{y}_n = \begin{pmatrix} 0 \\ 0 \\ \dots \\ 1 \end{pmatrix}.$$

For them, from (11.8) we obtain the corresponding gross output vectors:

$$\bar{x}_1 = \begin{pmatrix} s_{11} \\ s_{21} \\ \cdots \\ s_{n1} \end{pmatrix}, \quad \bar{x}_2 = \begin{pmatrix} s_{12} \\ s_{22} \\ \cdots \\ s_{n2} \end{pmatrix}, \ldots, \quad \bar{x}_n = \begin{pmatrix} s_{1n} \\ s_{2n} \\ \cdots \\ s_{nn} \end{pmatrix}.$$

Therefore, each element s_{ij} of matrix *S* is *the gross output of the i-th industry, necessary to ensure the output of a unit of the final product of the j-th industry.*

A matrix $A \geq 0$ is called **productive** if for any vector $\bar{y} \geq 0$ there exists a solution $\bar{x} \geq 0$ to equation (37.6). In this case, the Leontief model is called **productive.**

It turns out that there is no need to require the existence of a solution $\bar{x} \geq 0$ of equation (37.6) *for any vector* $\bar{y} \geq 0$. It is enough to establish the existence of such a solution for *at least one* vector $\bar{y} \geq 0$, as the following theorem shows, which we will accept without proof.

***Theorem* 37.1.** If for $A \geq 0$ and for some vector $\bar{y} \geq 0$, equation (11.6) has a solution $\bar{x} \geq 0$, then the matrix *A* is productive.

There are various performance criteria. Here are two of them.

The first criterion for productivity. A matrix $A \geq 0$ is productive if and only if the matrix $(E - A)^{-1}$ exists and is non-negative.

The second criterion for productivity. A matrix $A \geq 0$ is productive if the sum of the elements of any of its columns does not exceed unity:

$$\sum_{i=1}^{n} a_{ij} \leq 1.$$

37.2. Linear exchange model

Concepts of the eigenvector and eigenvalue of the matrix are applicable, in particular, to the analysis of the process of reciprocal buyings.

Let us consider the following question: what should the relationship between the budgets of countries be if they trade with each other so that it is mutually beneficial, i.e. there is practically no deficit for each of these countries. To answer this question, we consider a **linear model of exchange or a model of international trade.**

Let there be *n* countries. We denote their national budgets by x_1, x_2, …, x_n. Let a_{ij} be a share of the budget x_j,which the *j*-th country spends on the purchase of goods from the *i*-th country. We assume that the entire national budget of each country is spent only on the purchase of goods either within the country or outside it, i.e. fair equality:

$$\sum_{i=1}^{n} a_{ij} = 1 \quad , j = 1, 2, \ldots, n. \tag{37.9}$$

Consider a matrix composed of these coefficients a_{ij}:

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots. & \ldots. & \ldots & \ldots. \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix}. \tag{37.10}$$

It is called the **structural matrix of trade.**

In accordance with (37.9), the sum of the elements of any column of matrix *A* is equal to unity.

For the *i*-th country, the profit from domestic and foreign trade will be equal to

$$p_i = a_{i1}x_1 + a_{i2}x_2 + ... + a_{in}x_n . \quad (37.11)$$

The condition for the balance of trade is formulated as follows: the profit from the trade of each country should be no less than its national budget, i.e. trade must be balanced for every country: $p_i \geq x_i$ for all *i*, or

$$a_{i1}x_1 + a_{i2}x_2 + ... + a_{in}x_n \geq x_i , \; i = 1, 2, \ldots, n. \quad (37.12)$$

***Theorem* 37.2.** Equity-free trade condition is the following

$$p_i = x_i , \; i = 1, 2, \ldots, n.$$

Proof. Assume the opposite, i.e. $p_i > x_i$ for any *i*. Then the strict inequality holds:

$$\sum_{i=1}^{n} p_i > \sum_{i=1}^{n} x_i . \quad (37.13)$$

We write this equality taking into account (37.11):

$$\left(a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n\right) + \left(a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n\right) + ... +$$
$$+ \left(a_{n1}x_1 + a_{n2}x_2 + ... + a_{nn}x_n\right) > x_1 + x_2 + ... + x_n.$$

Grouping the terms, we obtain:

$$x_1\left(a_{11} + a_{21} + ... + a_{n1}\right) + x_2\left(a_{12} + a_{22} + ... + a_{n2}\right) + ... +$$
$$+ x_n\left(a_{1n} + a_{2n} + ... + a_{nn}\right) > x_1 + x_2 + ... + x_n.$$

It follows from (37.9) that all sums in parentheses are equal to unity. We get a contradiction:

$$x_1 + x_2 + \ldots + x_n > x_1 + x_2 + \ldots + x_n.$$

Therefore, strict inequality $p_i > x_i$ is impossible for any *i*. Therefore, all inequalities $p_i \geq x_i$, take the form of equalities:

$p_i = x_i$, $i = 1, 2, \ldots, n$. (37.14)

We introduce the vector of budgets:

$$\bar{x} = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix}.$$

Then the system of equalities (11.14) takes the form:

$$A\bar{x} = \bar{x}. \qquad (37.15)$$

This equation means that the eigenvector of matrix A corresponding to the eigenvalue $\lambda = 1$, consists of the budgets of countries conducting balanced trade. So, the problem was reduced to finding the eigenvector of the structural matrix of trade that corresponds to an eigenvalue $\lambda = 1$.

***Example* 37.1.** The structural matrix of trade of the three countries has the form:

$$A = \begin{pmatrix} 0{,}2 & 0{,}3 & 0{,}5 \\ 0{,}4 & 0{,}4 & 0{,}3 \\ 0{,}4 & 0{,}3 & 0{,}2 \end{pmatrix}.$$

Under what conditions is trade balanced in these countries?

Proof. We rewrite equation (37.15) in the form $(A - E)\bar{x} = \bar{0}$:

$$\begin{pmatrix} -0{,}8 & 0{,}3 & 0{,}5 \\ 0{,}4 & -0{,}6 & 0{,}3 \\ 0{,}4 & 0{,}3 & -0{,}8 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

The rank of this system is two. Solving it, we get

$$\begin{cases} x_1 = \dfrac{13}{12} x_3, \\ x_2 = \dfrac{11}{9} x_3. \end{cases}$$

Assuming x_3 = 36 (to avoid fractional numbers), we obtain a vector

$\bar{x} = (39, 44, 36)$,

which can be taken as an eigenvector.

So, the trade balance of these countries is achieved provided their budgets are in the ratio:

$x_1 : x_2 : x_3 = 39 : 44 : 36$.

Questions

1. What is the linearity hypothesis?
2. What form does the equation of linear interindustry balance have?
3. What is called the Leontief model?
4. Which matrix is called the total cost matrix? What is the economic meaning of this matrix?
5. Which matrix is called productive? In which case is the Leontief model called productive?
6 What are the criteria for matrix productivity?
7. What is the structural matrix of trade? What are the columns of this matrix characterized by?
8. What is the condition for balanced trade?

Table of Contents